DOCUMENT RESUME

ED 037 699 AL 001 139

AUTHOR	Tamati, Tuneo; Kurihara, Tosihiko
TITLE	A Semantic Approach to the Automatic Analysis of Japanese and English.
PUB DATE	Mar 68
NOTE	21p.; Paper given at the U.S.-Japan Joint Seminar on Computational Linguistics held at Honolulu, Hawaii, March 1968
EDRS PRICE	EDRS Price MF-$0.25 HC-$1.15
DESCRIPTORS	Algorithms, *Computational Linguistics, *English, *Japanese, *Machine Translation, Morphology (Languages), Phonology, Phrase Structure, *Semantics, Syntax, Transformation Generative Grammar
IDENTIFIERS	Dependency Grammar

ABSTRACT

In order to mechanize the processing of natural language, the linguist must make the machine interpret the meaning, or semantic content of the language, in some way or other. This means that the machine should extract not only syntactic but also semantic information from the source sentence through the analysis of it. In this paper, the authors describe their opinions as to the structure of language from the viewpoint of mechanical processing. Then two methods of introducing semantic information into the analysis process are proposed, followed by a brief description of the authors' method of English sentence analysis. Finally, reference is made to the system of machine translation. An example of Japanese sentence analysis is appended. (Author/FWB)

A SEMANTIC APPROACH TO THE AUTOMATIC ANALYSIS OF JAPANESE AND ENGLISH

by

TUNEO TAMATI and TOSIHIKO KURIHARA

Faculty of Engineering

Kyusyu University

Fukuoka, Japan

For presentation at the US-Japan Joint Seminar on CL

to be held at Honolulu, Hawaii

March 25-26, 1968

1. Introduction

In order to mechanize the processing of natural language, we have to make the machine interpret the meaning, or semantic content of the language, in some way or other. This means that the machine should extract not only syntactic but also semantic information from the source sentence through the analysis of it.

In this paper, we will describe our opinions as to the structure of language from the viewpoint of mechanical processing. Then we will propose two methods of introducing semantic information into the analysis process. Thirdly, we will give a brief description on our method of English sentence analysis. Lastly, we will refer to the system of MT. An example of Japanese sentence analysis is appended at the end.

2. Structure of language from the viewpoint of MT

The algorithm of sentence analysis and synthesis largely depends on distinctive features of language we take into account. The structure of language should be considered at each stratum, such as phonology, morphology, syntax and semantics.

Phrase structure grammar, dependency grammar, transformational grammar, etc. have been proposed so far as tools of language description. In the following, we will discuss that the kernel structure of language is well described by means of dependency grammar, and the internal structure of the syntactic unit by phrase structure rules, and the external, or transformational, structure by transformation grammar.

The syntactic unit, of which dependency grammar is composed, roughly corresponds to "the phrase" in the traditional grammar of English. And, in the case of Japanese, it is composed of two parts;

(a) uninflected parts: noun and pronoun n, stem of verb v,
stem of adjective a, adverb A, etc.

(b) inflectional parts: postposition ("josi") p,
endings of verb and adjective e.

Representing the grammatical functions of the syntactic unit as D(predicate) and M(modifier), they have the following correspondence:

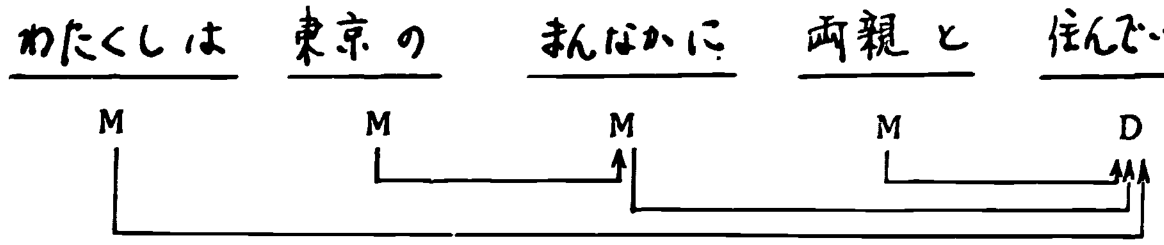
D ⟷ Ve or ae,

M ⟷ np or Ap,

where e and p might be either empty or a string of themselves. Thus syntactic units seem to be well described by phrase structure rules.

