# MORPHOLOGICAL ASPECT OF JAPANESE LANGUAGE PROCESSING

Kosho Shudo, Toshiko Narahara and Sho Yoshida

Department of Electronics
Fukuoka University
11, Nanakuma, Nishi-ku
Fukuoka-shi, 814 Japan

Department of Electronics
Kyushu University
6-10-1, Hakozaki, Higashi-ku
Fukuoka-shi, 812 Japan

A comprehensive grammatical model produced for analyzing the agglutinated structure which characterizes the Japanese language is presented. This model, which includes extensively idiomatic postpositional expressions as terminals, is quite effective for the development of the Japanese language processor receptive to a reasonable variety of sentential forms and applicable to relatively wide fields.

## Introduction

The following fundamental problems are still latent in most present systems of the natural language processing: (i)how to enable the system to have a higher quality processing that renders the output more feasible; (ii)how to broaden the applicable field of the system; and (iii)how to allow the system to accept more "natural" input sentences, including miscellaneous linguistic constructions. In order to remedy these problems, we will need not only far advanced A.I. researches on the knowledge representation or deduction, but also more elaborate studies on the surface structures of natural sentences from the engineering viewpoints. Among other things, the requirement for the linguistic approach on the engineering side is quite urgent for Japanese language processing, since we have no Japanese grammar which is extensive and definite enough for solving, especially, problem (iii).

The authors have been developing a Japanese language parser for a Japanese-English translation system on the following standpoints.

(1)*Wide coverage of the input forms;* We aim at a system which is powerful enough to accept with less exceptions the sentential forms which appear in the actual, colloquial and written texts (e.g. non-pre-edited sentences in technical papers).

(2)*Two-phase parsing;* The system first analyzes the local expression which is the syntactical and semantical unit constituting immediately the input sentence, and then analyzes the whole sentence by detecting the relationships between the units. The first phase, which corresponds to the morphological phase in the ordinary parser of the European language, is designed for analyzing not only the word's inflection but the "agglutinated" structure characterizing the Japanese language.
We attach much importance to the first phase which has a great influence on the overall performance of the system.

(3)*Elaborate preparation for the first phase;* In the first phase, we adopt an elaborate grammatical model that prescribes the internal structure of the above-mentioned units in detail. The extensive enumeration of postpositional expressions carried out in the model, among others, is quite effective for solving the problem (iii), since they determine the syntactical and semantical "framework" of the Japanese sentence. The inflection of the word can also be manipulated almost without exceptions in a relatively simple way in this model.

(4)*Matching of the first phase and higher phases;* Most of the atomic postpositional expressions enumerated in the model are idiomatic ones which should be treated without decomposing into words because of their definite and unitary meanings. This fact yields a good matching of the first phase and the higher semantical phases.

(5)*Disambiguation in the first phase;* A certain part of the polysemy of the postpositional expression can be reduced by the restriction for the co-occurence on the neighboring positions in the sentence. Our grammar for the first phase is designed to carry out disambiguation of this type. This is based on the idea that the syntactical and semantical structure ought to be disambiguated as early and as much as possible from the viewpoint of the system's total efficiency.

In this paper, the above mentioned grammatical model for the first phase of parsing, which may be called "pseudo-morphological" phase, is shown and the experimental system developed for the verification of its validity is outlined. After showing some operational examples and the result of the experiment, we conclude that our model is quite effective for Japanese language processing from the standpoints mentioned above.

## Japanese sentence, E-bunsetsu

The information to be extracted from the input sentence by the parser can be generally classified into following three types:

(a)the information of the concept which is ordinarily provided by the conceptual word (e.g. noun, verb, adjective);

(b)the information of the relationship between concepts;

(c)the supplementary information such as of

"tense","aspect","mood",etc.
Japanese is an agglutinative language and is very far from European languages from structural viewpoints, i.e. the information of type(b) or (c) is ordinarily given by the annex-expression agglutinated postpositionally to the conceptual expression which gives the information of type (a). We call the compound which consists of the annex- and conceptual expression E-bunsetsu[†]. The information of type(b) is given as the dependency relation, called kakariuke-relation between E-bunsetsus. A sentence consists immediately of E-bunsetsus positioned in a relatively free order except for a few constraints[††]. Because of this structural feature, we adopt the two-phase approach for the parsing of the Japanese sentence: the first phase for analyzing each E-bunsetsu; the second, for detecting the kakariuke-relational structure of the sentence.

