Petr Jirků

Towards an integrated theory of formal and natural languages

Kybernetika, Vol. 11 (1975), No. 2, (91)--100

Persistent URL: `http://dml.cz/dmlcz/124412`

KYBERNETIKA — VOLUME 11 (1975), NUMBER 2

Towards an Integrated Theory of Formal and Natural Languages

PETR JIRKŮ

A number of works dealing with the problem of semantics of natural languages, discussed from the point of view of mathematics, has appeared in recent years. The purpose of this article is to summarize some main results and to show reasons for constructing such a mathematical theory of language that simultaneously captures the most important internal properties of formal languages and fragments of natural ones. The most interesting results concerning this problem can be found in Montague [9] which will be also our main reference.

The author is deeply indebted to dr. P. Sgall and all his colleagues from the group of algebraic linguistics of the Centre of numerical mathematics (Charles Univ., Prague) for their most valuable suggestions and comments on this work.

1. INTRODUCTION

Mathematization

There are two aims of linguistic investigations in the area of syntax. General rules for forming sentences of some language ought to be constructed first, and, at the same time, a general theory of formal structure of language must be created independent on specific properties of certain languages. Both these problems have long been studied thoroughly in the case of formal languages by using mathematical means and methods, and many interesting facts discovered. On the other hand many difficulties occur when we deal with natural languages. Here the problem arises whether mathematical methods are adequate for studying such complicated objects as natural languages are. If we use mathematical means to describe a language, we must be aware of a necessary idealization of natural language. A considerable variability of natural language leads to an impossibility of a fully adequate description of all its possible aspects. Thus we shall deal with a certain abstraction from natural language and all our results concern this abstract mathematical model of language.

Such an approach to a study of a language as an ideal object is typical for mathematical linguistics. Since algebraic methods are predominantly used, our considerations that follow belong to the domain of algebraic linguistics.

Grammar

One of fundamental notions in mathematical linguistics is the notion of grammar. One thing about grammar: it is a method how to put right words into right places. This characterization of grammar is, of course, very vague, but suggests and idea that grammar has to be a device generating, if possible all, well-formed sentences of a language in question. Nevertheless, there is no common agreement in the conception of a well-formed sentence of a concrete natural language, for example of English.

According to Chomsky, a grammar of a language is a system of rules that expresses the correspondence between sounds and meanings in this language. But it is known that with the approach of Chomsky and his followers the theoretical problems of semantics still remain open for discussion. We would like to devote the present paper to some comments concerning the effort to solve these problems on the base of formal logic, where questions of semantics have been treated in a systematic way. Having in mind that we deal with a language in the written form in our paper, we can modify Chomsky's definition in the following way: A grammar is a system of rules that states how to delimit well-formed sentences from a set of all finite sequences of symbols of a given language not only in the syntactic (usual in formal languages), but also in the semantic respects.

2. PRELIMINARIES

Notation

Unless it is specified otherwise, we essentially follow the notation of Montague as well as his basic terminology. It is a standard terminology and notation commonly used in the theory of sets and in general algebra. In particular, we shall use small Greek letters $\alpha, \beta, \gamma, \ldots$ to refer to ordinal numbers. Thus the symbol $\langle a_\xi \rangle_{\xi < \alpha}$ denotes an α-place sequence of elements of the form a_ξ, $\mathscr{P}(X)$ denotes the power set of a set X. Special symbols will be explained in the text.

