[Proceedings of the *Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Colgate University, Hamilton, New York, August 14-16, 1985]

STRUCTURAL TRANSFORMATION IN THE GENERATION STAGE OF MU JAPANESE TO ENGLISH MACHINE TRANSLATION SYSTEM

MAKOTO NAGAO

Department of Electrical Engineering
Kyoto University
Sakyo, Kyoto, JAPAN

Presented at the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Colgate University, Hamilton, N.Y., August 14-16, 1985.

§ 1. Outline of Mu Machine Translation Project

1.1 Outline of the Project

The Japanese Government initiated the machine translation project in 1982, which aimed at a quick dissemination of scientific and technical information between Japanese and English. The project will end at March 1986.

The project is being advanced by the tight cooperation among the following four organizations. We, at Kyoto University, have the responsibility of developing the software system for the core part of the machine translation (grammar writing system and its execution system), grammar systems for analysis, transfer and synthesis, the detailed specification about the information to be written in the word dictionaries (analysis, transfer and generation dictionaries of all parts of speeches), and the manuals to work with for the construction of these dictionaries. Electrotechnical Laboratories (ETL) have the responsibility for the input and output of the texts for machine translation, morphological analysis and synthesis, and the work of constructing a dictionary of verbs and adjectives according to the working manuals prepared at Kyoto. Japan Information Center for Science and Technology (JICST) is in charge of noun dictionary and the accumulation of special terminological words of the science and technology. Research Information Processing System under the Agency of Engineering Technology is in charge of completing the total machine translation system by assembling the componential results of the other three participating organizations, and by adding man-machine interfaces for editing, updating grammatical rules and dictionary information.

The Japanese texts to be translated are the abstracts of the scientific and technical documents which are produced at JICST as the monthly journal "Current Bibliography on Science and Technology". Only the electronics and electrical engineering fields including computer science are treated at present in our project. The English texts to be translated into Japanese are the abstracts in INSPEC in the same fields. Sentential structures in abstracts are quite complicated compared to the ordinary sentential structures, with long nominal compounds, noun phrase conjunctions, mathematical and physical formulae, embedded long sentences, and so on. The analysis and translation of these sentential structures are far more difficult than the ordinary sentences, but we did not introduce the pre-editing stage, because we wanted to know the ultimate possibility of handling such difficult structures.

1.2 General Principles of Our Machine Translation System

The basic ideas in our system are the followings.(1)

(1) Use every surface and syntactic information. Write as detailed syntactic rules as possible. Develop a grammar writing system which can accept any sophisticated linguistic theories at present and future.

(2) Semantic information is introduced to help syntactic analysis, transfer and synthesis to be as accurate as possible. We aim at a well-balanced usage of syntax and semantics in the whole process of machine translation. Machine translation based on domain dependent semantic network will be fascinating and effective for the sentences from narrowly limited areas. However, the system of this kind can not cope with the complicated situations of wider world whose semantic description is almost impossible in reality. Therefore, we must primarily respect the syntax.

(3) Detailed linguistic phenomena are more like word specific than explainable in the general linguistic theory. Therefore word specific rules are to be accepted in the system. In our system such rules are written in the entry of lexical items, and are used with the priority to the general grammatical rules in all the phases of analysis, transfer and synthesis. This mechanism allows the quality of the system to be improved step by step by the accumulation of linguistic facts and word specific rules in the dictionary, and we can avoid the deadlock of the system's improvement.

(4) The system must have the ability not to fail by the imperfect analysis and unknown words, but to produce translation result even with imperfect sentential structure and unknown original words. This imperfect output will be much better for the post-editor than the system's failure.

1.3 Linguistic Framework

Machine translation process can be largely divided into the analysis, transfer, and synthesis, but we have three more additional stages to cope with the difficulties in the language pairs like Japanese and English, which are shown in Fig. 1. One is the pre-transfer loop, which converts the analyzed structure of a Japanese sentence into much more neutral deep structure representation. Another is the post-transfer loop, which converts the deep structure obtained by the transfer stage to much more appropriate internal representation of the target language. The third one is the structural transformation in the target language during the generation process to obtain a better stylistic expression in such cases as a noun of the "tool" case slot can never come to the subject position in Japanese, while it occurs very often in English. Top-heavy sentence in English will be replaced by the anticipatory subject "it". All these three additional stages are optional, and the system can produce target language sentences without them. But by the machine translation between the languages of completely different language families like Japanese and English, structural transformations at these stages are essential.

At the moment there are about sixty subgrammars for the analysis, and about 900 rewriting rules in total. Sentence generation process is also composed of a subgrammar network. The number of rewriting rules for the transfer and generation processes are about 800, and will be increased still more in the coming few months.

Dictionary contains about 16,000 items at present, and will be increased up to 100,000 items at the end of this project. Among these 36,000 items, verbs and adjectives are about 2000, adverbs are about 400, and the rest are mainly special terminological words. The information to be written in the dictionary entry is different to each part of speech, but generally the following kinds.

head word, number of characters of the word ending, Chinese character part, reading in Kana, variant, derivational words, related words, morphological part of speech, conjugation, prefixal information, area code, syntactic part of speech, subcategorization of part of speech, case patterns, aspect, modal, volition, semantic primitives, thesaurus code, cooccurrence information (adverb, predicative modifier), idiomatic expressions, degrees, degrees of nominality, etc.

Here one of the important information is the case patterns for verbs and nouns. We have distinguished more than 30 cases. Each case slot in a case pattern of a verb usage has the semantic information about the nouns which can come in that slot. The nouns has the corresponding semantic codes in its entry. We have distinguished more than 50 semantic primitives (codes). The noun dictionary also has the information about specific verbs which co-occur with it as shown in Fig. 2. The specific information of these categories is checked prior to the standard rule applications, and the default rules are applied at the last, to get some output which is better than nothing.

§2. Transfer Stage from Japanese to English

2.1 Annotated Dependency Structure

The intermediate representation we adopted as the result of analysis in our machine translation is the annotated dependency structure.(2) Each node has arbitrary number of features, such as part of speech, surface case, deep case, number, tense, semantic codes and so on. This tree representation is powerful and flexible for the sophisticated grammatical and semantic checking, especially when the completeness of semantic analysis is not assured and trial-and-error improvements are required at the transfer and generation stages.

We have three conceptual levels for grammar rules in the transfer and generation phases as well as in the analysis phase. (3)

lowest level: default grammar rules which guarantee the output of the translation process. The quality of the translation is not assured. Rules of this level apply to those inputs for which no higher layer grammar rules are applicable.

kernel level: main grammar rules which choose and generate a target language structure according to semantic relations among constituents which are determined in the analysis stage.

topmost level: heuristic grammar which attempts to get elegant translation for the input. Each rule has heuristic nature in the sense that it is word specific and it is applicable only to some restricted expressions which are found by a strong pattern matching function of GRADE, the grammar writing system of our machine translation system.(4)

The application of these rules are organized along the principle that "a better rule is used earlier."