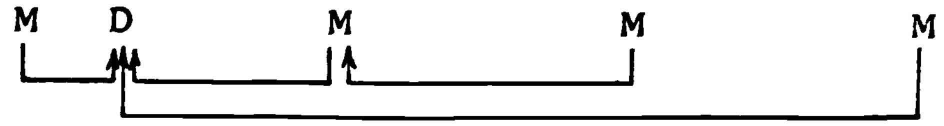
Then we will show the dependency structures of Japanese and English sentences.

(S1) Watakusi-wa Tokyo-no man-naka-ni ryosin-to sundeiru.



(The word order of three M's depending on D is optional.)

(S2) I live in the center of Tokyo with my parents.



Although there seems to be some difference in the language structure between Japanese and English, the dependency structures, or dependent-governor relations among syntactic units, are the same as shown in Fig. 2-1. Such a diagram constructs a D-tree (i.e. dependency tree).

It has been expected that the D-tree of any sentence in one language is almost the same with that in another language, if the sentence intends to convey the same meaning. We believe this is due to the fact that the D-tree has close relations with semantic structure of a language.

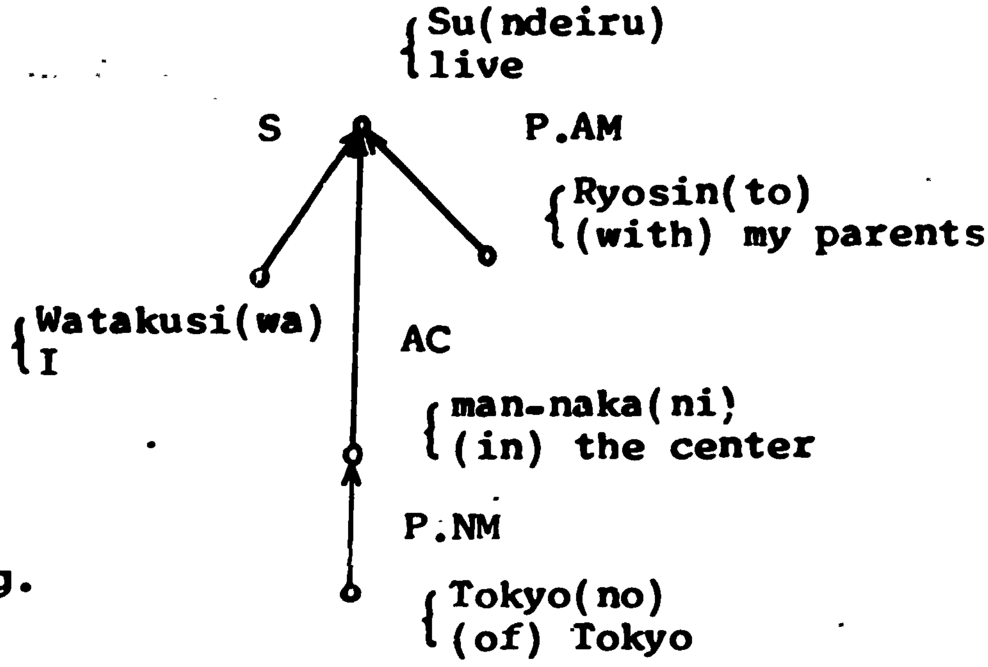


Fig. 2-1 D-tree

Suppose that the meaning of a sentence is represented as functional relations among semantic components of each word in the sentence. Then, a tree shown in Fig. 2-2 corresponding to (S1) and (S2) is supposed to be a description of a semantic structure of the source sentence. This tree is called "S-tree", and D-tree

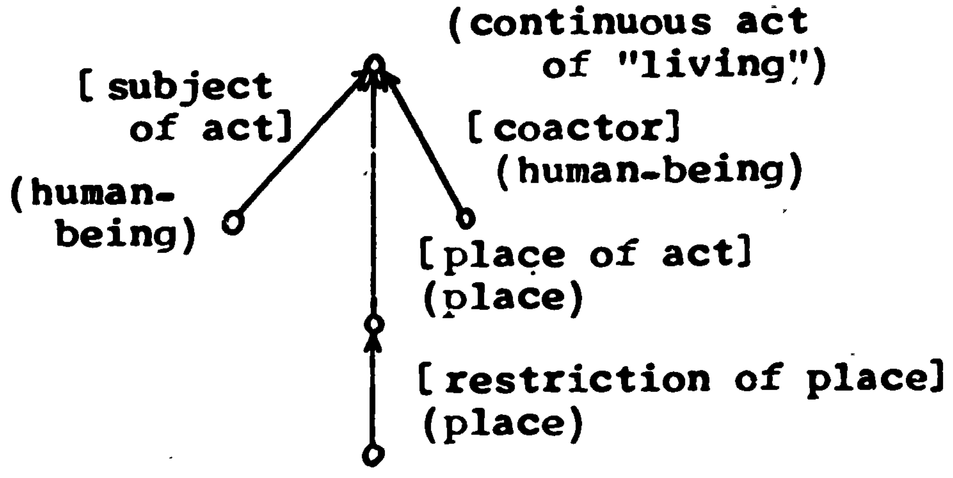


Fig. 2-2 S-tree

might be regarded as an approximation of the S-tree.