It is apparent that the extensive characterization of the E-bunsetsu yields the wide coverage of input sentential forms to the system. Specifically, the extensive enumeration of the annex-expressions will drastically broaden the range of acceptable input forms, since they make the syntactic and semantic "framework" of the sentence. However, the annex- or conceptual expression may itself be a compound of atomic expressions and is too multiformed to be enumerated extensively.

From these points of view, we have constructed a grammatical model for analyzing the E-bunsetsu by, first, enumerating extensively atomic expressions excepting most of the conceptual ones that are quite numerous; secondly, classifying them by the syntactic and partially semantic functions; thirdly, prescribing the connectability rules of atomic expressions within the E-bunsetsu.

## Atomic Expressions

The notion of "atomic expression" is the extended one of "word" so as to include the idiomatic word-string which has the unitary, self-supported meaning and the definite syntactic function. Though we often encounter such idiomatic strings in the sentence of everyday use, it has not been clarified exhaustively how many are needed for building up the natural sentence and how they can be used.

We have singled out the atomic expressions extensively excepting most of conceptual ones from approximately 12,000 sentences of technical papers and text-books of the senior high schools. Their rough categorization is shown in the following. (The number of the expressions is given in parentheses.)

### Annex-expressions

Atomic annex-expressions are classified into two kinds: relational expressions which provide the information of type(b); and co-predicative expressions which provide the information of type(c).

Relational Expressions(575). While the typical example of the relational expression is the particle in the conventional grammar, eighty percent of the relational expressions are idiomatic word-strings. For example, the word-string,'ni tsui te' is atomic and relational because it has a proper, undividable and self-supported meaning equivalent to that of the preposition,*'about'* in English in such context as 'Mary ni tsui te hanasu'*('talk about Mary')*. (The original meaning of the verb,'tsuku' is almost missing in the context.)

The atomic annex-expressions can be divided roughly into ten categories according to their abilities to indicate the kakariuke-relation. We denote these categories by $R_{NP1}$,$R_{NP2}$,$R_{NP3}$, $R_{PP1}$,$R_{PP2}$,$R_{PP3}$,$R_{NN1}$,$R_{NN2}$,$R_{NN3}$and $R_{PN}$. $R_{NP1}$, $R_{NP2}$ or $R_{NP3}$, for example, is a category of expressions which indicate the dependency of the nominal E-bunsetsu, N on the predicative E-bunsetsu, P. 'ni tsui te' mentioned above is included in $R_{NP1}$.

Co-predicative Expressions(348). The auxiliary verb in the conventional school grammar is typically co-predicative but ninety percent of the co-predicative expressions singled out are idiomatic. For example, the word-string,'ta hou ga yoi',which is equivalent to *'had better'* in English provides the information of the modality. These can be divided into seven categories,i.e. $A_{np1}$,$A_{np2}$,$A_{np3}$,$A_{pp1}$,$A_{pp2}$,$A_{pp3}$ and $A_{pp4}$ according to the functions of the connection and whether they can inflect or not. $A_{pp1}$, for example, represents a category of inflectable expressions each of which yields a predicative expression, p by connecting(agglutinating) to a predicative expression, p. The atomic expression, 'ta hou ga yoi' mentioned above is in $A_{pp1}$.

### Conjunctive Expressions(122)

Besides the traditional conjunction, many conjunctive, idiomatic expressions have been singled out as atomic ones. For example, the string 'sikasi nagara', wich is equivalent to *'however'* in English is conjunctive and atomic. The conjuctive expression is not annexational, but offers the information of type(b). There observed two categories: one, denoted by $C_1$, of expressions which can indicate both of the relation between two sentences and the relation

[†] The notion of "bunsetsu" in the conventional school grammar is well known as the unit for sentence construction. However, the unitary local structure in the real sentence used in every day life is often too multiformed to be analyzed with it. The notion of "E-bunsetsu", which is a fully extended version of "bunsetsu", was devised from the standpoints mentioned in the previous chapter.