We use deep case dependency structure as a semantic representation. Theoretically we can assign a unique case dependency structure to each input sentence. In practice, however, analysis phase may find out several alternative structures by the syntactic and semantic ambiguities. Therefore we use as an intermediate representation a structure which makes it possible to annotate multiple possibilities as well as multiple level representation. An example is shown in Fig. 3. Properties at a node is represented as a vector, so that this complex dependency structure is flexible in the sense that different interpretation rules can be applied to the structure.

Besides the ordinary grammatical rules which involve semantic checking functions, the grammar allows the reference to a lexical item in the dictionary. A lexical item contains lexical rules corresponding to its special grammatical usages and idiomatic expressions. During the transfer and generation stages, these rules are activated with the highest priority. This feature of using the lexical rules makes the system very strong and flexible for dealing with exceptional expressions. The improvement of translation quality can be achieved progressively by adding lexical rules as well as linguistic information and word usages in the dictionary entries.

Some heuristic rules are activated just after the standard analysis of a Japanese sentence is finished, to obtain a more neutral (or target language oriented) analyzed structure. We call this stage as the pre-transfer loop. Semantic and pragmatic interpretations are done in the pre-transfer loop. The more the heuristic rules are applied at this stage, the better will be the result. Figs. 4 and 5 are some examples.

§3. Word Selection in Target Language

Word selection in the target language is a big problem in machine translation. There are varieties of choices of translation for a word in the source language. Main principles adopted in our system are,

(1) Field restriction by using field code, such as electrical engineering, nuclear science, medicine, and so on.
(2) Semantic code attached to a word in the analysis phase is used for the selection of a proper target language word or a phrase.
(3) Sentential structure of the vicinity of a word to be translated is effective for the determination of a proper word or a phrase in the target language.

Table 1 shows examples of a part of the verb transfer dictionary. Selection of English verb is done by the semantic categories of nouns related to the verb. In the table, the number i attached to verbs like form-1, produce-2, is the i-th usage of the verb. When the semantic information attached to nouns is not available, the column indicated by Φ is applied to produce a default translation.

In most cases, we can use a fixed format for describing a word selection rule for lexical items. We developed a number of dictionary formats specially designed for the ease of dictionary input by computer-naive expert translators.

The expressive power of format-oriented description is, however, insufficient for a number of common verbs such as "する" (make, do, perform, ...), and "なる" (become, consist of, provide, ...) etc. In such cases, we can represent transfer rules directly by tree structures. An example is shown in Fig. 6. Every usage of a verb is listed up with their corresponding English sentential structures and semantic conditions.

This mechanism of the transfer stage bridges the gap between Japanese and English expressions. There are, however, still many odd structures after this stage, and we have to adjust the English internal representation into more natural ones. We call this part as post-transfer loop. An example is given in Fig. 7, where a Japanese factitive verb, SASERU, is first transferred to a English word "make", and then a structural change is made to eliminate it, to have a simpler and more direct expression. Another example is shown in Fig. 8, where a term corresponding to "the number of" is inserted in between "increase" and "car", because "a car" does not "increase".

Postpositions in Japanese generally express the case slots for verbs. A postposition, however, has different usages, and the determination of English prepositions for each postposition is quite difficult. It also depends on the verb which governs the noun phrase having that postposition.

Table 2 illustrates a part of *a* default table for determining deep and surface case labels when no higher level rule applies. This sort of tables are defined for all case combinations. In this way, we guarantee at least a literal translation to be given to an input. A better choice of a preposition depends on the usage of a verb, so that every usage of a preposition for a particular English verb is written in the lexical entry of the verb, and is used in the selection of preposition.

§4. Determination of Global Sentential Structures in Target Language

Global sentential structures of Japanese and English are quite different, and correspondingly the deep structure of a Japanese sentence is not the same as that of English. Fundamental difference from Japanese internal representation to that of English is absorbed at the (pre-, post-) transfer stages. But at the stage of English generation, some structural transformations are still required in such expressions as embedded sentential structures and complex sentential structures.

We classified four kinds of embedded sentential structures.

(i) A case slot of an embedded sentence is vacant, and the noun modified by the embedded sentence comes to fill in the slot.

(ii) The form like "N が V な N_2" = "(N_2の N_Iが V) なN_2". In this case the noun N, must have the semantic properties like parts, attributes, and action.

(iii) The third and the fourth classes are particular embedded expressions in Japanese, which have connective expressions like "場合 " (in the case of), "方法" (in the way that), "という" (in that), and so on.

An example of the structural transformation is shown in Fig. 9. The relative clause "why ..." is generated after the structural transformation.

Connection of two sentences in the compound and complex sentences is done according to the information in Table 3. An example is given in Fig. 10.

The process of sentence generation in English is as follows. After the transfer is done from the Japanese deep dependency structure to the English one, conversion is done from the English deep dependency structure to a phrase structure tree with all the surface words attached to the tree. The processes explained above are involved at this generation stage. The conversion is performed step by step from the root node of the dependency tree to the leaf as a top-down process. Therefore when a governing verb demands a noun phrase expression or a to-infinitive expression to its dependent phrase which may be a verbal phrase or a noun phrase, a proper structural change of the phrase must be performed. Noun to verb transformation, and noun to adjective transformation are often required due to the difference of expressions in Japanese and English. When we can not find out a verb or a noun corresponding to a noun or a verb respectively in the dictionary, we make the reference to a synonym word and see if it has a verb or a noun derived from the word. An example is shown in Fig. 11. The generation goes down from the root node until all the leaf nodes are converted to a phrase structure tree.

After this process of phrase structure generation, some sentential transformations are performed such as follows.

(i) When an agent is absent, passivization transformation is applied.

(ii) When the agent and object are both missing, the predicative verb is nominalized and placed as the subject, and such verb phrases as "is made", and "is performed" are supplemented. An example is shown in Fig. 12.

(iii) When a subject phrase is a big tree, the anticipatory subject "it" is introduced to avoid the top-heavy structure.

(iv) Pronominalization of the same subject nouns and the change of pronouns including deletion are performed.

(v) Duplication of a head noun in the conjunctive noun phrase is eliminated, such as, "uniform component and non-uniform component" —→ "uniform and non-uniform components".

There are many such structural transformations.

Another big structural transformation required comes from the essential difference between DO-language (English) and BE-language (Japanese). In English the case slots such as tools, cause/reason, and some others come to the subject position very often, while in Japanese such expressions are never used. The transformation of this kind is incorporated in the generation grammar such as shown in Fig. 14, and produces more natural English expressions. This stylistic transformation is very important in machine translation between Japanese and English.

