Barbara Starosta

ON A CERTAIN MODEL OF NATURAL LANGUAGE

Originally published as "O pewnym modelu języka naturalnego," Studia Semiotyczne 5 (1974), 147–157. Translated by Agnieszka Ostaszewska.

Introduction. Vocabulary and phrase. Language. Distribution Classes. Divisions family

INTRODUCTION

In works on logical language analysis, language is characterised most often by a set of primal, i.e. non-definable, expressions and by the rules of syntax, which include the rules for the generation of expressions and the rules of transformation of expressions. Generally speaking — language is a vocabulary and a set of rules. If we mark vocabulary as V and the set of rules as R, then language L may be presented as the following ordered pair: $L = \langle V, R\rangle$.

The above description of language is useful for construction of artificial languages and for the study of such languages. It is not a sufficiently effective tool for studying natural languages. In such a situation it is worth making an effort to try out other theoretical models of the language, hoping that one would find among them a more handy model.

In this paper I present a theoretical language model, proposed by a group of Soviet and Rumanian researchers. In particular, I have used the materials contained in the work of S. Marcus *Teoretiko-mnozhestvennye modeli yazykov* (1970), which contains an extensive bibliography. This model was discussed at experimental seminars of a group of IT researchers from the Warsaw University. This paper is a summary of part of that discussion, which pertained to the notional apparatus of algebraic linguistics, in particular to the selection of Polish terminology, a deeper explanation of particular terms, as well as examples and applications in the Polish language.

VOCABULARY AND PHRASE

Let us consider a set of expressions V. Its elements are simple expressions, i.e. such expressions, no part of which is itself an expression. As a result of joining these expressions we get compound expressions. The operation of joining, called concatenantion, shall be marked with symbol *. A set of compound expressions, or to put it simpler, of expressions, shall be marked with the letter V^*.

We assume that set of simple expressions V is finite and not empty. We will call this set the VOCABULARY. The elements of set V shall be called WORDS. If we limit our reflections to the written word only, then the elements of set V shall be simple expressions, separated by means of spaces or punctuation marks, e.g. expression of this chapter: $V = \{$consider, set, expressions, $\ldots\}$. The elements of set V shall be marked with small letters of the Latin alphabet, e.g. $V = \{x_1, x_2, \ldots. x_n\}$ or $V = \{a, b, c, \ldots m\}$. The elements of set V^* shall be marked with small letters of the Greek alphabet:

$$V^* = \{\alpha, \beta, \gamma, ...\}.$$

The operation of * is conjunctive, i.e. for each element of set V^*, $\alpha, \beta, \gamma \in V^*$.

1.2. $(\alpha^*\beta)^* \gamma = \alpha^*(\beta^*\gamma)$

The property determined by formula 1.2. makes it possible to put into brackets any group of components. We therefore may omit the brackets in our further considerations. Operation * is internal in set V^*, which means that as a result of the joining of two elements of set V^* we obtain an element of set V^*. In other words, by joining expressions, we also get expressions.

We assume that an element of set V^* is a neutral element of operation *, which shall be marked with the symbol ϵ. A neutral element is an expression which, when added to any other expression, does not result in a change of the latter. We shall write this as follows: if $\alpha \in V^*$, then

2.2. $\alpha^{*}\epsilon = \epsilon^{*}\alpha = \alpha$

A natural element of an operation which meets condition 2.2. shall be called unity.

A system composed of a set of expressions V^*, the operation of joining * and the separated element ϵ constitutes an algebraic creation called UNITY SEMIGROUP. The notation of this system is the following: $(V^*, {}^*, \epsilon)$. This does not mean, of course, that one may consider the semigroup to be the same thing as the set of elements thereof. The semigroup is a MATHEMATICAL STRUCTURE, richer than the set itself. This abbreviated notation will not however be misleading. In order to simplify the notation, we will also omit the operation sign *, appearing between expressions. This sign shall be replaced by a space. For example: may * semigroup * this shall be written down as: may ␣ semigroup ␣ this or as: may semigroup this.

Since each expression of set V^* is obtained from the elements of set V, vocabulary V shall be called a set of GENERATORS of semigroup V^*. In other words, set V generates set V^*, if for each $\alpha \in V^*$ there exist such x_1, $x_2, \ldots x_n \in V^*$ that:

3.2. $\alpha = x_1 * x_2 * \ldots * x_n$

As a result of joining set V, which has a finite number of elements, we get set V^* of an infinite, however countable, number of elements.

