TR-100

Basic Specifications of the Machine-Readable Dictionary

Hideo Miyoshi, Yuichi Tauaka, Toshio Yokoi,
(ICOT)
Toshio Ishiwata (Ibaraki Univ.),
Hozumi Tanaka (TIT),
Shinya Amano (Toshiba Corp.),
Hiroshi Uchida (Fujitsu Ltd.)
and
Takano Ogino (Institute of Behavioral Science)

January, 1985

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03) 456-3191~5
Telex ICOT J32964

Institute for New Generation Computer Technology

BASIC SPECIFICATIONS of the MACHINE-READABLE DICTIONARY

Miyoshi, Hideo [1]
Ishiwata, Toshio [2]
Amano, Shinya [4]
Ogino, Takano [6]
Tanaka, Yuichi [1]
Tanaka, Hozumi [3]
Uchida, Hiroshi [5]
Yokoi, Toshio [1]

January 30, 1985

1) Institute for New Generation Computer Technology
2) Ibaraki University
3) Tokyo Institute of Technology
4) Toshiba Corporation
5) Fujitsu Limited
6) Institute of Behavioral Science

ICOT Technical Report TR-100, 1985

ABSTRACT

The machine-readable dictionary system is being developed at ICOT. We are now in the first stage of development. In this paper, we describe the basic specification of machine-readable dictionaries. The purpose and the initial plan is described in chapter 2. The issues discussed by the committee and the solutions arrived at are listed in chapter 3. The structure and contents of forthcoming dictionaries are described in chapter 4. Items such as semantic information are left to the future. In chapter 5, the treatment of these issues in the second stage is mentioned. The research is being conducted by a committee consisting of ICOT researchers and other computer scientists and linguists.

CONTENTS

1. Introduction

2. Plan of Machine-Readable Dictionary
   - 2.1 Goal
   - 2.2 Preparatory Activities
     - 2.2.1 Lectures by Experts
     - 2.2.2 Examination of Existing Dictionaries
     - 2.2.3 Planning Machine-Readable Dictionary Development
   - 2.3 Design Plan
   - 2.4 Activity Plan

3. Issues Discussed by the Research Group
   - 3.1 Structure
   - 3.2 Entry Words
   - 3.3 Examples
   - 3.4 Items
   - 3.5 Miscellaneous

4. Basic Specifications of the Machine-Readable Dictionaries
   - 4.1 Structure of Dictionaries
   - 4.2 Contents of Dictionary Items
     - 4.2.1 Japanese Dictionary
     - 4.2.2 Japanese-English Dictionary
     - 4.2.3 English-Japanese Dictionary
     - 4.2.4 English Dictionary

5. Remaining Tasks
   - 5.1 Specification of Case Information
   - 5.2 Specification of Semantic Information

1 INTRODUCTION

In natural language processing by computer, the most fundamental and important task is the creation of large scale machine-readable dictionaries for the object languages. The dictionary not only forms the basic database for natural language understanding systems, knowledge base systems, and natural language interface technology, but also provide a lot of useful information for research in linguistics, psychology, philosophy, education and other fields [Walker 1984]. At ICOT, we are developing four dictionaries (Japanese, English-Japanese, Japanese-English, and English) and two thesauruses (Japanese and English), to be completed by the end of fiscal 1986. These dictionaries are referred to as general-purpose master dictionaries. They will be so large that any dictionaries for any application can be taken from these master dictionaries. The Research Group on Machine-Readable Dictionaries was for this purpose. The Research Group has been discussing research goals including the scale, structure, and contents of the dictionaries, as well as organization of future research. This paper consists of a report of these discussions in the Research Group and of ICOT's development plan.

2 PLAN OF MACHINE-READABLE DICTIONARY

2.1 Goal

The creation of large-volume machine-readable dictionary systems is indispensable for constructing a variety of application systems for natural language processing by computer. Indeed, it is among the most urgent issues facing such activities. Needless to say, the dictionary systems are not simply magnetic tape forms of existing printed dictionaries. They are closely linked to the other language processing systems such as language data bases, on-line information retrieval systems, editors, and so on.

ICOT's machine-readable dictionary system is being planned for use by various institutes, universities, and companies in Japan and abroad. For this purpose, the dictionary must have sufficient scale and contents for the demands of various application systems; and the copyright problem must be solved effectively.

There are two ways to develop a machine-readable dictionary system to meet these requirements. One is to take existing printed dictionaries as a basis and process them. The other is to develop a new dictionary from scratch. Whether we can complete a large-scale machine-readable dictionary system depends on how we make this choice.

2.2 Preparatory Activities

In 1983, the Machine-Readable Dictionary Study Group was organized to study the role of dictionary systems and how to create them. Twelve meetings were held in 1983. The activities of the group are described in the following subsections. The group was composed of the following researchers:

S. Amano	(Toshiba Corp.)
K. Furukawa	(ICOT)
S. Kuga	(Sharp Corp.)
H. Miyoshi	(ICOT)
Y. Nitta	(Hitachi Ltd.)
H. Tanaka	(Electrotechnical Laboratory)
H. Uchida	(Fujitsu Ltd.)
S. Yagi	(Matsushita Electric Industrial Co., Ltd.)
S. Yokoyama	(Electrotechnical Laboratory)

2.2.1 Lectures by Experts

We invited academics involved in linguistics and writing dictionaries to come and lecture to the group. The lectures were:

- Some Problems in Writing Dictionaries, by Prof. S. Mizutani. (Tokyo Women's College)

Some problems in developing real published dictionaries were presented from the viewpoint of an editor of a Japanese dictionary.

- Classification of Verbs, by Prof. M. Yoshida. (Kyushu University)

A lot of documents about the classification of verbs accumulated at Kyushu University were presented.

- Grammar Theories, by Prof. T. Ishiwata. (Ibaraki University)

Semantic information and case information are the most important items in machine-readable dictionaries. Various semantic theories in current linguistics were presented.

2.2.2 Investigation of Existing Dictionaries

We investigated the availability of some machine-readable dictionaries which have already been created or are under development both in Japan and abroad.

- Shinmeikai Japanese Dictionary (Sanseido) [Yokoyama 1977, 1984]

Originally, K. Fuchi developed the machine-readable version of this dictionary at Electrotechnical Laboratory. Strict proofreading has made this machine-readable dictionary faithful to the original book and reliable for applications [Ogino 1981a]. This version is used at a lot of institutes. The new version with thesaurus codes was also developed [Ogino 1981b].