§5. Some Problems Which Require Contextual Information and Knowledge

There are many problems which require contextual information and common sense knowledge for better machine translation. Some of them are pointed out in the followings, which are not yet incorporated into our machine translation system.

(1) Anaphoric and cataphoric expressions are often used in the ordinary texts, although such did not appear in the texts of abstracts of scientific and technical papers which were to be translated in our system. For the anaphoric usage we have a method which determines the referents reasonably well. We stack up in what we call anaphora stack the head nouns of the obligatory case slots of the sentences being analyzed. When a pronoun appears we take out the noun at the top of the stack and make semantic checking of the whole sentential meaning by replacing the pronoun with the noun. That is, we check whether the noun can occupy the slot of the pronoun in the sentence. If not, we take out the noun next to the top of the anaphora stack, and check the replaceability to the pronoun in the same way. This operation is performed until nouns of a few number of previous sentences are referenced. The use of the stack is to check the noun nearer to the pronoun in sequence.

In the case of the pronoun "you", distinction should be made between singular and plural by the analysis of the sentence where "you" appears. In the case of "they" we have to distinguish the semantic codes, human and non-human, from the meaning of the sentence "they" is contained. We can make reference to the verb which governs "they". If the verb has semantic code of human action, such as, think, write, and say, "they" must be human. In the case of "their X" we look for objects whose parts are X, such as "cars, and their wheels". Or, we look for objects which has some relations to X, such as "students, and their answer to a teacher" (students and teacher are related terms).

If a proper noun is not found by these processes, we will check the possibility of the cataphoric usage. But before going on to the cataphoric usage check, we have to see several other possibilities. One is the possibility that a part or the whole of a sentence can be grasped by a concept, and it is referred to by a pronoun. There is sometimes a case when a concept is explained by a series of sentences. In such a case we have to have a semantic network built up by the information of the set of sentences, and whose top node is the concept explained by the set. In this case the topical theme of the set of sentences is generally the concept which is referred to by the pronouns in these sentences, and it must be registered in the top of the anaphora stack. This means that a topical concept of a sentence must be pushed down at the latest, or the concept should come up to the top of the stack as a bubble comes up to the surface of a glass of water.

A pronoun in the cataphoric usage often appears in a sentence which starts a paragraph. In that case our stack for the anaphoric reference is vacant, and we have inevitably to go to check a cataphoric reference.

Anticipatory "it" can be detected by checking the syntactic structure such as "it is ... that ..." in a sentence.

(2) The determination of the function of articles is quite difficult. We have no reliable way of attaching "a", "the", or nothing to a noun in a sentence. As for the particular usage of "the" which refers to an object in the previous phrase like,

... a X .... the Y ... .

we can check whether X = Y, or X and Y have the relation of synonym. If there exists such a relation between X and Y, then we can translate this particular usage of "the" as "SONO" in Japanese. Thesaurus information is very useful in this respect. In the other usages of the definite article, we can just neglect them in the translation into

Japanese. In the translation from Japanese into English, however, we have a great difficulty of inserting definite and indefinite articles in proper place in the generated English.

(3) Ellipsis appears very often in Japanese sentences. In our machine translation system, omitted words are not recovered, but is avoided by using some particular English sentential styles which do not require these omitted words. Some of these mechanisms are explained in the previous section. We can use the case slot information to infer proper words for the omission. A case slot has a set of semantic codes, and a noun with any of the semantic codes is looked for in the anaphora stack. The noun with this condition is regarded as the referent for the ellipsis. However, this process is not so much reliable, and we do not like to use this mechanism very often.

§7. Conclusion

Machine translation does not necessarily require very deep understanding of the sentential meaning as it is believed in the AI circle. To achieve deep understanding we have to have very sophisticated, mechanisms of inference with huge amount of general knowledge of the real world. Still, we can not guarantee the correct inferencing by machine. For example, we have often a phrase like

SHOGAIKOKU	TONO BŌEKIMASATSU	NI	JYORYOKU	SURU
foreign countries	with conflict in trade		help	do

which is a kind of abbreviated expression„ We have to recover a much more precise Japanese expression to get a proper English translation. If the mental attitude of the speaker of the sentence is in favor of the improvement of foreign relation, the insertion of KAISHŌ (resolving) should be done as

(We) help resolve the conflict in trade with foreign countries.

But if the speaker's attitude is quite contrary, the insertion of JYOCHŌ (accelerate) will be done as

(We) help accelerate the conflict in trade with foreign countries.

In this way the interpretation comes out opposite by the speaker's/ receiver's attitude. It is quite difficult to know the speaker's attitude whether he/she is in favor of something or not from the text segments processed by computer.

Anyway, these sophisticated interpretations can often be left to the human being who reads the translated text. If the machine makes excessive inference in the wrong direction, that will be worse than doing nothing. In this sense machine translation system should not do too much extra linguistic inferences. We have to make effort to find out much more solid and reliable linguistic and non-linguistic information from the text itself before going in to the help of general "knowledge", which is vague and hard to be precise enough to be used for machine translation purposes.

References:

1) M. Nagao, J. Tsujii, J. Nakamura: Machine Translation System from Japanese into English: Mu-JE, JACL (forthcoming in 1985).
2) J. Tsujii, J. Nakamura, M. Nagao: Analysis Grammar of Japanese in the Mu-Project, Proc. COLING 84, 1984.7.
3) M. Nagao, T. Nishida, J. Tsujii: Dealing with Incompleteness of Linguistic Knowledge in Language Translation, Proc. COLING 84, 1984. 7.
4) J. Nakamura, J. Tsujii, M. Nagao: Grammar Writing System (GRADE) of Mu-Machine Translation Project and its Characteristics, Proc. COLING 84, 1984.7.

