AN AUTOMATIC TRANSLATION SYSTEM OF NON-SEGMENTED KANA SENTENCES INTO KANJI-KANA SENTENCES

Hiroshi Makino and Makoto Kizawa

Faculty of Engineering Science, Osaka University
Machikaneyama-cho, Toyonaka, Osaka 560, JAPAN

University of Library and Information Science
Yatabe-machi, Tsukuba-gun, Ibaraki-ken 305,
JAPAN

Summary

This paper presents the algorithms to solve the two main problems comprised in the automatic Kana-Kanji translation system, in which the input sentences in Kana are translated into ordinary Japanese sentences in Kanji and Kana : the segmentation of non-segmented sentences into Bunsetsu and the word identification from homonyms. Employing this algorithm, non-segmented Kana input sentences could be automatically translated into Kanji and Kana output sentences with 96.2 per cent success.

Introduction

In the computer processing of the Japanese language informations, the input method is much more difficult than in other Indo-European languages because thousands of kinds of characters in mainly two classes, Kanji(ideograms) and Kana(phonograms), are used together in writing regular sentences.

Conventional Japanese typewriters are equipped with least 2000 Kanji(Chinese characters) which are frequently used in daily use. A typewrite of this sort is difficult for us to handle and its typing speed is much lower than that of alphabetic typewriters because operators must look for characters one by one.

One of the most promising input methods to overcome this intrinsic input difficulty is Kana-Kanji translation system, in which all the sentences are input with Kana only using a regular 44-Key keyboard and then translated into regular Kanji-Kana sentences automatically in the computer.

The automatic translation system consists of two processes; the segmentation and the word identification processes.

The problems in Kana-Kanji translation

The problems in Kana-Kanji translation are:
(a) segmentation of input sentences.
(b) word identification from homonyms.

These problems are basic in the processing of Japanese sentences as language informations.

Japanese sentences in Kanji and Kana have no spaces between words as English ones do. However, in order to make the computer process Kana sentences easy, it would be necessary to put a space as a segmental symbol between words or some units in sentences. Therefore, some spacing methods, listed in Fig.1(concluding non-segmented sentence for convenience), was already adopted in Kana-Kanji translation systems.[1-3]

現在人類は勝れた目と指先の感覚を持っている。

(1) genzai jinrui ha sugure ta me to yubisaki no kankaku wo mot te iru.

(2) genzai jinrui ha sugure ta me to yubisaki no kankaku wo mot teiru.

(3) genzai jinruiha sugureta meto yubisakino kankakuwo motteiru.

(4) genzaijinrui ha sugu reta me to yubisaki no kankaku wo mot teiru.

(5) genzaijinruihasuguretametoyubisakinokankakuwomotteiru.

(1) segmented between words
(2) segmented between an independent word and a sequence of dependent words
(3) segmented between Bunsetsu
(4) segmented between Kanji and Kana
(5) non-segmented

Fig.1 Examples of segmentations in a Japanese sentence.

However, these pre-editing methods of word segmentation or unit segmentation are not only an too laborious for most of the Japanese people who are not accustomed in segmenting each sentence into words but also apt to be erroneous. It is, therefore, necessary in Kana-Kanji translation system to segment the Kana strings into words or other units automatically.

The number of different syllables in Japanese is much less than in English or in Chinese, while the number of Kanji is much more. Consequently, there are many groups of Kanji which have the same pronunciation. This fact makes word identification more difficult in Kana-Kanji translation since there is no one-to-one correspondence between Kanji and Kana. For example, Kana strings 'コウセン' corresponds to 25 words in an ordinary dictionary and a part of these are shown below.

Example.

