DEALING WITH INCOMPLETENESS OF LINGUISTIC KNOWLEDGE IN LANGUAGE TRANSLATION
— TRANSFER AND GENERATION STAGE OF MU MACHINE TRANSLATION PROJECT —

Makoto Nagao, Toyoaki Nishida and Jun-ichi Tsujii
Department of Electrical Engineering
Kyoto University
Sakyo-ku, Kyoto 606, JAPAN

1. INTRODUCTION

Linguistic knowledge usable for machine translation is always imperfect. We cannot be free from the uncertainty of knowledge we have for machine translation. Especially at the transfer stage of machine translation, the selection of target language expression is rather subjective and optional.

Therefore the linguistic contents of machine translation system always fluctuate, and make gradual progress. The system should be designed to allow such constant change and improvements. This paper explains the details of the transfer and generation stages of Japanese-to-English system of the machine translation project by the Japanese Government, with the emphasis on the ideas to deal with the incompleteness of linguistic knowledge for machine translation.

2. DESIGN STRATEGIES

2.1 Annotated Dependency Structure

The intermediate representation we adopted as the result of analysis in our machine translation is the annotated dependency structure. Each node has arbitrary number of features as shown in Fig. 1. This makes it possible to access the constituents by more than one linguistic cues. This representation is therefore 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.

2.2 Multiple Layer Grammar

We have three conceptual levels for grammar rules.

lowest level: default grammar which guarantees 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 which chooses and generates 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 bears heuristic nature in the sense that it is word specific and it is applicable only to some restricted classes of inputs.

2.3 Multiple Relation Structure

In principle, 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 fail or may assign a wrong structure. 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. 2. 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.

2.4 Lexicon Driven Feature

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

2.5 Format-Oriented Description of Dictionary Entries

The quality of a machine translation system heavily depends on the quality of the dictionary. In order to build a machine translation dictionary, we collaborate with expert translators. We developed a format-oriented language to allow computer-naive human translators to encode their expertise without any conscious effort on programming. Although the format-oriented language we developed lacks full expressive power for highly sophisticated linguistic phenomena, it can cover most of the common lexical information translators may want to describe. The formatted description is automatically converted into statements in GRADE, a programming language developed by the Mu-Project. We prepared a manual according to which a man can fill in the dictionary format with linguistic data of items. The manual guarantees a certain level of quality of the dictionary, which is important when many people have to work in parallel.

Fig. 1. Representation of analysis result by features.

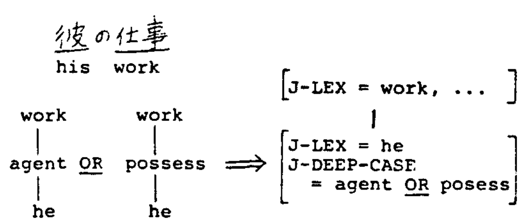


Fig. 2. An example of complex dependency structure.

3. ORGANIZATION OF GRAMMAR RULES FOR TRANSFER AND GENERATION STAGES

3.1 Heuristic Rule First

Grammar rules are organized along the principle that "if better rule exists then the system uses it; otherwise the system attempts to use a standard rule: if it fails, the system will use a default rule." The grammar rule involves a number of stages for applying heuristic rules. Fig. 3 shows a processing flow for the transfer and generation stages.

Heuristic rules are word specific. GRADE makes it possible to define word specific rules. Such rules can be invoked in many ways. For example, we can associate a word selection rule for an ordinary verb in a dictionary entry for a noun, as shown in Fig. 4.