- Concise English-Japanese Dictionary (Sanseido) [Nagao 1980]

The machine-readable version was developed at Kyoto University. There are two versions, one is faithful to the original book, the other is a structured version.

- Iwanami's Japanese Dictionary [Nishio 1982]

This machine-readable dictionary was developed at Iwanami Shoten Publishers for dictionary compilation.

- Longman's Dictionary of Contemporary English [Nagao 1982]

The machine-readable dictionary developed at Longman's Group Limited for dictionary compilation contains some semantic features.

- Dictionaries developed at JEIDA [JEIDA 1982]
The machine-readable dictionary developed at JEIDA — Japan Electronic Industry Development Association has two technical term dictionaries (accounting area and information processing area) as well as a Chinese character dictionary.

- Machine Translation Project of the Science and Technology Agency [Nagao 1983]
The machine-readable dictionary developed at the Machine Translation Project of the Science and Technology Agency is used for machine translation.

- Japanese lexicon developed at IPA [Murata 1982]
The machine-readable dictionary being developed at IPA — Information-Technology Promotion Agency will be used for general-purpose Japanese language processing.

These machine-readable dictionaries were found to be unsuitable as a basis for our general-purpose master dictionaries. Some cannot be processed freely owing to copyright constraints and some lack sufficient linguistic information.

2.2.3 Planning Machine-Readable Dictionary Development

In the Machine-Readable Dictionary Study Group, we discussed what kind of dictionaries will be necessary for machine translation and how they should be used. The following components were proposed for this purpose:

- Dictionary and Thesaurus for Semantic Analysis (Japanese, English)
- Transfer Dictionary for Machine Translation (Japanese, English)
- Post-Editing Dictionary (Japanese, English)
- Technical Term Dictionary
- On-Line KWIC System

The figures on the next page show the relations between subsystems at the development stage (Fig. 2.1) and at the application stage (Fig. 2.2.) This configuration is just for an application. We continued our discussion about more general configuration, because our final goal is to develop a general-purpose master dictionary system.

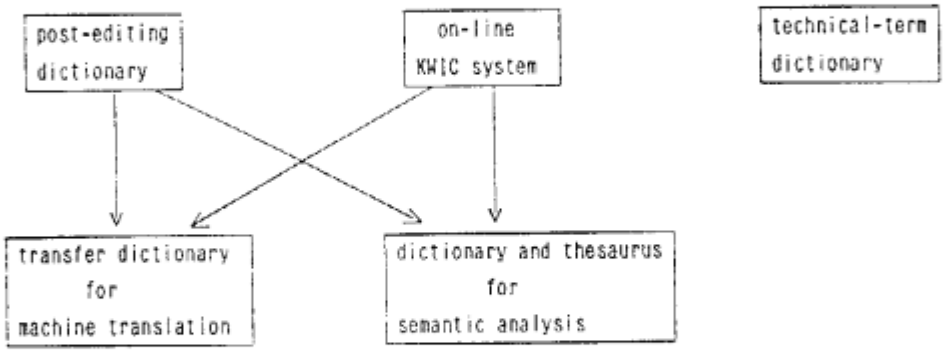

Fig. 2.1 Relations between the Subsystems

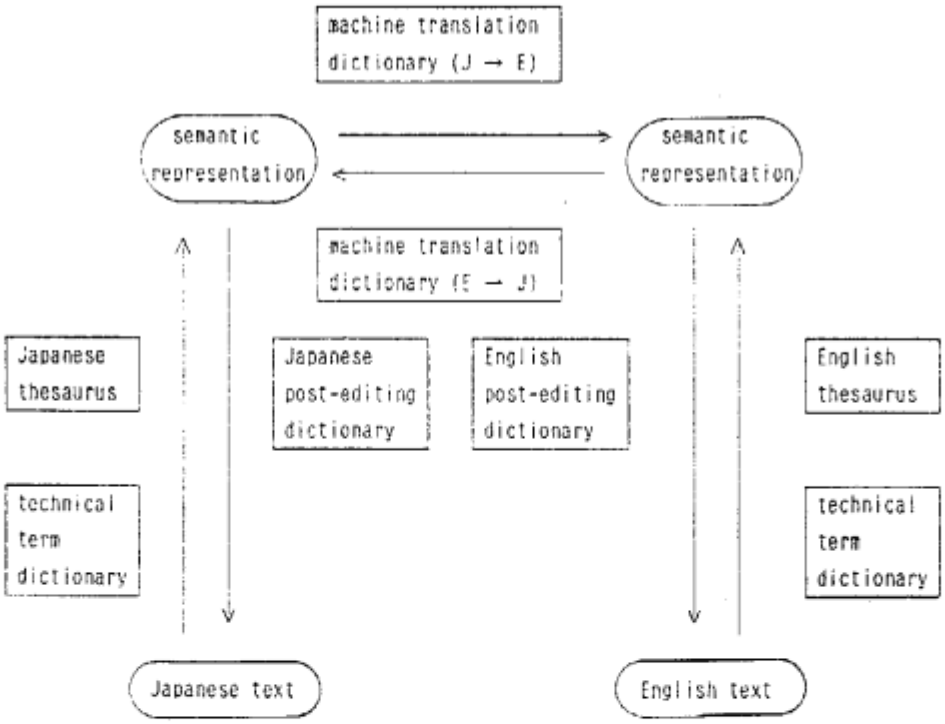

Fig. 2.2 Relations between the dictionaries in operation

2.3 Design Plan

The preparatory investigation showed us that we cannot use existing dictionaries because of copyright problems, and because none of the dictionaries developed by other projects satisfy our demands. Consequently we decided to design the specifications of the new dictionaries at ICOT and then place orders with the dictionary publishers to develop new dictionaries according to our specifications.

This will enable us to develop dictionaries with capacities sufficient to meet the demands of a wide variety of application systems. Japanese Government will hold the copyright through ICOT. Therefore, it is expected that all institutions will be able to use this dictionary system for study purposes.

We aim to design our machine-readable dictionaries to be so simple and basic that any user can easily access them for their applications. We will supply our dictionaries in magnetic tape form, in which each hierarchically structured item is expanded into sequential format. Thus, it will be left to users to transform such structures into more efficient index systems, to express items as list structures or Prolog terms, and so on.

2.4 Activity Plan

We decided that in the three year period we will develop these four dictionaries:

- Japanese,
- English-Japanese,
- Japanese-English, and
- English.