Kana	Kanji	a meaning
コウセン	交戦	a battle
	抗戦	a resistance
	鋼船	an iron ship

光線	a beam
公選	a public election
口銭	a commission
鉱泉	a mineral spring

The segmentation process

Bunsetsu

A Japanese sentence is composed of the sequences of syntactic units called Bunsetsu pronounced without pausing. Bunsetsu usually consists of two parts: an independent part and a dependent part. The independent part consists of an independent word or its derivative, and the dependent part consists of a sequence of dependent words, given as follows:

Bunsetsu=(independent part)·(dependent part)
independent part
=[prefix]·(independent word)·[suffix]
dependent part
=[dependent word]*

independent word=noun/pronoun/adverbs/
verb/adjective/verbal adjective/
attributive/conjuction/interjection

dependent word=auxiliary verb/particle or
postposition

Here, brackets indicate optionality, the asterisk indicates one or more repititions or non-existing and the slants indicate alternatives.

The independent words('jiritsugo') are divided into two main groups: inflected words which consist of verbs, adjectives and verbal adjectives('keiyodoshi'), and non-inflected words which consist of nouns, pronouns and others. On the other hand, the dependent words consist of particles and auxiliary verbs which have their inflections.

There are grammatical connectabilities between a preceding word and its succeeding word in Bunsetsu. This is explained using an example in Fig.2.

$$\underbrace{\text{ika}}_{\text{V}}\underbrace{\text{nakere}}_{\text{AUX}}\underbrace{\text{ba}}_{\text{P}}\underbrace{\text{nara}}_{\text{AUX}}\underbrace{\text{naka}}_{\text{AUX}}\underbrace{\text{tta}}_{\text{AUX}}\ \text{(had to go)}$$

V:verbs, AUX:auxiliary verb, P:particle

Fig.2 An example of Bunsetsu

An indicative form 'ika' of a verb 'iku' can be concatenated not only by inflectional form 'nakere' of auxiliary verb 'nai' in this example but also by all of inflectional forms of 'nai'. And the particle 'ba' is preceded by the conditional form of 'nai'. Thus, these properties are decided upon each inflectional form of the preceding word(if the word is an inflected word) and its succeeding word. These connectability features in Bunsetsu constitute the basis of the segmentation of Kana strings described in later sections.

The longest string-match method of two Bunsetsu

For segmentation, each independent word is, in the order of length, first separated by comparing the Kana strings with the vocabulary of a word dictionary, and is stored with the informations such as parts of speech and inflectional forms if necessary for further morhological analysis.

Then, the dependent words in the rest of the strings are recognized using the dependent-word list and grammatical connectabilities between the dependent word and the independent word are examined. This analysis is continued until no succeeding word is found in the successive Kana strings. Thus, the candidates of a Bunsetsu are extracted from Kana strings as below.

Example.

souiuzassiwo ... (a part of strings)

soui ...	(noun)
sou·iu ...	(adverb·auxiliary verb)
sou ...	(verb)

The same analysis as mentioned above is executed for the rest of the strings from which each candidate of Bunsetsu is separated.

Consequently, the sequence of two candidates of Bunsetsu is extracted from Kana strings, and then the Bunsetsu in the sentence is appropriately identified so as to make the total length of two consecutive strings of their candidates maximum. This algorithm decides only the boundary between two consecutive Bunsetsu. In other words, the preceding Kana strings and these constituents for the Bunsetsu are recognized. On the other hand, the decisions for succeeding Bunsetsu are tentative at this stage.

These processes named as the longest string-match method of two Bunsetsu[4] are executed sentence by sentence and at length the input sentences are converted into Bunsetsu and homonyms in Bunsetsu are stored. An example is illustrated in Fig.3.

souiuzasshiwo...

1) souiu zasshiwo...
2) soui...
3) soui iu...

Fig.3 Segmentation process of Kana strings by the longest string-match method of two Bunsetsu.

The successive candidates of Bunsetsu in 1) and 3) are compared since the succeeding Kana strings are not analyzed in 2). As the total length of two analyzed strings in 1) is longer than that in 3), the segmentation in 1), namely the Bunsetsu 'souiu' is decided as the result.

The proccessing of unknown words

The longest string-match method of two Bunsetsu is based on the grammatical characterristics of the words, and so is not applicable to unknown words to the word dictionary. Hence, it would be easily expected that the appearance of an unknown word in a sentence makes the segmentation impossible. Therefore, it is necessary in non-segmented sentences to take account of the processing of unknown words.