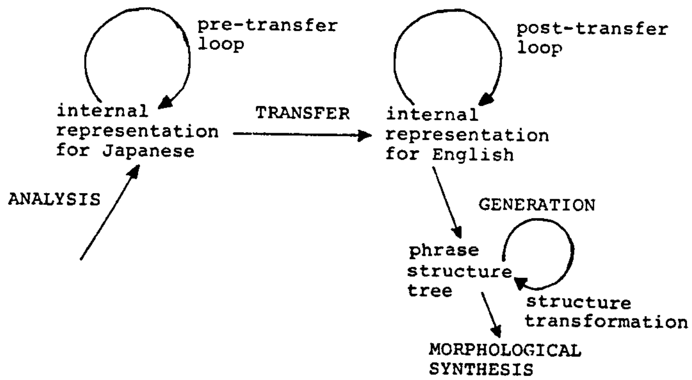
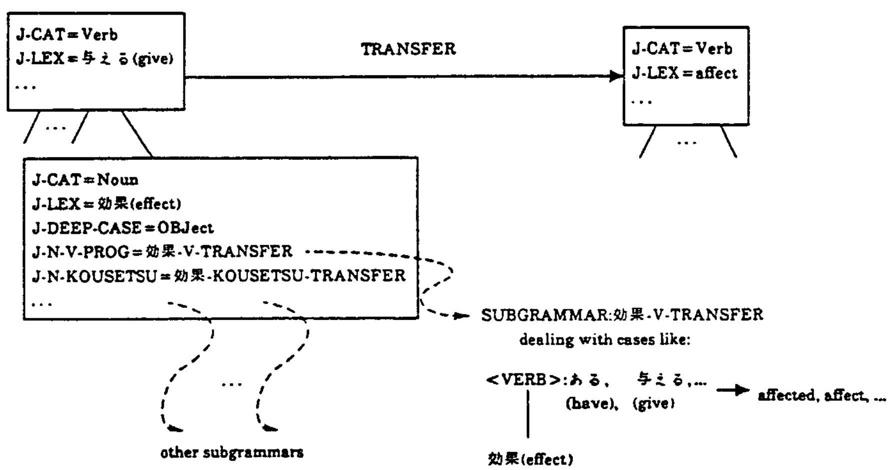
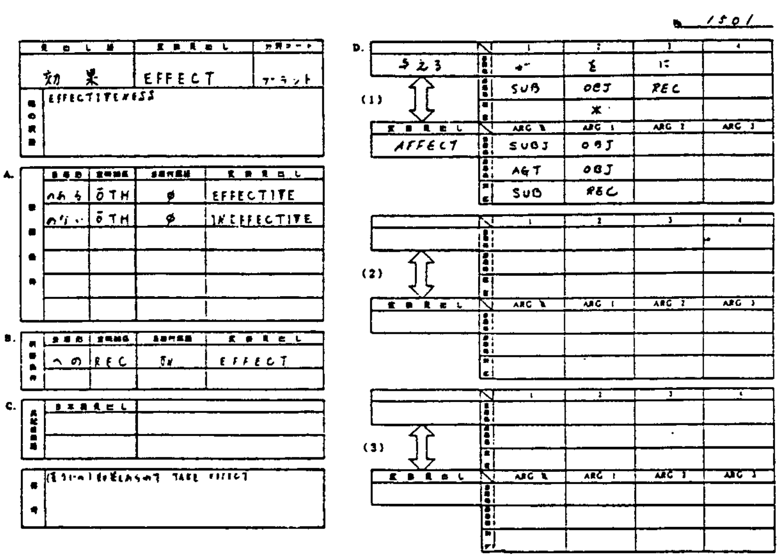


Fig. 3. Processing flow for the transfer and generation stages.

(a) Activating a Lexical Rule for a Noun "効果"(effect) from a Governing Verb "与える"(give).



(b) Form-Oriented Description of a Transfer Rule for a Noun "効果"(effect)



3.2 Pre-transfer Rules

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 such invocation the pre-transfer loop. Semantic and pragmatic interpretation are done in the pre-transfer loop. The more heuristic rules are applied in this loop, the better result will be obtained. Figs. 5 and 6 show some examples.

3.3 Word Selection in Target Language by Using Semantic Markers

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) Area 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 sometimes 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. The number i attached to verbs like form-1, produce-2 is the i-th usage of the verb. When the semantic information of nouns is not available, the column indicated by ϕ is applied to

Fig. 4. Lexicon-oriented invocation of grammar rules.

