Some Topics of Language Processing for Machine Translation

MAKOTO NAGAO,
JUN-ICHI TSUJII

Dept. of Electrical Engineering,
Kyoto University,
Kyoto, Japan

One of the most important problems in computerized processing of written Japanese texts is the input/output problem of Japanese characters. The authors propose a possible solution to this problem, and describe a machine-implemented dictionary, compiled on this basis. They also present a variant approach to the morphological analysis of Japanese, along with its application to translation of scientific document titles.

1. MACHINE-READABLE DICTIONARIES

Machine-Readable Japanese Dictionary

Algorithms for natural language processing must be verified by a large corpus of text data from various fields. The dictionary for the language processing programs, therefore, must be large enough to cover vocabularies of a variety of fields. Texts in a certain field usually contain both the technical terms, which are specifically used in the field, and common words, which are used in almost every field. A conventional dictionary which is currently published and daily used consists almost exclusively of common words. We have completed the computerisation of such a dictionary. The dictionary we adopted was *Shinmeikai Kokugojiten* published by Sanseido Publishing Co., which contains about 70,000 lexical entries (see Table 1).

Table 1

Content of the dictionary

Part of speech	Type of inflection	Number of entries
Transitive verb	Type 1	88
Transitive verb	Type 2	463
Transitive verb	Type 3	1 208
Intransitive verb	Type 1	57
Intransitive verb	Type 2	726
Intransitive verb	Type 3	1 351
Intransitive verb	Type 4	271
Intransitive verb	Type 5	12
Interjection		146
Prefix		51
Suffix		64
Adjective 1	Adjective type 1	626
Adverb		1387
Pronoun		119
Noun		46279
Adjective 2	Adjective type 2	1 103
Adjective 3	Adjective type 3	252
Nominalisation form of verb		6760

Note: Postpositions, inflectional postpositions and conjunctives are omitted from this table

Because of the particular Japanese writing method, special techniques have been devised to utilize the dictionary that are not necessary for English and other European languages. These techniques are also useful for the dictionaries of technical terms and terminology data banks in Japanese.

Data Base System for Japanese Dictionary

The main difficulty we encountered in developing the dictionary consultation system derives from the fact that lexical entries in Japanese usually have more than one spelling. They have basically two different forms of spellings, a Kana-spelling and Kanji-spellings. Moreover, we often find mixed forms of these two spellings in ordinary texts. We must be ,able to retrieve the lexical description of an entry by using any one of these spellings.

In our dictionary system, the intermediate indexing structures are provided for Kana-spellings and Kanji-spellings as shown in Figure 1. Other spellings, mixed

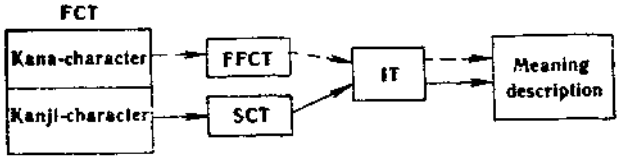


Fig. 1. Overall construction of the dictionary system: FCT — first character table, FFCT — first five characters table, SCT —second Kanji-character table, IT —item table

spellings, can be normalised into one of these two basic spellings by utilizing a certain auxiliary information. Each Kanji-character corresponds to three or four (or more) different Kana-spellings, and we can transform a mixed spelling into Kana-spellings by systematically changing the Kanji-character in the spelling into the corresponding Kana-strings. If the resultant Kana-spelling is found in the dictionary, it will be the correct one. We can also obtain Kanji-spellings from a mixed spelling by changing a Kana-string in mixed spelling into the corresponding Kanji-characters. Then we can utilize the index for Kanji-spellings to retrieve the lexical entry which has the given mixed spelling.

Another problem is the incorporation of the morphological analysis component into the dictionary consultation system. Because the inflection system for Japanese words is much richer than English, particularly because predicative words have many inflectional variants, the morphological analysis component is indispensable for the dictionary consultation system of Japanese.

Machine-Readable English-Japanese Dictionary

One of our ongoing projects is to computerize the English-Japanese dictionary, *New Concise English-Japanese Dictionary,* which contains about 90,000 lexical entries. The lexical entry for each English word contains not only Japanese equivalents (usually there exist more than one equivalent), but also:

(1) part of speech;
(2) pronunciation and stress position;
(3) plural forms (for nouns), inflected forms (for verbs);
(4) synonyms, antonyms, if they exist;
(5) compounds and derivatives, if they exist;
(6) usage examples of the word, and the translations in Japanese;
(7) idioms in which the word appears, and the translated equivalents of the idioms, etc.