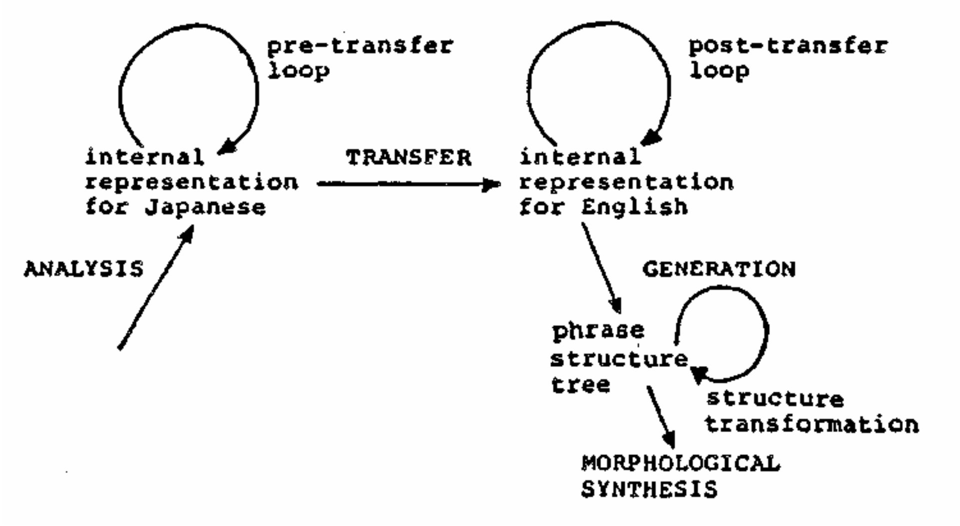
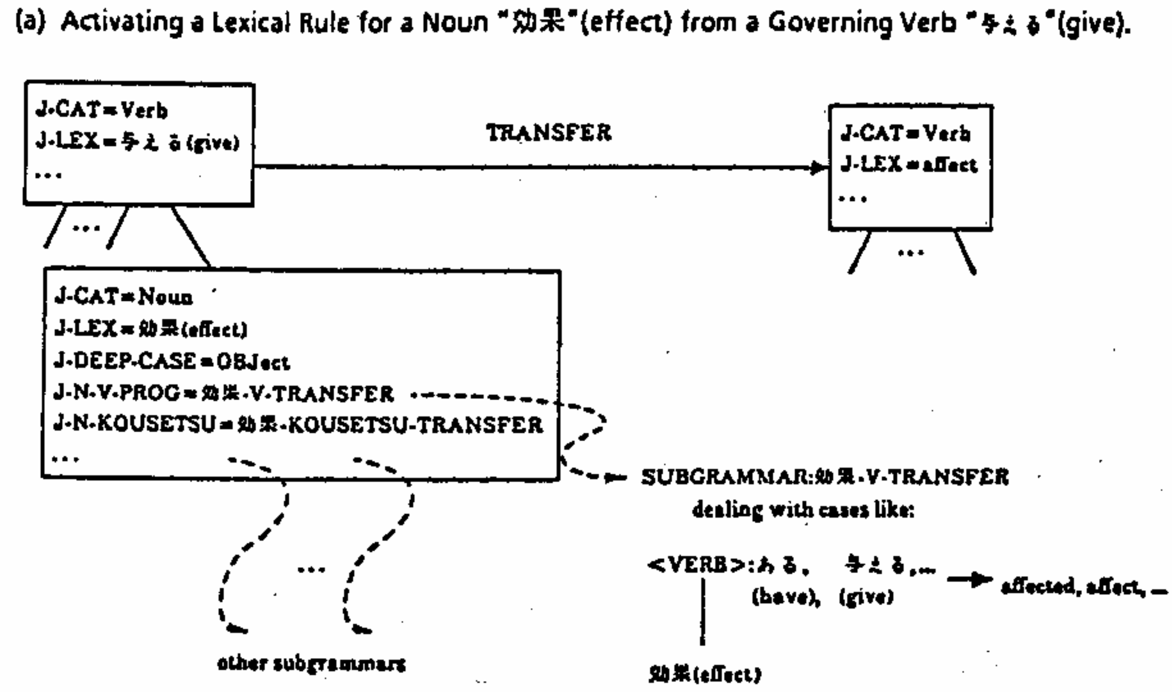


Fig. 1. Processing flow for the transfer and generation stages.



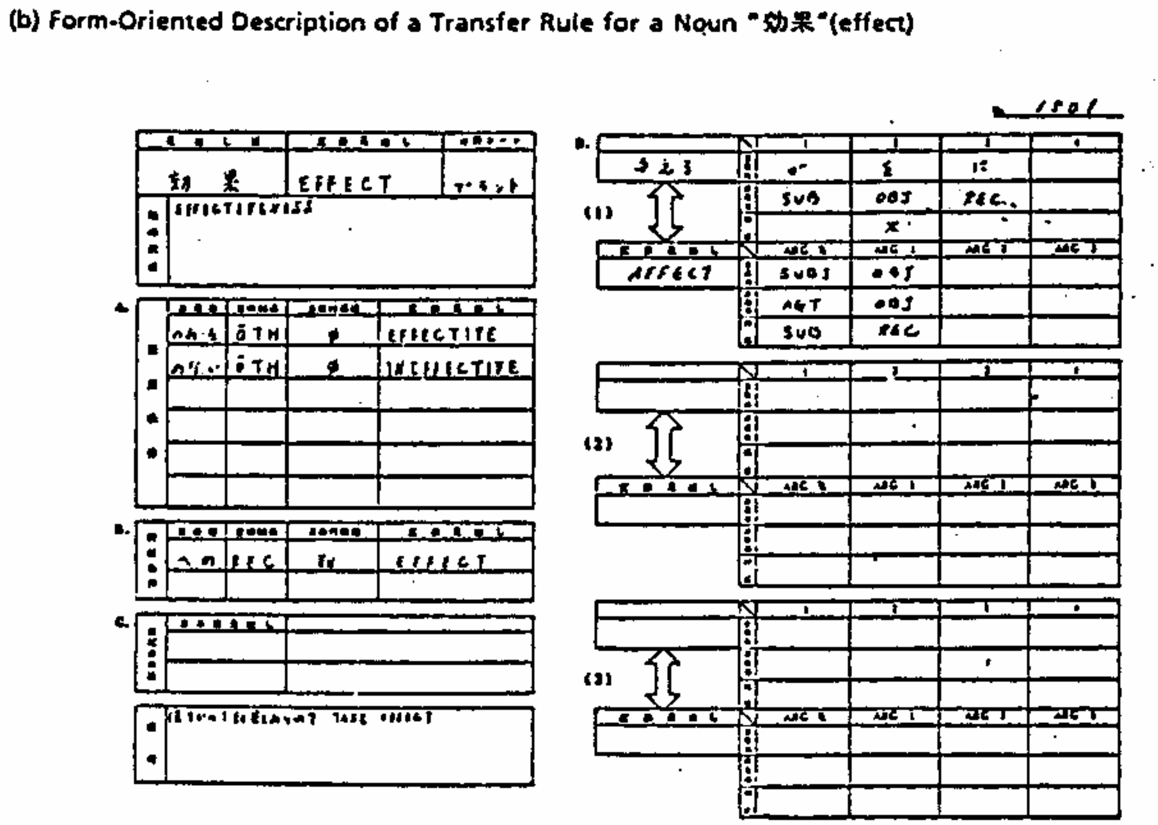


Fig. 2. Lexicon-oriented invocation of grammar rules.