For example, if $V = \{a, b, c\}$, the elements of V^* are, among others, the following expressions: *ab, aaa, abcd, a, cdee* and so on. We shall call the expressions from set V^* PHRASES.

Since each set V generates phrases in a unambiguous manner, we say that semigroup V^* is freely generated by set V or that we are dealing with a FREE semigroup V^* over set V. More specifically, we will say that semigroup V^* is freely generated by set V always, and only if for each

4.2.

$\alpha \in V^*$ and $x_i, x_j \in V^*$, $i = 1, 2, \ldots n$, $j = 1, 2, \ldots m$,

if

$\alpha = x_1\ x_2 \ldots x_n$

and

$\alpha = x_1\ x_2 \ldots x_m$

then $n = m$ and $x_i = y_j$ for $i = j$.

Therefore, each expression, according to 4.2., is determined in one way only.

Generally phrase a may be noted in the following manner:

$$\alpha = x_1\ x_2 \ldots x_k \ldots x_n = \prod_{i=1}^{n} x_i$$

The word x_k of expression α shall be marked as: $x_k = (\alpha)_k$, $1 \leq\ k \leq\ n$, and we call k a coordinate of expression α.

Another notion which we will introduce is the LENGTH of expression α. Length of an expression means the number of words constituting such expression. The length of expression α shall be marked as $\mid \alpha \mid$. If phrase α has n component words, then the length of this phrase is equal to n. Using the notion of the length of the phrase, we may utter a series of propositions, some of which, being quite obvious, will be presented without proofs.

Proposition 1. The length of a zero phrase is equal to zero, which we write down in the following way:

5.2. $\mid \epsilon \mid = 0$

Proposition 2. If the length of a phrase is equal to zero, then the phrase is a zero phrase.

6.2. $\mid \alpha \mid = 0 \Rightarrow\ \alpha = \epsilon$

Proposition 3. If the length of a phrase is equal to one, then such a phrase is an element of vocabulary V. In other words, simple expressions are expressions, which length is equal to one. Notation of this proposition looks like this:

7.2. $\mid \alpha \mid = 1 \Rightarrow\ \alpha \in\ V$

Proposition 4. The length of a phrase which is a concatenation of two phrases α and β is equal to the length of phrase α plus the length of phrase β.

8.2. $| \alpha\beta | = | \alpha | + | \beta |$

LANGUAGE

It is undisputed that from among all phrases which are possible to obtain from a finite number of simple expressions, only part is used in speech and in writing. It may be assumed that in fact we are dealing with a certain finite number of phrases selected from amongst all of the possible phrases. A distinction may be, with respect to the written word, whether a phrase can be encountered in publications. Criteria of such distinctions are interesting for a language researcher, yet they should not be the starting point, but quite to the contrary — the objective of the study of language is to find and describe these dependencies. In other words, the rules of generating and transforming selected phrases should be among the consequences of the assumptions adopted by the description of a language, and not among these assumptions themselves.

Using the notions introduced in chapter 2., and the notion of the selected phrase, we shall adopt the following definition of language L. If V^* is a free semigroup, generated by vocabulary V, and if Φ is a subset of set V^*, such that its elements are selected expressions, then LANGUAGE L shall mean an ordered pair $L = \langle V, \Phi \rangle$. If $\Phi = V^*$, ordered pair $\langle V, V^* \rangle$ shall be called a universal language.

In other words, we shall understand language as ordered pair $\langle V, \Phi \rangle$, whose first element is a set of words V, and the second element is the subset of the free unity semigroups, generated by vocabulary V, $\Phi \leq V$, It can be said with respect to every phrase from V^* whether it belongs to Φ or to V^* — Φ. As long as that does not result in misunderstandings, we will not distinguish between language L and the subset of selected phrases Φ. We shall say that phrase α belongs to language L, if and only if $\alpha \in \Phi$.

