PUC

Series: Monographs in Computer Science
and Computer Applications

Nº 7/70

TRANSFORMATIONAL GRAMMARS AS MODELS FOR NATURAL LANGUAGES

by

Sueli Mendes dos Santos

Computer Science Department - Rio Datacenter

CENTRO TÉCNICO CIENTÍFICO
Pontifícia Universidade Católica do Rio de Janeiro
Rua Marquês de São Vicente, 209 — ZC-20
Rio de Janeiro — Brasil

TRANSFORMATIONAL GRAMMARS AS MODELS FOR NATURAL LANGUAGES

It has been a concern for some linguistic theorists the elaboration of formal models apt to represent natural languages, i.e., the linguistic intuition of a native speaker-hearer of a natural language. The natural world presents the existence of structured languages, and one of the linguistis' intellectual pursuits has been the attempt at constructing formal analogues able to generate the structural descriptions of these languages. As the mathematical logicians intend to establish formal models, either axiomatized or not, powerful and valid enough to generate the mathematical structures, so the linguistic theorists would, according to the same persuasion, try to develop formal analogues to the natural languages.

The logicians have, however, the advantage of a consensus as to the raw material they must master, i.e., the mathematical corpus in itself. There is no such a consensus among the linguistic theorists. The very preliminary question as to how to acquire the data they will work with defines a methodological problem.

Given the absence of a satisfactory technique for gathering information on a native speaker's linguistic intuition. Chomsky has taken the

position that the linguistics must assume their capacity to grasp this intuition as a factual premise. Therefore, the linguistics must postulate the existence of an ideal native speaker whose learning processes and competence are then the topics to be explanined by the formal model.[1]

It is only natural that controversial questions had given place to controversial solutions, and Chomsky's position is by no means universally accepted. However the solution one adopts, there is still a further question as to the criteria any model has to meet to qualify as an adequate formal representation of a natural language. It is this last question that we will be concerned with in the following exercise.

Let us begin by clarifying what shall be understood. in this paper, by model and by linguistic theory. Formal models are generative grammars for natural languages. A linguistics theory for natural languages, on the other hand, formulates several possible generative grammars, that is to say, models.[2] The first set of criteria we have to deal with, then, refers to the requirements any linguistic theory must been to be considered an adequate representation of natural languages.

First, a linguistic theory has to be descriptively adequate, i.e., it must specify a collection of possible grammars for the language. In addition, it has to make available for each natural language at least one des-

criptively adequate grammar. Second, it has to have explanatory adequacy, that is to say, it has to establish criteria for selecting among the possible grammars the descriptively adequate one, on the basis of primary linguistic data. A linguistic theory satisfying the first criterion is called a descriptive linguistic theory; the fulfillment of the second criterion qualifyies a linguistic theory as an explanatory one.

Any descriptive linguistic theory can be further qualified according to its generative capacity. The generative capacity of a linguistic theory can be either weak or strong. A descriptive linguistic theory of a natural language has a weak generative capacity when its models or grammars are apt to generate the set of sentences belonging to that language. It has a strong generative capacity when, in addition, its models or grammars are apt to generate the structural descriptions of the sentences.

Turning now to the set of criteria that the grammars must satisfy to be an adequate model of a natural language, it can be said that, from a descriptive point of view, such a set is constituted precisely by those two capacities already mentioned, that is to say, adequate grammars must have a weak as well as a strong generative capacity.

On the other hand, any linguistic theory has an explanatory adequacy when it provides a measure for evaluating the grammars stemming from

it. This criterion however is not easy of being met. The primary linguistic data on the basis of which this evaluation would be performed is questionable. But in addition to it, there still is the requirement that the grammars generated by the linguistic theory have to be disposed is such way to allow a clear discrimination among them in relation to the degree of adequacy they posses.

These are some of the criteria presented by Chomsky in relation to the problem of formulating an adequate linguistic theory.[3] In order to elaborate the problem a step further, we will bring in another important criterion that any grammar must satisfy, if it is to be a model of a natural language. We mean the criterion of recursiveness, to the discussion of which we now turn.

The criterion of recursiveness requires of the whole set of sentences generated by the grammar to be recursive. We say that a set is recursive when it is possible by means of a mechanical device to enumerate all the elements of the set, and all the elements of its complements. A technical definition of 'enumeration' is 'to establish a one-to-one correspondence between the elements of the set, and a subset of the natural numbers'.

The ground for such a criterion becomes apparent when one recalls the very justification for the intellectual attempts at formulating models for natural languages. Such models intend to be models of the com-

Petence of native speakers. When we talk about the competence of an ideal native speaker, we are talking about someone who is able to, given any 'utterance', tell whether or not such an 'utterance' belongs to his native language. The supposition in here is that the native speaker has internalized as it were, one algorithm that makes him able to distinguish the set of well-formed sentences form its complement. In correspondence with this capacity of a native speaker, what the grammars must do is to make that algorithm explicit.