The Machine-Readable Dictionary Research Group was organized in June 1984 to carry out this project. The members are as follows:

Chairman

T. Ishiwata *	(Ibaraki University)

Members

S. Amano *	(Toshiba Corp.)
S. Ishizaki	(Electrotechnical Laboratory)
M. Kimura	(Institute of Behavioral Science)
K. Komorita	(Matsushita Electric Industrial Co., Ltd.)
Y. Kusanagi	(Tsukuba University)

Y. Nitta	(Hitachi Ltd.)
H. Satake	(The National Language Research Institute)
M. Sato	(JICST)
H. Tanaka	(Tokyo Institute of Technology)
A. Tsuruoka	(The National Language Research Institute)
H. Uchida *	(Fujitsu Ltd.)
T. Yokoi	(ICOT)

Observers

H. Miyoshi *	(ICOT)
T. Ogino *	(Institute of Behavioral Science)
Y. Tanaka *	(ICOT)
S. Yokoyama	(Electrotechnical Laboratory)

* Also members of the Machine-Readable Dictionary Task Group

The planning, investigation and actual research work are performed by the Task Group. Proposals and results produced by the Task Group are discussed later in the Research Group.

Apart from these activities of the Research Group, contracts were made between dictionary publishers and ICOT. Actually, ICOT entrusts the computer manufacturers with the development work. Each of these companies then organizes the work in collaboration with their respective associated dictionary publishing company.

The assignments of computer manufacturers and dictionary publishers are:

Dictionary	Manufacturer	Publisher
Japanese	Hitachi Ltd.	(undecided) *
English-Japanese	Toshiba Corp.	Gakken Co., Ltd.
Japanese-English	Fujitsu Ltd.	Obunsha Publishing Co.
English	Fujitsu Ltd.	Obunsha Publishing Co.

* The small scale Japanese dictionary for semantic analysis has already been developed by Hitachi Ltd. The development of the full Japanese dictionary will begin in 1985.

The development period is divided into two stages. In the first stage, the basic manuscripts without semantic information will be written by the authors according to our new specifications and will then be put onto magnetic tape. In this first phase, semantic information such as semantic marker or thesaurus code will also be

researched and developed. Although the Research Group makes decisions of scale and range of the entry words, selection of entry words and some other details are left to the authors.

As we understand the situation at this point, there are no dictionaries possessing logically hierarchical structure which is indispensable for machine-readable dictionaries. Further, it is unreasonable to expect many authors to write four dictionaries as a consistent system, but it is easy for machines to modify formats, compare dictionaries, and write cross-references after completion of the tentative machine-readable dictionaries. With these facts in mind, we are designing just basic specifications of the dictionaries as frameworks. The authors select entry words and fill each entry in the conventional manner. Finally, we compare dictionaries, take some statistics and modify the dictionaries if necessary

In the second stage, the semantic information will be added to the dictionaries, and then all four dictionaries will be merged into a complete system.

3 ISSUES DISCUSSED BY THE RESEARCH GROUP

3.1 Structure

• Common structure

It is not necessary for the four dictionaries to have a common structure at present. In the first stage, we give higher priority to the publishers' conventions for writing dictionaries. The structure of the four dictionaries will be modified into a common standard format in the second stage.

3.2 Entry Words

• Number of entry words

This dictionary system is intended to be general, that is, not to be restricted to a specific domain. For instance, when it is used to retrieve a judicial precedent, it must contain all the words used in daily life. Although a word occurring less than once in thirty thousand words appears rarely in ordinary texts, each dictionary will have to contain sixty or seventy thousand words. On the other hand, the dictionary of technical terms which is planned in another project will contain about ten million words.

• Standardization of entry words

Obviously, it is impossible to standardize the entry words for the English and Japanese dictionaries. It would seem possible, for example, for the English dictionary and the English-Japanese dictionary to have the same entry words. We assume that the English dictionary contains the definitions and other syntactic, semantic and pragmatic information of each English word, while the English-Japanese dictionary, as a transfer dictionary, lists the corresponding Japanese words and phrases for each English word in the complete machine-readable dictionary system. In the first stage, however, because our dictionaries will be developed based on the existing dictionaries, they cannot have the same entry words.

• Idioms

Idioms will have independent entries. Verbs and nouns in an idiom will be cross referenced from their original entries. Phrasal verbs are treated almost the same as idioms.

- English Words with multiple meanings

In the existing dictionaries, the words with multiple meanings are treated separately according to their etymologies. Some words such as 'ball' (spherical object/dance) are etymologically different and accidentally have the same forms. Usually different entries are assigned to these words. Other words have single entries containing several parts of speech as well as different meanings. Our dictionaries will conform to this style.

- Japanese Words with multiple meanings

In Japanese there are far more words with multiple meanings. Some words have a wide range of abstract meanings for which there are several Chinese character notations according to the different concrete meanings. The treatment of these kinds of words differs in existing dictionaries because the authors have different opinions on the breadth of meaning. We will leave decisions in this area to the authors of the Japanese dictionary.

3.3 Examples

- Purpose of examples

Examples are used for several reasons. When the case information is added to the verbs in the second stage of our project, these examples will be used to determine the semantic features of nouns associated with each verb. The examples are also used for choosing the appropriate equivalent word in machine translation and post-editing.

- Number of examples

The more examples it contains, the more substantial our dictionary system will be. Particularly, the examples in which sentence structure and case information are explicitly described for verbs are desirable.

- Closedness

Should every word in the examples appear as an entry? This is, of course, desirable, but we will not decide on this in the first stage. It is easy for the computer to look up all the examples and correct them after the manuscript of the dictionary has been completed.

- Semantic features

We take it as one of our principles that no piece of information in the existing

dictionary should be lost. Therefore, if the semantic feature is originally contained in an example sentence then it should be preserved.

e g.

A (house etc.) *be built of* B (material)

3.4 Items

- Reference words

Synonyms, antonyms and derivatives should be collected as reference words and linked to each other. This information is useful for making thesauruses and post-editing.

- Pragmatic information

Information such as usage labels or usage explanations are used for post-editing. Although this kind of information is written for humans, it should be preserved.

- Morphological information

Morphological information must be included even for regular inflections.

3.5 Miscellaneous

- Pronunciation and accent

Standard pronunciation of words must be described. There are two ways to express pronunciation: phonetic symbols and phonemic symbols. The choice will be left to the authors. In Japanese, kana notation also express pronunciation of a word, however, explicit description of pronunciation and accent is necessary for speech recognition and speech synthesis.