[††] When we let a string, $EB_1$ $EB_2$ ••• $EB_n$ be a sentence, each E-bunsetsu, $EB_i$ $(1 \leq i < n)$ must depend on only one of $EB_{i+1}$,•••,$EB_n$ without passing any $EB_j$ $(i<j)$ that governs at least one of $EB_1$, •••, $EB_{i-1}$. Moreover, $EB_n$ must be predicative.

between two E-bunsetsus; the other, denoted by $C_2$, of expressions which indicate exclusively the relation between two sentences.

Suffixal Expressions(403)

The conceptual expressions are too numerous to be enumerated exhaustively. In addition, it is difficult in the present state to settle the extensive rules for constructing the conceptual compound.

We have singled out only the suffixal constituents of the conceptual compounds that are used very frequently and have definite syntactic and semantic functions. These are classified roughly into seven categories, i.e. $S_{np1}$,$S_{np2}$, $S_{pp}$, $S_{nn1}$,$S_{nn2}$,$S_{nn3}$ and $S_{pn}$, by their functions. For example, $S_{np1}$, that includes such a string as 'de aru' being used quite frequently, is a category of expressions each of which constitutes a predicative conceptual expression,p when suffixed to a nominal conceptual expression,n. The conceptual compound of quantitative, temporal or locational meaning, e.g. '3 zi 15 hun' *('a quarter past three')* is sometimes exceptionally easy to be decomposed into constituents. A good many suffixal constituents of these compounds are included in $S_{nn1}$.

Adverbial Expressions(262)

The adverbial expressions fall into two categories, $D_2$, for the expression which is always used in cooperation with some other specific expression and $D_1$, for the rest. For example, 'kanarazusimo ••• (nai)' *('not necessarily')* is in $D_2$.

Adnominal Expressions(165)

The adnominal expression, such as 'subete no' *('all')* is similar to the adjective except that it is uninflective and used always attributively being located ahead of the nominal E-bunsetsu to be modified. The category of these expressions is denoted by T.

## Structure of E-bunsetsu

The structure of the E-bunsetsu can be characterized in the form of "transition net", since it has no complex embedded structures. Our structural characterization is based on prescribing the connection rules of the atomic expressions within an E-bunsetsu. It is shown in two stages in this chapter.

General Structure of E-bunsetsu

The general structure of E-bunsetsu is shown in Fig.1 using the above-mentioned categories and three traditional ones, $M_1$,$M_2$ and Y,representing for the noun, verbal-noun(i.e. noun called