By such a dependency grammar we can describe kernel structures, or simple sentences. There remain, however, some problems in dealing with transformational structures, namely complex and compound sentences, by means of such grammars. Therefore we introduce transformation grammar to process such structures.

3. Semantic Categorization

In order to make the machine interpret the meaning of sentences, we must establish a method of representing semantic information in the effectively enumerable form. In this connection we will suggest two methods, terminal categorization and conceptual categorization.

(1) Terminal Categorization

Supposing that N and P stand for syntactic units whose functions are M and D respectively, a sentence can, in general, be represented as a string of them as follows:

NPNNPNPPN.......

Semantic categorization can be done by means of the semantic connectability among these N's and P's.

Suppose we have $\{N_i\}$and $\{P_i\}$ which have such semantic relations as shown in Fig. 3-1. This can also be represented as formulas shown in (F.1). From these formulas we can construct a tree, supposing that P_j that has wider applications is placed above other ones in level. We can construct another tree about

N_i's by considering their connectability with P_j's. Such trees are shown in Fig. 3-2.

As a result, N_i's in the ( ) of (F.1) are represented by the underlined N_i, which is the uppermost concept of N's. Incidentally, connection rules [P , N] include information of syntactic function and mode (degree of necessity and other nuances) of N to P .

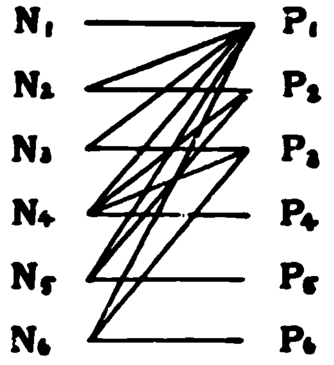
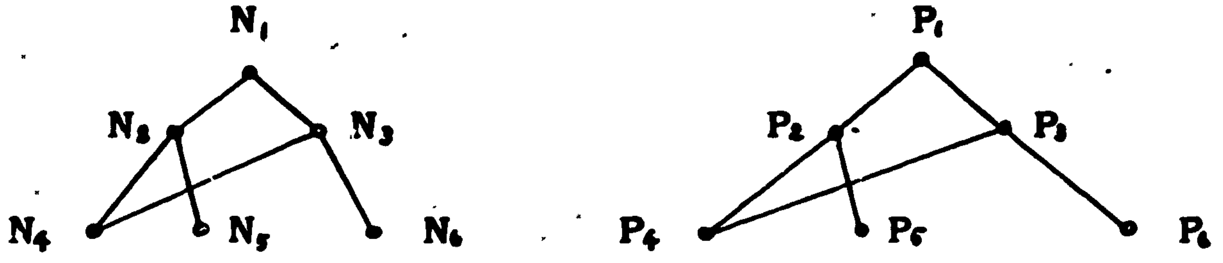


Fig. 3-1

$$\left.\begin{array}{l}
[P_1\,(\underline{N_1}, N_2, N_3, N_4, N_5, N_6)] \\
[P_2\,(\underline{N_2}, N_4, N_5)] \\
[P_3\,(\underline{N_3}, N_4, N_6)] \\
[P_4\,(\underline{N_4})] \\
[P_5\,(\underline{N_5})] \\
[P_6\,(\underline{N_6})]
\end{array}\right\} \cdots \quad (F.1)$$



(a) (b)

Fig. 3-2

(2) Conceptual Categorization

This method is based on the assumption that the meanings of words and sentences might correspond to the human concepts, From this standpoint, the meaning of a word is defined as the combination of semantic components that we assign to the word. It seems that this idea have some resemblance to the Osgood's semantic differentials. For example, the meaning of N_j is represented as follows:

$$[N_j : (C_{\alpha j} C_{\beta j} \cdots\cdots C_{\nu j})]$$

where C_x is a semantic component. If there exists N_j whose meaning is exactly C_j itself, it is represented as $[N_j ;(C_j)]$. In general, one concept corresponds to one of various combinations of C_j's, but possibly there may not exist a word corresponding to the concept. We can construct a tree-like structure of N's, if we regard that N_j which has semantic components common to N_{j_1} and N_{j_2} is the upper concept of N_{j_1} and N_{j_2}.