4 BASIC SPECIFICATIONS OF THE MACHINE-READABLE DICTIONARIES

4.1 Structure of the Dictionaries

The machine-readable dictionaries must satisfy several conditions. Dictionary items must be hierarchically structured. This is not a feature of the ordinary published dictionaries. Supplementary explanations, examples of entry words in context, sentence pattern information, and compound word information must be provided in sufficient detail. This linguistic information is used for investigating semantic information (thesaurus) and case information. The original plan of the structure of each master dictionary was made by the Machine-Readable Dictionary Task Group and then discussed in the Machine-Readable Dictionary Research Group. As a result of the discussions in the Machine-Readable Dictionary Research Group, the structure of each master dictionary was determined as shown in Figures 4.1, 2, 3, and 4.

- Reference Word and Supplementary Explanation

Reference words and supplementary explanations can be included either in the part of speech information section or in the semantic information section. Those which pertain to a specific meaning are given in the semantic information section, and those which are common to every meaning of the word are given in the part of speech information section. Those pertaining to a word which is only one part of speech and has only one meaning are given in the semantic information section.

- Variant Spelling

Variant spellings and forms are not treated separately; they are grouped under one entry. They are included in the variant spelling and form list of the one entry word.

e.g.

color and *colour*

- Homograph

Words which have the same spellings but different meanings are treated as different words. These are called homographs.

e.g.

ball[1]: any solid or hollow sphere as used in games

ball[2]: social gathering for dancing

• Homonym

Words which have the same pronunciation but different spellings and meanings are treated as different words. These are called homonyms.

e.g.

突く　*tsuku* (J.) — to thrust

付く　*tsuku* (J.) — to adhere

• Meaning

The meanings are distinguished in as many classes as possible. It is desirable that the distinctions be made clear by specifying associated sentence patterns, because they will be used to investigate the co-occurrence relation between words.

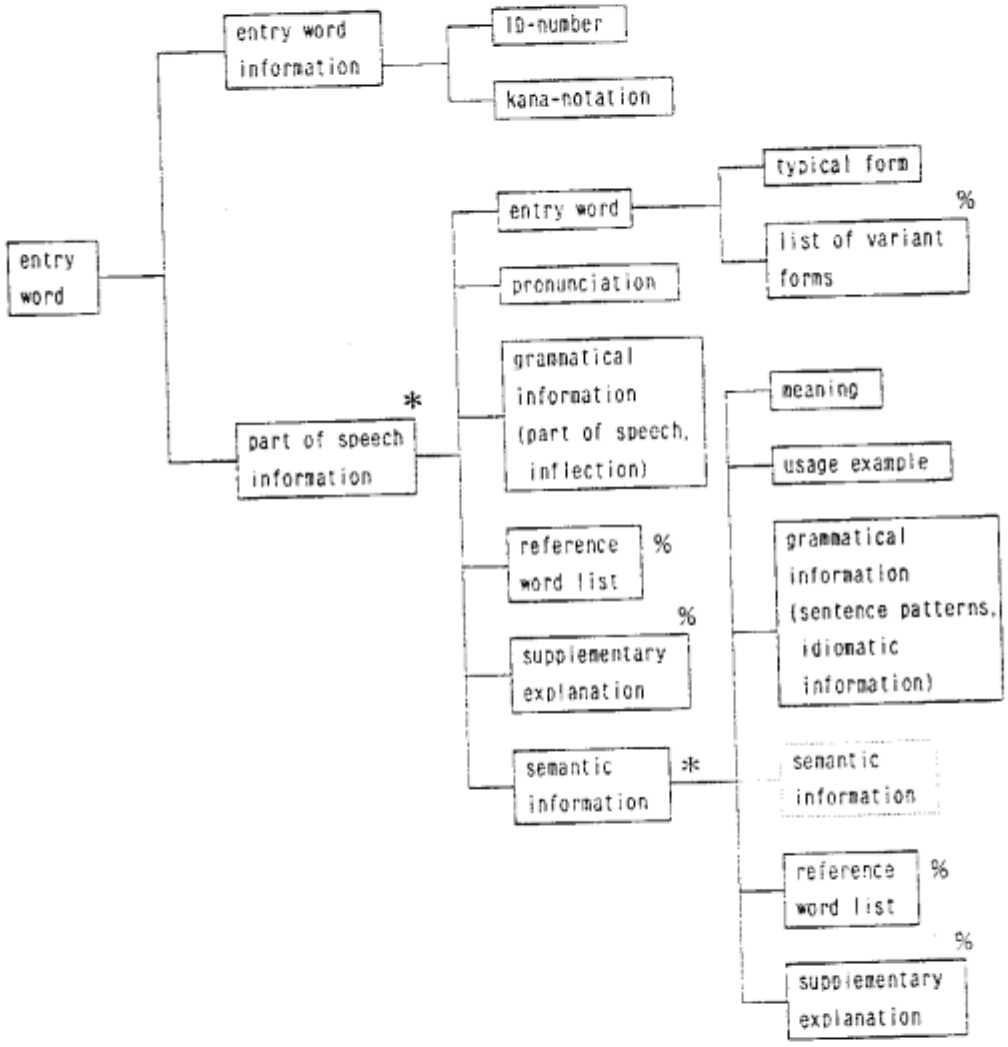

Fig. 4.1 The Structure of the Japanese Dictionary
(* ··· repeatable, % ·· omissible)