The dependent words are divided into two main groups by their connectability characteristics. One is the word class, named as A, that is preceded by nouns or non-inflected words. The other is the word class that is preceded by inflected words and is further sub-divided into four sub-classes, named as B, C, D and E, according to the preceding word conjugations which are of indefinite form, conjunction form, final form and conditional form, repectively. The dependent words and their classes of connectabilities are given in Table 1.

Table 1 Classification on connectability of dependent words.

words	class	words	class
no	A	ya	A
ni	A	u	B
te	C	nado	A
wo	A	dake	A
ha	A	zu	C
ta	C	demo	A
ga	A	yori	A
da	A	nagara	C
de	A	tara	C
to	A	n'	B
mo	A	tari	C
nai	B	shi	D
masu	C	rashii	A
kara	A	beki	D
desu	A	naku	C
he	A	bakari	A
ka	A	shika	A
ba	E	taru	A
made	A		

Now, suppose that the search for the word dictionary fails. Then, the word in the above dependent word list is searched for the rest of the strings without being segmented. If a dependent word is found and its preceding Kana corresponds to an inflected word-ending succeeded by it —— vowels of inflectional endings of indefinite, conjunction, final and conditional forms are '-a', '-i' or '-e', '-u' and '-e', respectively, then the dependent word is recognized and its succeeding Kana strings are analyzed morphologically as mentioned in the preceding section. Consequently, the dependent word sequences are extracted and utilized for next segmentation.

The word identification process among homonyms

As mentioned above, a part of words in input sentences is identified in grammatical or morphological analysis. But there are still many homonyms which have the same grammatical characteristics in general. Therefore, further word identification will need for syntactical and semantical analyses in a given sentence.

The usage dictionary

The usage dictionary contains the informations of word uses which play an important role on word identification from homonyms.

Informations of word uses would be divided into two groups: colloqual information of words such as derivatives, compound words and ideoms, and semantic informations such as "semantic pattern" representative of nouns and verbs.

Case relations accompanied with verbs in a sentence are explicitly marked with particles attached by nouns. Usually, the particles 'ga', 'wo' and 'ni' indicate nominative, objective and dative respectively, whose case relations are fundamental, and so these are called 'ga' case, 'wo' case and 'ni' case, respectively. Accordingly, the so-called case frame of each verb has been studied with an emphasis on these particles.

Example.

[watashi] ga aruke	[I] walk
[hon] wo yomu	read [book]
[mono] ni sawaru	touch [thing]

where, [x] means a semantic feature or semantic category of x.

One of difficulties of doing the work is the semantic classification of each word. To avoid this burden, the semantic category of each word is identified according to the system of "The Word List by Semantic Principles" edited by the National Language Research Institute, in which about 32,600 words are divided into 798 semantic categories.[5]

The particle 'ni' also occurs after locative noun which mean the location. However, it is empirically assumed that either locative nouns or dative nouns occur with each verb in a simple sentence. The example is given as follows,

[hito] ni [ie] ni itta

...said to the men to the house...
...went

The above example is unusual and this fact means that semantic features of nouns with 'ni' are derived from surface structures of sentences.

The case frame[6] of each verb is different, and so semantic categories of nouns and standard particles used as semantical "identifiers" are described in the usage dictionary.

Example.

Kaku	:	[hito] ga [ji] wo [kami] ni [dougu] de
write	:	HUMAN LETTER PAPER INSTRUMENT
iku	:	[hito] ga [basho] kara [basho] he
go	:	HUMAN LOCATION LOCATION

The particles 'de', 'kara', and 'he' with respective semantic categories are filled up in the usage dictionary in the above example.

For adjectives and verbal adjectives, semantic categories of nominative nouns are only filled up in the usage dictionary. The example is given as follows:

utsukushii	:	[hana] ga
beautiful	:	FLOWER
kireida	:	[hana] ga
pretty	:	FLOWER

Where, 'kireida' is a verbal adjective in Japanese which corresponds to an adjective in English. As a result, we have investigated "semantic pattern" for 3421 inflected words which consist of verbs, adjectives, verbal adjectives and verbs conjugated with 'suru' which are called 'sahenmeishi', since their word stems are regarded as nouns in Japanese. These words are extracted from the vocabulary frequency table edited by the National Language Research Institute.[7]

On the other hand, informations about nouns, namely, their derivatives composed with prefixes and suffixes, compound words and idioms are collected from an ordinary dictionary.[8] The example of a part of the usage dictionary is illustlated in Fig.4.