To utilize the dictionary by computer programs, we should first transform the description for each word into a certain formatted record. Each item in the above list should be located in the fixed position of the formatted record. A method for translating the printed dictionary into such formatted records is to pre-edit the dictionary manually. However, as human pre-editing requires a lot of time and tremendous manual efforts, we decided that we would input the dictionary information almost as it is, without any manual pre-editing. After the completion of the input, a utility program will verify the input and translate it into formatted records. In order to make the development of such a program easy, we first developed a universal data verifier/translator for text data based on finite state automaton. By using this system, we can encode the verification and translation program very easily for almost any unformatted body of complex information. Without such a utility program dictionary information will not be arranged into the well-formatted structure which allows easy access by computer program.

The input of the dictionary has been almost completed at present. The formatted data or dictionary was supposed to be completed by the end of 1979.

2. LINGUISTIC DATA FOR JAPANESE LANGUAGE PROCESSING

One of the important issues in natural language processing is the robustness of a system. We understand the concept of robustness as meaning how huge or large language texts are, which can be processed by the system. In order to facilitate natural language processing on a large scale, we should provide comprehensive linguistic data. The Japanese dictionary and the English-Japanese dictionary are such linguistic data. Besides these, we should also collect and computerize other kinds of linguistic data, and those linguistic data should be organized appropriately from computational points of view. Such Linguistic data are as follows:

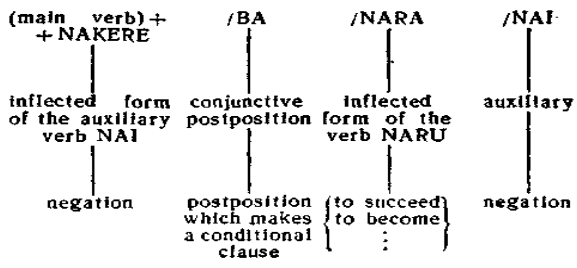
(1) Terminology data bank that contains specific terms in specific fields;
(2) List of inflectional suffixes of verbs, auxiliary verbs, and adjectives;
(3) List of function words in Japanese;
(4) List of suffixes and prefixes;
(5) List of complex Kana-strings which often appear in Japanese texts;
(6) List of sentential patterns.

(2), (3) and (4) are all for morphological analysis programs. We currently listed up 254, 150 and 70 items for these purposes respectively. We also have gathered other linguistic data to treat some exceptional cases in morphological analysis. In Japanese, content words such as nouns, verbs, adjectives and so on are usually written in Kanji-characters. On the other hand, function words such as postpositions, conjunctives, inflectional suffixes etc. are usually written in Kana-characters. Therefore, the type of character used gives us useful information for identifying words in written texts. However, there are many exceptions to these rules. Some function words are written in Kanji-characters and some content words are written in Kana-characters. To treat these exceptional cases, we should tabulate these exceptional words, and utilize them in the morphological analysis program. We currently have about 700 such exceptional words.

The list of complex Kana-strings (5) now consists of 2500 different Kana-strings. The units of words in different languages often disagree with each other. That is, a concept which is expressed by a single word in one language may be expressed by a complicated syntactic construction in the other language. This tendency is further emphasized if the two languages belong to quite different language families. Japanese and English are such a language pair. There are many English words whose concepts are expressed by phrases or other constructions in Japanese and vice versa.

Modals and aspects are usually expressed in English by auxiliary verbs such as *can, may, must, have -ed, be -ing* and in some cases, by certain specific constructions such as *it is impossible that S; it is necessary to VP.* Even some adverbs such as *necessarily* etc. can express modalities.

These expressions for modals and aspects are highly dependent on languages. In Japanese, these are expressed in quite different ways. They are usually marked by complex Kana-strings which follow predicative words in sentences. We can separate these Kana-strings into several component words by applying morphological rules of Japanese. For example, the complex Kana-string 'NAKEREBANARANAI' which marks the modality of 'necessity' or 'obligation' can be analysed into a sequence of words such as



So the whole meaning of this complex Kana-string is approximately 'If not (s), (it) will not { work / succeed / go well }'

Therefore, it is almost impossible to make the correspondence between the above Kana-string in Japanese and the English word 'should' which is the translation of 'NAKEREBANARANAI'. This shows that it is much better to handle a longer Kana-string as a whole rather than divide it into parts of speech.

