[MT Summit IV, July 20-22, 1993, Kobe, Japan]

# EDR Electronic Dictionary

Yoichi Takebayashi
Japan Electronic Dictionary Research Institute, Ltd. (EDR)
Mita-Kokusai Building Annex, 4-28, Mita 1-chome Minato-ku, Tokyo 108
Phone: +81-3-3798-5521; FAX: +81-3-3798-5335
Email: yoichi@edr5r.edr.co.jp

## 1. Introduction

Advances in computing environments have facilitated many application systems using various media, including text, graphics and speech. In office or personal daily applications, natural language plays a central role in these media, and the amount of accessible document data containing natural language information is rapidly increasing. Therefore, it has become increasingly necessary that computers have the ability to process and understand natural language. To generate and understand spoken or written language, a computer requires linguistic knowledge about the meanings of words and their usage.

In the past ten years, much effort has been devoted to building large-scale knowledge-bases for real-world intelligent systems, including machine translation systems and intelligent word processors [1-7]. Since 1986, we have been developing a large-scale linguistic knowledge-base, the EDR Electronic Dictionary, which has the following features:

(1) Large-scale, covering all the vocabulary used in ordinary writing

(2) Aimed at general-purpose applications and not biased toward particular application systems or algorithms

(3) Provided with the knowledge-base required for practical semantic analysis

(4) Highly objective, based on a large volume of text

(5) Highly generalized across different languages and application fields

This paper describes our approach to building the large-scale EDR Electronic Dictionary and summarizes its structure, characteristics and specifications.

## 2. Structure of the EDR Electronic Dictionary

### 2.1 Structure of the EDR Electronic Dictionary

The EDR Electronic Dictionary is composed of four types of dictionaries (Word Dictionary, Concept Dictionary, Cooccurrence Dictionary and Bilingual Dictionary) and the EDR Corpus, as shown in Figure 1.

The Word Dictionary is divided into a General Vocabulary Dictionary and a Technical Term Dictionary. The General Vocabulary Dictionary is further divided into Japanese and English General

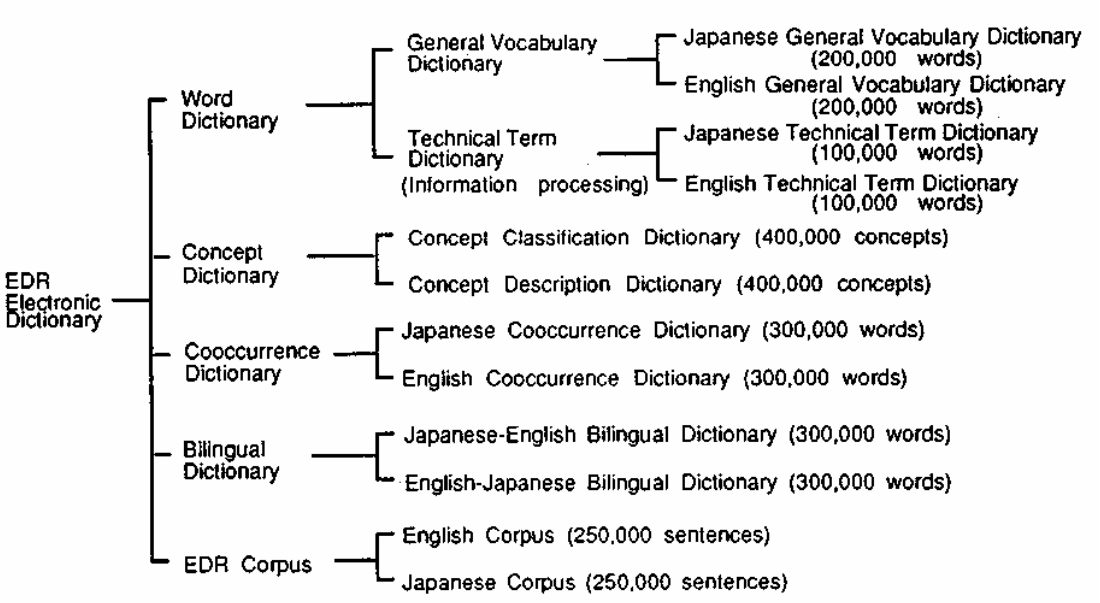


Figure 1. Structure of the EDR Electronic Dictionary

Vocabulary Dictionaries of 200,000 words each. The Technical Term Dictionary is available both in Japanese and English, covering the field of information processing, each containing 100,000 words. The Word Dictionary contains grammatical properties and concepts represented by each word, which would serve to ascertain the syntactic structure of a sentence.

The Concept Dictionary contains information concerning the 400,000 concepts defined in the Word Dictionary and is divided into Concept Classification and Concept Description Dictionaries. The Concept Dictionary provides computers with information necessary to understand concepts listed in the Word Dictionary.

The Cooccurrence Dictionary is created for each language. Japanese and English Cooccurrence Dictionaries contain 300,000 words each. They provide information on word usage to select the appropriate word when the computer tries to generate a sentence.

The Bilingual Dictionary is classified by the direction of translation. Japanese-English and English-Japanese Dictionaries list 300,000 headwords each and provide information on the correspondences between Japanese and English headwords.