P_i has the meaning that it establishes a mutual relation among N_i's. This can be represented as follows:

$$[P_i : e_i d_j (C_{\alpha j} C_{\beta j} \cdots\cdots C_{\mu j}) d_k (C_{\alpha k} C_{\beta k} \cdots\cdot)\cdots]$$

where d_j, $d_k \cdots$ are syntactic functions (with mode) by which the succeeding combination of semantic components are connected to P_j, and e_i is the remaining meaning of P_j.

Supposing the semantic components $C_{\alpha j} C_{\beta j} \cdots\cdot C_{\mu j}$ that P_i has for some function d_i are included in N_j, P_i is connectable to N_j and its inferiors in the tree.

Suppose that the meaning of a sentence is defined as semantic connections of concepts of words. Then a formula:

$$N_j P_i \rightarrow [P_{is(j)} ;e_i d_j (C_{\alpha j} C_{\beta j} \cdots C_{\nu j}) d_k (C_{\alpha k} C_{\beta k} \cdots)\cdots] \quad (\nu \geq \mu)$$

denotes a meaning of a sentence. The difference between $P_{is(j)}$ and P_i is due to the semantic components $C_{(\mu+1)j} \sim C_{\nu j}$. Supposing that $C_{\delta j}$, for example, out of $C_{\alpha j} \sim C_{\mu j}$ is not contained in N_j, N_j is not connectable to p_i. But $P_{is(j)}$ makes sense if $C_{\delta j}$ is rejected. This is expanded interpretation of P_i in composing a sentence. Then, such a sentence as;

Neko-ga hon-wo yomu.

(A cat reads a book.)

becomes interpretable. The fewer the common semantic components become, the more nonsense the sentence becomes.

(3) An Example

For reference, we will show an example of our semantic categorization of nouns in Table 3-1. This has been obtained through manual analysis of Japanese texts. Under each numbered head in the table we have extracted about 50 items, which we call "semantic categories". In general, a noun belongs to more than one semantic category, and many nouns might belong to the same semantic category. Therefore those words that belong to the same category can be arranged in arborization.

For predicative words, we can categorize in the similar way. At present we have about 150 items of semantic categories.

human being			
	concrete noun	1.	topological properties of human body and its constituents •physical properties (color, dimensions, etc.) of the constituents •purposes of use of the constituents
	abstract noun	1.	mental function acts
		2.	collected body (both defined hierachically)
animals		0,	species (its topological properties, its constituents) •physical properties of the constituents •purposes of use of the constituents •other definitions
		1.	collected body

plants	0.	species (its topological properties, its constituents) • physical properties of the constituents • purposes of use of the constituents • other definitions
	1.	collected body
products	0.	sorts • topological properties • purposes of the goods of the sort • physical properties
	1.	collected body
natural objects	0.	sorts • topological properties • the nature of the thing of the sort
	1.	collected body
places	0.	topological properties
time	0.	past, present, future, passage of time

Table 3-1 Semantic Categorization of noun

Incidentally, varieties of postpositions, or "josi" in Japanese, are troublesome obstacles to the establishment of semantic and syntactic connections between N's and P's. In this regards we summarized the functions of "josi" which modify predicative words together with nouns into the following eight words:

"ga (が)", "wo (を)", "to (と)", "ni (に)", "de (で)",
"yori (より)", "made (まで)", "missing"

We are making a similar attempt for predicative words in the case they play the role of predicative modifier. These dependency rules might be used also for the case of noun modification.

4. Sentence Analysis

We intend to introduce semantic information into our system of sentence analysis by such methods as mentioned above. Here we will give a brief description of our system taking an example of English sentense:

"We live in the house in the suburbs
which he built last year." .....(F.2)

(1) Word Processing:

At this stage as much linguistic information as possible are extracted from the source sentence by the word-for-word comparison between source sentence and word dictionary. They consist of informations of stem, ending, semantic components, target equivalents, and so on. And idiomatic phrases are marked also at this stage.

