MATHEMATICAL MODEL OF LINGUISTIC ANALYSIS

JERZY POGONOWSKI

Adam Mickiewicz University, Poznań

We construct in this paper a general mathematical model of a wide class of linguistic theories which we call linguistic analyses. The definition of a linguistic analysis will be given below. We assume that the reader is familiar with fundamental notions from set theory and the theory of models (see for example Chang and Keisler 1973, Kuratowski 1976). The study of language consists of the establishment of linguistic units, their description and the determination of relations between these units. By a *linguistic analysis* we mean an investigation of language which uses theoretical-linguistic notions only. Of course, some pure linguistic notions are formed on the basis of external (with respect to linguistics) sciences — for example classification of sounds determined by their acoustic or physiological features. The construction of the notional apparatus, however, precedes the linguistic theory (here: the linguistic analysis). The latter investigates the structure of language with the help of fixed and connected notional apparatus.

We assume that the following statements are true:

1. Concrete utterances (of arbitrary length) are the only data for any linguistic analysis.
2. The decomposition of utterances into constituent segments is a principle in the linguistic analysis under consideration.
3. Relations between linguistic units form a basis for decomposition of utterances.
4. Any linguistic analysis distinguishes levels in language. Two utterances belong to the same level if and only if the relations between their segments are of the same type.
5. For any two consecutive levels of language utterances which belong to one of them are treated as combinations of utterances from the second level.

6. Each utterance is a concrete individual object.

Compare these statements with Hjelmslev's postulates for linguistics theory (Hjelmslev 1953).

Shortly speaking, the matter of any linguistic analysis lies in the association of a theoretical construction with any utterance. Let us call these constructions *analyzed texts*. Every analyzed text consists of a set of elements and a family of relations between these elements.

We use the following notations:

— Str_ω denotes the class of all finite structures of type ω
— $\bar{\bar{X}}$ denotes the cardinal of the set X
— $\mathscr{P}(X)$ denotes the family of all subsets of X
— $|\mathfrak{A}|$ denotes the domain of the relational structure $\mathfrak{A}$
— $\mathfrak{A}\models\varphi[x]$ means that x satisfies φ in $\mathfrak{A}$
— $f[X]$ denotes the image of the set X under the function f
— if R is an equivalence relation on X then $X/_R$ denotes the family of all R — equivalence classes

We are ready to construct a mathematical model of linguistic analysis.

Let $\Omega_1,\ldots,\Omega_k$ $(k>1)$ be a sequence of (finite) relational types. We introduce the abbreviation: $\Omega=<\Omega_1,\ldots,\Omega_k>$ The following definition is a basis for all further investigations:

Definition 1.

A sequence $<S_1,\ldots,S_k>$ of sets of non-empty relational structures is called an *Ω-analysis*, if the following conditions are satisfied:

1. $S_i\subseteq \mathrm{Str}_{\Omega i}$, for all i such that $1\leqslant i\leqslant k$
2. $0<\bar{\bar{S}}_k\leqslant\chi_0$
3. for every i, $1\leqslant i<k$, and every $\mathfrak{A}\in S_i$ there exists $\mathfrak{B}\in S_{i+1}$ such that $\mathfrak{A}\in|\mathfrak{B}|$
4. for every i, $1\leqslant i\leqslant k$, and any $\mathfrak{A}$, $\mathfrak{B}\in S_i$: if $\mathfrak{A}\neq\mathfrak{B}$ then $|\mathfrak{A}|\cap|\mathfrak{B}|=0$
5. for every i, $1\leqslant i<k$: if $\mathfrak{A}\in S_{i+1}$ then $|\mathfrak{A}|\subseteq S_i$