In the above definition the notion of the selected phrase is the most disputable. Sometimes it is treated as the equivalent of a sensible sentence. The latter does not however belong to notions which have been sufficiently precisely defined. The notion of a sentence itself, according to Z. Klemensiewicz (1969: 213), has had ca. one hundred different definitions. I would like to draw your attention to the fact that the notion of a selected phrase is broader than the notion of a sentence, since among others it also includes descriptions. E.g. *the man who climbed Mount Everest* is a selected phrase, but is not a sentence. The situation is similar with titles of books, chapters, short warning signs, etc., which are also selected phrases. In this sense, every element of set V, i.e. each word, may appear as a selected phrase of a length equal to one.

We will also omit the discussion about whether the set of selected phrases should only include those phrases, which have appeared in speech and writing, or also such phrases, which may appear in the future (Quine 1969: 79). We assume that in practice it is possible to determine a set of simple expressions V and a set of selected phrases of the natural language. It is however necessary to add additional assumptions to this proposition. One of these assumptions is the necessity for separating a FRAGMENT of the studied language, and the necessity to carry out an analysis of this selected fragment. The matter of selection of a representative fragment is a separate issue, which I will not tackle in this paper. I only wish to note that in the case of a natural language, which is characterised by a spontaneous, unplanned and changing process of creation, general rules may be found in statistically determined fragments of such a language. One always needs to remember, however, that a language of this kind always contains exceptions to the determined rules. Having made these assumptions, there are no obstacles carrying out research, either of the most broadly understood natural language, for example of the Polish language, or of the vocabulary and selected phrases of the epic poem *Pan Tadeusz,* or of the language of the works on geology or theoretical physics. One may also analyse one page of a printed text, for example the page on which the above sentences are written.

In a given ethnical language, Polish in our case, one may find phrases which indisputably belong to said language, and phrases which raise certain doubts. For example the phrase *koń ciągnie wóz* [*a horse pulls a wagon*] is considered to belong to the Polish language, whereby *łopata koszula lub* [*shovel shirt or*] is not considered to belong to $L_{p.}$We do not include into $L_p$such a phrase as: *mówił mówił siadł lub lub lub lub lub* [*spoke spoke sat or or or or or*]. However, such phrases as *fruwające krzesło pije ciągnący wóz* [*a flying chair is drinking a pulling wagon*] belong to $L_p$and may be encountered in poetic works. Doubts arise in the case of expressions such as: *wczoraj pójdzie do kina* [*he will go to the movies yesterday*] or *liczby pomarańczowo śpią* [*numbers sleep in an orange manner*]. We assume that elements of set Φ are selected phrases which are COMMONLY assumed to be such. We will only consider INDISPUTABLE selected phrases, narrowing, should a need arise, the group of Polish speakers to, for example, a group of chemists. Disputable phrases will be included in a set $V^* — \Phi$.

DISTRIBUTION CLASSES

The assumption that the object of our study is a language determined as the ordered pair $\langle V, \Phi\rangle$ is sufficient to provide a classification of the expressions of vocabulary V. Namely, we are able to divide vocabulary V into classes, which we will describe as distribution classes. This division shall be characterised in the following manner: two words, a and b, a, $b \in V$ belong to the same distribution class, if for each pair of phrases α and β belonging to language L, phrase $\alpha\ a\ \beta$ belongs to language L, always and only if phrase $\alpha\ b\ \beta$ also belongs to this language. We say that expressions a and b are exchangeable in a phrase α ... β. The ordered pair of expressions $\langle\alpha, \beta\rangle$ shall be called the context. We say that an expression is admissible in context $\langle\alpha, \beta\rangle$, if expression $\alpha\ a\ \beta$ belongs to language L.

The method of the division of language units, which in our case are simple expressions, into distribution classes has been applied for many years now. The first supporters of this method include the European structuralists from the glossematic school (Malberg1969: 230-259; Hjelmslev 1961) and representatives of the American structural linguistics, i.a. Z. S. Harris (1968). The differences between them boiling down to varying descriptions of set V.

The notion of distribution classes may be defined in various propositional apparatuses. Apart from the one described above, I also propose an equivalent description in the language of the set theory or of mathematics. In chapter 5 I will be using descriptions of the latter as they are simpler and shorter.

In the language of the set theory, the distribution classes may be described in the following manner: in a given set V we are describing the relation of exchangeability, which we shall mark with letter W.