(2) Word Group Processing:

Word groups, or syntactic units of dependency rules, are constructed here by immediate constituent rules. As a result information of stem and ending of a word group, that of word order in the target language, and that of transformation are extracted for each word group. For (F.2), we get:

We	live	in the house	in the suburbs
PRN2 {Nom}	VRB {Pres}	NOU1 {Prp}	NOU1 {Prp}

which	he	built	last year.
PRN {Acc}	PRN2 {Nom}	VRB {Past}	ADV1 {/}

(F.3)

(3) Dependency and Transformational Analyses :

At this stage the search is performed for all the alternative dependency connections among word groups in a sentence and the syntactic roles of each word group. For this purpose the analysis process is devided into two parts; one is for the analysis of simple sentence, and the other for the analysis of transformational structure. The former is called Dependency Processing I (abbrev. DPI), and the latter DPII.

At the stage of DPI, the search is started at the beginning of a sentence, and all the alternative dependency connections within the range of kernel structure are looked for. The semantically compatible ones are selected out of them consulting the semantic dictionary called "function pattern dictionary", and they are recorded together with their syntactic roles.

At the stage of DPII, transformational structure is analyzed on the basis of the results obtained so far. Function pattern dictionary is consulted also at this stage.

For a semantic consultation in the course of the analysis of (F.2), for example, such function pattern rules as shown below would be necessary:

```
[live ; (animal) —— (place) (time) ]
           S          AC     P.AM

[build;  (human  —— (thing          (place) (time) ]
          being)     constructed)
           S             O           P.AM    P.AM
```

where semantic components are assigned in ( ), and a symbol attached under semantic components stands for the syntactic function it performs in the function pattern.

From the viewpoint of isolated binary relation, the phrase "in the suburbs" may depend on both phrases "live" and "in the house". But the phrase "in the house" is isolated from the rest, if the phrase "in the suburbs" is decided to depend on "live". The antecedent of "which" is turned out to be "house" at the stage of DPII by consulting the semantic dictionary. Thus we obtain such a D-tree as shown in Fig. 4-1.

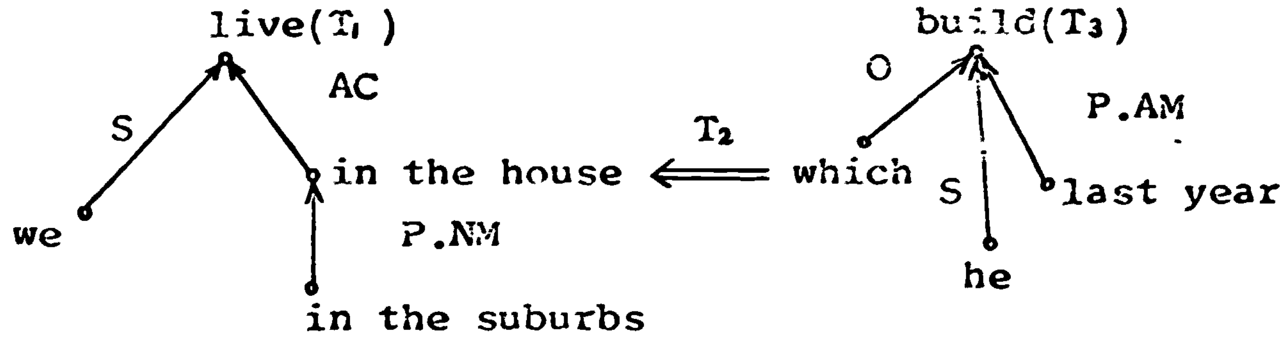


Fig. 4-1 D-tree

(T_i: transformational information)

Analysis procedure for Japanese sentence will be given in the Appendix.

5. Mechanical Translation

It has thus become clear that our D-tree bears all the syntactic and semantic information available as a result of the analysis. The D-tree of a sentence in one language is, in general, somewhat different from that in another language because of the difference in the ways of expression and in the syntax. However we might regard that a target sentence composed on the basis of D-tree of source sentence is the first approximation of the exact translation.

In addition to this, semantic categories or semantic components may probably, we believe, make easier the selection of appropriate target equivalent out of multi-meanings.

Moreover, "Lists of idiomatic expressions" should, we believe, be prepared for the improvement of the approximation by means of adopting the D-tree as an intermediate structure.

Incidentally, we shall show a procedure of our principle of target sentence synthesis for the D-tree shown in Fig. 4-1.

First we transform every node of the tree from English to Japanese. Thus we obtain a tree shown in Fig. 5-1.