We will clarify now the linguistic sense of the above definition. The integer k equals the number of language levels, which are investigated by the linguistic analysis in question. S_0 k depends on the number of kinds of considered linguistic units. Every relational type Ω_i consists of a finite number of predicates. The sets Ω_i are determined by relations between linguistic entities. In other words: every set Ω_i consists of predicates, which can be treated as names for relations between linguistic entities on the defined linguistic levels. For example, the ordering of words in a sentence is a relation between words. The sets $S_1,\ldots,S_k$ are formed in the following way. We associate with every analyzed text a relational structure. The power of the domain of this structure equals the number of linguistic units in the considered text. Simultaneously, the relations of this structure correspond to the relations between the linguistic units in the considered text. Every set S_i consists of

relational structures obtained from all texts (from a fixed, defined level) in the way described above. For example, if the set S_{i+1} corresponds to the set of all sentences then the set S_i corresponds to the set of all words used in these sentences (under the assumption that sentences consist of words). This example explains the meaning of conditions 3 and 5. Condition 1 becomes clear when we remember that every text consists of a finite number of elements. Condition 4 expresses the fact that every utterance (and every analyzed text, respectively) is treated as a unique, unrepeatable object. Condition 2 will be explained below. At the moment, it suffices to know that the assumption $\bar{\bar{S}}_k<\chi_0$ corresponds to the situation where only a finite number of texts is considered and the assumption $\bar{\bar{S}}_k=\chi_0$ expresses an infinite, countable set of possible texts.

Let us consider some concrete examples.

Example 1. (J. A. Szrejder 1975:154)

Let $k=2$. Assume that $\Omega_1=0$ and $\Omega_2=\{P_1, P_2\}$ where P_1, P_2 are two-argument predicates.

Let us consider two levels in Russian, namely those of sentences and words. We are interested in two relations between words in sentences:

1. The ordering of words in a sentence
2. The relation of concord.

We associate with every Russian sentence (analyzed with respect to the relations mentioned above) a relational structure $\mathfrak{A}=\langle|\mathfrak{A}|, R_1, R_2\rangle$ in the following way:

— $|\mathfrak{A}|$ has as many elements as many words are in a considered sentence
— R_1 is a relation corresponding to the ordering of words in a considered sentence
— R_2 is a relation corresponding to the relation of concord

All structures, constructed in the way described above form the set S_2. Finally, the set S_1 is defined as follows:

$$S_1=\bigcup_{\mathfrak{A}\in S_2}|\mathfrak{A}|$$

As an example we will take the sentence from A. S. Pushkin poetry:

"И сатана пристав, с веселием на лике

a_1 a_2 a_3 a_4 a_5 a_6 a_7

лобзанием своим насквозь прожет уста

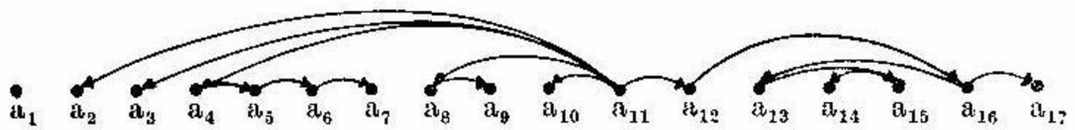
a_8 a_9 a_{10} a_{11} a_{12}

в предотельскую ночь лобзавщие Христа"

a_{13} a_{14} a_{15} a_{16} a_{17}

We have then:

$$\mathfrak{A}=\langle\{a_1, a_2, \ldots, a_{17}\}, R_1, R_2\rangle$$
$$a_i R_1 a_j \equiv j=i+1 \qquad 1\leqslant i, j\leqslant 17$$



Here an arrow between a_i and a_j means that $a_i\ R_2\ a_j$. This example could be developed. We can assume that the considered relations have some general properties. Then we can obtain some theorems about the Ω-analysis under investigation (see Szrejder 1975:154 ff. for this matter).

Example 2.

This example will be more general than the previous one. Let some concrete linguistic analysis be given. Assume that this analysis distinguishes the language levels of:

- sequences of sentences
- sentences
- words
- morphs
- phones

Any sequence of sentences consists of sentences, any sentence consists of words and so on. Suppose that the linguistic analysis under consideration gives a list of relations between linguistic entities on each language level. In this case we can construct the corresponding Ω-analysis in the following way:

1. We associate a relational structure with any sequence of sentences. The power of the domain of this structure equals the number of sentences in the considered sequence. Relations in this structure correspond to the linguistic dependencies between sentences in the sequence under consideration.