Another kind of complex Kana-strings in Japanese is the strings that indicate 'cases'. The case relationships between noun phrases and verbs are usually marked by a relatively small set of postpositions in Japanese. However, certain complicated constructions which contain certain verbs sometimes mark 'cases'. The expression "NIYOTTE", for example, contains the inflected form of the verb 'YORU'. The verb 'YORU' has lost its original meaning in this string, and the string as a whole plays the role of a case postposition which marks the instrumental case.

We can find many such complex Kana-strings in ordinary Japanese texts. Currently, we have listed up about 2,500 such expressions. By using these expressions in the morphological analysis phase, we can reduce the complexity of input sentences. Further processings such as syntactic and semantic processings work on these reduced input sentences.

The list of sentential patterns (6) is for syntactic and semantic analysis of Japanese sentences. Many researchers are convinced that most of the sentential concepts are organized around predicative concepts which have certain case structures. Then the nominal concepts (usually expressed by noun phrases) in a sentence are related to a predicative concept (expressed by a verb or adjective). The problem of which argument or 'case' of the verb is filled in by what noun phrase is determined by the syntactic and semantic information. Therefore the semantic description of the noun phrase and the verb is very important. However, the syntactic or surface-level knowledge (that is, the knowledge under what syntactic environment the verb may appear and under what syntactic environment the verb and the noun phrase may be related to each other by which 'case' relation) is also very useful for the analysis.

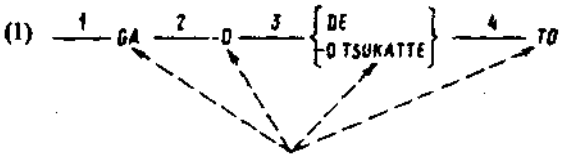
J. Bresnan, a linguist at MIT, proposed a grammatical framework in which lexical description Tor each verb plays the central role. In her framework, each verb is accompanied by:

(1) a description of the syntactic environments under which the verb appears;

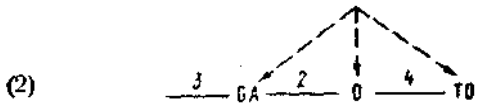
(2) the interpretation rules which show how to extract the deep, semantic relationships between the verb and the noun phrases from a given syntactic environment.

The syntactic environment in English can be represented naturally by an ordinary phrase structure tree, because most of the syntactic information in English is conveyed by the surface word ordering. A phrase structure tree can preserve the surface word order in natural manner. However, in Japanese the syntactic information is not conveyed by the surface word ordering, but instead, it is marked explicitly by the postpositions that follow noun phrases. Moreover, because a postposition can usually express several different deep cases in different syntactic environments, a case postposition can not be interpreted by itself. The interpretation of * postposition depends on the whole syntactic construction or environment: what other postpositions appear in the same sentence, and what verb governs the whole sentence. Furthermore, a certain verb strongly suggests the existence of certain postpositional phrases. Therefore, we are now accumulating the surface word usages or surface case patterns for Japanese verbs. We call the patterns 'sentential forms of the verb'. A single verb may have more than one sentential pattern. An example is shown in Figure 2. The same attempt is also being made by another group (Defence Agency of Japan). They restrict their domain to diplomatic documents. In their experimental system, 5600 Japanese verbs are classified into 576 groups depending on their sentential patterns.

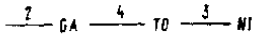
SHITE1-SURU *(to specify)*
surface sentential patterns



case postpositions or
complex Kana-strings indicating cases



(3) if the verb SHITEI-SURU *(to specify)* is followed by the postpositional expression TEARU (this expression transforms a verb which describes an action into a verb which describes a state), the sentential patterns



Note: The numbers 1, 2. 3 and 4 roughly correspond to the deep cases 'agent', 'object', 'instrument', and 'result', respectively.As you can see, the surface case marker or case postposition 'GA' is the example (1); (2) and (3) represent different deep cases. That means the interpretations of the postposition 'GA' depend on the whole sentence patterns.