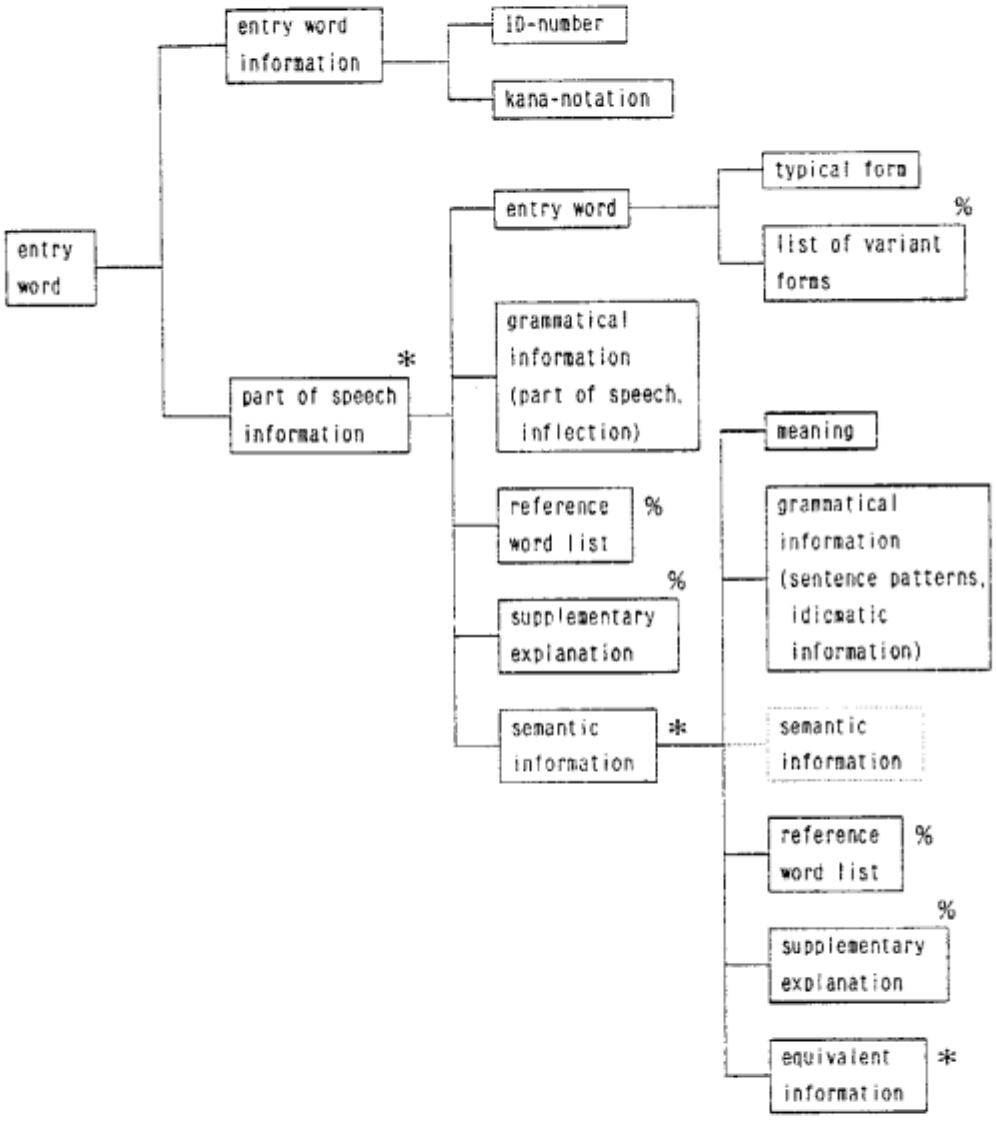

Fig. 4.2 The Structure of the Japanese-English Dictionary
(* ··· repeatable, % ··· omissible)

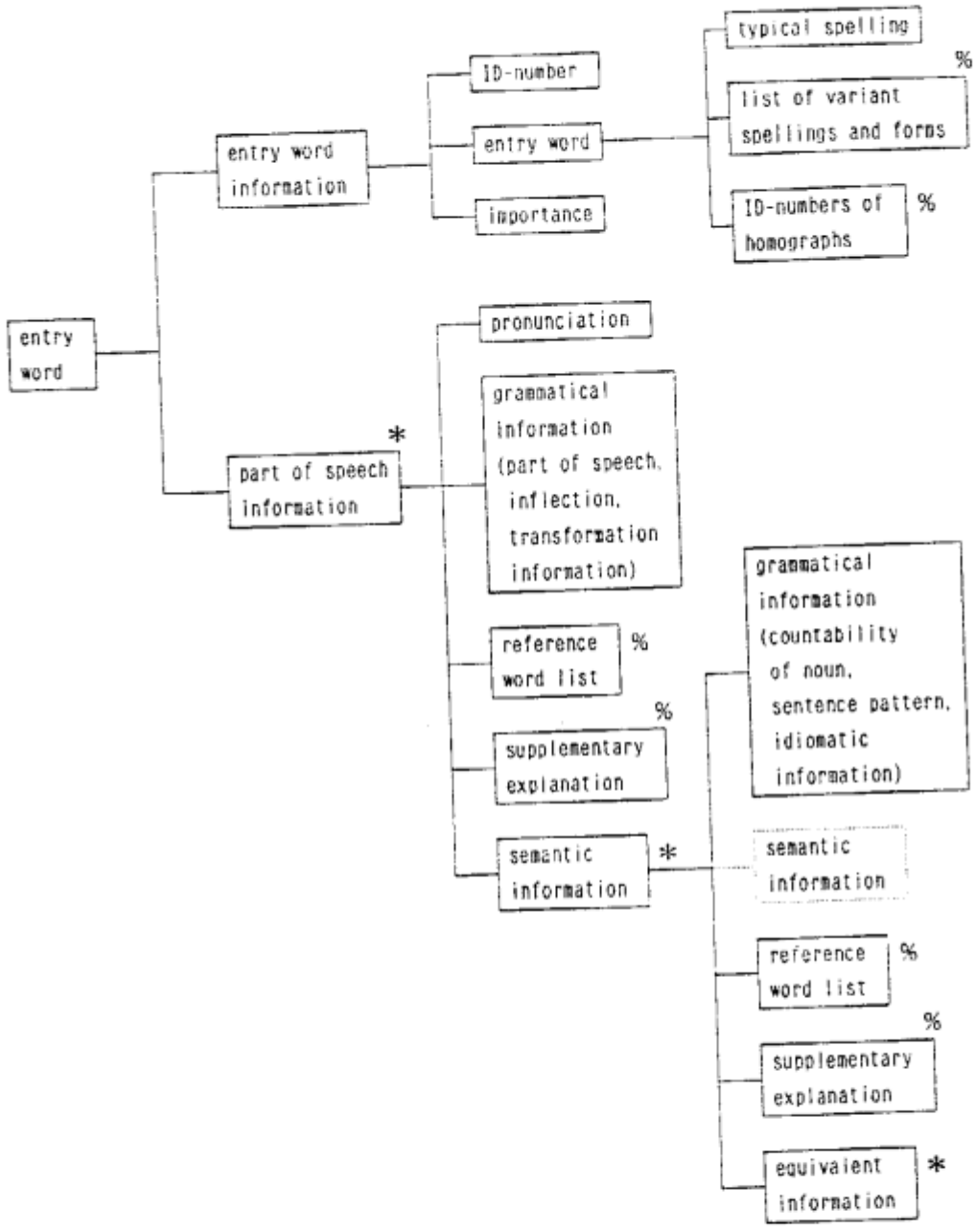

Fig. 4.3 The Structure of the English-Japanese Dictionary
(* ··· repeatable, % ··· omissible)