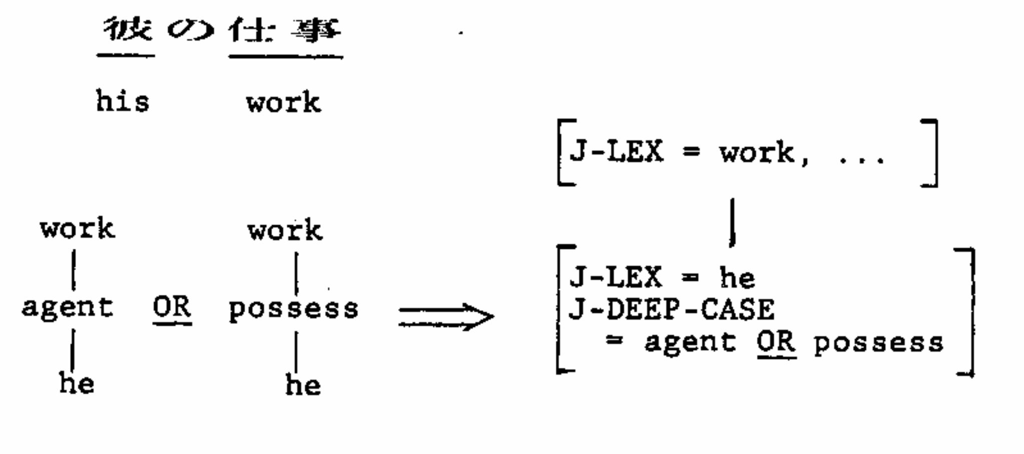
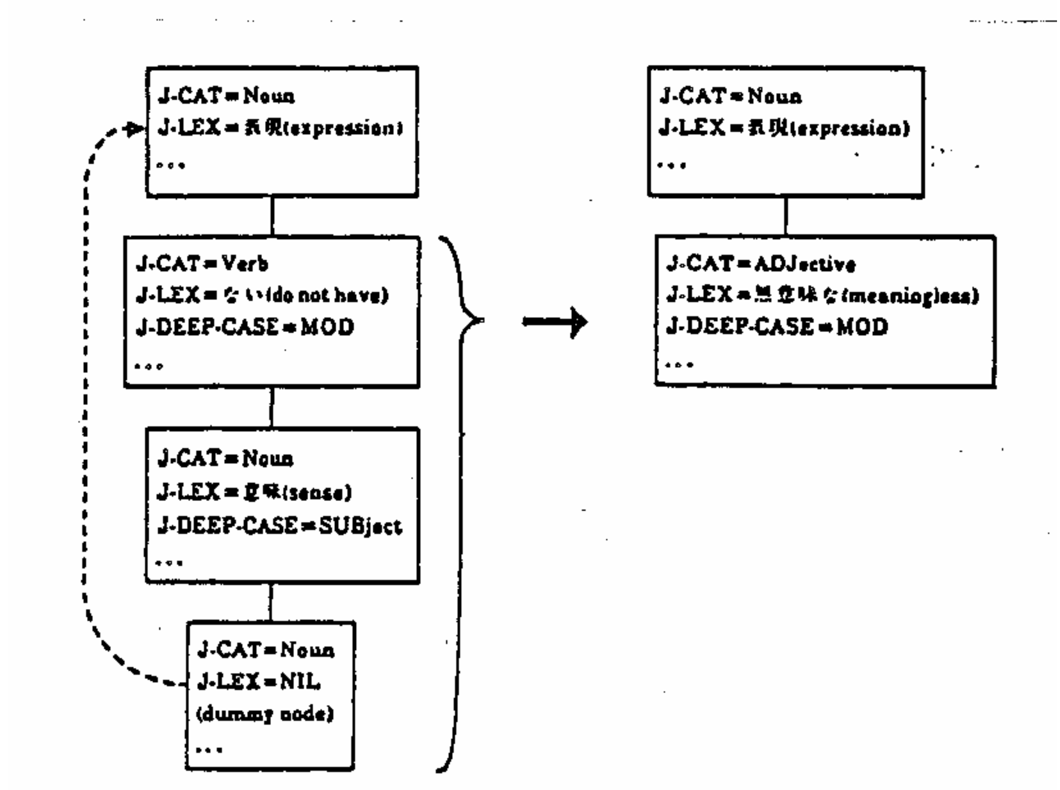


Fig. 3. An example of complex dependency structure.



"expression which does not have sense" —> "meaningless expression"

Fig. 4. An example of a heuristic rule used in the pre-transfer loop.

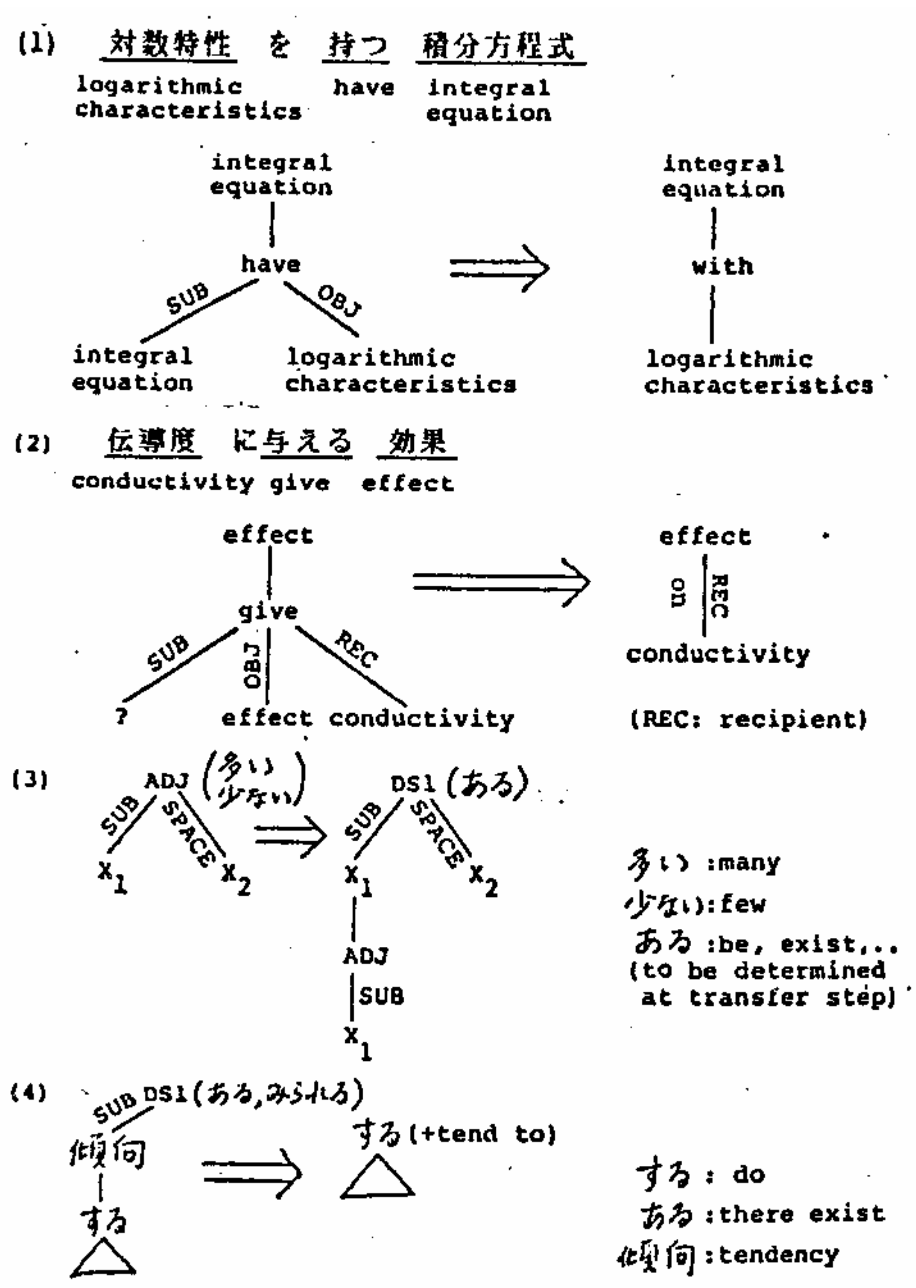


Fig. 5. Examples of pre-transfer rules.