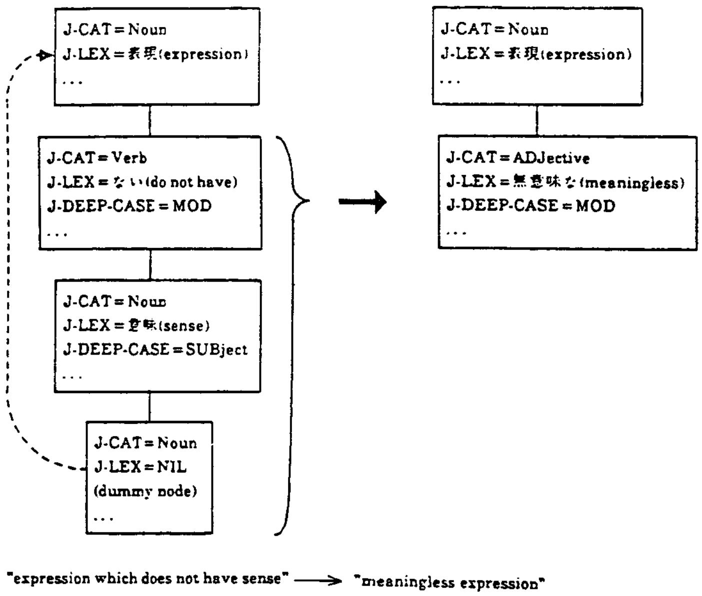


Fig. 5. An example of a heuristic rule used in the pre-transfer loop.

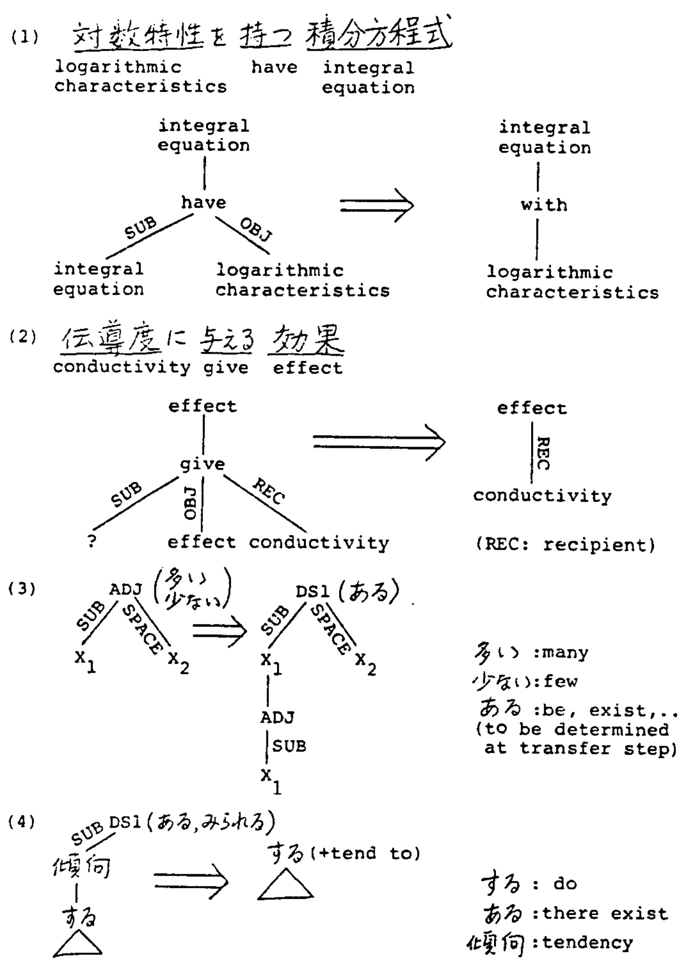


Fig. 6. Examples of pre-transfer rules.

produce a default translation.

In most cases, we can use a fixed format for describing a translation 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 encode transfer rules directly by GRADE. An example is shown in Fig. 7. Varieties of usages are to be listed up with their corresponding English sentential structures and semantic conditions.