Basic algebraic notions and theorems

An algebra is a system $\mathscr{A} = \langle A, F_i \rangle_{i \in \Gamma}$, where A is a non-empty set, Γ is an index set of any sort, and each F_i is an operation on A. Two algebras are similar if they have the same number of same-place operations. Two similar algebras are homomorphic if there is at least one homomorphism between them. A homomorphism is such a mapping between two similar algebras that preserves operations. A notion of algebra is equivalent to the notion of a closure operator. A mapping $C : \mathscr{P}(A) \rightarrow \mathscr{P}(A)$ is a closure operator on A if it is reflexive, monotonous, transitive and compact. A set $X \subseteq A$ is called a generator set of $\mathscr{A}$ if $C(X) = A$. A generator set X of $\mathscr{A}$ is a free base of $\mathscr{A}$ if for every similar algebra $\mathscr{B} = \langle B, G_i \rangle_{i \in \Gamma}$ there is a homomorphism from A into B extending every mapping of the generator set X into B. An algebra is called free if it has at least one free base. It is a well-known fact that for every set X there is a free algebra of every given type.

3. INTERPRETATION OF MONTAGUE'S THEORY OF UNIVERSAL GRAMMAR

Definition of language, syntax

In this section we reproduce the fundamental definitions of syntactic notions which concern mathematical idealization of natural language. First we introduce the more simple notion of disambiguated language. A disambiguated language is to be understood as a system $L_d = \langle \mathscr{A}, X_\delta, S, \delta_0 \rangle_{\delta \in \Delta}$ such that $\mathscr{A} = \langle A, F_i \rangle_{i \in \Gamma}$ is a free algebra of expressions, X_δ are subsets of expressions determining categories of basic expressions indexed by some index set Δ, S is a set of syntactic rules and δ_0 is an index of the category of basic declarative sentences. Operations in the algebra of expressions are structural operations and a domain A of this algebra forms a set of all proper expressions, i.e. all the expressions that we obtain by repeated application of structural syntactic rules. Thus, the free algebra of expressions is generated by the union of all categories of basic expressions. A triplet $s = \langle F_i, \langle \delta_\alpha \rangle_{\alpha < \beta}, \delta_s \rangle$ is a syntactic rule of the level δ_s if F_i is an operation of algebra $\mathscr{A}$ and δ_α, δ_s are elements of the index set Δ. A disambiguated language L_d generates a class of syntactic categories SC provided that it is the smallest class of subsets of algebra $\mathscr{A}$ such that a) for each index $\delta : X_\delta \subseteq SC_\delta$, where X_δ is a category of basis expressions, b) each syntactic rule s leads to the syntactic category of its level. It is not difficult to see that every disambiguated language generates exactly one class of syntactic categories. Categories of basic expressions of a disambiguated language, roughly speaking, correspond with parts of speech.

A language in general is defined as a pair $L = \langle L_d, R \rangle$ where L_d is a disambiguated language defined above and R is a binary relation with its domain included in A. This relation characterizes an ambiguity and a synonymy in a language.

An interpretation for the language $L = \langle L_d, R\rangle$ is a system $I = \langle B, G_i, f\rangle_{i\in\Gamma}$, where $\mathscr{B} = \langle B, G_i\rangle_{i\in\Gamma}$ is an algebra similar to the algebra of expressions $\mathscr{A} = \langle A, F_i\rangle_{i\in\Gamma}$ of the language L and f is a function that maps the union of all categories of basic expressions X_δ into B. Thus, a domain of the algebra $\mathscr{B}$ represents a set of all meanings that may be prescribed by the interpretation I, G_i are semantic operations corresponding with structural operations F_i and the function f assigns meanings to basic expressions. A meaning assignment for language L determined by the interpretation I is the unique homomorphism from $\mathscr{A}$ to $\mathscr{B}$ extending the mapping f.

Now we are able to distinguish meaningful expressions from the set of all finite sequences of symbols of an alphabet. A finite sequence of symbols is a meaningful expression provided it is a member of the union of all categories of expressions.

Some fundamental linguistic definitions

Notions of weak and strong synonymy can be also defined by using previous definitions. The meaningful expressions a, a' are strongly synonymous in L with respect to the interpretation I if for every $\delta \in \Delta$ the set of homomorphic images of those expressions that are in the relation R with a equals to the set of homomorphic images of those expressions that are in the relation R with a'. The meaningful expressions a, a' are weakly synonymous in L with respect to the interpretation I if sets of their meanings equal, i.e. sets of elements of algebra $\mathscr{B}$ that are prescribed by I to a and a'.