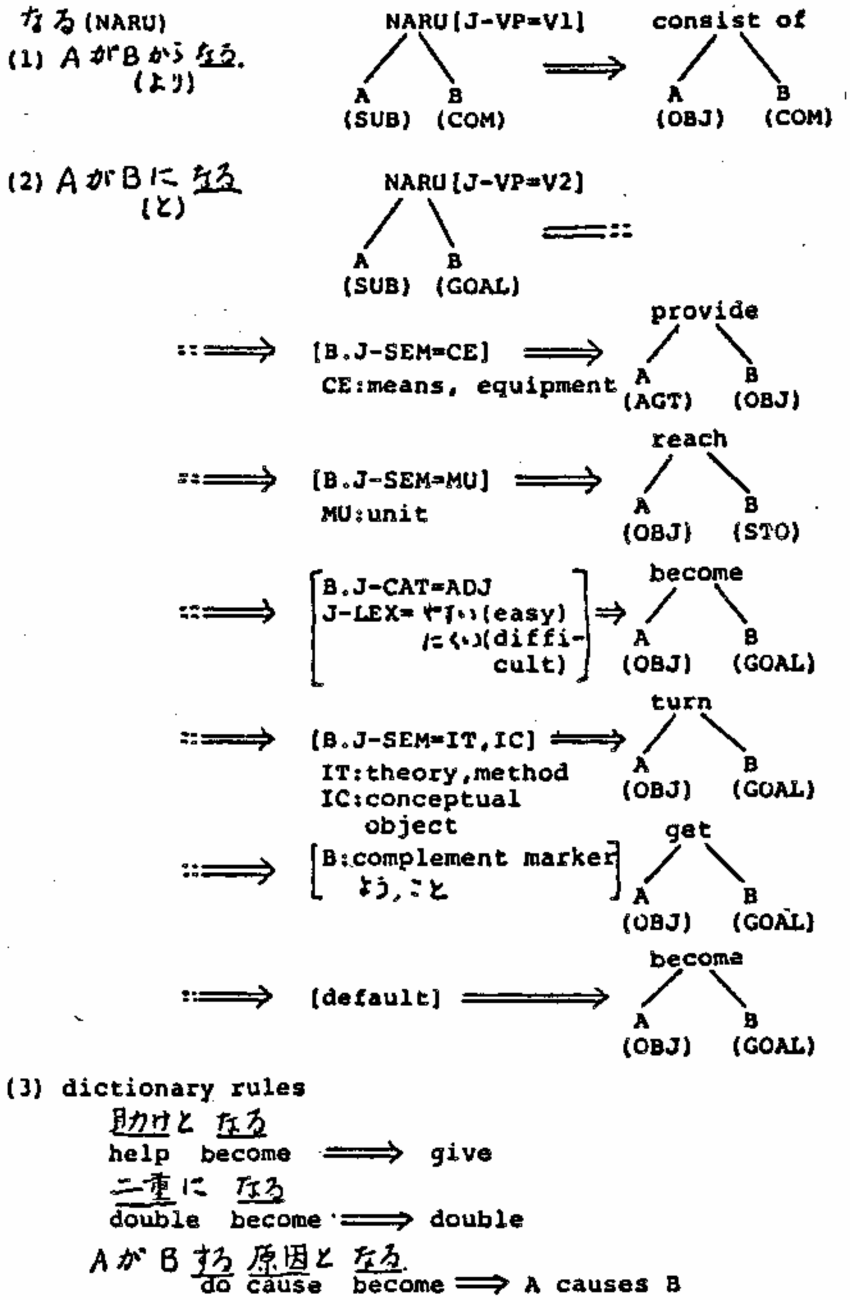


Fig. 6. An example of dictionary transfer rules of popular verbs.

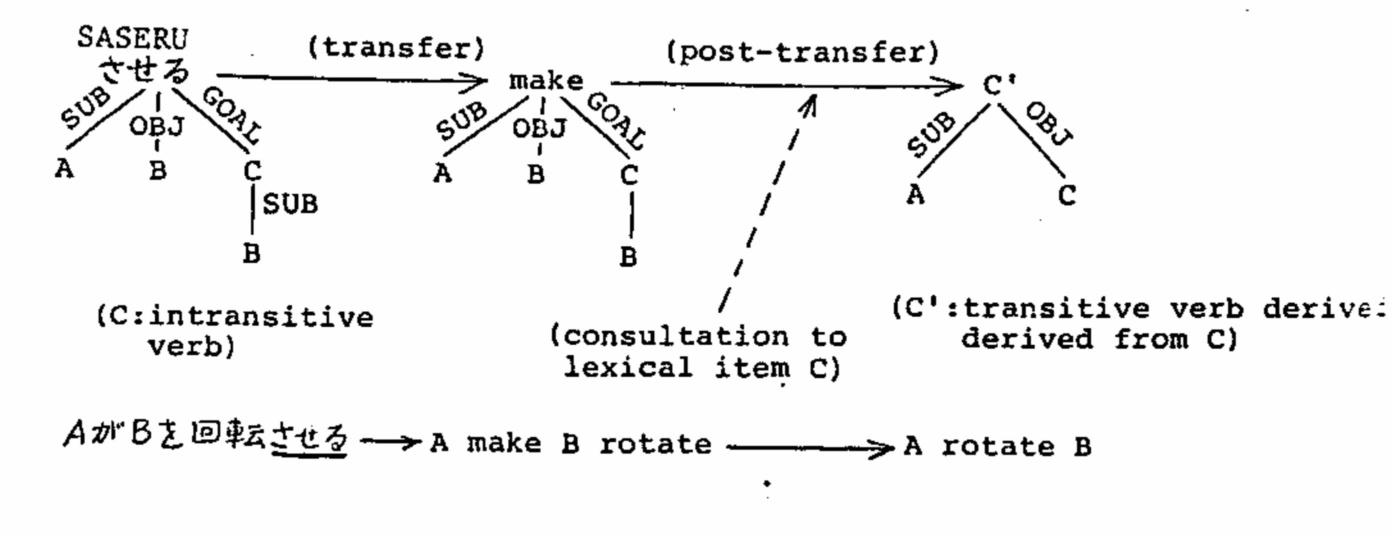


Fig. 7. An example of post-transfer rule application.