3.4 Post-Transfer Rules

The transfer stage bridges the gap between Japanese and English expressions. There are still many odd structures after this stage, and we have to adjust further more the English internal representation into more natural ones. We call this part as post-transfer loop. An example is given in Fig. 8, where a Japanese factitive verb is first transferred to English "make", and then a structural change is made to eliminate it, and to have a more direct expression.

4. GENERATION PROCESS

4.1 Translation of Japanese Postpositions

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 combination. In this way, we confirm at least one translation to be assigned to an input. A particular usage of preposition for a particular English verb is written in the lexical entry of the verb.

4.2 Determination of Global Sentential Structures in Target Language

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
			∅	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
			∅	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
			∅	raise-1	X raise Y

↑ Semantic marker for X/Y

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

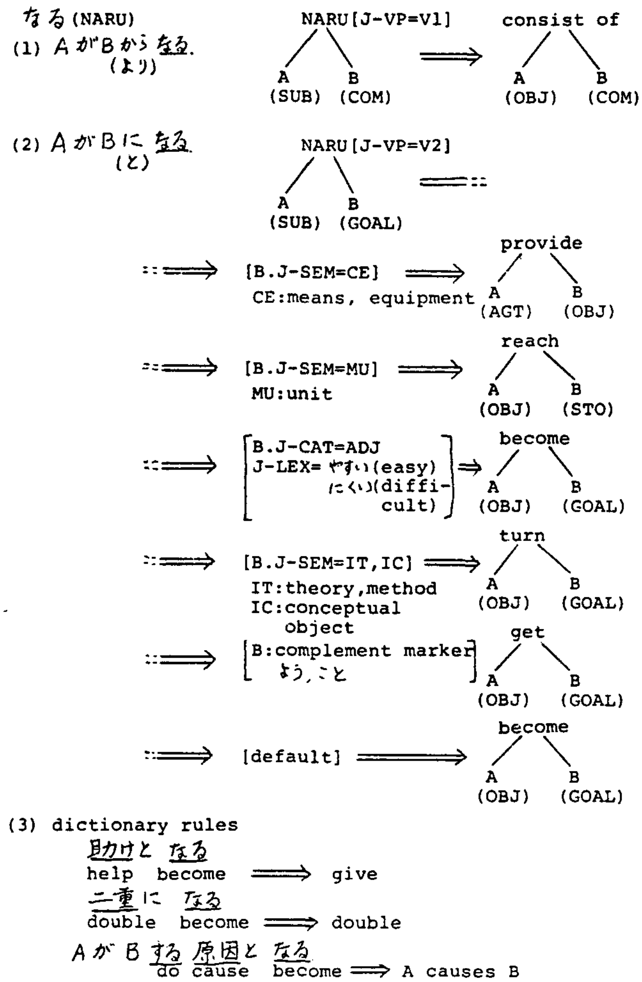


Fig. 7. An example of dictionary transfer rules of popular verbs.

Grobal sentential structures of Japanese and English are quite different, and correspondingly the internal 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 cases as (a) embedded sentential structure, (b) complex sentential structure.

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 the slot.

(ii) The form like "N_1が V な N_2" ≡ " (N_2 の N_1 が V) な N_2". In this case the noun N_1 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 the connecting 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 Table 3. An example is given in Fig. 10.

4.3 The Process of Sentence Generation in English

After the transfer is done from the Japanese deep dependency structure to the English one, conversion is done to a phrase structure tree with all the surface words attached to the tree. The processes explained in 4.1 and 4.2 are involved at this generation stage. The conversion is performed top-down from the root node of the dependency tree to the leaf. Therefore when a governing verb demands a noun phrase expression or a to-infinitive expression to its dependent phrase, the structural change of the phrase must be performed. Noun to verb transformation, and noun to adjective

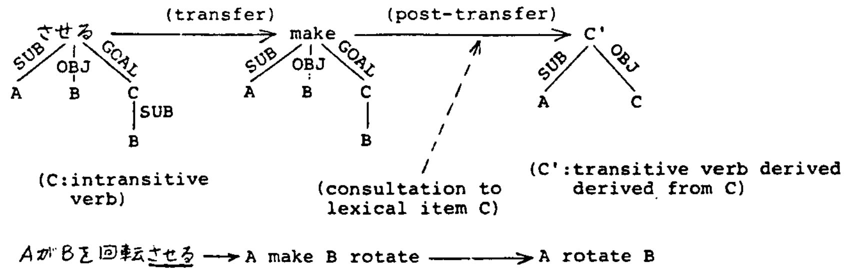


Fig. 8. An example of post-transfer rule application.