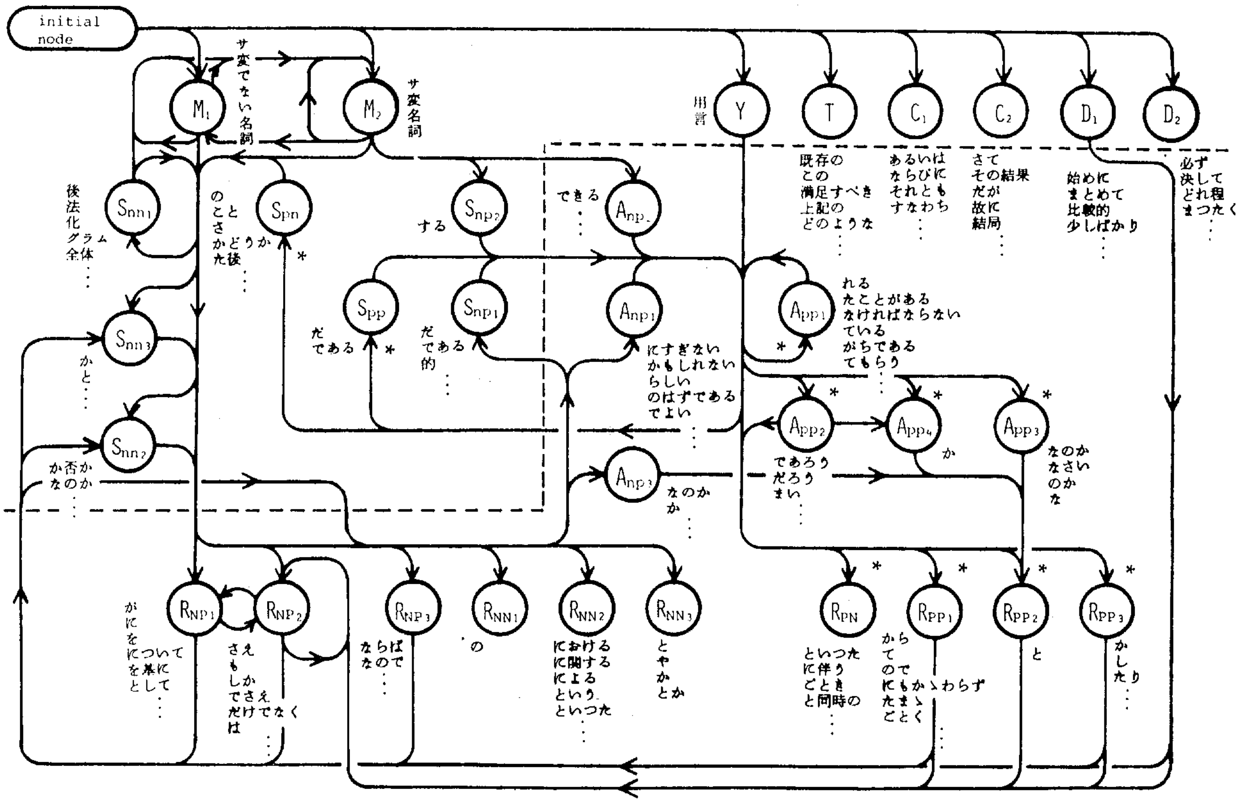


Fig. 1. Connection Graph Describing the General Structure of E-bunsetsu

"sahen-meishi") and yougen(i.e. verb, adjective, adjective verb), respectively. In Fig.1, nodes represent the categories and arrows denote that expressions in starting nodes can be immediately followed(agglutinated) by those in ending nodes. The E-bunsetsu can be analyzed, though roughly, by starting at the initial node and tracing a path in the figure. Each node is the acceptable node for the E-bunsetsu. The conceptual expression corresponds to a path terminating at a node located above the dotted line. The syntactic and semantic function of the E-bunsetsu can be estimated by recognizing the terminating node in the path.

Generality of Characterization. In order to verify the generality of the characterization shown in Fig.1, we have inspected approximately 1,500 actual sentences in technical papers by segmenting each sentence into E-bunsetsus applying the above rules. Table 1 shows the results of the inspection. From this, it came out that our enumeration of annex-expressions is almost sufficient and all of the sentences inspected can be segmented into E-bunsetsus if we newly register and classify the expressions missing in the enumeration into existing categories. In addition, it turned out that the idea of the E-bunsetsu, which elucidates a

Table 1. Results of Inspection

| | |
|---|---|
| number of atomic expressions missing in the enumeration: | |
| annex- | 6 |
| conjunctive | 21 |
| suffixal | 0 |
| adverbial | 49 |
| adnominal | 25 |
| unclassifiable | 0 |
| number of: | |
| sentences , n | 1,532 |
| bunsetsus , $n_1$ | 23,432 |
| E-bunsetsus , $n_2$ | 20,118 |
| $n_1/n$ | 15.3 |
| $n_2/n$ | 13.1 |
| $(n_1-n_2)/n_1$ | 0.14 |
| total appearances of: | |
| annex-expressions , $n_3$ | 10,124 |
| compound annex-expressions , $n_4$ | 1,655 |
| $n_4/n_3$ | 0.16 |

Table 2. Paradigm

| | form | | negative-conjectural form | | | adverbial form | | standard form | adnominal form | subjunctive form | imperative form | | stem |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | type | code | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B |
| verb type | 5-vowel, I-type | 0 | a | a | o | i | $i^*$ | u | u | e | e | | ex.kik |
| | 5-vowel, T-type | 1 | a | a | o | i | $t^*$ | u | u | e | e | | ex.okor |
| | 5-vowel, Q-type | 2 | a | a | o | i | $q^*$ | u | u | e | e | | ex.sin |
| | 5-vowel type | 3 | a | a | o | i | i | u | u | e | e | | ex.kes |
| | 1-vowel type | 4 | ε | ε | yo | ε | ε | ru | ru | re | ro | yo | ex.tozi |
| | S-type | 5 | i | e | iyo | i | i | uru | uru | ure | iro | eyo | ex.s |
| | K-type | 6 | o | o | oyo | i | i | uru | uru | ure | oi | | ex.k |
| | W-type | 7 | wa | wa | o | i | t | u | u | e | e | | ex.tiga |
| adjective-type | | 8 | ku | | karo | ku | kat | i | i | kere | | | ex.yo |
| adjective verb type | NA-type | 9 | | | | ni | | | na | | | | ex.kirei |
| | NO-type | A | | | | ni | | | no | | | | ex.hodo |
| | ε-type | B | | | | ni | | | ε | | | | ex.onazi |
| T-type | | C | | | aro | | | a | a | | | | ex.t |
| D-type | | D | | | aro | e | at | a | | | | | ex.d |