Item / Index	prefix	suffix	compound word, idiom	case			
				ga	wo	ni	others
労働	重	際,者,省	委員，運動 基準金庫組合	[人]			デ [場所]
気			利く，する 付く,付ける				
読む				[人]	[図書]		
呼ぶ				[人]	[人]		

Fig.4 A part of the usage dictionary

The parsing

After segmenting sentences into Bunsetsu, the parsing phase begins, in order not to take out so-called tree structures but to extract the syntactic relations between Bunsetsu or words. The parsing of the sentence is executed on the basis of the Kakariuke relations(something like the dependency relations) between Bunsetsu. The Kakariuke is the term in Japanese traditional school grammar.

Characteristics of Kakariuke relations in a sentence are given as follows:

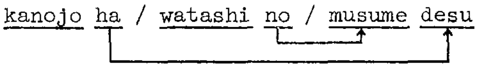
(1) A final word or an inflectional form in a Bunsetsu decides what kinds of words to modify, on the other hand each of the independent words decides how to be modified.
(2) Each Bunsetsu as a dependent always appears before its governor in a sentence.
(3) Kakariuke relations between any two Bunsetsu do not cross with each other in a sentence.

For simplicity of the parsing, we adopted the following two assumptions that would be correct in most sentences.

(4) A Kakariuke relation is decided on the smallest distance between a dependent and its probable governors.
(5) Each Bunsetsu can be a dependent of only one Bunsetsu appearing after it except the Bunsetsu at the end of a sentence.

The relations among Bunsetsu are searched taking account of the following three factors: five conditions mentioned above, final word as a dependent and an independent word class as a governor. The term noun phrase is used for Bunsetsu in which an independent part is a noun, and similarly a verb phrase for Bunsetsu consisting of a verb and its dependent part. But, for the phrase of the form of a noun and some of auxiliary verbs, which are called as copulas ('desu', 'da' etc.), it is necessary to regard the phrase as a predicate in a sentence.

Example



kanojo ha / watashi no / musume desu

(She is my daughter.)

In the above example, an underline denotes a word and a slant does a segmental symbol between Bunsetsu. An arrowed line denotes the Kakariuke relation between Bunsetsu. Usually, the Kakariuke relation between Bunsetsu, 'watashino' and 'musumedesu', is determined by the particle 'no' and the noun 'musume', on the other hand the relation between 'kanojowa' and 'musumedesu' is determined by the particle 'ha' and the auxiliary verb 'desu'.

The pre-processing for the word identification

In Japanese, the different semantic relations are reduced to the same syntactic relations of verbs with nouns intermediated by particles in active voice as in passive voice. The passive or causative voice is represented explicitly by the attachment of auxiliary verbs ('reru, rareru') or auxiliary verbs('seru, saseru') to inflectional forms of verbs. Accordingly, the semantic normalization is necessary in the cases below.

(i) passive:
N1 ga N2 ni V+reru(or rareru).
→ N2 ga N1 ni V.
(ii) causative:
N1 ga N2 ni N3 wo V+seru(or saseru).
→ N2 ga N3 wo V.

where N1, N2 and N3 denotes a noun and V denotes a transitive verb. The auxiliary verbs (reru and seru) are used for the consonant conjugation verbs(godan katsuyo doshi), on the other hand the auxiliary verbs(rareru and saseru) for the vowel conjugation verbs(ichidan katsuyo doshi).

The meaning of independent part which consists of an independent and a suffix is substituted for the meaning of its suffix. Similarly, the meaning of the number that consists of the set of the numeral plus counter is representative of the meaning of its counter.

Example

[nihon+jin] → [jin]
[100+nin] → [nin]

where,'jin'and'nin' are a suffix and a counter, respectively that mean the word "human".

