20

1965 International Conference on
Computational Linguistics

SYNTACTIC ANALYSIS IN THE CASE OF HIGHLY INFLECTING LANGUAGES

D. Varga

Computing Centre of the
Hungarian Academy of Sciences

53, Uri u., Budapest I., Hungary

ABSTRACT

The paper discusses the two main methods based on the dependency grammars and on PS grammars used in syntactic analysis of natural languages. In the case of highly inflecting languages the PS analysis has the main disadvantage that they considered syntactically homogeneous categories the number of rules to be applied increases rapidly. The paper proposes the method of partial decomposition into morphemes in order to increase the efficiency of the rewriting rules, so that the problems of rection and agreement can be solved for highly inflecting languages.

1. Language analysis needs an approach to language different from the generation of the sentences of a given language:

1.1. In the case of analysis one has to reckon with the fact that because of the restricted accuracy of the way language data are designated we can often attain our aim /i.e. the establishing of the real structure of the sentence considered/ only after the testing of several alternatives, i.e. it is impossible to solve the raised problems directly, without returns. We have not at our disposal at every stage of the analysis the information that would make a clear-cut decision possible with respect to the path to be followed in the next stages of the analysis. This is why it can be said that analysis depends to some extent on the previous history of the analysis. This requirement, however, does not necessarily lead to the reformulation of the rules but may come to the fore in a new way of their application or their order of application[1]. Of course, it has to be ensured the correctness of the analysis that the correct structure can be obtained by testing in all cases..

1.2. If we are interested in the problem not only from the side of theory but also that of its practical applicability, then we have to ensure the optimalization of the way the correct structure is revealed.The optimalization of analysis is related - in many respects - to the requirement of simplicity in language theory. Of course the method to be applied is not independent of the typological properties of the language under consideration,and this applies, above all, to the optimalization.

1.3. If we aim at the analysis of natural languages our main requirement may be much less stringent than the requirement of generative grammars. Generative grammars, quite reasonably, consider as a principal requirement that any grammar should generate all sentences of a given language and only these. An analogous stipulation is not necessary in the case of analysis since we may assume that we want to analyze only impeccable sentences. /In the case of artificial languages - for instance, in the case of programming languages - the situation is quite different: it is a basic requirement that the analyzator should be able to distinguish the syntactically impeccable strings from the incorrect ones, i.e. that he could disclose the syntactic faults./

Now the question is that what kind of methods or which combination of methods may lead to the recognition of the structure or structures of any syntactically impeccable sentence, within optimal time and with especial regard to highly inflecting languages.

2. With respect to the non-inflecting or only weakly inflecting languages there is a useful method for analysis, namely, the reversed application of the so-called rewriting rules. Besides its

simplicity, this method offers quite a few advantages, firstly, it is based on the mathematically well-formalized phrase structure grammars, secondly, from a linguistic point of view, it is related to the IC grammar that has been elaborated for the analysis of natural languages.

In the case of inflectional languages, however, the application of such rules meets with a difficulty which is due to the fact that the application of such rewriting rules means the processing of symbols assigned to the categories of syntactically homogeneous elements. The number of the categories consisting of such syntactically homogeneous elements is very high in these languages and each additional category increases the number of the rewriting rules by so many rules as there are different structures in which the category in question may occur. The number of rules would amount, for instance, in Russian to about 30-40 thousand, which diminishes the applicability of the system considerably.

The excessive increase of the categories is mostly due to the fact that the classifications according to the different points of view may occur independently of each other. If m different basic categories were needed according to one aspect of classification and n different categories according to another aspect then, taking into account both aspects, $m \cdot n$ different basic categories would be called for. If, for instance, the classification of substantives according to rection needed seven basic categories, the classification according to the cases 6 basic categories, and the classification according to the numbers 2 different categories then - instead of a single substantival /N/ category - $7 \cdot 6 \cdot 2 = 84$ categories would be necessary. It is easy to see that should we take into consideration the differences between male and female, animate and inanimate, let alone the semantic categories, then we would obtain a completely unmanagable apparatus.