Fig. 2. Examples of sentential patterns

3. MACHINE TRANSLATION SYSTEM FOR TITLES OF SCIENTIFIC DOCUMENTS

By starting the research into machine translation systems, we first selected an easy domain, that is, the titles of scientific documents. They have no contextual information at all and the words in this domain are usually less ambiguous than those in other domains. We restrict our domain to a more limited area of electrical engineering documents. We collected one hundred thousand titles in this area. Our ultimate goal is to translate them correctly. We think the amount of one hundred thousand texts is large enough to check the robustness of the system. After one year of research, we have corn-

pleted a prototype system. Now, we are going to develop the second, augmented version of this prototype system.

One of the important research objectives of this project is to develop a practical system which is feasible from an economical point of view. The system should be able to treat large numbers of vocabularies. Therefore, we can not incorporate into our system a sophisticated semantic analysis component which requires detailed semantic descriptions for words. The dictionary which we have at present contains 9,800 items in the field of electrical engineering, but it is still not large enough. We are adding new lexical items and revising the dictionary descriptions through experimentation. The outline of the system is as follows:

Step 1. Morphological analysis of words in English titles.

Step 2. Dictionary look-up: Because many technical terms or concepts are expressed by a set of words, not by a single word, we listed those expressions in the dictionary. By looking up this dictionary, certain sequences of words are reduced into single words and treated in the following processing as if they were single lexical units. Some idiomatic expressions are also processed at this step.

Examples: Settling time→(n)
time varying→(adj)
perpendicular to→(prep)
based on→(prep)

Step 3. Processing of local 'AND' construction: In general, it is very difficult to determine the scope of the conjunctive phrases. However, certain specific conjunctive constructions are easily analysed by considering only the local syntactic environment. At present, only such conjunctive phrases are processed in order to make the succeeding processing easier.

Step 4. Processing of consecutive noun-noun constructions: In scientific documents, especially in their titles, many complex concepts are often expressed by long concatenated noun — noun sequences. In some cases, adjectives or adverbs are also imposed in this sequence. It may require a very sophisticated semantic analysis to determine the internal structures of such sequences. Therefore, in our present system, the- internal structures of such noun — noun sequences are not analysed. We only identify that such constructions exist We provided a transition network only to accept such noun — noun constructions, and not to produce the corresponding internal structures. At the generation stage, such noun—noun constructions in English will be translated into corresponding noun—noun constructions in Japanese. Fortunately, the word orders in such constructions are almost the same in English and Japanese.

Examples: extremely thin SiO films→films (n)
high current arc discharges→discharges (n)

Step 5. Determination of Japanese word orders by skeleton patterns: As mentioned before, English and Japanese have almost the same word orders in long noun—noun constructions. However, the global surface word orders are completely different. For example, prepositional phrases in English usually modify the preceding nouns or verbs. In Japanese, these modifying phrases should precede the words which are modified. Moreover, in some cases the same word orders in English should be translated into different word orders in Japanese, depending on the semantic contents of the words.

Examples:

measuring temperature→ON DO SOKUTEI
(temperature) (measuring)
measuring device→ SOKUTEI SOCHI
(measuring) (device)

The determination of an appropriate word order is very difficult.

The input titles have been systematically reduced into simpler forms by Steps 1, 2, 3 and 4. These simplified forms, called 'skeletons', consist only of the words that are relevant to the determination of the global word order.

Examples: Input: Industrial and scientific techniques for measuring field effect mobility
Skeleton: Techniques for measuring mobility
Skeleton pattern: (n) (prep) (ing) (n)
Input: Evaluation of high quality phosphor screens for image tubes
Skeleton: Evaluation of screens for tubes
Skeleton pattern: (n) (prep) (n) (prep) (n)
Input: An automated general purpose test system for solid state oscillators
Skeleton: System for oscillators
Skeleton pattern: (n) (prep) (n)

The skeletons thus extracted from the titles will be matched against the skeleton patterns in the dictionary. Corresponding to each skeleton pattern, several different word orders are prepared to generate different Japanese title structures. Which order is adequate for the current title is determined by considering the following two factors:

(1) Flexible idiom, if any.

Examples: Application of N to N
Technique for -ing N

(2) Semantic contents of words: Very shallow semantic descriptions are given in the dictionary. Nouns are classified into 5 groups, verbs are accompanied with their case patterns, and so on. These semantic descriptions are used to calculate semantic preference of each word order.

Step 6. Morphological generation of Japanese titles.

At present, about 70% of English titles can be translated into Japanese, and 80% of translated Japanese titles are evaluated as good translation. Most of the failures are caused by lack of skeleton patterns or lack of lexical entries. We are now accumulating skeleton patterns which appear in the titles of scientific documents, and augmenting the system by adding the patterns and the lexical items.