larger structure than a "bunsėtsu", is quite effective for reducing the load of the second phase of the parser because it causes fourteen percent decrease of the number of immediate constituents of the sentence. Moreover, the rate of appearance of the atomic relational expressions which are originally compound was found to be sixteen percent. These facts assure the generality of the characterization to a reasonable extent.

## Detailed Structure of E-bunsetsu

In the course of the development of the natural language system, it is a fundamental and crucial problem how much the grammatical rule should be elaborate or how much the syntactic and semantic structure of the sentence should be disambiguated within the grammatical phase of the processing. We think it profitable for increasing the total efficiency of the system to disambiguate them as much and as early as possible. From this point of view, we try to do it in the phase of analyzing the E-bunsetsu by refining the characterization of Fig.1 without destroying its grammatical features and generality.

Inflectional Endings. The word-inflection of Japanese language is closely related to the agglutination of words. The connection represented in Fig.1 by the asterisked arrow should be restricted by the inflectional type and inflectional form of the preceding expression, which is inflectable.

While subcategorizing the inflectable expressions by their inflectional types, we gave respective expressions in the ending nodes of the asterisked arrows a dictionary entry denoting what inflectional types and forms it can be connected to. The inflectional form is known by detecting the ending. Table 2 shows the paradigm. The asterisked letter in the table is a euphonical one by which the final letter of the stem may be replaced. 'ε' represents an empty ending.

This paradigm (and the experimental system described in the next chapter) is based on a way of expressing Japanese characters by English letters which is devised from the viewpoints of mechanical processing.

Subcategorization. We subcategorized some of the annex-expressions by their detailed agglutinative functions using a formal algorithm[†].

It should be noted that the homonymous expression whose meanings have individual agglutinative functions was categorized duplicatively into different categories according to respective functions. These expressions' meanings, therefore, can be disambiguated by checking the agglutinative structure of the E-bunsetsu.

Table 3. Outline of Subcategorization

| original category | | number of expressions | number of subcategories |
|---|---|---|---|
| relational | $R_{NP1}$ | 153 | 24 |
| | $R_{NP2}$ | 63 | 31 |
| | $R_{NP3}$ | 13 | - |
| | $R_{NN1}$ | 1 | - |
| | $R_{NN2}$ | 118 | - |
| | $R_{NN3}$ | 4 | - |
| | $R_{PN}$ | 40 | - |
| | $R_{PP1}$ | 149 | 4 |
| | $R_{PP2}$ | 1 | - |
| | $R_{PP3}$ | 38 | 3 |
| co-predicative | $A_{np1}$ | 37 | 5* |
| | $A_{np2}$ | 4 | 2* |
| | $A_{np3}$ | 2 | - |
| | $A_{pp1}$ | 298 | 12* |
| | $A_{pp2}$ | 15 | - |
| | $A_{pp3}$ | 4 | - |
| | $A_{pp4}$ | 1 | - |
| suffixal | $S_{nn1}$ | 288 | 25 |
| | $S_{nn2}$ | 6 | - |
| | $S_{nn3}$ | 3 | - |
| | $S_{pn}$ | 92 | 10 |
| | $S_{np1}$ | 11 | 4* |
| | $S_{np2}$ | 1 | - |
| | $S_{pp}$ | 2 | 2* |
| conjunctive | $C_1$ | 35 | - |
| | $C_2$ | 87 | - |
| adverbial | $D_1$ | 180 | - |
| | $D_2$ | 82 | - |
| adnominal | T | 165 | - |
| noun | $M_1$ | - | 5 |
| | $M_2$ | - | - |
| yougen | Y | - | (14*) |

Suffixal expressions were also subcategorized mainly by their semantical functions in order to decompose limited types of the conceptual compounds in the experimental system.