3. Dependency grammars have been elaborated mainly to circumvent the difficulties raised by inflectional languages. It is interesting to note in passing that in the Soviet Union this conception prevails even today in the groups engaged in ma-

chine translation. According to dependency grammar we have to consider the category of the distinguished word form as a representative of a complex category in each case a rewriting rule is applied. In this way the concretness of the categories is maintained. Lastly the predicate represents the whole sentence, standing as it does at the top of the tree diagram.

At first glance a dependency grammar seems to exhibit quite a few advantages from the point of view of highly inflecting languages. This advantages may be summarized as follows.

(i) It traces back the relations within the sentence to the relations between concrete word forms. In this way the establishment of the sentence structure is traceable back to the establishment of the relations between concrete words, i.e. to the examination of micro-structures.

(ii) In the case of highly inflecting languages where the relatinns between words come to the fore through their outer form, namely through the form of agreement and rection, the information obtained in this way may be used immediately for finding out the sentence structure.

(iii) On the basis of the direct relations between words the analysis may start at any point: at the top of the tree diagram or at the bottom or in the order given by the words of the sentence.

(iv) No difficulty in principle is encountered in a dependency grammar analysis in the uniform handling of continuous and discontinuous structures. /These structures are rather frequent in highly inflecting languages, due to the fact that they have more effective means at their disposal than word order for expressing relations between words./

In spite of these advantages dependency grammars have not solved the problems definitively as it has turned out that these advantages are only of a rather restricted character.

Ad (i). It may happen that the examination of the relation between two words does not provide enough information for further analysis. The statement of complementary conditions is rather difficult in these cases and can be done most cases only by an ad hoc adjusment.

Ad (iii). Although it is possible to begin the analysis at the top of the dependency tree, such an analysis demands either a rather laborious testing process or the storing of a grat amount of information. / It is illuminating from this point of view to follow the development of predictive analyses beginning with the original conception of Ida Rhodes up to the variant elaborated by Kuno-Oettinger-Plath. According to Rhodes the analysis is to be carried out on the basis of dependency grammar, beginning at the top of the dependency tree. The new version of dependency grammar is based thoroughly on the conception of IC grammars. As is known, the main defect of the earlier version was caused by the fact that when longer sentences were to be analyzed the predictions to be stored increased in an excessive way./

Ad (iv). In principle it would be possible to analyze all possible cases of the discontinuous structures but such a full analysis seems to be unattainable in the forseeable future. / Kulagina's main endavour is aimed at excluding on the basis of a preliminary analysis those constructions that cannot be further expected and making possible this analysis equal to the full analysis[2]/. In practice the analysis is always carried out on the basis of some simplifying conditions or hypotheses concerning chiefly the decomposition of sentences or the relations of some structures /projectivity/.

4. Different methods have been proposed to circumvent the difficulties raised by the IC grammar analysis. Chomsky tackles these proposals /proposals of Harris,Matthews, Stockwell, Anderson, Schachter, Harman and others/ in his paper submitted to the Magdeburg conference; he concludes,"the problem of remedying this defect in PSG is clearly very much open, and deserves much further study" [3]. With respect to Russian it is Plath who has recently elaborated an ingenious indexing and index-transmission system which sets out to ensure the many-sided applicability of the rules and the transmission of the information from one symbol to another. Chomsky points to the fact that the indexing of

categories and the introduction of complex symbols means essentially the application of a special type of transformational grammar. Undoubtedly, the pure methods have not yielded the expected results in the analysis of natural languages. Chomsky himself suggests a compromise with respect to similar difficulties that arise in generative grammars. Practically it goes about a new dimension,neglected so far, namely about the paradigmatic lavel. Chomsky posed the alternative straightforwardly : either one should accept the decomposition into morphemes or opt for the paradigms. He himself pronounced in favour of the paradigmatic conception.

Chomsky has been led to this decision by the complexity of the morphemes. However, it should be added that quite different questions arise in the case of agglutinative languages where the inflectional morphemes generally serve to express a single grammatical function. So, for instance, in Hungarian házaknak =

= ház + ak + nak

house + Pl + Dat

If we take account of this structure of words the decomposition into morphemes seems more justified.