2. We associate a relational structure with any sentence. The power of the domain of this structure equals the number of words belonging to the considered sentence and its relations correspond to linguistic connections between words in the sentence.

3, 4, 5. Words, morphs, and phones are treated in a similar way as above.

Immediately from def. 1 some simple theorems follow. We omit proofs, which are rather technical.

Theorem 1.

For every i such that $1\leqslant i<k$ the following equality holds:

$$\bigcup_{\mathfrak{A}\in S_{i-1}}|\mathfrak{A}|=S_i$$

Theorem 2.

Let $1\leqslant i<k$. Then relation $\sim_i$ defined by $\mathfrak{A}\sim_i\mathfrak{B}$ if and only if there is $\mathfrak{B}\in S_{i+1}$ such that $\mathfrak{A}\in|\mathfrak{C}|$ and $\mathfrak{B}\in|\mathfrak{C}|$ is an equivalence relation on S_i.

Corollary 1.

The equivalence classes of the relation $\sim_i$ are of the form $|\mathfrak{A}|$ where $\mathfrak{A}\in S_{i+1}$.

We can apply the above results for defining the functions f_i, which will be useful in further considerations:

$$f_i: S_i\to S_{i+1}$$
$$f_i(\mathfrak{A})=(\iota\mathfrak{B})\,[\mathfrak{A}\in|\mathfrak{B}|]$$

One can define in an easy way the function $\hat{f}_i$, which associates with each relational structure $\mathfrak{A}$ its domain $\mathfrak{A}$:

$$\hat{f}_i: S_{i+1}\to\mathscr{P}(S_i)$$
$$\hat{f}_i(\mathfrak{A})=|\mathfrak{A}|$$

There funtions are connected by the condition:

$$\text{for every } \mathfrak{A}\in S_{i+1},\quad f_i[\hat{f}_i(\mathfrak{A})]=\{\mathfrak{A}\}$$

Theorem 3.

a) if $\bar{\bar{S}}_k<\chi_0$, then for every i such that $1\leqslant i\leqslant k$: $\bar{\bar{S}}_i<\chi_0$

b) if $\bar{\bar{S}}_k=\chi_0$, then for every i such that $1\leqslant i\leqslant k$: $\bar{\bar{S}}_i=\chi_0$

It is well known that we can associate some formal languages with relational structures of a fixed type. Namely, let $L(\Omega_i)$ denote a first-order language for which the set Ω_i forms the set of non-logical constants. Our next goal is the investigation of relations between the sets S_i.

Theorem 4.

Let φ be a formula with one free variable from the language $L(\Omega_{i+1})$. Define the relation $\sim_\varphi$ for all $\mathfrak{A}, \mathfrak{B}\in S_i$ by the condition:

$\mathfrak{A}\sim_\varphi\mathfrak{B}$ if and only if $(f_i(\mathfrak{A})|=\varphi\,[\mathfrak{A}]\equiv f_i(\mathfrak{A})|=\varphi\,[\mathfrak{B}])$. Then $\sim_\varphi$ is an equivalence relation on S_i.

Intuitively speaking, $\mathfrak{A}$ and $\mathfrak{B}$ are in the relation $\sim_\varphi$ if and only if $\mathfrak{A}$ and have the property φ. Theorem 4 can be of course generalized:

Theorem 5.

Let ψ be a formula with two free variables from the language $L(\Omega_{i+1})$. Take $\mathfrak{A}, \mathfrak{B}, \mathfrak{C}, \mathfrak{D}\in S_i$ such that $\mathfrak{A}\sim_i\mathfrak{B}$ and $\mathfrak{C}\sim_i\mathfrak{D}$. Define the relation $\approx_\psi$:

$\langle\mathfrak{A}, \mathfrak{B}\rangle\approx_\psi\langle\mathfrak{C}, \mathfrak{D}\rangle$ if and only if $(f_i(\mathfrak{A})|=\psi\,[\mathfrak{A}, \mathfrak{B}]\equiv f_i(\mathfrak{C})|=\psi\,[\mathfrak{C}, \mathfrak{D}])$.