4. MORPHOLOGICAL ANALYSIS OF JAPANESE

In English and some other European languages, words can be easily recognized, because they are usually separated from each other by spaces or punctuation marks in texts. However, there are many languages in the world in which there are no such definite separators, and words in texts are simply juxtaposed. Japanese is one such language.

The procedure for morphological analysis of Japanese consists of the following components:

(1) detection of boundaries between pause groups,
(2) analysis of inflections or consecutive Kana-strings,
(3) testing the grammatically of the connection of adjacent words in pause groups,
(4) dictionary look-up,
(5) segmentation of compound words composed of Kanji-characters.

Morphological analysis is usually performed as the first step of the whole language processing task. Errors in this step may have a damaging effect on the following language processing. Therefore, our morphological analysis procedure gives all possible interpretations, if they exist. This procedure utilizes various kinds of linguistic data listed in Section 2.

Tables 2 and 3 show the results of the morphological analysis when it is applied to sentences from an elementary chemistry text book.

Table 2

	Single correct result		Plural results are obtained. Correct one is contained.		Failure	
	Word	Pause group	Word	Pause group	Word	Pause group
Before consultation of dictionary	73.0	61.4	20.9	33.2	6.1	5.4
After consultation of dictionary	87.0	74.8	9.5	20.0	3.5	5.2

Note: 500 PG's are processed by the procedure.

Table 3

Average processing time (seconds)

	Per a sentence	Per a PG	Per a word
Without dictionary consultation	4.6	0.43	0.17
With dictionary consultation	35.3	3.32	1.30

Correct rate of analysis (%)

Note: The processing times include the time for MT I/O.

A set of utility programs are also provided to augment the procedure. By using these utility programs, we can obtain various summary information of the analysis results. For example, we can obtain the following:

(1) List of failures (snap shots of the analysis can also be obtained);

(.2) List of independent words (nouns, verbs, adjectives, etc.) which are not registered in the lexicon of independent words: Because many independent words are domain dependent, it is impossible to have a comprehensive dictionary of independent words. Our morphological analysis procedure can work without such a comprehensive dictionary. In fact, it can even work without any dictionary of independent words. Our algorithm can estimate the part-of-speech of an unknown word from the Kana-string which follows it (most of the independent words in Japanese are written in Kanji-characters) ;

(3) Frequency of each word in the given text.

We can augment the morphological analysis procedure by adding new entries to the dictionary or by revising the morphological rules from the information of these summaries.

5. INPUT/OUTPUT OF JAPANESE CHARACTERS

Input/Output Terminal for Kanji-Characters

Rapid progress of recent technology has enabled us to input/output Kanji-characters comparatively easily. Our research group has such a terminal connected to the centra] computer of Kyoto University. The system diagram for this terminal is shown in Figure 3.

Main Characteristics

Display size: 125 mm × 250 mm (12 inches)
Number of characters on the screen:
40 characters × 20 lines = 800 characters (for Kanji)
80 characters × 20 lines=1600 characters (for alphabets)
Size of display memory: 40 characters × 50 lines (Kanij)
80 characters × 50 lines (alphabet)
Character size: 5.3 mm × 5.3 mm (Kanji)
3.5 mm × 2.6 mm (alphabets)

Each Kanji-character is displayed by 24 × 24 dots
Each alphabetic character is displayed by 16 × 10 dots
Number of dot-patterns in the ROM: 2560 characters
Number of dot-patterns in. the floppy disc:
3000 characters × 2 = 6000 characters
Transfer rate from/to the main computer:
9600 baud (max), 1200 baud (at present)
Display modes: Normal mode, reverse mode, flicker mode, blink mode.

Because a microprocessor is built in this terminal, many flexible editing facilities are prepared in the local mode. A file system is also provided on the floppy disc. Therefore this terminal can also be used as a stand-alone system for producing private documents. We are writing a Japanese text editor for it. The terminal will become a good word processor for Japanese.

Software for Manipulating Kanji-Characters

Until recently it was not so easy for computer systems to input/output Kanji-characters. Most computer systems have been designed to manipulate only alphanumeric characters. Especially high-level languages such as PL/1, COBOL, LISP etc. have no provisions for manipulating Kanji-characters. We had to revise such programming languages to accept Kanji-characters. Now we have a Kanji-LISP system, that was originally implemented at Utah University and revised by us to handle Kanji-characters. Because we use PL/1 for morphological analysis and low-level processing of Japanese, we have developed PL/1 software packages to manipulate Kanji data.