The numerical outline of these refinements of the categories is given in Table 3. The asterisk in the table implies the subcategorization based on the inflectional type.

Refined Connection Rules. The connection rules were refined by using the finally obtained categories that amount to 142. The number of these rules is approximately 3,600.

Table 4 shows some examples of the rule and of the expression.

[†] i.e. to partition the set, $E = R_{NP1} \cup R_{NP2} \cup R_{PP1} \cup R_{PP2} \cup R_{PP3}$ by the following relation, $R$ into equivalence classes.

for $\forall x, y \in E$ $(xRy \underset{d.}{\leftrightarrow}$ for $\forall w_1, w_2 \in E$ $((x*w_1 \leftrightarrow y*w_1) \wedge (w_2*x \leftrightarrow w_2*y)))$, where $a*b \underset{d.}{\leftrightarrow}$ "a can be agglutinated by b"

Table 4. Examples of Refined Rules

| subcategory | examples of expressions (their meanings) | succeedable categories |
|---|---|---|
| R01 | 'ga'(AGENT,OBJ-1,•••),'no'(AGENT,•••) | |
| R02 | 'wo'(OBJ-1,•••) | R37,R38,R55,••• |
| R19 | 'wo motii te'(INSTRUMENT),'ni tui te'(OBJ-1,SITUATION),••• | R36~R38,R55,R56,••• |
| R23 | 'ni tui te'(NUMBER-2•RATE,•••),'atari'(NUMBER-2•RATE,•••) | |
| R27 | 'he'(DIRECTION),'made'(DIRECTION),••• | R36~R38,R55,R56,••• |
| R36 | 'ha'(AGENT[THEME],OBJ-1[THEME],•••) | |
| R37 | 'mo'(AGENT[ADDITION],•••),'mo mata'(AGENT[ADDITION],•••),••• | |
| R38 | 'koso'(AGENT[STRESS],OBJ-1[STRESS],•••) | R01,R02,R19,R36,••• |
| R55 | 'made'(AGENT[STRESS-OTHER],OBJ-1[STRESS-OTHER],•••) | R01,R02,R19,R36,••• |
| R56 | 'dake'(AGENT[LIMITATION],•••),••• | R01,R02,R19,R27,R36,••• |
| R62 | 'made'(AGENT[T-POINT],•••),••• | R01,R02,R19,R36~R38,••• |

## Experiment

### Overview

The Japanese sentence is ordinarily written in kana(phonetic) letters and Chinese(ideographic) characters without leaving a space between words. From the viewpoint of machine-processing, however, it is preferable to express clearly the units composing the sentence in such a way as to leave a space between every word as in English. We have no standard way of spacing the units though the need for this has been demanded for a long time. Supposing tentatively that a sentence is written in English letters with a space between each E-bunsetsu, we have developed an experimental system which decomposes the input E-bunsetsu into atomic expressions using the refined rules and decides its function. The system is overviewed as follows:

(1) The system consists of five components: a program; a dictionary of atomic expressions; a table of the connection rules; a paradigm; and a table of euphonical rules(not mentioned in this paper);
(2) Each entry expression is given one or more triple of the information in the dictionary. A triple consists of a code of the (refined) category such as A48 or R56, a code of the inflectional condition of the connection, and a code of the meaning;
(3) As to the inflectable expression, the dictionary includes only its stem;
(4) E-bunsetsu is decomposed from left to right on it by the "longest-match method" and all possible analyses are tried in the "depth-first" manner;
(5) The category code such as M13 or Y05, of the noun or yougen is used in the input and dictionary for the actual expression in it.

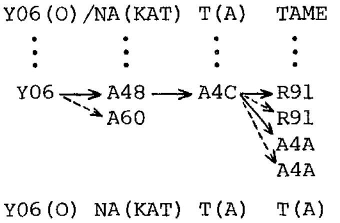
### Operational Examples

Operational examples follow. The string of letters parenthesized in the output description is the inflectional ending and '/' denotes the boundary between the conceptual expression and the annex-expression detected by the system. The arrows in the following illustration show the string of categories which corresponds to a leftmost substring of the input and is assured to be successful by both of the connection rules of the category level and the inflectional conditions given in the dictionary. On the other hand, the dotted arrow shows that the connection is allowed by the rule of the category level but not by the rule of the inflectional level.