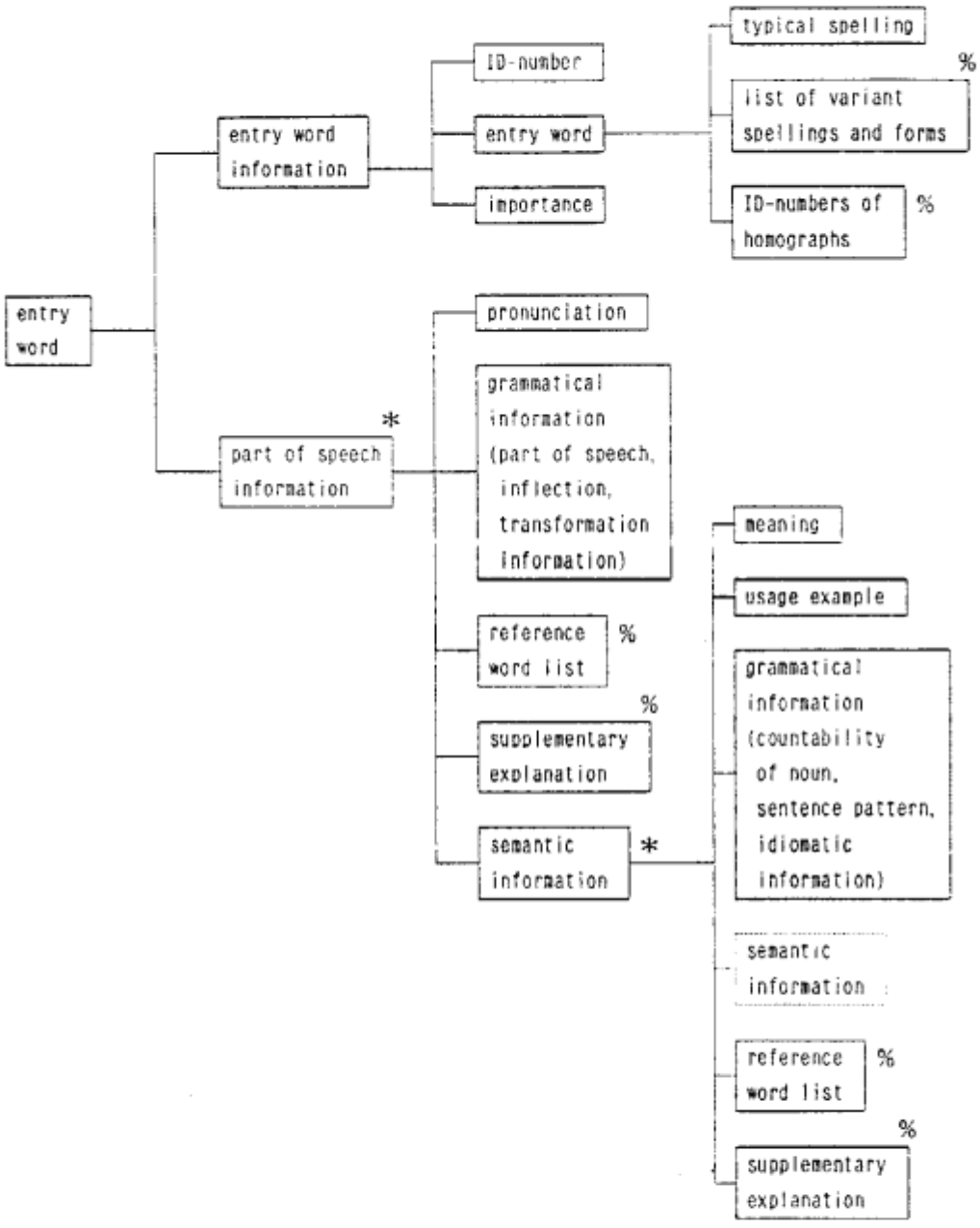

Fig. 4.4 The Structure of the English Dictionary
(* ··· repeatable, % ··· omissible)

4.2 Contents of Dictionary Items

4.2.1 Japanese Dictionary

(1) Entry Word Information

- **ID number**
  a unique number attached to the entry word
- **kana notation**
  kana notation of the entry word

(2) Part of Speech Information (*repeatable*)

- **entry word**
  - **typical form**
    the most generally used form
    e.g.
    取り扱う (to treat)
  - **list of variant forms** (*omissible*)
    written forms of the same Japanese word using different declensional kana
    e.g.
    取扱う, 取りあつかう (to treat)
- **pronunciation**
  pronunciation and accent
- **grammatical information**
  part of speech and inflection
- **reference word list** (*omissible*)
  - **type**
    reference word types
  - **ID number**
    ID number of the reference word
  - **reference word**
    entry word of the reference word

  [note] reference word types are as follows:

  (A) other part of speech form
  other part of speech form produced by inflection

e.g.

繰り上げる (to advance) ⟶
繰り上げ (n. advance)

悲しい (sad) ⟶ 悲しさ (sadness)

(B) compound word

compound word containing the entry word

e.g.

曇り (cloud) ⟶ 曇り空 (cloudy sky)

(C) idiom

idioms which contain the entry word

e.g.

伝家 (heirloom) ⟶
伝家の宝刀 (heirloom sword)

- **supplementary explanation** (*omissible*)
  - **field labels**

    e.g.

    (医), (化), (数), etc.
  - **usage labels and notes**

    e.g.

    (古), (俗), (比喩), etc.
  - **other explanations**

    e.g.

    継手 （金属・木材などの）接合部分
- **semantic information** (*repeatable*)
  - **meaning**

    a meaning of the entry word
  - **usage example**

    usage examples of the entry word in context
  - **grammatical information**
    - **sentence patterns, idiomatic information**

      formulation of specific usage examples of a sentence or a phrase with specific adverbs or nouns

      e.g.

      場所を取る (to keep a place)

      年を取る (to get old)

栄養を取る　(to take some nourishment)
（中）に（液体）を注ぐ　(to pour some fluid into)
（ある地位・役・状態など）に付く
(to take up one's post)

- **reference word list for a particular meaning** (*omissible*)
  - **type**
  - **ID number**
  - **reference word**

  [note] reference word types are as follows:

  (A) other part of speech form

  (B) compound word

  (C) idiom

  (D) synonym

  e.g.
  神社，お宮　(shrine)

  (E) antonym

  e.g.
  成功　(success) ⟶ 失敗　(failure)

  (F) reference

  e.g.
  音韻　(phoneme) ⟶ 音声　(sound)

  [note] Synonyms, antonyms, and references are cross-referenced to one another.