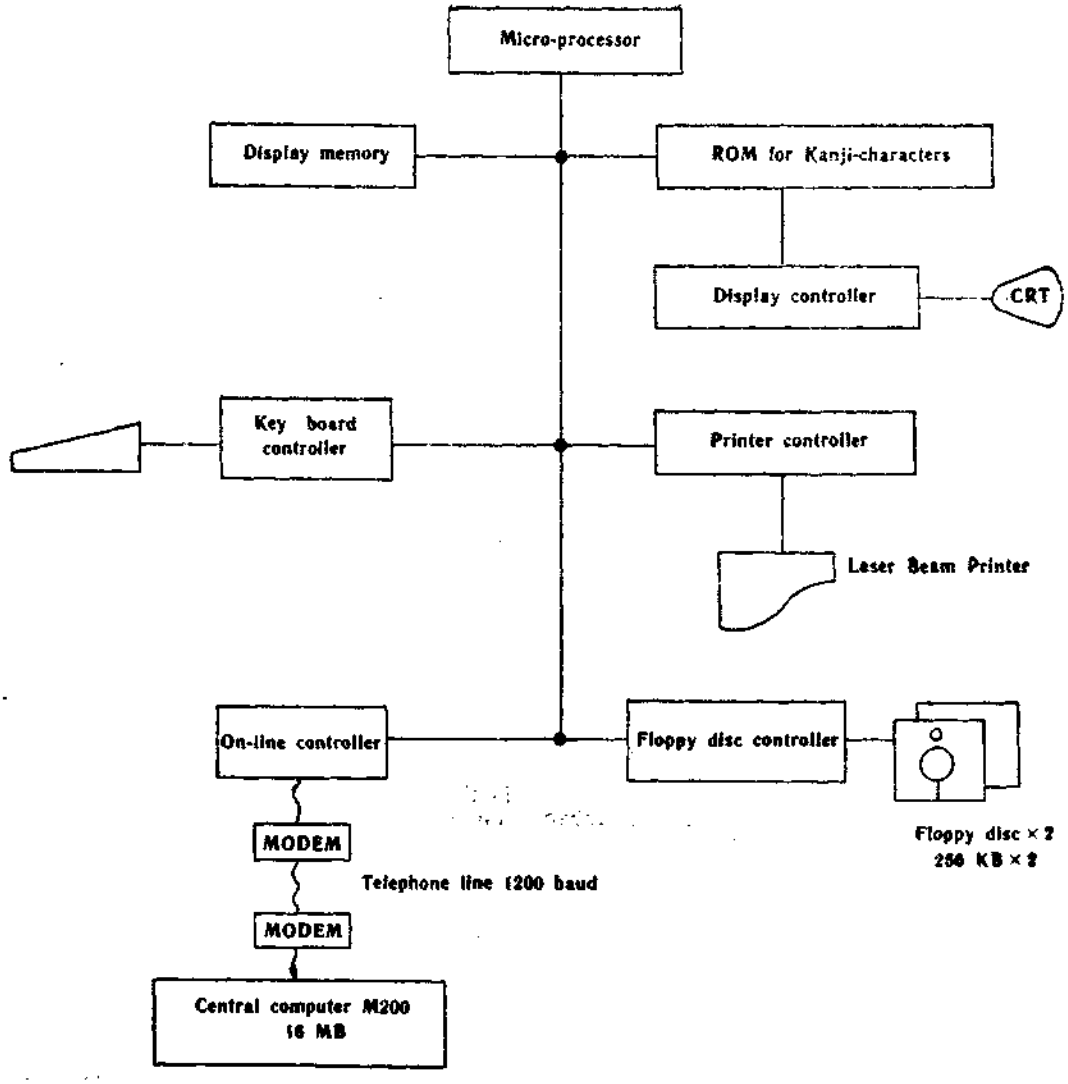


Fig. 3. Input/output terminal for Kanji-characters

We have also developed a flexible editor to store, revise and manipulate Kanji-data. Especially this editor has a facility of defining the record format of stored data. Many useful commands are provided for this purpose. For example, SORT is the command to sort the records according to the values in the specified columns, and KEYFILE is for extracting records that have specified keys in the specified columns and for creating a new file of them. These commands are very useful for managing the various linguistic data described in Section 2.