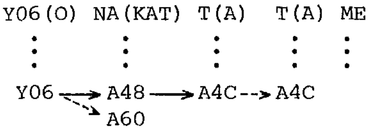
Example 1.

```
input = Y06ONAKATTATAME (来なかったため)
output :
segmentation = Y06(O)/NA(KAT) T(A) TAME
categories = Y06 A48 A4C R91
function = P MODIFYING P
```

Without checking the refined rules ( of two levels: the category level, and inflectional level), the following two decompositions would have been obtained.



1-1



1-2

While the decomposition 1-1 is successful, 1-2 was rejected because the auxiliary verb,'ta' is prohibited from being connected to the preceding auxiliary verb,'ta' by the inflectional rule. The triples given in the dictionary to 'tame' are as follows:

{R91; "connectable to adnominal forms of all types"; CAUSE•REASON };
{R91; "connectable to adnominal forms

```
              of verb types";  PURPOSE };
  {A4A; "connectable to adnominal forms
              of all types";  CAUSE•REASON };
  {A4A; "connectable to adnominal forms
              of verb types";  PURPOSE }.
```

In 1-1, since the inflectional type of 'ta' is not verbal, the second and fourth triples are not acceptable. In addition, the third one is unavailable since the ending form of the input E-busetsu results to be a stem, and inadequate. Finally, only the first one was accepted and at the same time the meaning of 'tame' was disambiguated.

Example 2.

```
input = Y08SANIMOTODUITESIKA(大きさに基づいてしか)
output :
segmentation = Y08 SA/NIMOTODUITE SIKA
categories = Y08 S45 R19 R42
function = N MODIFYING P
```

Without using the rules, the following three kinds of decompositions would have been possible.

```
Y08      SA/NIMOTODUITE  SIKA                2-1
 :        :        :       :
Y08 —> S45 ——> R19 ———> R42
                           R92

Y08      SA  NIMOTODUITE  SI  KA             2-2
 :        :        :       :   :
Y08 —> S45 ——> R19       R95

Y08      SA  NIMOTODUITE  S(I)  KA           2-3
 :        :        :        :    :
Y08 —> S45 ——> R19        S35
```

The atomic expression, 'si' in 2-2 and 'si' in 2-3, which are understood as a conjunctive verb, and a suffixal expression, respectively, can not be connected to 'nimotoduite'.

Example 3.

```
input = M14TEKINANODEHANAI.(効果的なのではない.)
output :
segmentation 1 = M14 TEKI/NANODEHANA(I)
categories 1 = M14 S29 A48
function 1 = P IN THE SENTENCE-FINAL POSITION
segmentation 2 = M14 TEKI(NA) NO/DEHANA(I)
categories 2 = M14 S29 S47 A18
function 2 = P IN THE SENTENCE-FINAL POSITION
```

The result was twofold according to two sorts of interpretations of 'no':the first one is to understand it has no special meaning; the second, it is a suffixal variant of the noun, 'mono'*('thing')*. There exist latently following eight different decompositions but only 3-1 and 3-6 were accepted by the rules.

```
M14    TEKI/NANODEHANA(I)                    3-1
 :      :          :
M14 —> S29 ————> A48
                  A18

M14 TEKI NANODE HA NA(I)                     3-2
M14 TEKI NANOD(E) HA NA(I)                   3-3
M14 TEKI(NA) NODEHANA(I)                     3-4
M14 TEKI(NA) NODE HA NA(I)                   3-5
M14 TEKI(NA) NO/DEHANA(I)                    3-6
 :    :       :    :
M14 —> S29    R70     A18
          \   R01   ↗ A48
           ↘  S47 /
M14 TEKI(NA) NO DE HA NA(I)                  3-7
M14 TEKI(NA) NO D(E) HA NA(I)                3-8
```

As for 3-6, it was understood that the atomic expression,'no' was not a particle(R70) which indicates a kakariuke relation between two nominal E-bunsetsus or a particle(R01) of the meaning of AGENT, but a suffixal expression(S47) which nominalizes the predicative expression.