Then $\approx_\psi$ is as equivalence relation on the set

$$\{\langle\mathfrak{A}, \mathfrak{B}\rangle\in S_i^2: \mathfrak{A}\sim_i\mathfrak{B}\}.$$

The linguistic sense of the above theorems is explained by the following example.

Example 3.

Take Ω-analysis from example 1.

a) let $\psi(v_2, v_3)$ be a formula from $L(\Omega_2)$ of the form

$$\bigvee_{v_1}(P_2(v_1, v_2)\ \&\ P_2(v_1, v_3))$$

The corresponding relation $\approx_\psi$ holds between those pairs of sturctures from S_1 which are associated with words governed through a common word.
We have in notations of ex. 1:

$$\langle a_2, a_3\rangle \approx_\psi \langle a_3, a_4\rangle$$
$$\langle a_{10}, a_{12}\rangle \approx_\psi \langle a_{13}, a_{14}\rangle$$

b) let $\psi(v_2, v_3)$ be a formula:

$$P_1(v_2v_3)\&\neg\bigvee_{v_1}(P_1(v_2v_1)\&P_1(v_1v_2))$$

Then the relation $\approx_\psi$ holds between elements of the set S_1 which are associated with consecutive words.

As we have seen, we are able to speak about connections between two language levels with the help of relations like $\sim\varphi$ or $\approx_\psi$. For a fixed linguistic analysis we can represent a large number of relations between linguistic entities with relations of the form considered above. It is easy to give further examples of relations of this type. Namely, let φ be a formula with one free variable from the language $L(\Omega_{i+1})$. For all $\mathfrak{A}, \mathfrak{B}, \in S_i$ define relations:

$\mathfrak{A}R_1\mathfrak{B}$ if and only if $(f_i(\mathfrak{A})\models\varphi[\mathfrak{A}] \to f_i(\mathfrak{B})\models\varphi[\mathfrak{B}])$
$\mathfrak{A}R_2\mathfrak{B}$ if and only if $\mathfrak{A}R_1\mathfrak{B}\ \&\ \neg(\mathfrak{B}R_1\mathfrak{A})$

The linguistic sense of the above relations can be easily imagined.

The linear character of language is often emphasized. Elements of speech are linearly ordered: words in sentences, phones in morphs, etc. On the other hand, linear ordering of elements plays a less important role in the analysis of big utterances (for example sequences of sentences). Having this in mind we define now some special class of Ω-analyses.

Definition 2.
Let $i_1, \ldots, i_s$ be some of numbers $1, \ldots, k$ where $1 \leqslant s \leqslant k$ and if $m \neq n$, then $i_m \neq i_n$. By *an $i_1, \ldots, i_s$-linear Ω-analysis* we mean an Ω-analysis for which the following condition holds: 6. For every j such that $1 \leqslant j \leqslant s$ there is binary predicate $<_{i_j}$ in Ω_1 such that $<_{i_j}$ is interpreted as linear ordering in all structures from S_{i_j}.

We will show that in the case where some Ω-analysis is $i+1$—linear we can associate some interesting algebraic structures with the set S_{i+1} (under some assumptions about S_i). Let h be an equivalence relation on S_i, defining without use of predicates from Ω_{i+1}. Denote the canonical mapping induced by hby π_h. We have then:

$$\pi_h : S_i \to S_i/_h$$
$$\pi_h(\mathfrak{A}) = \{\mathfrak{B}\ \varepsilon\ S_i : \mathfrak{A}\ h\ \mathfrak{B}\} = [\mathfrak{A}]_h$$

Here $[\mathfrak{A}]_h$ denote the h-equivalence class containing $\mathfrak{A}$.