Let I be an interpretation for L, I' interpretation for L'. The expressions a, a' are interlinguistically synonymous if they are meaningful expressions in L and L', respectively, and they both have the same sets of meanings in I and I', respectively.

It is obvious that all previous notions are essentially syntactic. Especially an interpretation of a language says nothing about extralinguistic objects, it only distributes meanings among well-formed expressions. Only when some additional conditions are satisfied, an interpretation of language can be treated as a device generating referentional relations, i.e. relations among expressions and meanings in the sense of extralinguistic objects.

4. THEORY OF SENSE AND THEORY OF REFERENCE

The main prupose of this section is to study a special class of possible interpretations of a language in the sense mentioned above. It will be a class of semantic interpretations usually understood as relational structures. We therefore suppose that terms used in a language refer to non-linguistic entities which are called extensions. These extensions are ontological entities built in an agreement with the formal structure of a language, i.e. individual expressions (constants) refer to individual entitites of ontology, one-place predicates refer to classes of entities, two-place predicates to binary relations, etc. These extensions then play the role of an argument which forms a base for decision about truth-values of statements. This requirement will be exactly expressed in the definition of Fregean interpretations.

However, a notion of intension of an expression is closely connected with the notion of extension. Since a semantic theory of language is traditionally constructed on an extensional basis, our aim is to show that there are good reasons to construct semantics on an intensional basis, i.e. as a theory of sense.

Denotation

First, following Montague's explanation, we define a class of possible semantic types. Let e, t, s be mutually distinct atomic types. The class T of all possible semantic types is the smallest class such that e (type of individual objects) and t (type of truth-values) are elements of T; for every two types σ, τ the ordered pair $\langle\sigma, \tau\rangle$ (type of function from objects of type σ into objects of type τ) is an element of T; for every type τ an ordered pair $\langle s, \tau\rangle$ (type of a sense corresponding to object of type τ) is an element of T.

Let E and $\mathcal{I}$ be two disjoint non-empty sets. The set E is intended as a set of ontological entities, the set $\mathcal{I}$ as a set of possible worlds. If both these sets are fixed, we can define for every semantic type τ a set of possible denotates D_τ by the following:

$$\begin{aligned} D_\tau &= E && \text{for} \quad \tau = e\,, \\ D_\tau &= \{0, 1\} && \text{for} \quad \tau = t\,, \\ D_\tau &= D_\varrho^{D_\sigma} && \text{for} \quad \tau = \langle\sigma, \varrho\rangle\,, \\ D_\tau &= D_\varrho^{I} && \text{for} \quad \tau = \langle s, \varrho\rangle\,. \end{aligned}$$

Note. More precisely we should write $D_{\tau,E,\mathcal{I}}$ to register relation to given ontological entities and to the given set of intensions.

Now, for every semantic type τ we define an abstract set of possible meanings of an expression using a specific set J which will be interpreted as a set of contexts of use. For the time being we suppose nothing about the set J. Some comment will be given later. Let J be a non-empty set disjoint with E and $\mathcal{I}$, then the set of possible meanings of type τ is $M_\tau = D_\tau^{\mathcal{I} \times J}$.

Note. Analogously to what has been said above, we should write precisely $M_{\tau,E,\mathcal{I},J}$.

The set $M = \bigcup_{\tau \in T} M_\tau$ is called the set of all possible meanings connected with the language L.

Let δ_0 be an index – a specific element of the set Δ. A set $X_{\delta_0} \subseteq A$ (where A is the domain of an algebra of expressions) is called a basic set of declarative sentences.

Hence, we can define a type assignment function of language L as a mapping $\Sigma : \Delta \rightarrow T$ such that $\Sigma(\delta_0) = t$. Therefore, we require only that this function assigns to every basic declarative sentence the semantic type belonging to truth-values.