The dependent part composed of more than two dependent words are substituted for a dependent word representing a case in order to consult the usage dictionary in next steps.

Example

Tokyo·he·mo itta → Tokyo·he itta

(went to Tokyo, too) (went to Tokyo)

Word selections from homonyms

Word selections from homonyms an executed using both colloqual informations and informations about cases with verbs.

Word selection based on noun-to-verb relation

Word selections from homonyms are executed particle attached to each noun. At that time, each particle is converted into the "standard particle" in the preprocessing phase. And so, each semantic category of homonyms (nouns) is compared with the corresponding semantic category code in the usage dictionary, and the most matched word is selected. When homonyms are verbs, the verb and the nouns as case elements of the verb are selected taking account of the numbers of case found in the sentence. The nouns related with verbs intermediated by the particle 'no' are referred to the nominative nouns. As it is assumed that the noun attached by copulas such as 'desu' are in the synonymous relation to nominative nouns, each pair is selected from homonyms.

As it is difficult to estimate the case relations between verbs and nouns modified by their verbs because of no occurrence of particles, the reference to the case elements not identified yet are tried. In the example below, the words 'hon' are examined whether they are nominative or objective elements of the verb 'morau'.

Kare ni moratta hon (book received from him)

hon ga morau

hon wo morau

Word selection based on noun-to-noun relations

For the Bunsetsu composed of prefixes and/or suffixes and independent words, the derivative is decided according to their prefixes and suffixes in the usage dictionary.

When the successive nouns are found, each registration is examined, and the registered word in the usage dictionary is selected if any.

Informations as for idioms are also, referred for nouns and verbs in the Kakariuke relation because their words are identified in colloqual expressions. In the sequence of two nouns, either of which is 'sahenmeishi', it is often assumed that the semantical relation between two nouns is based on the case relation because 'sahenmeishi' also have the characteristics as verbs.

Example

jouhou shori (information processing)

jouhou wo shorisuru (... process informations...)

The semantic category of alternative nouns 'jouhou' are compared with semantic categories of case elements of a verb 'shori + suru' are so '情報' is selected from homonyms(上方,乗方, etc.)

As it is assumed that two nouns intermediated by the conjunctive particles('to', 'ya', 'dano', 'nari', etc.) are in the relation of the same or similar semantic categories.

The pair of nouns is selected, whose semantic category codes are close to each other. A synonym and antonym are included in the same semantic category as shown in the following example.

Example

Sensei to reiju

(absolutism and slavery)

The most frequent word is selected for homonyms undetermined by the analysis of word uses.

Implementation

Dictionaries

The dictionaries for this Kana-Kanji translation system are given in Table.2 with a brief explanation.

(a) The independent word dictionary
The contents consist of sequential numbers, indexes of Kana, Kanji representation, numbers of Kanji, inflectional forms, word frequency, semantic category and information for dictionary search.
This dictionary has about 8000 independent words chosen from "Vocabulary and Chinese Characters in Ninety Magazines of Today."[7]

(b) Connection matrix
The connectability between preceding words and succeeding words in Bunsetsu is represented by the matrix, in which each row corresponds to the preceding words or their conjugations and each column to the succeeding words. Each element takes the value of 1 or 0, and 1 stands for that words of row are connectable to the succeeding words of the column.
The size of this matrix is 154 X 108.

(c) The table of inflectional word endings
For analyzing three inflected words(verbs, adjective and verbal adjectives), their conjugations and their correspondences to each row of connection matrix are listed, because these occur before dependent words in Bunsetsu.

(d) The dependent word list
This list consists of dependent word (particles and inflectional forms of auxiliary verbs) and their correspondence of rows and columns of the connection matrix.

(e) The prefix, the suffix and the counter dictionaries
These dictionaries include 47 prefixes, 311 suffixes and 141 counters, respectively, and also their Kanji representations. Moreover, the suffix and the counter dictionaries include their semantic category codes.

(f) The dependent list for segmentation
The dependent list consists of the words and their classes listed in Table 1.

(g) The usage dictionary
This dictionary have contents such as in Fig.2.