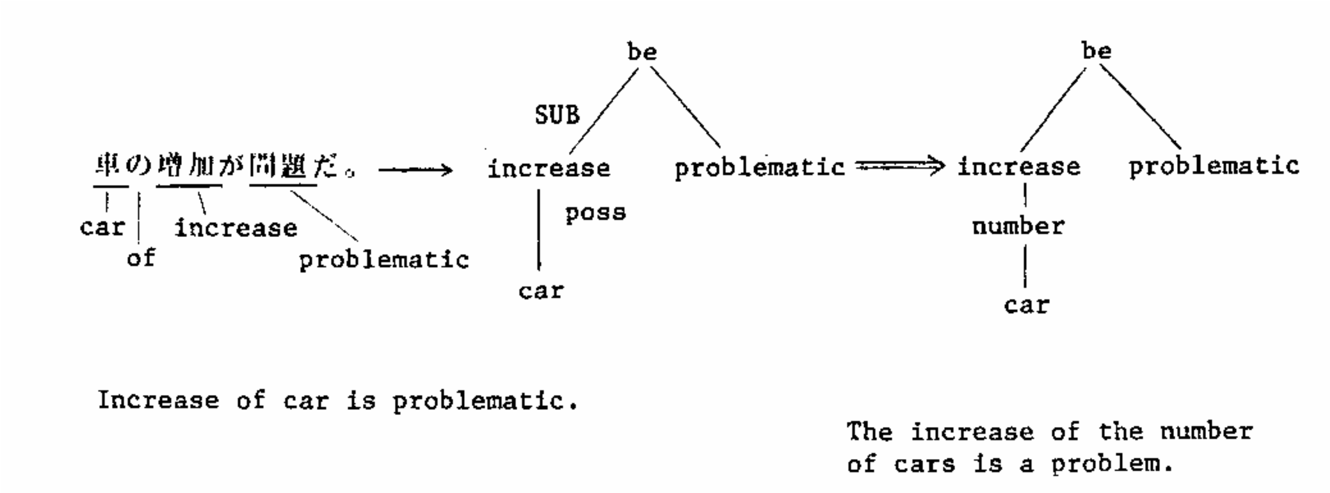


Fig. 8. An example of structural change at the post-transfer loop.

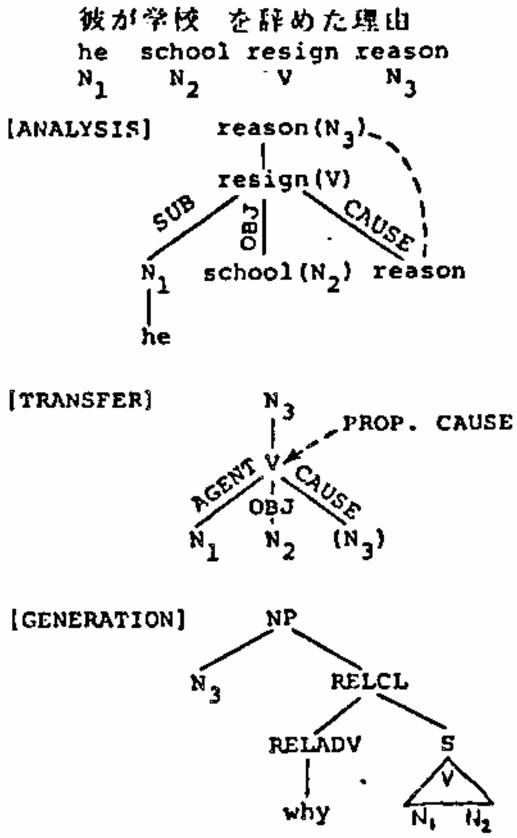


Fig. 9. Structural transformation of an embedded sentence of type 3.

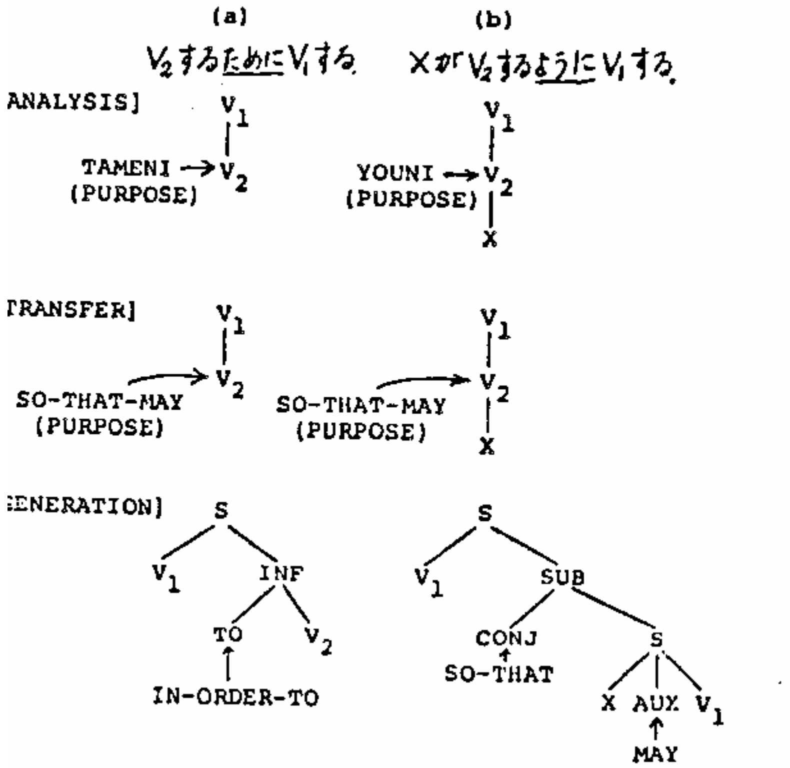
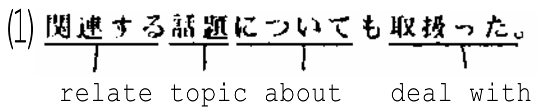


Fig.10. Structural transformation of an embedded sentence.



(2) ( ) dealt with also the related topics.

