An Experiment on Phrase Structure Analysis by an Electronic Computer

KIYONO, Takeshi; SAKAGUCHI, Akira

An Experiment on Phrase Structure Analysis by an Electronic Computer

By

Takeshi KIYONO* and Akira SAKAGUCHI*

(Received July 8, 1963)

This paper is a report on the syntactical analysis of the English language by an electronic computer. In this experiment a tree system is used in recognizing the phrase structure of English.

The experimental results by the electronic digital computer KDC-1 are also included in this report.

1. Introduction

Mechanical translation is one of the most important problems in the applications of electronic computer techniques, actively studied and experimented in many parts of the world. A remarkable feature of natural languages, which we usually use as the means for transmitting various informations, is the fact that their system (grammatical structure) cannot be described as a complete system of metalinguistic language. For this reason, it is very difficult to process natural languages using electronic computers. But if, in the future, computers will be more developed and have such advanced functions as man has, the study of mechanical translation may be expected to produce fruitful results.

In general, any language is studied from two points of view: syntax and semantics.

Of course, semantical treatment of languages is necessary in mechanical translation. But at the present stage, we have not yet had methods describing semantical informations quantitatively. And this problem is a very important one which should be thoroughly investigated by the experts in various fields. On the other hand, the syntax of a natural language can be processed mechanically to a fair extent, though it has more redundant structure than certain artificial languages. However, it should be noticed that the syntax and semantics of a natural language are closely related to each other.

* Department of Electronics

In the following sections, one of the methods of the phrase structure analysis and the experimental results are shown.

2. Phrase Structure of the English Sentences

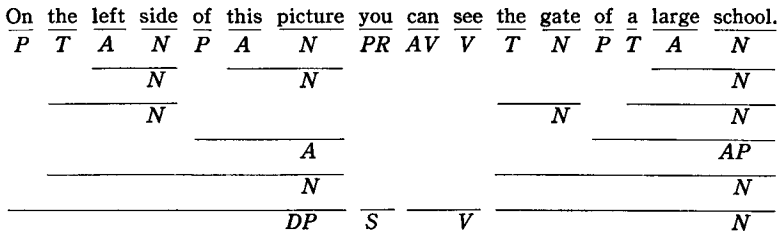
Language translation, in general, requires two processes, analysis and synthesis. As previously mentioned, this paper is concerned with the analysis of the English language. In this section the phrase structure of English is briefiy described.

The fundamental form of declarative sentences in English is shown as follows:

$$S \quad V \quad +$$

S is a noun equivalent which represents the subject of a sentence. The noun equivalent is a noun, a noun phrase, or a noun clause. V is a verb phrase and + represents a complement or object. $V+$ represents a predicate of a sentence. Usually, many adjuncts are attached to the above fundamental form. It is almost always necessary that given sentences which are strings of words be decomposed into the form shown above.

Now, in order to see the structure of an English sentence, we show one typical example below.



As this example shows, an English sentence has a partially recursive structure. However, because of the multiplicity of meanings and functions which a word has, it is very difficult to discover rules for computing the syntactical structure of the English language.

We then choose the following three phrases as the fundamental elements of a sentence:

1) elementary noun phrase [ ]
2) elementary verb phrase { }
3) adjunct phrase ( )

Examples of these phrases are shown below.

[We] {have (just) been using} [an algebra book] (in [a special way])

This experiment is devoted to discriminating these fundamental phrases (in particular, elementary noun phrases).

3. Tree System and Recognition of Elementary Noun Phrases

In this section, we introduce one method of recognizing the fundamental phrases and describe the experiment by our electronic computer, KDC-1.

A sentence of a natural language is a string composed of words. Each word of a sentence belongs to a class or several classes (in other words, a part or parts of speech). It is assumed that a sentence is given in the forms of a string of class marks, and that the class mark of each word is uniquely decided. The class marks used here are shown in Table 1. We then scan a sentence in the right-to-left or in the left-to-right direction to recognize the elementary phrases

Table 1.

Class Marks			
.	N	Noun	2000
,	PR	Pronoun	2500
;	T	Article	5000
:	P	Preposition	5500
?	CC	Co. Conjunction	7000
!	CS	Sub. Conjunction	7500
"	RP	Relative Pron.	8000
„	RD	Rel. Adverb	8500
‘	IP	Int. Pron.	9000
’	ID	Int. Adverb	9500
——	IA	Int. Adjective	9250
—	AV	Aux. Verb	3500
	V	Verb	3000
	I	Interjection	6000
	A	Adjective	4000
	AD	Adverb	4500

mentioned in the preceding section. These are assumed to be recognized in the following order: noun phrases, adjunct phrases, verb phrases.
Elementary noun phrases are marked off with ( ) when they are discriminated.

The general procedure of discriminating the elementary (noun) phrases is as follows: Strings of class marks admitted in any phrases are assumed to be given in the form of a tree system, for instance, as in Fig. 1.

