[From: *Term banks for tomorrow's world: Translating and the Computer 4*.
Proceedings of a conference ... 11-12 November 1982, ed. Barbara Snell. (London: Aslib, 1983)]

TERMINOLOGY BANKS AND DICTIONARIES IN JAPAN AND THEIR COMPUTER PROCESSING

Dr. Hirosato Nomura
Musashino Electrical Communication Laboratory, NTT, Tokyo, Japan

1. INTRODUCTION

Recently, in Japan, the importance of terminology banks and dictionaries has been understood and emphasised and some projects have been started. But there are problems in making terminology banks and dictionaries because of the specific characteristics of the Japanese language (large character set, no inserted blanks between words, etc.). In this article, we describe current activities on terminology banks and dictionaries in Japan, problems in making and using these data in Japanese, and some techniques to solve these problems.

2. CURRENT ACTIVITY IN JAPAN

In Japan most of the daily work is supported by electrical equipment. Especially language processing by computer is a very common phenomenon. Office automation is transforming the office, and numerous electrical appliances are being introduced in offices to do jobs including language processing. Though terminology banks and dictionaries are basic elements in natural language processing, until recently not much work has been done on them. In fact, in Japan, there are only a few terminology banks and general dictionaries which are suitable for use by computer. Of course there are many small dictionaries for experimental systems in laboratories and dedicated dictionaries for Japanese word processors, but they are inadequate for advanced use. So we are looking forward to the preparation of large terminology banks and dictionaries to utilise this equipment more fully.

2.1 Terminology Banks

The distinction between terms and usual words is not always clear. Terminology consists of words which are used in a different meaning from the usual use in a specific field, or which are mainly used in some specific field. We can distinguish terms from usual words to some extent by their frequency in a specified field. Most Japanese terms are nouns or compound nouns. [A term may be a word or group of words assigned to a concept. Terminology is a collection of terms representing a system of concepts pertaining to a special field. (Ed. with acknowledgement to ISO and Infoterm)].

JICST (Japan Information Centre of Science and Technology) made a thesaurus of scientific and technological terminology in 1975, and this is used in a literature retrieval service by computer. The ISO (International Organisation for Standardisation) domestic committee produced guidelines for the establishment and development of monolingual thesauri. Several small terminology banks of each kind were made in several projects. Tanaka conducted an experiment to collect the terms in English and Japanese titles stored in JICST's files.

2.2 Dictionaries

Several dictionaries have been made to satisfy several purposes (word processing, machine translation, etc.) in each project. Also, several machine readable dictionaries made originally for human access are available (Sanseido's Shin Meikai Kokugo Jiten, Sansiedo's New Concise English Japanese Dictionary). A fairly large dictionary for Kana Kanji conversion was made at Kyushu University. IPS (Information-Technology Promotion Agency) is going to make a dictionary, which will contain a large amount of semantic information (case frames for verbs, semantic categories, aspectual categories for verbs, etc.). This dictionary is mainly intended for machine translation use. Also, the Science and Technology Agency started a machine translation project in 1982, and began to make a terminology bank and a dictionary for this purpose.

3. PROBLEMS

3.1 Characteristics of the Japanese Language

In making terminology banks and dictionaries, there are many problems because of the characteristics of Japanese. Japanese is distinguished from European languages by the following characteristics:

1. agglutinative

2. word order relatively free but predicate (verb, adjective, and adjective verb) placed at the last position in a sentence,

3. large character set (sentences are written in mixed style),
   ideographic characters (Kanji, i.e. Chinese characters)
   Phonetic characters (Hirakana, Katakana and Romaji),

4. no spaces inserted between words,

5. many homonyms but few homographs

6. easy to make compound words
   e.g. compound noun: KAKAI-HON'YAKU (machine translation)
   compound verb: YOMI-HAJIMERU (begin to read)

7. many adopted words from foreign languages
   (usually written in KATAKANA)

3.2 Input Methods

To put Japanese texts into computers, several input methods and devices have been developed. Development of useful input devices is also important from the viewpoint of making dictionaries. There are two types of input method: typing and pattern recognition. Typing methods are classified into two categories: character selection and code conversion, and used in several word processor systems. Pattern recognition methods are classified into two categories: character recognition and speech recognition. An online writing character recognition method is now used in practice. Kanji optical character readers (OCR) and voice typewriters are being intensively studied. Kanji OCR is currently being used in a few groups.

The Kana-Kanji conversion method is interesting and is being used in several representative word processor systems. The Kana-Kanji conversion method was originally proposed by Kurihara in 1964(1).

In the Kana-Kanji conversion method, there are several input modes such as the following:

1. does not place blanks between Bunsetsus,
   where a Bunsetsu consists of an independent word and adjuncts,
2. places blanks between Bunsetsus,
3. designates Kanji parts,
4. places blanks between words,
5. places blanks between an independent word and adjuncts.