Example 4.

```
input = M20DEKINAKUNARUTO(判断できなくなると)
output :
segmentation 1 = M20/DEKI() NAKUNAR(U) TO
categories 1 = M20 A24 A41 R92
function 1 = P MODIFYING P
segmentation 2 = M20/DEKI() NAKUNAR(U) TO
categories 2 = M20 A24 A41 R94
function 2 = P MODIFYING P
```

The decomposition was unique but the interpretation of 'to' was twofold as follows.

```
M20 / DEKI()  NAKUNAR(U)  TO
 :      :          :       :
M20 —> A24 ————> A41      R03
                     \ ↘  R19
                      \↘  R92
                        ↘ R94
                          R72
                          S90
```

In the first interpretation, 'to' is a conjunctive particle of the meaning,ASSUMPTION, and in the second, it is a particle of the meaning, QUOTATION. This ambiguity is, therefore, quite reasonable.

Results of Experiments

We show the results of experiments made for 162 E-bunsetsus in Table 5 and 6. The average number of atomic expressions composing an E-

Table 5. Ambiguity of Decomposition

| number of decompositions | number of E-bunsetsus |
|---|---|
| zero<br>(not decomposable) | 1 |
| one | 158 |
| two | 3 |
| more than or<br>equal to three | 0 |

Table 6. Ambiguity of Category Sequence

| number of category sequences per a single decomposition | number of decompositions |
|---|---|
| 1 | 145 |
| 2 | 12 |
| 3 | 1 |
| 4 | 3 |
| 5 | 2 |
| 8 | 1 |

bunsetsu fed to the system has been 4.8. The ambiguities of both the decomposition and the category sequence have been reduced sufficiently. Most of the ambiguities left by the system have been quite reasonable in the sense that further reductions of them would require more detailed information from the outside of the E-bunsetsu. In addition, the ambiguities to be left to higher phases of parsing for reduction have not been reduced by the system.
As exemplified in Example 1., the disambiguation of the atomic expression's meaning is carried out by selecting the triple of functional information given in the dictionary. Nine percent of the entry expressions are given plural triples and then their meanings can be reduced by our rules on the bases of its structural surroundings in the E-bunsetsu.

## Conclusions

Extending the domain of input sentential forms of the natural language processing system enables, in principle, the system to manipulate more precice or delicate meanings and to communicate with men more naturally. Our grammatical model presented in this paper is so comprehensive that the local structures of colloquial and written sentences actually used in everyday life can almost always be analyzed with it.
It is also elaborate enough to reduce the syntactic and semantic ambiguities of the local structure. It should be noted that the local structure analyzed by our grammar plays a quite important role in the Japanese language processing because it is not only a larger structure which can include idiomatic strings of words than a bunsetsu, but also a syntactic and semantic unit for sentence construction.
Every atomic expression, which is the smallest component of the sentence, has been chosen to have undividable and self-supported meaninqs. Though we have not mentioned it in detail in this paper, we have already settled extensively the meanings of annex-expressions by classifying them.

## REFERENCES

[1] K.Shudo:"On Machine Translation from Japanese into English for a Technical Field", Information Processing in Japan,14 (1974).

[2] K.Shudo,H.Tsurumaru & S.Yoshida:"A Predicative Part Processing System for Japanese-English Machine Translation"(in Japanese), the Trans. of the IECE of Japan,J60-D,10 (1977) --- Abstract in English,E60-D,10 (1977).

[3] K.Shudo,T.Fujita & S.Yoshida:"On the Processing of Annexational Expressions in Japanese", the Proc. of the 7th International Conference on Computational Linguistics (COLING 78) (1978).

[4] K.Shudo,T.Narahara & S.Yoshida:"A Structural Model of Bunsetsu for Machine Processing of Japanese"(in Japanese), the Trans. of the IECE of Japan, J62-D, 12 (1979) --- Abstract in English, E62-D, 12 (1979).

[5] K.Shudo:"Studies on Machine Processing of Japanese Using a Structural Model of Bunsetsu"(in Japanese), the Bulletin of the Institute for Advanced Research of Fukuoka University, 45 (1980).