Arrows in this tree indicate the direction of scanning. Class marks in the first column from the right are those permitted to be the last class mark of

the elementary phrase under consideration. Now we assume a string of class marks is given as follows.

$$C_0\, C_1\, C_2 \cdots\cdots C_{n-1}\, C_n \cdots\cdots$$

where C_n $(n=1, 2, 3 \cdots\cdots)$ is a class mark which has replaced a word of a sentence. First, C_n is compared with the class marks in the first column of the tree system. If C_n coincides with any one of these class marks, a closing bracket,), is placed at the right of C_n as follows.

$$\cdots\cdots C_{n-1} C_n) \cdots\cdots$$

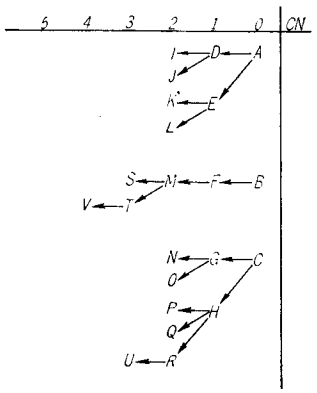


Fig. 1.

C_{n-1} is then compared with the class marks in the second column which are attached to the class mark in the first column corresponding to C_n. A similar procedure is continued until an opening bracket is placed before some class mark. In the above example, if C_{n-1} did not coincide with any class mark of the second column, we would have placed an opening bracket, (, before C_n. Thus it follows that one phrase is recognized, apart from the problem of whether it is correct. Once we enter the tree at the first column we continue to test, following definite branches. If one phrase is discriminated, the same procedure is repeated for the remaining parts of the string.

We experimented on the recognition of elementary noun phrases in the English sentences.

First, we explain the method of producing automatically a tree system by an electronic computer. A tree system is made in a way similar to the method of the recognition of phrases, using many examples of previously recognized strings. The flow chart of the program for producing the tree system is shown in Fig. 2. This program at the same time counts the frequency of occurences of the particular kinds of elementary noun phrases. A tree system which was produced in such a way by our computer is shown in Fig. 3. The connection between class marks or a branch is expressed by the location of the drum memory or the first four digits. These digits represent the location in the memory where there exists the class mark with which the relevant class mark connects. And the next three digits represent the frequencies of the noun phrase which terminates at this point of the tree.

We then made an experiment on the recognition of elementary noun phrases using the tree system produced by the above method. The examples obtained

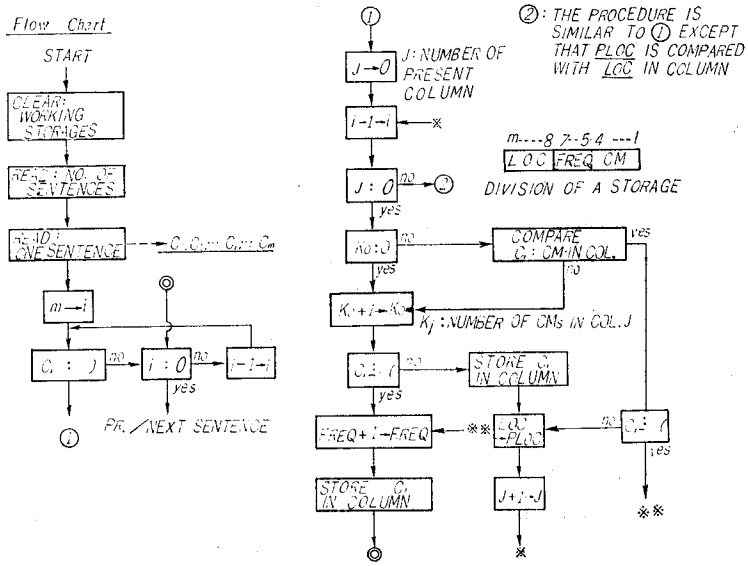


Fig. 2.

δ83010001000 ‖ 4δ

0000 035 2000 ‖
0000 021 2500 ‖
0000 000 4000 ‖

δ83010001100 ‖ 4δ

1000 024 5000 ‖
1000 019 4000 ‖
1002 100 5000 ‖

δ83010001200 ‖ 4δ

1101 002 5000 ‖
1101 002 4500 ‖
1101 001 4000 ‖

δ83010001300 ‖ 4δ

1201 002 5000 ‖

δ83010004240 ‖ 4δ

0003 ‖ 0003 ‖ 0003 ‖ 0001 ‖

TOTAL=0107

Fig. 3(a). This tree was made using 107 noun phrases.