3.3 Kana-Kanji Conversion

As for the Kana-Kanji conversion, homonym disambiguation is the hardest problem of this method. Some words have several dozen homonyms. Mori reported that there are about 1.7 homonyms per word in a 40,000 word dictionary(2). To resolve the problem of homonyms, the following methods are being developed or studied:

1. selection by interaction,

2. using grammatical information, that is, examining the concatenation restriction of morphemes,
   e.g. KOUSYOU NA — noble (correct),
   negotiation, sing loudly (false),
   "NA" can be a conjugation of an adjective verb, but cannot follow a noun.

3. using semantic information, that is, checking the semantic relation between words, using semantic categories, using discourse information, and so on,
   e.g. ISSEN WO KANUSU -- draw a line (correct)
   draw a battle (false)

e.g. SENSEI GA IKU — a teacher goes (correct)
an oath goes (false)
(data) agent GA place E IKU
(agent goes to some place)

4. utilisation of frequency information,
using short-term memory,
frequency in the ongoing processing text,
using long-term memory,
frequency in general

So, dictionaries must contain the information relevant to the above processes. But many unresolved problems remain regarding the use of semantic information.

The following problems about particular types of Bunsetsus must also be resolved in Kana-Kanji conversion. These problems are closely related to dictionary entries. Because combinatorial explosion occurs, all entries for 1, 2 and 3 below cannot be contained in the dictionary. On the other hand, words for 4 must be contained in the dictionary of a competent Kana-Kanji conversion system.

1. compound words,
e.g. KIKAI-HON'YAKU (machine translation)

2. adjuncts,
e.g. OOGATA-KA (tendency of enlargement),
SHIN-HOUAN (a new bill),

3. numerals,
e.g. DAI15KAI (15th),

4. proper nouns
e.g. place name, person's name, company's name.

3.4 Problems in making Terminology Banks

There are several problems in making terminology banks. Here, for example, we present three problems.

1. The method of collecting terminology from Kanji parts of materials is usually used. But this method does not necessarily succeed because terms sometimes consist of Kanji mixed with HIRAKANA and KATAKANA.
e.g. GO I SUU (the number of words)
SOUTAI ADORESU (a relative address),
where underlined parts are written in HIRAKANA and KATAKANA respectively.

2. Terms are sometimes made up of usual words in compound forms,
e.g. TAIKI OSEN (air pollution)

3. Many terms are loan words from foreign languages. These are usually written in KATAKANA, but the expression is not uniquely defined.
e.g. KONPYU-TA (computer)
KONPYU-TA-

3.5 Construction of Terminology Banks and Dictionaries

When we make terminology banks and dictionaries they must be considered as a subsystem in a total system. First of all, the target total system is analysed and designed, and the role of the dictionary in the total system must be clarified. We must decide what kind of algorithm or dictionary can be used to satisfy our requirements.

To make a dictionary satisfying the specified requirements, a systematic approach (analysis, conceptual design, detail design, analysis of job flow, making manuals, organisation of workers, performance, test, evaluation, etc.) is required.

We must decide the following:

1. words to be collected
2. collection method, source of collection
3. morpheme of word (unit, character set, code, etc.),
4. data element,
5. flow of jobs
6. plans for maintenance

Deciding on the word units in Japanese is a very difficult problem because of compound words and adjuncts.

There are two methods of making dictionaries: the analytic and the "Gestalt" methods. The analytic method is appropriate when the purpose is rather narrow and clear. Though it takes a great deal of time, and requires a fairly large amount of knowledge and experience of the subject, the completed work is very useful and not wasteful. On the other hand, in the "Gestalt" method we collect words from a variety of sources. We can make a dictionary routinely in a relatively short time, but it may not be adequate for the purpose. It is necessary to use these two methods properly.

We must provide a good manual to enable the compiler to make a unified and consistent dictionary. Preparing a good manual for using this dictionary is also an important part of the job.

4. COMPUTER PROCESSING AND EQUIPMENT

4.1 Two examples

We present here two case studies of making a thesaurus and for terminology. One is JICST's thesaurus (3), and the other is a terminology bank made by Tanaka(4).

JICST's Thesaurus:

1. Name: JICST Scientific and Technological Terminology Thesaurus
2. Investment: 100 persons * 6 years * 0.1 = 720 man-months, cost 17,000,000 yen
3. Completion Date: August 1975, later revised four times,
4. Purpose: improvement of retrieval efficiency by using terms in information storage and retrieval,
5. Subject: Science and Technology,
6. Number of Words: (first version) 29,000 descriptors and 4,800 non-descriptors,
7. Source: high frequency words out of 170,000 words collected from 2,100,000 indexed scientific articles over a period of 6 years,
8. Code: JICST's Kanji code, 2,418 characters,
9. Data Elements: entry, phonetic transcription in Kana, categories, statistical information, comments and related information.