Table 2 List of dictionaries

(a) The independent dictionary
(b) The connection matrix
(c) The table of inflectional endings
(d) The dependent word list
(e) The prefix, the suffix and the counter dictionaries
(f) The usage dictionary

The system

The automatic Kana-Kanji translation system was inplemented on FACOM 230-45S equipped with 256 kilobyte memory. The programs in PL/I consist of 17 sub-programs.

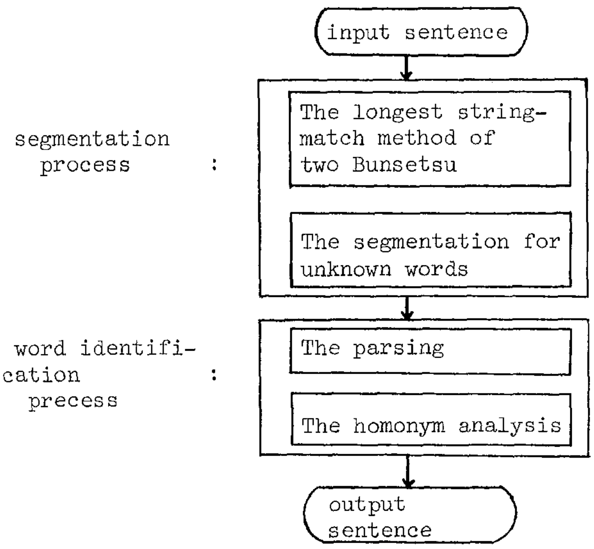

input sentence
segmentation process :
The longest string-match method of two Bunsetsu
The segmentation for unknown words
word identification precess :
The parsing
The homonym analysis
output sentence

Fig. 5 The flow of Kana-Kanji translation

Input sentence :

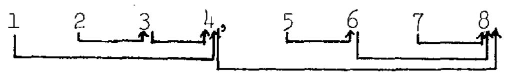
カレハシンブンキジヲヨンデ，ジョウキョウノ
ヘンカニキガツイタ.

(I) Segmentation process

1)	2)	3)	4)	5)	6)
彼は /	新聞 /	生地を /	呼んで， /	状況の /	変化に /
		記事	読ん	上京	返歌
		木地			

7)	8)
木が /	付いた.
気	着い
期	突い

(II) Parsing



(III) Output sentence

彼は新聞記事を呼んで，状況の変化に気が付いた.

Note: Words are arranged in their frequency order in (I). Arrowed lines denote the Kakariuke relation between Bunsetsu.

Fig. 6 An example of Kana-Kanji translation process.

An input sentence is first segmented in Bunsetsu, and second Kana homonyms in Bunsetsu are identified, consequently transformed into Kanji and Kana sentence. These processes are executed alternatively in a sentence as illustrated in Fig.5.

An Example of Kana-Kanji translation process is illustrated in Fig.6.

(I) in Fig.6 shows segmented Bunsetsu and homonyms and (II) shows Kakariuke relations between Bunsetsu, on the basis of that relations in (II),

case relation:	(記事を) ,	(読んで)
idiom :	(気が) ,	(付いた)
compound word:	(新聞) ,	(記事を)
'sahenmeishi':	(状況の) ,	(変化に)

each word is selected from homonyms. At a result, the output sentence is acquired in (III).

Experimental Result

In order to evaluate translation efficiency, 2592 Bunsetsu in 214 sentences were chosen from various literatures, magazines, articles etc.

Results of the experiment is shown in Table 3.

Table 3 Experimental result

	segmentation	translation
correct	98.8 %	96.2 %
error	1.2 %	3.8 %

Translation errors are classified into segmentation errors and word selection errors.

Segmentation errors are divided into errors caused by the longest string-match method of two Bunsetsu, unknown word and grammatical incompleteness, whose examples are denoted at (1), (2) and (3) in Table 4, respectively.

Errors by the longest string-match method of two Bunsetsu occurred on seven boundaries of Bunsetsu in the data.