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.

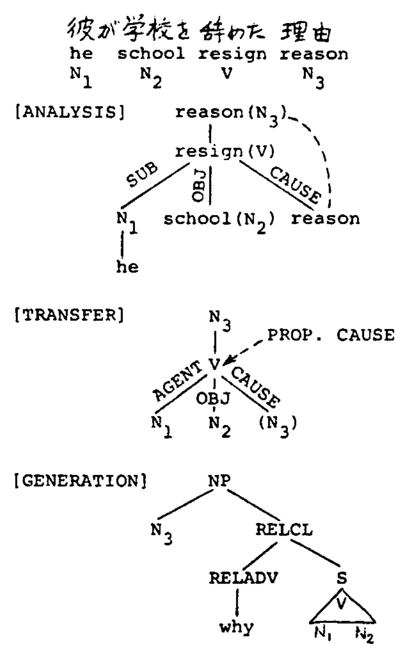


Fig. 9. Structural transformation of an embedded sentence of type 3.

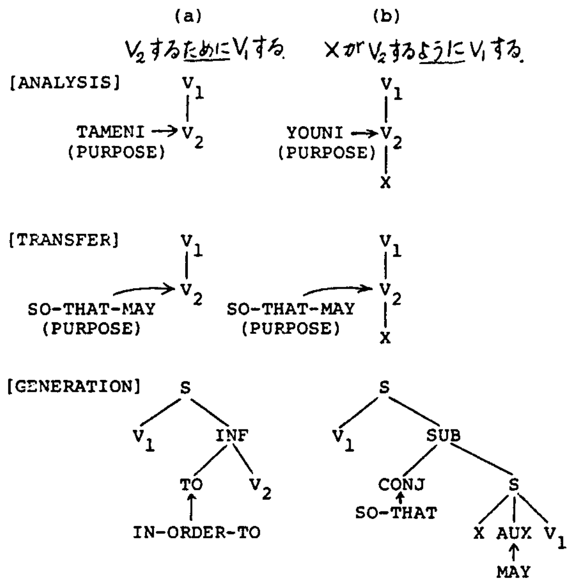


Fig. 10. Structural transformation of an embedded sentence.

transformation are often required due to the difference of expressions in Japanese and English. This process goes down from the root node to all the leaf nodes.

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

(i) When an agent is absent, passive 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.
(iii) When a subject phrase is a big tree, the anticipatory subject "it" is introduced.
(iv) Pronominalization of the same subject nouns is done in compound and complex sentences.
(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".
(vi) Others.

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. 11, and produces more English-like expressions. This stylistic transformation part is still very primitive. We have to accumulate much more linguistic knowledge and lexical data to have more satisfiable English expressions.

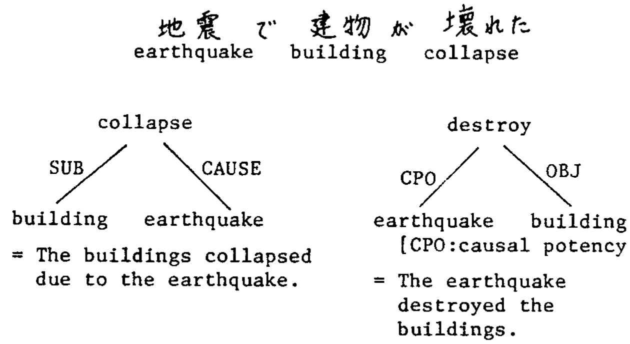


Fig. 11 An example of structural transformation in the generation phase.