Taking into consideration the aspects of the syntactic analysis an intermediary solution offers itself: with the aid of common rewriting rules /without increasing their number essentially/ a considerable part of the syntactic relations may be detected if we decompose the sentence - but only partially - into morphemes, i.e. if we separate the case category from the basic category. This means that we may use the same symbols for the designation of cases of substantives, adjectives , pronouns etc. and it is necessary to decompose the corresponding categories. On the other hand, the case category is handled separately, the role of which is a syntactic one in the first place. Last but not least it facilitates the separation of case and gender - number which is important in the processing of relative pronouns.

A similar situation can also be produced artificially in the case of the machine translation of nonagglutinative languages. As in machine translation the morphological analysis precedes the syntactic one, in practice there are no difficulties to transform the occuring word forms on the basis of the morphological analysis carried out previously in such a way that the grammatical information becomes explicit and so the word forms are rendered "agglutinative".

To find out the rection we have usually to take into consideration the following factors:

a./ the category of the construing word stem;
b./ the case ending of the construed word;
c./ the category of the construed word stem.

It is, however, unnecessary to consider the case ending of the construing word. E.g.

руководитель кафедрой $N^{instr.}_{nom.} + N_{instr.} \longrightarrow N_{nom.}$
руководителя кафедрой $N^{instr.}_{gen.} + N_{instr.} \longrightarrow N_{gen.}$ (*)
руководителю кафедрой $N^{instr.}_{dat.} + N_{instr.} \longrightarrow N_{dat.}$

By separating the case ending and by placing it before the word have instead of (*) a single rule:

$$N^{instr.} + instr. + N \longrightarrow N$$

The rection can be examined by means of simple context-restricted phrase structure rules:

$$A + \alpha + N \longrightarrow N / \alpha \ \text{——}$$
$$A + \delta + N \longrightarrow N / \delta \ \text{——} \qquad \text{etc.}$$

The decomposition into morphemes can also be used with respect to the participles and the infinitive. Consequently the problems connected with the rection of participle as verbal derivate may be handled separately from the problems connected with the participles as secondary parts of speech being embedded in the structure of the sentence.

5. The advantages of dependency grammars derived from the fact that they could draw conclusions with respect to the type of the relations taking no account of the arrangement of the words in the structure of the sentence only by examining separate concrete words. With respect to some local units the same holds in the case of an IC analysis as well. Such local examinations can be used as input information to further analysis, and on the other hand, they may effect the reduction of the number of the possibilities to be considered.

1. A typical local problem is represented by the morphological analysis which means /in common parlance/ the determination of the grammatical properties of separate words.

2. As local problem may be considered, for instance, the agreement of the substantive with the immediately preceding adjective/s and/or preposition in Russian. /The risk to make a mistake

is minimal, although it is not entirely unlikely because of the adjectives that may be used as substantives too:

В столовой девушке дали обед.

Such preliminary examination of compatibility is of great importance in MT because hereby the number of case homonymies may be reduced essentially.

3. We place the examination of the possibilities of extension or of the realization of these possibilities among the local problems, at least insofar as it provides preliminary information for the analysis. The number of these possibilities is limited and is characteristic of the language under consideration. First, in what direction and second, what kind of grammatical and lexical methods may be used for the extension, the continuation of a word or structure. It is highly revealing to examine how a given structure can be extented starting from a single sentence kernel /i.e. not from several full sentences/. So, for instance, in English:

> Sometimes a decision to compute is followed by a process of selecting the particular kind of computing machine best suited for the given problem.

or

> The designer should be careful in choosing circuit designs that he not build in additional difficulties with a choice of a particular circuit in an attempt to eliminate other difficulties.

The same grammatical relations would be expressed in Hungarian or in Russian in entirely different ways. /We would have full clauses instead of participles in Hungarian, in Russian the participles would be replaced by substantives derived from verbs/.