Tanaka's terminology bank

His main subject is automatic segmentation of Japanese sentences. He has also studied making terminology banks from English and Japanese title sentences in JICST's files. He has tried making a terminology bank by the method of finding the part in Japanese titles which corresponds to automatically collected English terminology, but there is a difficulty in Japanese which comes from the writing style with no spaces inserted between words.

His method was as follows. First, he automatically extracted English terms from the English titles. In this process, he used the part-of-speech information. Then in the corresponding title, he found the Japanese term corresponding to the extracted English term.

He also studied a number of technical terms generally used in a specific field, and in the case of unknown words encountered in handling the English terms, he proposed a synthesis of new Japanese terms by juxtaposing Japanese words corresponding to each of constituent elements of the term, using an English-Japanese dictionary.

4.2 Contents of the Dictionary

The contents of dictionary vary with the purpose of the system. Even in order to make powerful kana-kanji conversion systems, a fairly large amount of semantic information is required. In a system which pursues semantic processing, semantic information in the dictionary is very important.

A dictionary must contain certain grammatical information such as the part of speech, and the conjugation of the verb, hypernonym, hyponym, synonym, antonym etc. Further, idiomatic expressions, sentence examples, etc. may also be included.

4.3 Format of Dictionary

When a dictionary for human use is stored in a computer, it is not structured and remains in its printed form. We can separate it into blocks, and each block corresponds to one piece of information. To use the dictionary more efficiently, highly structured formats have to be adopted. Network and Frame represent such formats.

The network structure was originally introduced by Quillian(5) to represent the meaning of English words. He aimed to represent associative relations between English words in terms of network structures. Later, this network format was used to represent not only the meaning of words but also knowledge and a variety of meanings such as the meanings of sentences, etc.

A frame system is a framework to represent knowledge in the field of artificial intelligence. This idea was proposed by Minsky(6). The dictionary can be regarded as a part of knowledge. In our experimental machine translation system, all knowledge including the contents of a dictionary is represented in Frame format(7).

4.4 Searching Technique in a Dictionary

When a dictionary becomes very large, it is necessary to use auxiliary memories. So it is important to reduce access time to the secondary memory. To satisfy this requirement, several techniques such as hashing, binary search, using a balanced tree as data structure, etc. were proposed. Hidaka et al. proposed extending the balanced tree so that all words which are substrings starting at the leftmost of a given character string can be retrieved in a few accesses(8).

4.5 Equipment

We need several different types of equipment to construct and use terminology banks and dictionaries. As input devices, we can use tablet, typewriter, etc. Further we can use a Japanese editor, KWIC (key word in context), word processor, etc. as tools. KWIC is often used to collect words or terms from a mass of materials.

5. CONCLUSION

In this article, we have described the current activities in making terminology banks and dictionaries in Japan, problems in making and using these data, and their computer processing. The necessity for natural language processing is increasing, and terminology banks and dictionaries are becoming more significant as basic data. We can say that the quality and quantity of terminology banks and dictionaries are critical in deciding the performance of language processing systems, and to make a language processing system more powerful, we must study the semantic information contained in these data and efficient methods of storing and retrieving them.

REFERENCES

(1) KURIHARA, T. and KUNISAKI, Y. On the transformation process of phonetic sentences into ideographic sentences in Japanese (in Japanese), Bulletin of the Faculty of Engineering, Kyushu University, Vol. 39, no. 4, 1967.

(2) MORI, K. and KAWATA, T. Kana-Chinese translation (in Japanese). Journal of the Information Processing Society of Japan, Oct. 1979.

(3) Report on trends in natural language processing techniques and systems (in Japanese). Japan Information Processing Development Centre, 1982.

(4) TANAKA, Y. Automatic extraction of terminologies (in Japanese). Memo of the Special Interest Group on Computational Linguistics, No. 25, Information Processing Society of Japan, 1981.

(5) QUILLIAN, M.R. Semantic memory. In M. Minsky (ed), Semantic information processing, MIT Press, 1968.

(6) MINSKY, M. A framework for representing knowledge. In P Winston (ed), The Psychology of Computer Vision, McGraw-Hill, 1975.

(7) SHIMAZU, A., NAITO, S. and NOMURA, H. Japanese language semantic analysis in machine translation system LUTE (in Japanese). Memo of the Special Interest Group on Natural Language Processing, Information Processing Society of Japan, No. 33, 1982.

(8) HIDAKA, T., INANAGA, H. and YOSHIDA, S. Construction of Japanese dictionary by the Extended B-Tree Method (in Japanese). Memo of the Special Interest Group on Natural Language Processing, Japan Information Processing Society, No. 33, 1982