δ83010001000 ‖ 4δ

0000 139 2000 ‖
0000 074 2500 ‖
0000 000 4000 ‖

δ83010001100 ‖ 4δ

1000 080 5000 ‖
1000 083 4000 ‖
1002 006 5000 ‖
1000 004 2000 ‖

δ83010001200 ‖ 4δ

1101 024 5000 ‖
1101 003 4500 ‖
1101 006 4000 ‖
1101 000 7000 ‖
1103 002 4000 ‖
1103 001 5000 ‖
1100 001 4000 ‖

δ83010001300 ‖ 4δ

1201 003 5000 ‖
1203 001 4000 ‖
1200 002 4000 ‖
1202 001 4500 ‖
1204 001 5000 ‖

δ83010004240 ‖ 4δ

0003 ‖ 0004 ‖ 0007 ‖ 0005 ‖

TOTAL=0431

Fig. 3(b). This tree was made using 431 noun phrases.

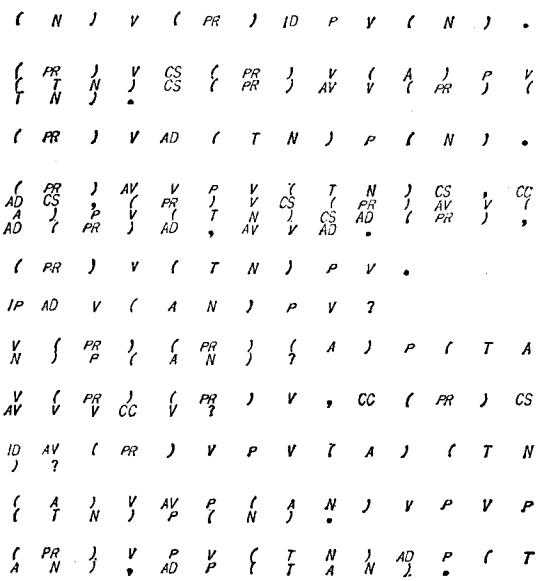


Fig. 4(a). Above examples were produced by use of the tree in Fig. 3(a).

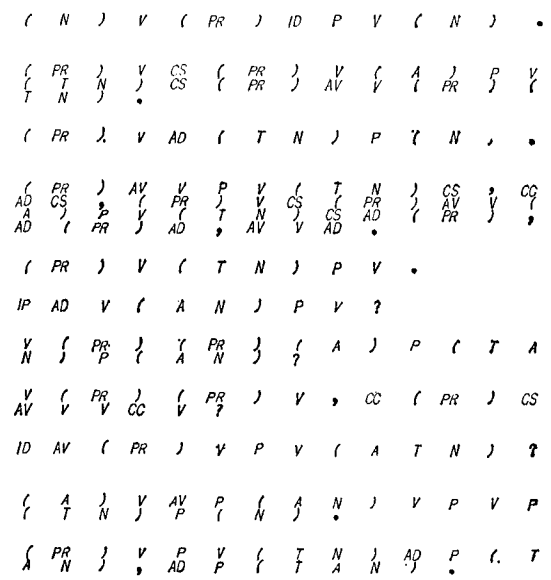


Fig. 4(b). Above examples were produced by use of the tree in Fig. 3(b).

in this experiment are shown in Fig. 4 and the original sentences of them are shown at the end of this report. The flow chart of the program for executing this experiment is omitted as it is similar to the flow chart in Fig. 2.

4. Discussion

As shown by the examples, we could correctly recognize the elementary noun phrase in most cases, although the method itself is very simple. However, it is a very important fact that the class mark of each word in a sentence is assumed to be uniquely decided. Therefore the problem of the unique decision is difficult but rather interesting for us. This problem appears at various levels of the analysis. In some cases we may be forced to treat a language semantically at the same time.

Now we enumerate a few problems arising in the above analysis:

1) (A) is mistakenly recognized as the noun phrase. This is because (T+A) was included in the examples used for producing the tree system.
2) Multiple functions of a word
3) Treatment of the N---N N sequence

The ambiguity caused by these problems cannot be eliminated only by knowing local informations.

Acknowledgement

The authors wish to express their acknowledgement to Mr. T. Tsuda for his discussions.

Also they would like to thank to the members of Kyoto Univ. Computing Room for allowing them the use of the KDC-1.

Reference

1) A.K. Joshi : Computation of Syntactic Structure, Chap. 32, Part 2, Information Retrieval & Machine Translation, Vol. 3, Advances in Documentation and Library Science.

1) Mathematics teaches us how to solve puzzles.
2) Everyone knows that it is easy to do a puzzle if someone has told you the answer.
3) That is simply a test of memory.
4) You can claim to be a mathematician if, and only if, you feel that you will be able to solve a puzzle that neither you, nor anyone else, has studied before.
5) That is the test of reasoning.
6) What exactly is this power of reasoning?
7) Is it something separate from the other powers of our minds?
8) Is it something fixed, or something that can be trained and encouraged?
9) How do we come to possess such a power?
10) Mathematical reasoning does at first sight seem to be in a class by itself.
11) It seems to find a place neither in the experimental sciences, nor in the creative arts.

(from MATHEMATICIAN'S DELIGHT by W. W. SAWYER)