These definitions being given, a definition of Fregean interpretation can be formulated. An interpretation $I = \langle B, G_i, f \rangle_{i \in \Gamma}$ of language L is a Fregean interpretation if

1) $B \subseteq M$ (i.e. the domain of interpretation is a subset of the set of all possible meanings),
2) for every $\delta \in \Delta$ and $a \in X_\delta$, $f(a) \in M_{\Sigma(\delta)}$,
3) if $\langle F_i, \langle \delta_\xi \rangle, \delta \rangle$ is a syntactic rule of language $L$and $b_\xi \in M_{\Sigma(\delta_\xi)}$, then $G_i(\langle G_\xi \rangle) \in$ $\in M_{\Sigma(\delta)}$.

This definition guarantees that Fregean interpretations satisfy our requirement so that extensions must correspond with formal structure of language, i.e. they form usual relational structure.

Consequence operation

We want to conclude our considerations about the theory of reference by the definition of semantic model. It is of course, necessary to create a general method which can give necessary and sufficient truth conditions for declarative sentences of a given language. Therefore, the notion of semantic consequence operation must be formulated without dependence on specific properties of language and adequately to the direct syntactic method of proof. Such a consequence operation can be defined as a special closure operator (see [6]). Let A be a non-empty set (e.g. a set of formulas); a mapping $C : \mathscr{P}(A) \rightarrow \mathscr{P}(A)$ is a consequence operation on A if

$$X \subseteq C(X),$$

$$X \subseteq Y \rightarrow C(X) \subseteq C(Y),$$

$$C(C(X)) \subseteq C(X)$$

for every set $X, Y \subseteq A$.

These three conditions are common for both consequence operations, i.e. for syntactic and semantic ones. Moreover, the syntactic consequence satisfies a condition

$$C(X) = \bigcup_{\substack{Y \subseteq X \\ \mathrm{Fin}(Y)}} C(Y),$$

where Fin (Y) denotes that Y is a finite set. A subset X of a set of formulas is called closed if $C(X) = X$. Then we can define conversely a consequence operation using a system $H \subseteq \mathscr{P}(A)$ such that it includes a set H and with every class of sets belonging to H it includes their intersection, or a set of all such closed sets that include a set X.

This definition of the consequence operation permits an immediate comparison of two consequence operations from the point of view of inferential strength. It is not difficult to show that a class of all consequence operations on a given set with a partial ordering defined by inferential strength is a distributive lattice with zero and unit elements. This, of course, holds for an arbitrary set of formulas.

Let Mod be a class of abstract relational structures intended as models of given formula (or theory) in language L. Supposing that extensions of compounds are only functions of their atoms (this is certainly correct for any formal language), we could consider a class of meaning assignments of a language L determined by model $\mathcal{M} \in \mathrm{Mod}$, i.e.

$$\mathcal{S}_{\mathcal{M}} = \{\Sigma_{\mathcal{M}};\ \mathcal{M} \in \mathrm{Mod}\}\ , \quad \mathcal{S}_{0,\mathcal{M}} = \{\Sigma_{0,\mathcal{M}};\ \Sigma_{0,\mathcal{M}} = \Sigma_{\mathcal{M}}/X_{\delta_0}\ \&\ \mathcal{M} \in \mathrm{Mod}\}\ .$$

Then $\mathcal{S}_{0,\mathcal{M}}$ is a class of mappings from declarative sentences into $\{0, 1\}$, where $\Sigma_{\mathcal{M}}/X_{\delta_0}$ denotes the mapping $\Sigma_{\mathcal{M}}$ partialized on the set X_{δ_0}. We define $\mathcal{S} = \bigcup_{\mathrm{Mod}} \mathcal{S}_{\mathcal{M}}$, $\mathcal{S}_0 = \bigcup_{\mathrm{Mod}} \mathcal{S}_{0,\mathcal{M}}$. The system $\mathcal{S}_0$ then forms a base for indicating consequence relations among declarative sentences. This follows immediately from the theorem of mutual definability, see [7].