To meet all of those criteria is far form an easy question for any grammar. Most of the grammars that have been proposed satisfy one of another type of criterion, but not all of them. Let us take, for instance, the unrestricted rewriting systems. These systems are extremely powerful, being able to generate an arbitrary number of natural or artificial languages. It has been proved, that such systems are equivalent to the class of partial recursive functions, i.e., the class of Turing Machines. Therefore, these systems are recursively enumerable, but not recursive. This means that they have power in excess.

The context-free grammar constitute another class of examples. They have been proved to be recursive but they do not have any generative capacity, not even the weak generative capacity. It is accepted, in general, tha the phase-structure grammar formulated to this date do not have the capacity to generate the phrase structures of the sentences of the natural languages. Whence the

necessity, which gave birth to the formulation of transformational grammars, of formulating another type of grammars,with more generative power. But before discussing the requirement of recursiveness in relation to the transformational grammars, let us make clearer the distinction between these and the phrase-structure grammars.

Definition: A phrase-structure grammar is a four-tuple

$G = \langle V, W\ P, \sigma \rangle$ where:

1 - V is an alphabet;

2 - W is a subset of V;

3 - P is a finite set of ordered pairs (u, v) where $u \in ((V-W)^* - \{\varepsilon\})^4$ and $v \in V$

4 - $\sigma \in V-W$

The elements of V-W are the non-terminal variables, and the elements of W are the terminal variables. The ordered pairs of P are the rewriting rules, and are denoted by $u \rightarrow v$. The elements of W are the primitive terms of the language. The elements of V-W are the synthatical categories of the language. In P is the set of rules that will permit the derivation of the sentences of the languages from the grammatical classes. "σ" represents the synthatical class of 'sentences'.

In general, phrase-structure grammars set up a correspondence between phrase-structures and sentences of the language, but they are subjected to some of the following restriction.

a)- Rules derived from some phrase-structure grammars are of the string-replacement kind, that is to say, they allow the replacement of the symbols to the left of the arrow in (1) if and only if the actual symbols occur in the conclusion of a derivation;

b)- They are, in addition, expansion rules, that is, the string on the right of the arrow has to have a lenght equal to or greater than the string on the left. The string to the left of the arrow, however , may consist of more than one symbol. The string to the right of the arrow may consist of one or more elements but must consist of at least one symbol. Finally, the rewritten element may not be empty.

c)- The rules permit only one replacement at each time.

d)- No rules of the form u → u concatenated to v or u → v concatenated to u are allowed.

e)- No rules having the effect of permutation of elements are allowed.

The fact that these restrictions prevent deletions and permutations makes possible the reconstruction of the history of the derivation of a sentence. In addition, the restriction upon the lenght of the derived expression, i.e., the requirement that the lenght of the derived expression has always

to be equal to or greater than the lenght of the premise makes it easy to prove that the set of sentences of a language generated by a context-free grammar, and also by a context-sensitive grammar, is recursive.

We can sketch the proof that a context-sensitive grammar is recursive in the following way: Given the element v we know, by the lenght restriction, that there are only finitely many ways of deriving v. Then, we can set up a mechanical device that screen all possible derivations of strings of symbols of lenght equal to of greater than v, effectively checking whether v is among the derived strings of symbols.

These restrictions show why this type of phrase-structure grammars, although able to generate a recursive set of sentences, does not posses descriptive generative capacity to the extent required of a model for natural languages. Hence the necessity of formulating a different type of grammar.

The most basic distinction between a phrase-structure grammar and a transformational grammar consists in the fact that, while the former sets up correspondences between phrase markers and sentences of the language , the latter establishes correspondences between phrase markers and phrase markers. Furthermore, the restrictions placed upon the rules of a phrase-structure grammar do not hold true in relation to the rules of a transformational grammar. Rules of a transformational grammar do allow for the deletion of terms, for the permutation of terms, for conjunctions, and for the embedding of strings

of expressions one into another.

It is this resilience of the rules of the transformational grammars what gives them their generative strenght. The derivation of new sentences, on the other hand, becomes more economical within the framework of a transformational grammar, to the extent that on the basis of a more restricted set of sentences it becomes possible to derive an unlimited number of sentences. The full extension of this feature of the transformational grammars, is understood when one recalls that, by using only rewriting rules, one would have to derive independently each one of the sentences, no matter how much similar or correlated they were to each other.

Once one has the notion of transformational grammar it becomes possible to reformulate the definition of generative grammar for natural languages. The new conception of a generative grammar will be the con - ception of a transformational grammar, which will have two main components: the base component - the rewriting rules; and the transformational components - the transformational rules.