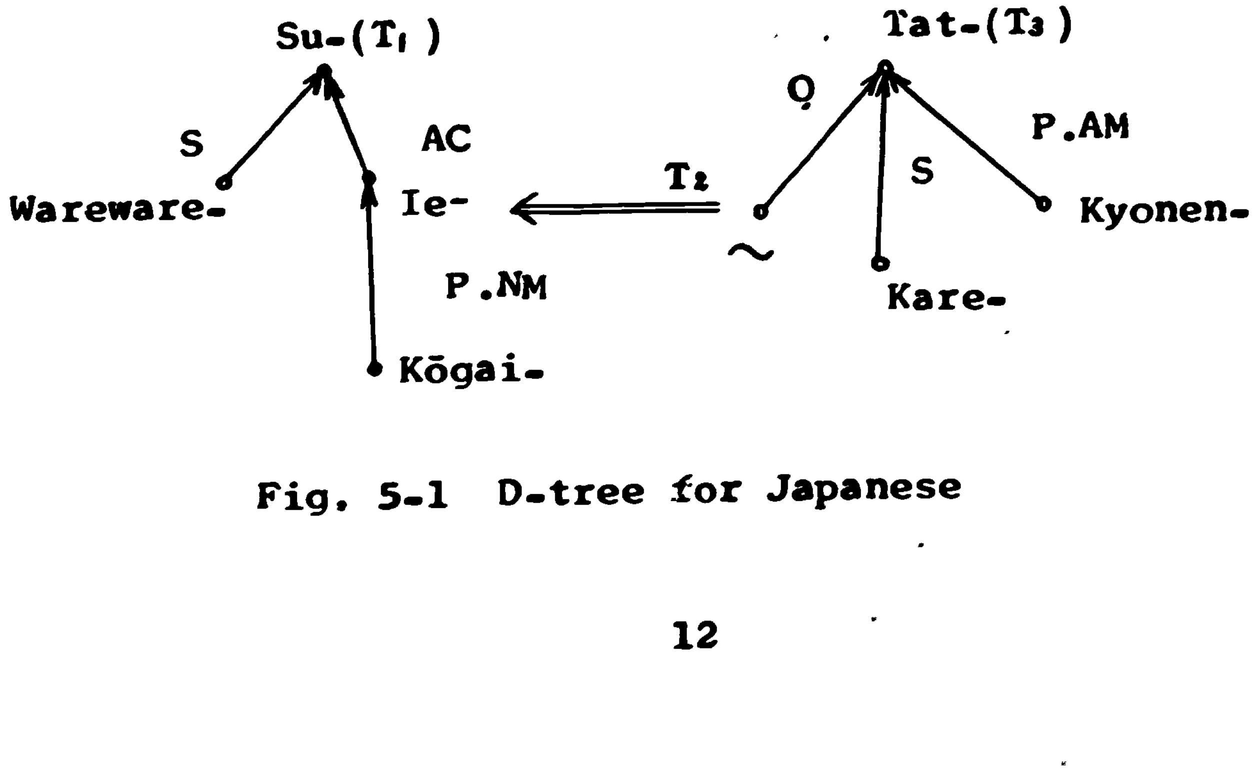


Fig. 5-1 D-tree for Japanese

where $T_1 = \tau_1$(present) τ_2(positive) τ_3(state) τ_4(active)
$T_2 = \tau_5$(sentencial noun modifier)
$T_3 = \tau_1$(past) τ_2(positive) τ_3(__) τ_4(active)
S : subject, O : object,
AC : adverbial complement ·
P.NM : phrasal noun modifier

Then we transform the D-tree into a string, which satisfy the following Japanese conventions:

G_{J_1} : The governor of the tree or the sub-tree is put to the last of its dependents.

G_{J_2} : Word order among dependents of the same governor is optional.

Transformational information T_i is also put into Japanese in the course of synthesis. Thus we obtain:

T_1 {Wareware-(S) T_2{T_3{Kare-(S) Kyonen-(P.NM) ~ (O) Tat-($\not{t}_3$)}
Kōgai-(P.NM) Ie-(AC) Sun-($\not{t}_2$)}}

= T_1{Wareware-wa T_2{#Kare-wa Kyonen Ie-wo Tateta#}
Kōgai-no Ie-ni Sun-($\not{t}_2$)}

= T_1{Wareware-wa # Kare-ga Kyonen Tateta #
Kōgai-no Ie-ni Sun-($\not{t}_2$)}

= Wareware-wa Kare-ga Kyonen Tateta
Kōgai-no Ie-ni Sundeiru.