- **supplementary explanation for a particular meaning** (*omissible*)
  - **field labels**
  - **usage labels and notes**
  - **other explanations**

4.2.2 Japanese-English Dictionary

(1) Entry Word Information

- **ID number**
  a unique number attached to the entry word
- **kana notation**
  kana notation of the entry word

(2) Part of Speech Information (*repeatable*)

- **entry word**
  - **typical form**
    the most generally used form
  - **list of variant forms** (*omissible*)
    written forms of the same Japanese word using different declensional kana
- **grammatical information**
  part of speech and inflection
- **reference word list** (*omissible*)
  - **type**
    reference word types
  - **ID number**
    ID number of the reference word
  - **reference word**
    entry word of the reference word

[**note**] reference word types are as follows:

(A) compound word
a compound word which contains the entry word
e.g.
美容 (beauty culture)
美容院 (a beauty salon)
美容師 (a beauty artist)

(B) idiom
a phrase which contains the entry word
e.g.
腹を立てる (to get angry)
茶を立てる (to make tea)
ふろを立てる (to prepare a bath)

- **supplementary explanation** (*omissible*)
  - **field labels**

    e.g.

    (医), (化), (数), etc.
  - **usage labels and notes**

    e.g.

    (米), (英), (俗), (比喩), etc.
  - **other explanations**

    grammatical or semantic explanations of usage, background knowledge, usage, etc.
- **semantic information** (*repeatable*)
  - **meaning**

    The meanings are given in Japanese also along with the English equivalents associated with them.

    e.g.

    たてる： 起す set up; put up
    建築する build
    敬う respect
  - **grammatical information**
    - **sentence patterns, idiomatic information**

      formulation of specific usage examples of a sentence or a phrase with specific adverbs or nouns

      e.g.

      たてる （会社などを） to organize (company etc.)
      （顔や義理を） to save one's face
  - **reference word list for a particular meaning** (*omissible*)
    - **type**
    - **ID number**
    - **reference word**

    [note] reference word types are as follows:

    (A) compound word

    (B) idiom
  - **supplementary explanation for a particular meaning** (*omissible*)
    - **field labels**
    - **usage labels and notes**

- **other explanations**

- **equivalent information** (*repeatable*)
  - **English equivalent**
  - **translation examples**

    corresponding Japanese and English expressions

4.2.3 English-Japanese Dictionary

(1) Entry Word Information

- **ID number**
  a unique number attached to the entry word
- **entry word**
  - **typical spelling**
    the most generally used spelling (hyphenated)
  - **list of variant spellings and forms** (*omissible*)
    When the entry word has more than one spelling, spellings except the typical one are listed here.
    e.g.
    color ⟶ colour
    ax ⟶ axe
  - **ID numbers of homographs** (*omissible*)
    cross-reference of homographs of the entry word
- **importance**
  the level of importance

(2) Part of Speech Information (*repeatable*)

- **pronunciation**
  accent and pronunciation
- **grammatical information**
  - **part of speech**
    a part of speech of the entry word
  - **inflection**
    plural form of noun; past, past participle, present participle, and third person forms of verb; comparative and superlative of adjective
  - **transformation information** (*omissible*)
    If the entry word is a transformed form of a basic word, then the ID number of the basic word and the relation between them are given.
    e.g.
    feet: foot の複数形

- **reference word list** (*omissible*)
  - **type**

    reference word types
  - **ID number**

    ID number of the reference word
  - **reference word**

    entry word of the reference word

  [**note**] reference word types are as follows:

  (A) other part of speech form

  other part of speech form produced by inflection, and affixed form (including same part of speech form)

  e.g.

  nature —— natural
  know —— knowledge
  incorrect —— incorrectly, incorrectness
  easy —— uneasy

  (B) specific noun form

  a noun which has specific meanings when capitalized or preceded by *the*

  e.g.

  highness: Highness 殿下
  devil: the Devil 魔王

  (C) compound word

  a compound word which contains the entry word

  e.g.

  coast: Coast Guard 沿岸警備隊

  (D) idiom

  a phrase which contains the entry word

  e.g.

  build: build in
  keep: keep … to oneself

- **supplementary explanation** (*omissible*)
  - **field labels**

    e.g.

    (医), (化), (数), etc.
  - **usage labels and notes**

e.g.

(米), (英), (俗), (比喩), etc.

- **instruction for usage**

  e.g.

  clothes: 複数扱い. many, much のいずれとも
  用いるが、口語では many のほうが普通.

- **semantic information** (*repeatable*)
  - **grammatical information**
    - **countability of noun**

      distinction between countable noun and uncountable noun
    - **sentence pattern, idiomatic information**

      formulation of specific usage examples of a sentence or a phrase with specific adverbs or nouns

      e.g.

      (rob A of B) A（人）のB（物）を奪う

      (harmful to) に有害な

      (be built of A) （建物）がA（材料）でできている
  - **reference word list for a particular meaning** (*omissible*)
    - **type**
    - **ID number**
    - **reference word**

    [note] reference word types are as follows:

    (A) other part of speech form

    (B) specific noun form

    (C) compound word

    (D) synonym

    e.g.

    hard ═══ difficult

    (E) antonym

    e.g.

    hard ⟷ easy

    (F) reference

    e.g.

    cow ⟶ bull, calf, heifer, steer
  - **supplementary explanation for a particular meaning** (*omissible*)
    - **field labels**

- **usage labels and notes**
- **instruction for usage**
- **supplementary explanation on usage**

  e.g.

  put: （副詞を伴って）を置く

- **explanation of synonyms**

  e.g.

  injure: *wound* は「戦傷．切傷を負わせる」
  hurt は「傷みを伴う怪我を負わせる」
  injure は一般語

- **equivalent information** (*repeatable*)
  - **Japanese equivalent**
  - **translation examples**

    corresponding English and Japanese expressions

4.2.4 English Dictionary

(1) Entry Word Information

- **ID number**

  a unique number attached to an entry word

- **entry word**
  - **typical spelling**

    the most generally used spelling (hyphenated)

  - **list of variant spellings and forms** (*omissible*)

    When the entry word has more than one spelling, spellings except the typical one are listed here.