Model, points of reference

We proceed now to formulate the notion of model for natural languages. Here we must keep in mind one very important internal property of natural languages (which is accepted by majority of linguists), namely that it is not possible to assign truth-values to the overwhelming majority of grammatical sentences of a certain natural language without specifying some additional delimitations which are not usually explicitly expressed. This fact may be explained by means of the following example. Let us consider the English sentence:

"The wind blows"

Such a sentence cannot be true or false unless we specify at least two coordinates of this event, namely its time and place. After that specification this sentence will be informationally equivalent to a more complete assertion, i.e. "The wind is blowing now in Prague". Here the time specification means the time-point at which the sentence is pronounced.

It is clear that more complicated examples can be introduced and a high variety of coordinates must be investigated. Some authors distinguish in addition to time and place such coordinates as for example speaker coordinates (to account for such sentences as "I am a teacher"), audience coordinates (to account for sentences as "You are a teacher"), indicated – objects coordinates (to account for sentences as "He is a teacher", "Those things are big"), etc. From these examples we can see that truth-values of declarative sentences must be relative to certain delimitations as it was mentioned above. Dana Scott, who originated this approach, calls a sequence of coordinates, required for the assignment of a truth-value to a given statement, an index. An extensive discussion about the variety of possible coordinates is presented in Ö. Dahl [4]. This treatment has been accepted by many linguists, since it makes it possible to account for deictic or indexical elements in a way deeper than that of Chomskyan grammars.

In his recent work Montague [9], developing some of these ideas, has separated those coordinates which delimit context of use from other necessary specifications. A very instructive example of this idea can be cited from Dahl: "What an expression such as *the leader of the British Labour Party* or in general any description of the form "the individual that has the property F" refers to depends on two things. First, one must supply a point in time t such that the individual has the property F at t. Second, one must know something about what the world is like at t, namely who has the property F at t. In other words, the extension of the description depends on what world it is uttered in. The meaning of the description, in Montague's use of the term, then, is what is constant in the description regardless of the context of use and the world. The sense, on the other hand, is that which we obtain if we in addition to the meaning specify a context of use. For example, *the leader of the British Labour Party*, used in 1972, will have the same sense as the more complete description *the leader of the British Labour Party in 1972*, although their meanings differ." Now it is evident that meanings are functions of two arguments, i.e. they are mappings from possible coordinates and contexts of use into extralinguistic objects. Moreover, in a more detailed analysis, possible coordinates may be regarded as n-ary vectors. Here, components of those vectors are all coordinates which must be necessarily specified for actual delimination of a meaning. It is obvious that such a treatment of possible coordinates expresses an old Leibniz's idea of possible and actual wordls, an idea which has later influenced conceptions of intensional semantics so significantly.

According to previous considerations the product $\mathscr{I} \times J$ will be called a set of points of reference. A model $\mathscr{M}$ for language L is then $\mathscr{M} = \langle \mathscr{B}, \langle i, j \rangle \rangle$, where $\mathscr{B}$ is a Fregean interpretation for L and $\langle i, j \rangle \in \mathscr{I} \times J$ is a point of reference for $\mathscr{B}$. Evidently, i is here associated with actual world, j with actual context of use. A denotation function for language L (determined by model $\mathscr{M} = \langle \mathscr{B}, \langle i, j \rangle \rangle$) is such a function h that maps expressions into meanings so that $h(\xi) = g_{\langle i,j \rangle}(\xi)$. Here g is a meaning assignment function for language L and interpretation $\mathscr{B}$.

Now, having a certain consequence operatin, we can define when a sentence φ of language L is true. A sentence $\varphi \in L$ is true with respect to model $\mathscr{M} = \langle \mathscr{B}, \langle i, j \rangle \rangle$ if and only if there exists a declarative sentence ψ in L_d such that $\psi \; R \; \varphi$ and $h(\psi) = 1$, where R is the ambiguity relation in L and h is a meaning assignment determined by the model $\mathscr{M}$.