6. Conclusion

The most difficult problems in mechanizing the processing of a language would be how we could make semantic information computable in a machine, and how we should introduce it into the process of sentence analysis. There would be various levels in realizing the object as shown in Fig. 6-1.

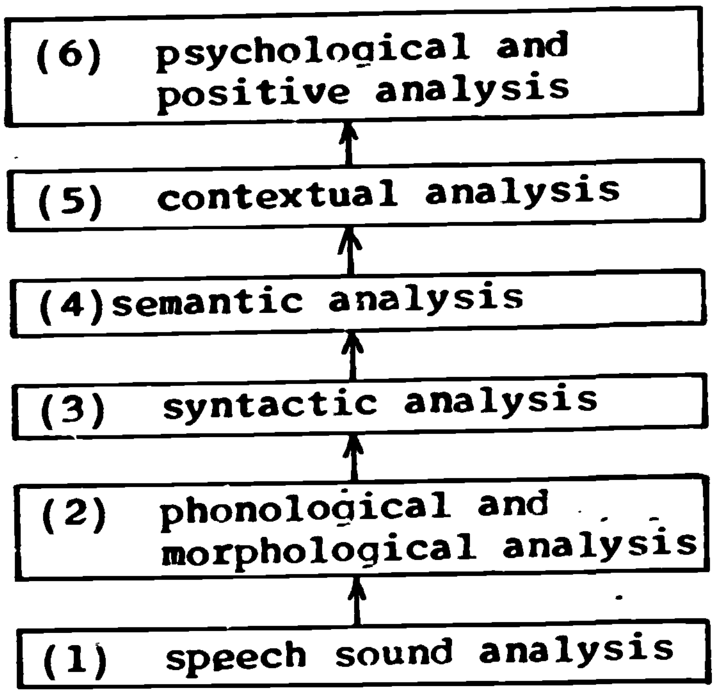

(6) psychological and positive analysis
(5) contextual analysis
(4) semantic analysis
(3) syntactic analysis
(2) phonological and morphological analysis
(1) speech sound analysis

Fig. 6-1 Various levels of sentence analysis

The method we have described so far is an approach at the level of (4), where semantic connections of words is defined within a kernel sentence.

In the similar way, however, we can consider semantic relations between the predicates, or between kernel sentences. Then we can expect to go further to the levels of (5) and (6).

References

Tuneo Tamati; "Application Phrase Structure Language to Mechanical Translation" ("Phrase Structure Language の 機械翻訳への応用 ") in Japanese : Symposium 15-6, The Institute of Electric and Communication Engineers of Japan, April, 1967.

(2) Tosihiko Kurihara and Sho Yoshida; "On the Automatic Analysis of Japanese sentence" (″日本語文の分析について″) in Japanese: A report to the CL research committee of IPSJ, 39p, Jan. 20, 1968.

(3) Yoshihiro Ishihara, Yae Fukushima and Tuneo Tamati; "An experiment of MT from English into Japanese (I) ～ On a method of automatic analysis of English sentences ～" (″英日機械翻訳実験(I)—英語の文章分析の方法について″) in Japanese: Proceedings of the 20th Joint Conference of Kyusyu branches of four Institutes of Electric field. pp191~194. Nov. 1967.

(4) Naoyuki Okada and Tuneo Tamati; "Automatic Extraction of Semantic Information and Classification of Words by Learning Process" (″学習による単語の意味情報抽出および意味分類の一方法″) in Japanese: A report to the IT research committee of IECEJ, 20p, April, 1967.

Appendix A

This appendix describes our procedure of Japanese sentence analysis using an example. This procedure resembles fairly well to that of English sentence analysis notwithstanding the difference in their apparent structures. This may be due to the fact that there would be some underlying features that are common to both languages. But we will not go into any detailed discussion, which was already reported to MT research committee of IPSJ.

The basic procedure is the following:

(I) Reading of a source sentence;

Suppose that the following text is fed into our system with segmentation as shown in the example.

(Example)

Oto ga kūki no sindō de aru koto wa, dokusya syokei no yoku siru tokoro to omou ga, sikasi wareware ga hanasi wo si tari, ongaku wo tanosin dari si te iru oto wa wareware ga riyo suru kūki sindō no goku itibubun ni sika sugi nai.

(The English version of the example, for reference)

The readers, the author believes, know very well that the sound is the vibration of the air. The speech sound which we utter and the sound of music which we enjoy are, however, only the small part of the air vibration which we utilize.

(II) Determination of word class;

Word class is assigned to each word by consulting a word dictionary. The results are shown under the sign "II" in the Table A.1.