    e.g.

    color ⟶ colour

    ax ⟶ axe

  - **ID numbers of homographs** (*omissible*)

    cross-reference of homographs of the entry word

- **importance**

  the level of importance

(2) Part of Speech Information (*repeatable*)

- **pronunciation**

  accent and pronunciation

- **grammatical information**
  - **part of speech**

    a part of speech of the entry word

  - **inflection**

    plural form of noun; past, past participle, present participle, and third person forms of verb; comparative and superlative of adjective

  - **transformation information**

    If the entry word is a transformed form of a basic word, then the ID number of the basic word and the relation between them are given.

    e.g.

    feet: *pl* of foot

- **reference word list** (*omissible*)
  - **type**

    reference word types

- **ID number**

  ID number of the reference word

- **reference word**

  entry word of the reference word

[note] reference word types are as follows:

(A) other part of speech form

other part of speech form produced by inflection, and affixed form (including same part of speech form)

e.g.

nature —— natural

know —— knowledge

incorrect —— incorrectly, incorrectness

easy —— uneasy

(B) specific noun form

a noun which has specific meanings when capitalized or preceded by *the*

e.g.

highness: Highness

devil: the Devil

(C) compound word

a compound word which contains the entry word

e.g.

coast: Coast Guard

(D) idiom

a phrase which contains the entry word

e.g.

build: build in

keep: keep ··· to oneself

- **supplementary explanation** (*omissible*)
  - **field labels**

    e.g.

    (med), (chem), (maths), etc.

  - **usage labels and notes**

    e.g.

    (US), (GB), (slang), (fig), etc.

  - **instruction of usage**

e.g.

clothes: *pl* (no *sing*; not used with numerals)

- **semantic information** (*repeatable*)
  - **meaning**

    a meaning of the entry word
  - **usage example**

    usage examples of the entry word in context
  - **grammatical information**
    - **countability of noun**

      distinction between countable noun and uncountable noun
    - **sentence pattern, idiomatic information**

      formulation of specific usage example of a sentence or a phrase with specific adverbs or nouns

      e.g.

      rob sb/sth *of* sth

      be built of/out of parts/materials
  - **reference word list for a particular meaning** (*omissible*)
    - **type**
    - **ID number**
    - **reference word**

    [**note**] reference word types are as follows:

    (A) other part of speech form

    (B) specific noun form

    (C) compound word

    (D) synonym

    e.g.

    hard ═══ difficult

    (E) antonym

    e.g.

    hard ⟷ easy

    (F) reference

    e.g.

    cow ⟶ bull, calf, heifer, steer
  - **supplementary explanation for a particular meaning** (*omissible*)
    - **field labels**
    - **usage labels and notes**

- **supplementary explanation on usage**

  e.g.

  put: (*with adverb*) to move sth to a stated place

- **instruction of usage**
- **synonym explanation**

  e.g.

  injure: *wound* — to damage the body esp by a weapon
  hurt — to cause a person to feel pain
  injure — a generic word

5 REMAINING TASKS

5.1 Specification of Case Information

Verbs can be classified by verb patterns. Moreover, for each verb, there exist some restrictions on the semantic feature of the subject, direct object, indirect object, and so on. For example,

N(+hum) *inform* N(+hum)
N(+hum) *inform* N(+hum) *of* N(+abs)
N(+hum) *inform* N(+hum) that-clause
N(+hum) *inform* N(+hum) interrogative clause.

This kind of information will be added to each verb in the second stage. As mentioned before, the usage examples themselves in our dictionaries will also be used to investigate such information. We have been researching on this information, since the basic specifications were fixed.

5.2 Specification of Semantic Information

Computational description of meaning has not been established yet. This is still a research theme. So we cannot fully describe the word in machine-readable manner. To make up for it, we will use thesaurus codes or semantic features. We are discussing in the Research Group which information we should adopt from the viewpoint of field of application, and developing cost and time. We have begun our research by investigating several existing thesauruses to find appropriate classification of concept and meaning for general-purpose thesaurus.

ACKNOWLEDGEMENT

The authors would like to thank the members of the Machine-Readable Dictionary Research Group. The main part of this paper is taken from the discussions in the Research Group. We also wish to express our thanks to Kazuhiro Fuchi, Director of ICOT Research Center, who, as the pioneer in this research area, provided us with the opportunity to pursue this development in the Fifth Generation Computer Systems Project at ICOT.

REFERENCES

[JEIDA 1982] JEIDA (ed.), *Investigation of Japanese Language Information Processing Technology* (in Japanese), 57-c-439, 1982.

[Murata 1982] Murata, K., Muraki, S., *Lexicon of Japanese Language Information Processing* (in Japanese), Proc. of 25th IPSJ National Conference, 1982.

[Nagao 1980] Nagao, M., *Handbook of the Machine-Readable New Concise English-Japanese Dictionary* (in Japanese), Kyoto Univ., 1980.

[Nagao 1982] Nagao, M., Nakamura, J., Hatazaki, K., Fujita, K., *The Application of Longman's Dictionary Database to Machine Translation* (in Japanese), NLWG preprints of IPSJ, 1982.

[Nagao 1983] Nagao, M., *Outline of the Machine Translation Project of the Science and Technology Agency* (in Japanese), NLWG preprints of IPSJ, 1983.

[Nishio 1982] Nishio, M., et al (ed.), *Iwanami's Japanese Dictionary* (in Japanese), Iwanami Shoten Publishers, 1982.

[Ogino 1981a] Ogino, T., *Development of a Japanese Dictionary Database* (in Japanese), IBS Research Report, 1981.

[Ogino 1981b] Ogino, C., Ogino, T., Fuchi, K., Tanaka, H., Yokoyama, S., *Adding Thesaurus Codes to the Japanese Dictionary Database based on 'Bunrui-Goihyo' ('Word List by Semantic Principles')* (in Japanese), NLWG preprints of IPSJ, 1981.

[Walker 1984] Walker, D. E., *Machine-Readable Dictionaries,* Proc. of Coling84, 1984.

[Yokoyama 1977] Yokoyama, S., *Preparation for the Database Management of a Japanese Dictionary* (in Japanese), Bul. Electrotechnical Laboratory, Vol. 41, No. 11, 1977.

[Yokoyama 1984] Yokoyama, S., Ogino, T., *Documentation for a Japanese Dictionary Database* (in Japanese), Bul. Electrotechnical Laboratory, Vol. 48, No. 8, 1984.