Exchangeability relation W, as an equivalence relation determined on a non-empty set V, makes a DIVISION of this set into subsets of expressions which are in some way equivalent. The subset of elements of set V, equivalent to a given set α, shall be called the exchangeability relation abstraction class. Distribution classes are exchangeability relation abstraction classes. The family of exchangeability relation abstraction classes in set V is a subset of this set. This division shall be marked with the letter S. Abstraction classes, i.e. segments of subset S, shall be marked as $S\ (a)$, $S\ (b)$, $S\ (c)$, etc. $\mathrm{S} = \{\mathrm{S}(a), \mathrm{S}(b) \ldots\}$.

In mathematical language the above considerations have the following form: the exchangeability relation determined on set V determines a classification in this set. This set of distribution classes shall be called the QUOTIENT SET of set V, determined by the exchangeability relation W. We shall mark this set with symbol V/W, $S = V/W$. Each equivalence class is determined by one of its elements, which is a REPRESENTATIVE of this class. *S(a)* is a class in which each element is exchangeable in the same contexts as element *a*. Element *a* is a representative of this class.

Let us consider an example from natural language analysis. I assume that in this language set V is a finite set of MOST of the used primary expressions, and phrases are those expressions which are considered to be correct by a specified group of people. Division S combines the elements of vocabulary V in such subsets that each element of the vocabulary belongs to one and one class only. And so, for example S (*cat*) = {*cat, dog, king...*}. S (*cat*) means a set of all words exchangeable in the same contexts with the word *cat*. For example: *I hit a ... with a stick and it came back, I see a .... on the window, the house of ....* etc. These are the context, which after introduction in the place of dots of any word from the S (*cat*) set shall remain selected phrases of the Polish language. Similarly S (*or*) = {*or, and...*}. In contexts: *big... small*, *happy full ... green rock*, etc. expressions from set S (*or*) are interchangeable.

If vocabulary V does not contain homonymic or polysemic expressions, i.e. to put it generally, equivalent expressions, then it is possible to carry out the division into distribution classes. In the natural language equivalent expressions constitute, however, a considerably large group and cannot be omitted. For example: S (*cat*) = {*cat, dog, major...*} but also S (*cat*) = {*cat, house...*}. In the contexts *there is no..., he was looking for the ... long in the night*, etc. interchangeable are expressions from the set S (*cat*) = {*cat, house...*}. Similarly S (*lilac*) = {*lilac, bush, tulip ...*}, or S (*without*) = {*without, instead, apart from...*}.

We assume that vocabulary V does not contain equivalent expressions. Homonyms and polysems shall be marked by additional symbols, e.g.: cat_1, cat_2. $S(cat_1) \neq S(cat_2)$.

The division of the set of simple expressions into distribution classes constitutes a relatively good characteristic of language with poorly developed inflection forms. In Polish and other Slavic languages an additional difficulty is caused by the fact that expressions in those languages have various cases, genders, persons. A set of inflection forms of particular expressions, covering all of their declinations and conjugations is called a paradigm. We say that Slavic languages are paradigmatic. This must be reflected in the description of the language.

DIVISIONS FAMILY

The notion of a division into paradigms cannot be presented within the frameworks of the hitherto described model of the language. This model needs to be somewhat enriched. We shall then consider a family of divisions from a non-empty set V. We shall mark the divisions from this set as K_1, K_2 ... The family of divisions shall be marked with letter K, $K = \{K_1, K_2 \ldots K_m\}$. The number of divisions of set V is finite.

In other words, we assume that there exists a family of quotient sets of set V, determined by equivalence relations R_1, $R_2 \ldots R_m$. In such a case: $K_1 = V/R_1$, $K_2 = V/R_2$ ... $K_m = V/R_m$.

Each division K_i, $i = 1, 2 \ldots m$, is a family of subsets of set V, meeting the following conditions:

1. Each element of K_i is a non-empty set of set V, which we write down in the following manner:

1.5. $\bigwedge_{A \in K} A \subset V \Rightarrow A \neq \emptyset$

2. Two different elements of set K_i are disjunctive sets.

2.5. $\bigwedge_{A,B \in K} A \neq B \Rightarrow A \cap B \neq \emptyset$

3. Each element of set V belongs to a certain element of division K_i. If $K_i = \{A_1, A_2 \ldots A_k\}$

3.5. $A_1 \cup A_2 \cup \ldots \cup A_k = V$

Elements of the family of divisions K are found in various types of divisions of set V. One such division has been discussed in the previous paragraph. This is the division into the distribution classes S, $S \in K$. Among these divisions we can distinguish two particular divisions, which we shall mark as K_∇ and K_Δ. Division K_∇ is such a division, whose sole element is the entire set V $K_\nabla = \{V\}$. That division is the THICKEST division of set V. The second distinguished division is the family of one-element subsets of set V, $K_\Delta = \{\{x\}\}$, $x \in V$. This division is marked sometimes with letter E and we say that E is the FINEST division of set V. On any two divisions of set V, K_i, $K_j \in K$ we may say that:

$K_i = K_j$ or $K_i \subset K_j$ or $K_j \subset K_i$

Set K is ordered into a relation of including. If subdivision K_i is contained in subdivision K_j, $K_i \subset K_j$, we say that division K_i is INSCRIBED in set

K_j, or that set K_i is FINER than set K_j. We than may say that set K_j is DESCRIBED on subdivision K_i, or that it is THICKER than division K_i.

Apart from two divisions K_∇ and K_Δ, we distinguish one more division, which we interpret as the division of set V into paradigms. We shall mark this division with the letter P. Each expression from set V belongs to one class and one class only. We shall mark these classes as $P(x_i)$, $x_1 \in V$. $P(x_i)$ means a class of those elements from set V, whose representative is x_i. Set $P(x_1)$ is a set of various inflexion forms of a word x_i. For example P (*kot*) = {*kot, kota, kotom, kocie*...}, P (*mówił*) = {*mówił, mówiła, mówić*}, P (*na*) = {*na*}, P (*zamówiło*) = {*zamówiła, zamówić, zamówiłem*}. One class includes all the words with the same morpheme and affix.

If $a = x_1\ x_2\ ...\ x_k$ is a phrase from set V^*, $a \in V^*$, then sequence $P(x_1)$ $P(x_2)$ ... $P(x_k)$ shall be called P — the structure of phrase a and shall be marked as $P(a)$.

4.5. $P(a) = \prod_{i=1}^{k} P(x_i)$

For example when a = *poszedł lub kino*, $P(a) = P$ (*poszedł*) P (*lub*) P (*kino*), whereby P (*poszedł*) = {*poszedł, poszła, poszło*...}, P (*lub*) = {*lub*} P (*kino*) = {*kino, kina, kinom, kinami*...}.

If $P(a)$ is P — structure of phrase a, $a \in V^*$, then phrase a is an element of structure $P(a)$, $a \in P(a)$.

5.5. $a \in V^*$ and $P(a) \Rightarrow a \in P(a)$

Let $V = \{a, b, c, d\}$, $P(a) = \{a, b\} = P(b)$, $P(d) = P(c) = \{c, d\}$, i.e. $P = \{\{a, b\}, \{c, d\}\}$, $a = abca \in V^*$, then $P(a) = P(a)\ P(b)\ P(c)\ P(a)$ $= \{a, b\}\ \{a, b\}\ \{c, d\}\ \{a, b\} = \{a\ a,\ a\ b,\ b\ b,\ b\ a\}\ \{c\ a,\ c\ b,\ d\ a,\ d\ b\} = \{a\ a\ c\ a,\ a\ a\ c\ b,\ ...\ a\ b\ c\ a\ ...\ b\ a\ d\ b\}$. Phrase $a\ b\ c\ a$ belongs to $P(a)$.

If phrase a is a selected phrase, $a \in \Phi$, we shall call the structure of this phrase P a SELECTED structure. For example P — the structure of phrase: *książka leży na stole* [*a book is on a table*], is a selected phrase: P (*książka leży na stole*) = P (*książka*) P (*leży*) P (*na*) P (*stole*).

The above considerations may be presented in a slightly different form. Let us consider set $P = \{P_1, P_2 \ldots P_k\}$ and the operation of concatenation determined by this set. The set of all P — structures is any semigroup generated by set P. We shall mark this set as P^*. Each P — structure is an element of P^*. For example let $P = \{P_1, P_2, P_3\}$, From set P we get the following P — structures: P_1P_2, $P_1P_2P_3$, $P_1P_3P_1P_3$ etc. From among all of the P — structures which are possible to obtain a part constitutes selected structures. Let P_1= P (*książka*) P_2= P (*lub*) P_3= P (*stół*). The phrase *książka lub stół* [*a book or a table*] belongs to selected phrases. The structure $P_1P_2P_3$= P (*książka*) P (*lub*) P (*stół*) belongs to selected structures.