Let F_h be a free semigroup generated by the set $S_i/_h$. Assume that con-

sidered Ω-analysis is i+1-linear. For every $\mathfrak{A} \in S_{i+1}$ there exists exactly one structure $\mathfrak{A} \restriction \{<_{i+1}\}$ which is a reduct of $\mathfrak{A}$ to the type $\{<_{i+1}\}$. Define:

$$S_{i+1} \restriction \{<_{i+1}\} = \{\mathfrak{A} \restriction \}<_{i+1}\{ : \mathfrak{A} \in S_{i+1}\}$$

There is a map $r_i : S_{i+1} \to S_{i+1} \restriction \{<_{i+1}\}$ defined by:

$$r_i(\mathfrak{A}) = \mathfrak{A} \restriction \{<_{i+1}\}$$

If $\mathfrak{A} \in S_{i+1} \restriction \{<_{i+1}\}$, $|\mathfrak{A}| = \{a_1, \ldots, a_n\}$ (assume that elements of $|\mathfrak{A}|$ are ordered in the way shown above) then define:

$$\pi_h^*(\mathfrak{A}) = \pi_h(a_1) \cdot \pi_h(a_2) \cdot \ldots \cdot \pi_h(a_n)$$

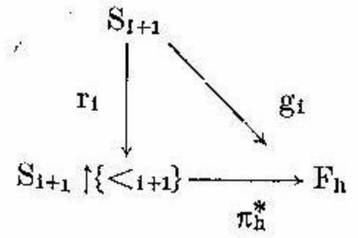
In this way we obtain a map:

$$\pi_h^* : (S_{i+1} \restriction \{<_{i+1}\}) \to F_h$$

If we define

$$g_i(\mathfrak{A}) = \pi_h^*(r_i(\mathfrak{A})) \text{ for all } \mathfrak{A} \in S_{i+1}$$

then the following diagram commutes:



The function g_i induces some interesting relations in S_{i+1} and in $S_i/_h$:

1. for $\mathfrak{A}, \mathfrak{B} \in S_{i+1}$ define

$$\mathfrak{A} \sim g_i \mathfrak{B} \text{ if and only if } g_i(\mathfrak{A}) = g_i(\mathfrak{B})$$

Then $\sim g_i$ is an equivalence relation on S_{i+1}

2. Let $a \in S_i/_h$. Define:

$$D(a) = \{b \in S_i/_h : c, d \in F_h(c \cdot a \cdot d \in g_i[S_{i+1}] \equiv c \cdot b \cdot d \in g_i[S_{i+1}])$$

Then the family $\{D(a) : a \in S_i/_h\}$ is a partition of $S_i/_h$. This partition corresponds to the distribution partition investigated in algebraic linguistics (see for example Marcus 1967).

Example 4.
Assume that some fixed linguistic analysis distinguishes levels of words and sentences. Let sentences be treated as linear sequences of words. Finally, let h correspond to the relation of homophony between words. Equivalence classes of the relation h are exactly the objects of the vocabulary. The function g_i is a formal counterpart of a procedure of forming (written) sentences.

We conclude this paper with some general remarks. It is easy to show how

the traditional distinction between syntagmatic and paradigmatic relations is reflected in our model. To see this assume that some fixed language level is considered. Let S_i corresponds to this level. Every relational structure $\mathfrak{A} \in S_i$ is associated with some analyzed text from the level under consideration. It is easy to see (ex. 1) that relations in this structure (realizations of predicates from Ω_i) correspond to syntagmatic relations between appropriate language entities.

We constructed the function g_i with the help of the relation h, defined without the use of predicates from Ω_{i+1}. Relations of this type (determined on S_i and defined without use of predicates from Ω_{i+1}) correspond to paradigmatic relations which hold between texts from the i-th language level.

If some linguistic analysis investigates a sequence of paradigmatic relations on i-th language level, then we can extend our model by adding to it a "paradigmatic part". Namely, we can treat S_i as a relational structure:

$$\langle S_i, R_2, \ldots, R_{n_i} \rangle$$

where R_j are relations on S_i corresponding to investigated paradigmatic relations.

REFERENCES

Chang, C. C. and J. H. Keisler. 1973. *Model theory*. Amsterdam-London-New York: North Holland Publishing Company.

Hjelmslev, L. 1953. *Prologomena to a theory of language*. Bloomington: Indiana University Press.

Kuratowski, K. and A. Mostowski. 1976. *Set theory*. Warszawa: PWN.

Marcus, S. 1967. *Algebraic linguistics, analytical models*. New York:. Academic Press.

Szrejder, J. A. 1975. *Równość, podobieństwo, porządek*. Warszawa: Wydawnictwo Naukowo-Techniczne.