5. SUMMARY

This paper described a number of strategies we employed in the transfer and generation stages of our Mu system to make the system both powerful and fault-tolerant. As is mentioned above, our system has many advantages such as the flexibility of the generation process, the utilization of strong lexical information. The system is in the course of development in collaboration with a number of computer scientists from computer industries and expert translators. Some of the translation results are attached in the last, which show the present level of the translation system. Progressive improvement is expected in the next two years.

ACKNOWLEDGEMENTS

We acknowledge the members of the Mu-Project, especially, Mr. S. Takai(JCS), Mr. Y. Fukumochi (Sharp Co.), Mr. T. Ishioka(JCS), Miss M. Kume (JCS), Mr. H. Sakamoto(Oki Co.), Mr. A. Kosaka (NEC Co.), Mr. H. Adachi(Toshiba Co.), Miss A. Okumura(Intergroup), and Miss A. Okuda(Intergroup) who contributed greatly for the implementation of the system.

REFERENCES

[1] M. Nagao: Machine Translation Project of the Japanese Government, a paper presented at the workshop between EUROTRA and Japanese machine translation experts, held in Brussels on November 24-25, 1983.

[2] J. Nakamura, et al.: Grammar Writing System (GRADE) of Mu-Machine Translation Project and its Charactersitics, Proc. of COLING 84, 1984.

[3] J. Tsujii, et al.: Analysis Grammar of Japanese in the Mu-Project — A Procedural Approach to Analysis Grammar —, ibid.

[4] Y. Sakamoto, et al.: Lexicon Features for Japanese Syntactic Analysis in Mu-Project-JE, ibid.

[5] J. Tsujii: The transfer Phase in an English-Japanese Translation System, Proc. of COLING 82, 1982.

Sample outputs as of April, 1984 are attached in the next page.

Sample outputs:

SENTENCE E82060010_5 LOAD OK.
"相電位回路のパラメータは測定結果の数学的確率論的方法による処理で得られる当該電気設備の設計電圧に基づいて選択すべきであると述べた。"

It was mentioned that the parameter of a phase-voltage circuit should be selected based on the design voltage of electrical facilities which can be obtained by the processing of a measurement result by mathmatical statistical method.

SENTENCE E82060011_5 LOAD OK.
"保護対策として電気固有抵抗の低減の重要性を指摘し，その静電電圧計での計測を述べる。"

The importance of the reduction of the electrical resistivity is pointed out as the protective measure, and the measurement in this electrostatic voltmeter is described.

SENTENCE E82060012_2 LOAD OK.
"有限要素法による軸対称問題の磁界解析。"

The magnetic field analysis of a problem of axisymmetry by a finite element method.

SENTENCE E82060014_6 LOAD OK.
"E，H偏波についての放射輝度温度の計算式を得た。"

The calculation formula of radiation brightness temperature about E and H waves was obtained.

SENTENCE E82060015_8 LOAD OK.
"位相ゆらぎ分散と相関関数決定誤差の関係を調べ，連続媒質内での実験結果と比較。"

The relationship of a phase fluctuation variance and the correlation function determination error is examined, and compared with the test result in the continuous medium.

SENTENCE E82060017_2 LOAD OK.
"平面成層ランダム不均質媒質内でのビームの多重散乱。"

The multiple scattering of a beam in plane stratified random nonhomogeneous medium.

SENTENCE E82060020_3 LOAD OK.
"真空中の誘電体板に入射する平面波の回折により生じる表面波の複素振幅と等価表面電流分布を積分方程式の数値解により見出す。"

The complex amplitude and the equivalent surface current distribution of the surface wave formed due to the diffraction of the plane wave emited to the dielectric plate in the vacuum are found by a numerical solution of an integral equation.

"方位方向に磁化したフェライト層で覆われた理想伝導体の円筒で作られた制御系の特性波を取り扱った。"

A characteristic wave of the control system made out of the cylinder of the ideal conductor covered by the ferrite layer magnetized in a orientation direction was dealt with.