(Ⅲ) Word group forming;

A string of words which are syntactically clear to be put together is put together to form a "word group". When such strings of words as the following;

(a) Noun and the succeeding "josi", "jodōsi" (auxiliary verb), and "hojo dōsi" (supplementary verb)

(b) Predicate and the succeeding "josi", "jodōsi", and "hojo dōsi"

become clear to be combined, they are put together and are numbered as shown under the sign"Ⅲ" in the Table A.1.

(IV) Dependency search;

Using semantic and syntactic information, dependencies between word groups are looked for. And all the alternatives are recorded in a Matrix expression as shown in the Table A.2.

(V) Assignment of the sign "I" to mark the word that have nothing to do with any other words in the column of Matrix expression. This is shown under the sign (V) in the Table A.1.

(VI) To mark the section where the dependencies can be decided uniquely;

If only one dependency exists in the row of Matrix expression, the dependency is established immediately. And the section is underlined as shown under the sign (VI) in the Table..A.1.

(VII) Final determination of dependency;

The processing is started at the end of the sentence, from the bottom to the top. The following informations are referred to at this stage:

(a) binary dependency relations at the levels of both syntax and semantics.

(b) n-ary ($n \geqq 3$) dependency relations in the level of only syntax.

(c) crossing of dependency does not take place in Japanese with the following exceptions:

- concord of adverb and verb, such as "ketsite∼ nai" (never).
- some special cases.

As a result we can reject almost all possibilities of wrong connection, and get correct answers enclosed with circles as shown in the Table A.2.

	Oto	ga	kūki	no	sindō	de	aru	koto	wa,	dokusya	syokei	no
(II)	NOU	PCS	NOU	PCS	NOU	VAU	VRB	NFO	PAD	NOU	NPR	PCS
(III)	1		2		3			4		5		
(V)												
(VI)	1		2		3			4		5		

	yoku	siru	tokoro	to	omou	ga,	sikasi	wareware	ga
(II)	ADV	VRB	NFO	PCS	VRB	PCJ	COJ	NPR	PCS
(III)	6				7		8	9	
(V)							|	|	
(VI)	6				7		8	9	

	hanasi	wo	si	tari,	ongaku	wo	tanosin	dari	si	te	iru
(II)	NOU	PCS	VRB	PCJ	NOU	PCS	VRB	PCJ	VRB	PCJ	VSU
(III)	10		11		12		13		14		
(V)											
(VI)	10				11		12				

	oto	wa	wareware	ga	riyō-suru	kūki	sindō	no	
(II)	NOU	PCS	NPR	PCS	VRB	NOU	NOU	PCS	
(III)	15		16		17	18			
(V)									=
(VI)	13		14		15				

	goku	itibubun	ni	sika	sugi	nai.
(II)	ADV	NOU	PCS	PAD	VRB	VAU
(III)	19	20			21	
(V)	|					
(VI)	16					

NOU:	Noun
NPR:	Pronoun
NFO:	Formal noun
VRB:	Verb
VSU:	Supplementary verb
VAU:	Auxiliary verb
ADV:	Adverb
COJ:	Conjunction
PCS:	Case postposition
PAD:	Adverbial postposition
PCJ:	Conjunctive postposition

Table. A.1

		Oto ga	kūki no	sindō dearu	koto wa	dokusya syokei no	yoku siru tokoro to	omou ga	sikasi	wareware ga	hanasi wo	sitari	ongaku wo	tanosin dari	siteiru	oto wa	wareware ga	riyō suru	kūkisindō no	goku	itibubun nisika	suginai
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Oto ga	1			Ø								/			/			/				/
kūki no	2			Ø	/						/				/	/					/	
sindō dearu	3				Ø						/		/			/			/		/	
koto wa	4						Ø	/							/			/				/
dokusya syokei no	5						Ø				/		/	/	/			/			/	
yoku siru tokoro to	6							Ø														
omou ga	7											/		/	/			/				Ø
sikasi	8													/	/		/	/				Ø
wareware ga	9											Ø		/	/			/				/
hanasi wo	10											Ø		/	/			/				
sitari	11													/	Ø							
ongaku wo	12													Ø	/			/				
tanosin dari	13														Ø							
siteiru	14															/	/		/		/	
oto wa	15																	/				Ø
wareware ga	16																	Ø				/
riyō suru	17																		Ø		Ø	
kūkisindō no	18																				Ø	
goku	19																				Ø	
itibubun nisika	20																					Ø
suginai	21																					

Table. A.2