Considering now that the transformational grammar, as defined above, will be posit as a model for natural languages, a precise formulation of its rules becomes necessary, if we are to verify whether this model satisfy the adequacy criteria previously defined.

Supposing that the rules of transformational were given a formulation in such a way that they would have an unlimited flexibility, then they would also have a generative capacity probably as powerful as the unrestricted rewriting systems. As a consequence, it would seem correct to say that the set U of phrase structures generated by the transformational grammar, in likely formulations, is not recursive.

In an attempt at overcoming the problem of vagueness of the usual characterizations of the transformational grammars, Chomsky proposes several restrictions upon the transformational rules, in order to make them more precise and simpler.[5] As a lateral consequence of Chomsky's main interest, the restrictions originally devised as a means of reducing the vagueness of the transformational rules may be considered as a promising starting point in the direction of proving that the set U is recursive. We now turn to the discussion of this possibility.

A first important restriction established by Comsky refers to the elimination of the quantifiers in the formulation of the conditions for applying the transformational rules. This rectriction makes it possible to establish, in terms of Boolean rules, the criteria determining the domain of the transformations. The point is higly important because, were the transformational rules to use quantifiers essentially, it would be possible to apply Church's undecidability result to prove that the set U is not recursive.

A second restriction establishes that the only deletions allowed are those making possible the recuperation of the deleted element. The possibility of remaking backwards the derivation becomes then more feasible, being in this way tentatively incorporated, by the transformational grammars, an important characteristic of the phrase-structure grammars.

A third restriction states that rules permitting permutations of elements are to be dropped, and replaced by other kinds of transformations such as substitutions, deletions, and adjunctions.

This set of restrictions aims specially at blocking the formation of non-grammatical sentences by appying the transformational rules. For instance, the deletion of an element that cannot be recovered, or the irrestrict use of the relativization, may lead to some non-legitimate transformations, such as the well know example of the phrase- "John hurts John" - being transformed into - "John hurts himself" - , which is a legitimate transformation, and "the boy hurts the boy" being transformed into "the boy hurts himself", which is not.

However, making sure that those restrictions had been established and that they are sufficient to prove that deviant sentences cannot be derived from the rules, does not warrant that a device for deciding whether a given sentence is deviant or not exists. The transformational grammars with this set of restrictions, can be called, although not in a technical sense,

"consistent", but they cannot be called recursive. The question whether or not we need more restrictions, and of what kind, remains therefore open.

Another important modification in the transformational grammars suggested by Chomsky is the incorporation of context restrictions by transformational grammars.[6] This modification leads to the use of context-free languages as component basis of the transformational grammars. It was precisely this type of transformational grammars, using context-free rewriting rules as component basis, and a set of restricted transformational rules, that have been proved to be equivalent to the family of recursively enumerable sets, by Seymour Ginsburg and Barbara Partee. But remains open the problem of knowing whether by joining a new set of restrictions to the this type of transformational grammar it would meet the criteria of recursiviness.

NOTES

1 - Chomsky, N., Aspects of the Theory of Syntax, p. 19

2- In this paper we will be concerned only with the synthatical components of the grammar, not with the semantical or the phonological aspects of it.

3- Chomsky, op, cit., p. 18-37.

4- Where ϵ is the empty string, and * is a denotation for the set of words over an alphabet V.

5- Chomsky, op. cit., p. 128-147

6- Chomsky, op. cit., p. 139

7- A Mathematical Model of Transformational Grammars - Ginsburg. Seymor, Partee, Barbara-System Development Corporation - Santa Monica - California report nº 21 - TM 38/ 048/ 00

BIBLIOGRAPHY

Bach, E., "An Introduction to Transformational Grammars", Holt Rinehart & Winston, 1964.

Chomsky,N."Synthatic Structures", Mouton, Paris, 1968 (first printing 1957)

Chomsky,N."Aspects of the Theory of Syntax" The MIT Press, 1965.

Chomsky,N."Formal Properties of Grammars", in R. D. Luce, R. Busch, and E. Galanter (eds), Handbook of Mathematical Psychology, vol. II pag. 323-418, Wiley, 1963.

Chomsky,N. and Miller, G. A., "Introduction to the Formal Analysis of Natural Languages", in R. D. Luce, R. Busch, and E. Galanter (eds) op.cit. vol. cit., pp. 269-322.

Ginsburg,S., "The Mathematical Theory of Context Free Languages", McGraw Hill, 1966.

Ginsburg,S., "Lectures on Context Free Languages" , in Arbib, A.M. Algebraic Theory of Machines, Languages, and Semigroups, Academic Press, 1968,

Swanson,J.W., "An Unresolved Problem in Transformational Grammar", in The Journal of Philosophy.