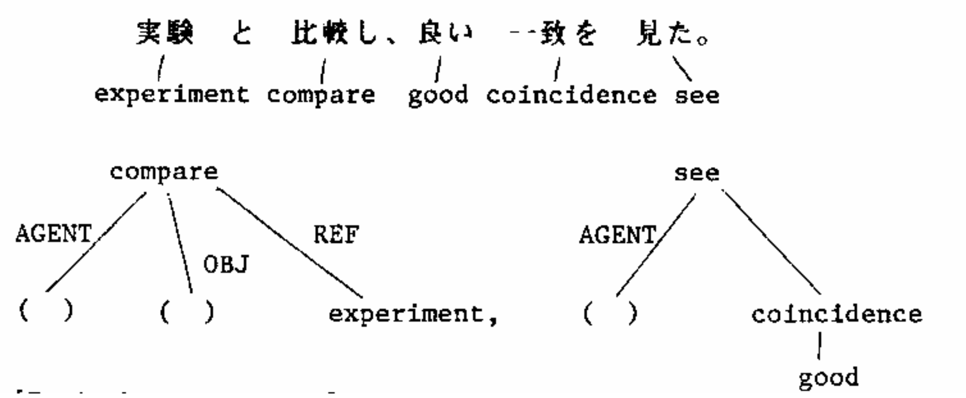
(3) Related topics were also dealt with.

(4) Handling was done also about the related topics.

Fig. 11. An example of transformations in the generation process.

Original Japanese (1) has no subject. (2) is- a literal translation. (3) is a passive form of (2). Another direction is the nominalization of the verb phrase "deal with". But the dictionary entry of "deal with" has no nominal form. Therefore the system sees the entry of synonym of "deal with", and finds out the word "handle", and find out the nominalization "handling". This noun is then put on at the subject position and a sentence (4) is generated with a nominal verb "be done".

[Japanese Input]



[English Translation]

The comparison is made with experiments, and good coincidence is obtained.

Fig. 12. An example of a sentence without subject and object.

Nominalization of "compare" is performed. "see" is changed to "obtain" by the consultation of the lexical entry of “coincidence”

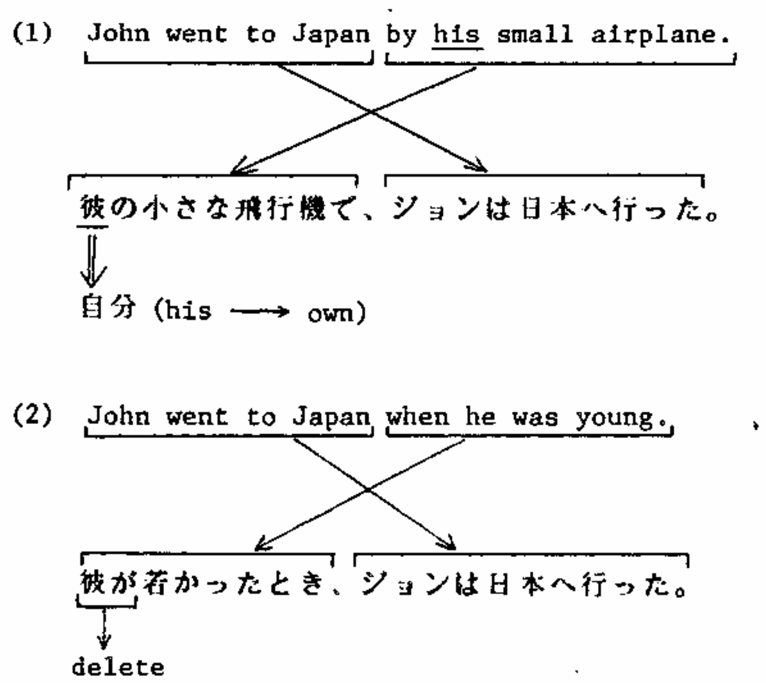


Fig. 13. Translation of pronouns.

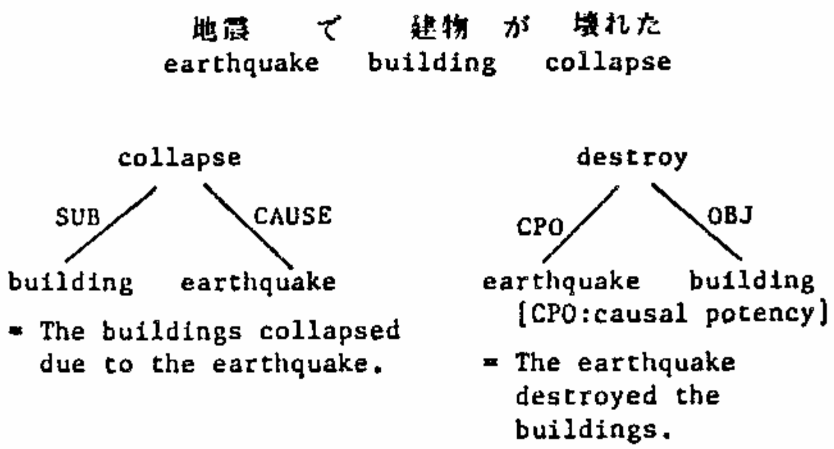


Fig. 14. An example of structural transformation in the generation phase.

生ずる	Xが生ずる	X	non-living substance structure	form-1	form X(obj
			social phenomena	take place	X take place
			action,deed,movement reaction	occur-1	X occur
			standard,property state,condition relation	arise-1	X arise
			ø	produce-2	produce X
	XがYを生ずる	Y	non-living substance structure	form-1	X form Y
			phenomena,action	cause-1	X cause Y
			ø	produce-2	X produce Y
上げる	XがYを上げる	Y	property	improve-1	X improve Y
			measure ·	increase-2	X increase Y
			ø	raise-1	X raise Y

↑ Semantic marker for X/Y

Table 1. Word selection in target language by using semantic markers.

J-SURFACE-CASE	J-DEEP-CASE	E-DEEP-CASE	Default Preposition
に(ni)	RECipient	REC, BENeficiary	to (REC — to, BEN — for)
	ORIgin	ORI	from
	PARticipant	PAR	with
	TIMe	Time-AT	in
	ROLe	ROL	as
	GOAl	GOA	to
	...	...	...

Table 2. Default rule for assigning a case label of English to a Japanese postposition "に" (ni).

JAPANESE SENTENTIAL CONNECTIVE	DEEP-CASE	ENGLISH SENTENTIAL CONNECTIVE
RENYO	TOOL	BY -ING ..
(-SHI)TE	TOOL	BY -ING ..
RENYO	CAUSE	BECAUSE ..
(-SHI)TE	"	"
-TAME	"	"
-NODE	"	"
-KARA	"	"
-TO	TIME	WHEN ..
-TOKI	"	"
-TE	"	"
-TAME	PURPOSE	SO-THAT-MAY
-NONI	"	"
-YOU	"	"
-YOU	MANNER	AS-IF
-KOTONAKU	"	WITHOUT -ING ..
-NAGARA	ACCOMPANY	WHILE -ING ..
-BA	CIRCUMSTANCE	WHEN ..
....	....	....

Table 3. Correspondence of sentential connectives.