On the other hand, word selection errors are apparently due to the uses of word frequencies. However, the true causes of errors are due to incompleteness of homonym analysis. They are given as follows; not taking account of the segmentical relation underlying between nouns formed with the noun phrase pattern"noun + 'no' + noun", not identifying the meaning of pronoun in context, not identifying the ambiguities between case relations and other semantic relations, for example, such as adverbial relation for verbs, Their examples in the data are illustrated in (4), (5) and (6) of Table 4, respectively. Appendix shows examples of the segmented sentences and the corresponding sentences in Kanji and Kana.

Table 4 Examples of errors

	Erroneous	Correct
1)	確か似合った	確かに合った
2)	ポツポツミエテイテ	ポツポツ見えていて
3)	泣き香の	無きかの
4)	時刻の主権	自国の主権
5)	これを犯しては	これを冒しては
6)	左官に使われ	さかんに使われ

* Katakana shows the segmentation based on dependent word only.

Conclusion

We have proposed new approach for two main problems: segmentation of sentences into Bunsetsu and homonym analysis, in automatic Kana-Kanji translation, which should be basic linguistic problems. Moreover, an experimental system was constructed to make sure of their efficiency. As a result of experiments 96.2 per cent of the whole Bunsetsu in input sentences were seccessfully translated into Kanji where they should be.

For promoting applicabilities of this system, we are going to prepare the dictionary including about 30,000 words in daily use.

The difficulties in Kana-Kanji translation is based on ambiguities about the utterance, accordingly, further studies on understanding sentences would be needed for overcoming these difficultes.

Acknowledgements

We would like to thank Mr. Masakazu Okada for his cooperation in this work.

The research described in this paper was partially supported by the Ministry of Education Science and Culture in 1979.

References

[1] I. Aizawa and T. Ebara, "Machine Translation System of 'Kana' Presentations to 'Kanji-Kana' Mixed Presentations." NHK. Tech. Res., pp. 261-98(1973).

[2] Y. Matsushita, H. Yamazaki and F. Sato, "Kana Alphabet to Kanji Converting System." JOHOSHORI, Vol. 15, No. 1, pp. 2-9(1974).

[3] H. Makino, M. Kizawa and Y. Katsube, "Transformation of Kana-input into Kanji-presented Sentence." JOHOSHORI. Vol. 18. No. 7, pp. 656-63(1977).

[4] H. Makino and M. Kizawa, "Automatic Segmentation for Transformation of Kana into Kanji" Trans. of Inf. Proc. Society of Japan, Vol. 20. No. 4, pp. 337-45(1979).

[5] The National Language Research Institute, "The Word List by Semantic Principles" p. 362, SYUEI SYUPPAN, Tokyo, Japan (1973).

[6] C. J. Fillmore, "The Case For Case" in Universals in Linguistic Theory, Holt, Rinehart and Winston, New York (1968).

[7] The National Language Research Institute, "Vocabulary and Chinese Characters in Ninety Magazines of Today" p. 321, SYUEI SYUPPAN, Tokyo, Japan (1962).

[8] K. Kindaichi edited, "SHIN-MEIKAI KOKUGO JITEN", SANSEIDO, TOKYO (1971).