We understand PARADIGMATIC language as the following ordered triad $\langle V, \phi, P\rangle$, where V is a set of primary expressions, Φ is a set of selected phrases and P is a SUBSET of set V. We assume that apart from the vocabulary and the selected phrases we also have given, by a certain interpretation, sets of all forms of a given word. The division of vocabulary V does not depend on the fact which phrases are recognized in a given language as selected phrases.

Let us consider set of expression V. According to the definition of language provided above, we adopt a division of this set into classes of paradigms. Then, each words from set V belongs to one of the classes from set P. E.g. P (*dom*) = {*dom, domowi, domu...*}, when V is the vocabulary of the natural Polish language. We also have given a set of selected phrases Φ. Having the above data we may determine division S into distribution classes. E.g. S (*dom*) = {*dom*, *stół*, *pies*} [S (*house*) = {*house*, *table*, *dog*}].

Each word of language $L = \langle V, \phi, P\rangle$ belongs to one class from set P, as well as to one class from set S. In other words, each element of set V belongs to a new subdivision of set V, obtained as a result of applying both subdivisions S and P of the operation of CROSSING. Generally speaking, if we have two non-empty subdivisions of set V, $P = \{P_1, P_2 \ldots P_k\}$ and $S = \{S_1, S_2 \ldots S_m\}$ we generate all products $P_i \wedge S_j$, $i = 1, 2 \ldots k$, $j = 1, 2 \ldots m$ elements of the first subdivision by the elements of the second subdivision. These products are called the COMPONENTS of set V due to divisions P and S. The components are mutually disjunctive and together exhaust set V. The family of NON-EMPTY components $P \wedge S$ is a SUBDIVISION of set V. This division shall be marked as I. $I = P \wedge S$, $I \in K$.

For example, we characterise language L in the following manner:

$V = \{a, b, c, d\}$, $P(a) = \{a\}$, $P(b) = \{b\}$, $P(c) = \{c, d\}$, $\Phi = \{a\ b, c\ b, a\ d, c\ d\}$.

In view of the data above, division S shall look as follows:

$S(a) = S(c) = \{a, c\}$, $S(b) = S(d) = \{b, d\}$
$S(c) \wedge P(a) = \{a, c\} \wedge \{a\} = \{a\}$
$S(c) \wedge P(b) = \{a, c\} \wedge \{b\} = \varnothing$

$S(c) \wedge P(c) = \{a, c\} \wedge \{c, d\} = \{c\}$
$S(b) \wedge P(a) = \{b, d\} \wedge \{a\} = \varnothing$
$S(b) \wedge P(b) = \{b, d\} \wedge \{b\} = \{b\}$
$S(b) \wedge P(c) = \{b, d\} \wedge \{c, d\} = \{d\}$

As a result of crossing two subdivisions S and P, we get a new division of set V, $I = \{\{a\}, \{b\}, \{c\}, \{d\}\}$. In this case, this division is equal to subdivision K_Δ, which is the family of one-element subsets of set V. As a result of crossing two subdivisions, we obtain a finer division than the component divisions.

The operation of crossing may be performed on all elements of set K. As a result of this operation we obtain new division of set V. These divisions are also elements of set K. Generally we may say that set K is closed in view of the operation of crossing.

Set K is also closed to the operation which is sometimes called the union of subdivisions or the joining of subdivisions (in French: somme; see: Félix 1973: 13). In Polish there is no relevant word for a description of these kinds of operations on sets. Let us present this operation with an example:

Language L is characterised in the following manner:

$v = \{a,b,c,d,e\}$
$P(a) = \{a\}$
$P(b) = \{b\}$
$P(c) = \{c,d,e\}$
$\Phi = \{ab, ba, abc, bac, de\}$.

In view of the above data, S is a division as follows:

$S(a) = \{a,b\} = S(b)$
$P(c) = \{c\}$
$P(d) = \{d\}$
$P(e) = \{e\}$.

As a result of a union, we obtain the following division I:

$I(a) = I(b) = \{a, b\}$
$I(c) = I(d) = I(e) = \{c, d, e\}$.