4. Semantic information may also be used for the reduction of the possibilities in the case of a partial analysis of ambiguous structures. /In case of no ambiguity it makes no sense to use semantic information if we assume that the input sentences are impeccable not only grammatically but also semantically, cf.1.2/.Notice that the constructional homonymy extending over the whole sentence is rather unusual, we have, however, frequent cases of ambiguous structures within sentences. So, for instance, in Russian the string "вследствие других законов сохранения, а также особенностей взаимодействия частиц"

may have 7 different bracketings, i.e. 7 different structures. If there are several syntactically ambiguous structures in the sentence then it would be unnecessary to carry out a new syntactic analysis for each of them: if we can localize the ambiguous structure the production of all possible sentence structures is merely a matter of combination.

The mentioned local problems need not be incorporated into the main program, i.e. the proper syntactic analysis. A considerable part of them may be carried out either previously or simultaneously with the morphological analysis, while other problems may be solved as subsidiary operations, in each case separately, when some rules are applied, if necessary.

6. The crucial point of the syntactic analysis of the whole sentence /i.e. not of the form of the rules, but of the strategy of their application/ is the problem where to begin the analysis, i.e. at which word of the sentence [4]. Lees says with respect to the order in which the transformational rules must be applied, that one has to begin with

the constituent sentence that is embedded deepest and that further transformations can only be applied to matrix sentences previously "satisfied". This holds - mutatis mutandis - with respect to the simplest structures, word groups as well. /Namely, assuming that we begin with the analysis of the given string to be examined,i.e.from the bottom of the tree. The other possibility is to begin from the top of presupposed tree diagram , i.e. with the hierarchy of the given system of rules. This path has been followed in predictive analysis/. A basic problem is the determination of the structure that is embedded deepest in some other structures. If we have succeeded in determining this structure then we could obtain the analysis of rather complicated sentences by a stepwise processing of the embedded structures in a rather simple way. Naturally, if it is wanted that an erroneous step should not destroy the whole analysis the different possibilities must be remembered by the algorithm. A suitable algorithm worked out by Bálint Dömölki[5] could be used with only slight alterations for the analysis meeting the above requirement.

We can considerably diminish the number of the unnecessary blind alleys by taking into consideration the type of the language under consideration. As to Russian, for instance,the right recursive rewriting rules prevail. There is a right recursivity, for instance, in the case of substantival complements connected with substantives,adjectives,participles or the participial constructions embedded in each other etc. According to Yngve's terminology we can say that a considerable part of the Russian structures are of the progressive type.As a consequence, the tree diagram of the sentence is in most cases characterized by right-branching /or at least this holds for some subtrees of most structures/ In this case, however, we arrive at the deepest part of the right-branching tree in the simplest way if we begin the analysis at the end of the sentence. To put it differently, if we consider the sentence structure given by a bracket expression then in the case of progressive languages we have often a case of the brackets accumulating at the end but not at the beginning of the sentence. To take a simple example, we have in Russian such sentences as

(Вы (знаете (много (теорем (о пределах)))))

If we began the analysis at the beginning of the sentence, we should have to try connecting quite a few words and structures that are in fact separated by brackets, that is that are not connected with each other. If we start, however,at the end of the sentence and embed the obtained symbol corresponding to the structure discovered till that moment into subsequent structures we can arrive at the correct analysis of the whole string more quickly and with less effort.

Bibliography

[1] Cf. Abraham, S., Some Questions of Phrase Structure Grammars, Computational Linguistics IV. /forthcoming/

[2] Кулагина, О. С. Использование машин в исследованиях по машинному переводу, Проблемы кибернетики IO, pp. 205-2I3

Вакуловская, Г. В., Кулагина, О. С., Об одном способе анализа текста, Проблемы кибернетитки I2, pp.233-7

[3] Chomsky, N., Categories and Relations in Syntacted Theory, mimeographed, FN 7, 1964

[4] Varga, D., Yngve's Hypothesis and Some Problems of the Mechanical Analysis, Computational Linguistics III, pp.47-74

[5] Dömölki, B., An Algorithm for Syntactic Analysis, Computational Linguistics III, pp.29-46