We conclude this paragraph by a definition of a logically true sentence, which is based on a class of logically possible models of L. A class Mod of logically possible models for L is

$$\mathrm{Mod} = \{\mathcal{M};\ \mathcal{M} = \langle \mathcal{B}, \langle i, j\rangle\rangle \,\&\, I(\mathcal{B})\}\ ,$$

where $I(\mathcal{B})$ means that $\mathcal{B}$ is such an interpretation that respects the formal structure of language Lin the previous sense. Now a sentence φ is logically true if it is true in all logically possible models.

These last definitions make sure that many standard useful notions of mathematical logic can be reformulated into the language of Montague's universal grammar. However, the theory of universal grammar is not the only theoretical framework for an intensionalization of semantics.

Another formalization

In [10] P. Tichý* constructs a formal system based on Church's calculus of λ-conversion, which uses (analogously to Montague's system) a set of semantic types called here class of type symbols. The class of type symbols is the smallest class containing 1) atomic types: o (type of truth-values), ι (type of individuals) and μ (type of possible worlds); 2) for every two types α, β an ordered pair $(\alpha\beta)$ (the type of a mapping from objects of type β into objects of type α). It is not difficult to see that we can establish a correspondence among semantic types in Tichý's and Montague's conception and we can show that there is no essential difference in semantic considerations with both of these systems.

5. CONCLUSIONS

It follows from the preceding paragraphs that it is possible to construct a theory which integrates fundamental properties of formal languages and of fragments of natural ones, and, at the same time, such a theory can be regarded as a framework for an adequate explication of basic semantic notions – sense and meaning. Although many authors have tried their best in this domain, only relatively small fragments of natural language were described satisfactorily. Thanks to the discussions, referred to in the present paper, it seems that if we deal with declarative sentences only, no considerable difficulties occur. But it is not known how to apply results obtained above to sentences of other types, namely to interrogative, exclamatory, imperative, etc. Thus we suggest that the value of theoretical study presented here should be found rather in an exact formulation of vague linguistic notions. At present, of course, we

* This article was recommended to me by dr. P. Materna.

cannot exclude that further investigations will show an inadequacy of those mathematical means used here for more comprehensive fragments of natural language. On the other hand, the author is sure that the reason for mathematization of study of a natural language is just in the fact that it plays an important explanatory role.

(Received March 25, 1974.)

REFERENCES

[1] Yehoshua Bar-Hillel: Language. In: Scientific Thought, Mouton/Unesco, The Hague 1972, 107—128.

[2] Noam Chomsky: Syntactic Structures. Mouton, The Hague 1957.

[3] Noam Chomsky: Studies on Semantics in Generative Grammar. Mouton, The Hague 1972.

[4] Östen Dahl: On Points of Reference. Logical grammar reports of Univ. of Göteborg, 1972.

[5] Gottlob Frege: Über Sinn und Bedeutung. Zeitschrift für Philosophie und philosphische Kritik *100* (1892), 25—50.

[6] N. Jardine and C. J. Jardine: Model Theoretic Semantics and Natural Language. Presented on conference on formal semantics in Cambridge, 1973.

[7] Petr Jirků: Theory of Logical Consequence. Acta Universitatis Carolinae, in press.

[8] Richard Montague: Pragmatics and Intensional Logic. Synthese *22* (1970), 68—94.

[9] Richard Montague: Universal Grammar. Theoria *XXXVI* (1970), 373—398.

[10] Pavel Tichý: An Approach to Intensional Analysis. NOÚS *V* (1971). No. 3.

Dr. Petr Jirků; Matematické středisko biologických ústavů ČSAV (Mathematical Centre of Biology — Czechoslovak Academy of Sciences), Budějovická 1083, 142 20 Praha 4, Czechoslovakia.