Appendix.

```
ニッポン　コクミンハ，セイトウニ　センキョサレタ
コッカイニオケル　ダイヒョウシャヲ　ツウジテ　コウ
ドウシ，ワレラト　ワレラノ　シソンノタメニ，ショコ
クミントノ　キョウワニヨル　セイカト，ワガクニ　ゼ
ンドニ　ワタッテ　ジユウノ　モタラス　エイタクヲ
カクホシ，セイフノ　コウイニヨッテ　フタタビ　セン
ソウノ　サンカガ　オコルコトノ　ナイヨウニスルコト
ヲ　ケツイシ，ココニ　シュケンガ　コクミンニ　ソン
スルコトヲ　センゲンシ，コノ　ケンポウヲ　カクテイ
スル.
ソモソモ　コクセイハ，コクミンノ　ゲンシュクナ　シ
ンタクニヨルモノデアッテ，ソノ　ケンイハ　コクミン
ニ　ユライシ，ソノケンリョクハ　コクミンノ　ダイヒ
ョウシャガ　コレヲ　コウシシ，ソノ　フクリハ　コク
ミンガ　コレヲ　キョウジュスル，コレハ　ジンルイ
フヘンノ　ゲンリデアリ，コノ　ケンポウハ，カカル
ゲンリニ　モトヅクモノデアル.
ワレラハ　コレニ　ハンスル　イッサイノ　ケンポウ，
ホウレイ　オヨビ　ショウチョクヲ　ハイジョスル.
ニッポン　コクミンハ，コウキュウノ　ヘイワヲ　ネン
ガンシ，ニンゲン　ソウゴノ　カンケイヲ　シハイスル
スウコウナ　リソウヲ　フカク　ジカクスルノデアッテ，
ヘイワヲ　アイスル　ショコクミンノ　コウセイト　シ
ンギニ　シンライシテ，ワレラノ　アンゼント　セイゾ
ンヲ　ホジショウト　ケツイシタ.
ワレラハ　ヘイワヲ　イジシ，センセイト　レイジュウ，
アッパクト　ヘンキョウヲ　チジョウカラ　エイエンニ
ジョキョショウト　ツトメテイル　コクサイ　シャカイ
ニオイテ，メイヨ　アル　チイヲ　シメタイト　オモウ，
ワレラハ　ゼンセカイノ　コクミンガ　ヒトシク　キョ
ウフト　ケツボウカラ　マヌガレ，ヘイワノウチニ　セ
イゾンスル　ケンリヲ　ユウスルコトヲ　カクニンスル.
ワレラハ，イズレノ　コッカモ，ジコクノコトノミニ
センネンシテ　タコクヲ　ムシシテハ　ナラナイノデア
ッテ，セイジ　ドウトクノ　ホウソクハ，フヘンテキナ
モノデアリ，コノ　ホウソクニ　シタガウコトハ，ジコ
クノ　シュケンヲ　イジシ，タコクト　タイトウ　カン
ケイニ　タトウトスル　カッコクノ　セキムデアルト
シンズル.
ニッポン　コクミンハ　コッカノ　メイヨニ　カケ，ゼ
ンリョクヲ　アゲテ　コノ　スウコウナ　リソウト　モ
クテキヲ　タッセイスルコトヲ　チカウ.
```

(1) Kana sentences in automatically segmented Bunsetsu

```
日本国民は，政党に選挙された国会における代表者を通
じて行動し，我等と我等の子孫のために，諸国民との共
和による成果と，我国全土に渡って自由のもたらす恵沢
を確保し，政府の行為によって再び戦争の参加が起こる
ことの内容にすることを決意し，ここに主権が国民に存
することを宣言し，この憲法を確定する．そもそも国政
は国民の厳粛な信託によるものであって，その権威は国
民に由来し，その権力は国民の代表者がこれを行使し，
その福利は国民がこれを享受する．これは人類普遍の原
理であり，この憲法は，かかる原理に基づくものである．
我等はこれに反する一切の憲法，法令及び詔勅を排除す
る．日本国民は，恒久の平和を念願し，人間相互の関係
を支配する崇高な理想を深く自覚するのであって，平和
を愛する諸国民の公正と信義に信頼して，我等の安全と
生存を保持しようと決意した．我等は平和を維持し，専
制と隷従，圧迫と偏狭を地上から永遠に除去しようと勤
めている国際社会において，名誉ある地位を占めたいと
思う．我等は全世界の国民が等しく恐怖と欠乏から免れ，
平和のうちに生存する権利を有することを確認する．我
等は，いずれの国家も，時刻のことのみに専念して他国
を無視してはならないのであって，政治道徳の法則は，
普遍的なものであり，この法則に従うことは，時刻の主
権を維持し，他国と対等関係に立とうとする各国の責務
であるとシンズル．日本国民は国家の名誉に掛け，全力
を上げてこの崇高な理想と目的を達成することを誓う．
```

Note: Underlined words are in error.
Katakana denotes no analized strings.

(2) Output sentences in Kanji and Kana

Output examples (The preamble in the Constitution of Japan)