The EDR Corpus consists of a Japanese Corpus and an English Corpus, each containing 250,000 sentences. A large number of examples on usage are provided along with linguistic data collected as a result of sentence analysis. The Corpus was originally prepared for the development of the EDR Electronic Dictionary, but it is also useful in various research on natural language processing.

## 2.2 Guidelines for the Development of the EDR Electronic Dictionary

In order to provide a large-scale linguistic knowledge-base, we have designed the EDR Electronic Dictionary under the following guidelines:

(1) The dictionaries that handle surface information are completely separated from the dictionaries that handle semantic information. Surface information, which is heavily dependent upon a particular language, is stored in the Word Dictionary, while semantic information is stored in the Concept Dictionary.

(2) Correspondence between concepts and headwords is established in the Word Dictionary.

(3) Information which is dependent upon specific grammatical rules and algorithms is excluded, and listings of information on words and the concepts they represent are based solely on a large volume of text.

The aim of guideline (1) is to share voluminous semantic information among various dictionaries in each language.

Guideline (2) provides a means of access from the Word Dictionary to the Concept Dictionary and allows concept identifiers to be used as an interface to link the dictionaries of each language. This guideline also permits language dictionaries to be developed independently of each other. Thus, surface information, which is heavily language-dependent, is stored in the Word Dictionary, while semantic information encoded in the Concept Dictionary is shared among particular language dictionaries. Information on set phrases and their corresponding expressions in other languages is kept apart from the Word Dictionary. The former information is used to compile the Cooccurrence Dictionary, and the latter the Bilingual Dictionary.

Guideline (3) ensures that our EDR Dictionary under development is universally applicable rather than being restricted to a specific system, and that the Dictionary has a broad scope for future application and development. At the same time, the guideline serves to alleviate the difficulty of guaranteeing the consistency and accuracy of information in the development of large-scale dictionaries.

## 3. Word Dictionary

The Word Dictionary provides morphological, syntactic and semantic information required to process natural language[8]. The Word Dictionary consists of a General Vocabulary Dictionary and a Technical Term Dictionary. The word entries of the General Vocabulary Dictionary have been selected from words typically used in daily life and from technical terms as well as proper nouns in common use. The word entries of the Technical Term Dictionary have been selected from specified fields of technology. The basic role of the Word Dictionary is to provide information on relations between words and concepts, and to provide grammatical attributes that hold in these relations.

The Word Dictionary is a set of word entries that contain headword information, morphological information, syntactic information, semantic information and supplementary information (Figure 2). Morphological information includes information on word (morpheme) entries and the juncture of morphemes. Morpheme entries are used in morphological analysis and morpheme generation.

Information on junctures dictates the morphemes which can be used in conjunction with a given morpheme.

Syntactic information, generally called grammatical attributes, includes parts of speech such as nouns, verbs and adjectives, as well as surface cases.

Semantic information consists of concept identifiers and concept illustrations for the concepts corresponding to various meanings of a word, and serves as an interface to the Concept Dictionary.

The principal features of the Word Dictionary are:

(1) Surface level information in the Word Dictionary is separated from deep (semantic) level information in the Concept Dictionary.

(2) Surface information is independent of grammatical rules and algorithms.

(3) Sufficient vocabulary for general writing is listed.

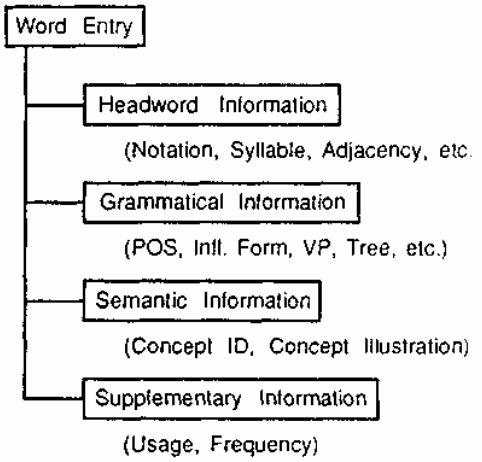


Figure 2. Structure of Word Entry

【Notation】 (Syllable)(Adjac.)[Gram. Attr.]<Concept ID>Conc. Illustr.

【dictionar】
(dic/tion/ar/) (Noun Beginning with Cons., Noun Stem Followed by *y/ies*)
[Common Noun; Uninflected; *y/ies* Inflection; Countable; Neuter]
<3d1db7 > 文字を類別して集めた書物 a book which classifies and lists characters

【dictionar】
(dic/tion/ar/) (Noun Beginning with Cons., Noun Stem Followed by *y/ies*)
[Common Noun; Uninflected; *y/ies* Inflection; Countable; Neuter]
< 0b8332 > any repository of knowledgeor information

【consult】
(con/sult) ( Verb Beginning with Cons., Verb Stem Followed by *s* (3 PS Pres) / *ed* (Past))
(Verb; Base; *s/ed* Inflection; Direct Object; DO = Noun; Specific Adverbial Phrase Required; ADP = Preposition + Noun; about; on]
< 0b5174 > to ask the advice or opinion of

Figure 3. Examples of the Word Dictionary

Figure 3 shows examples of the Word Dictionary. Further details are given in the technical guide [9].

## 4. Concept Dictionary

The Concept Dictionary contains knowledge on the concepts which are linked to the Word Dictionary [10-12]. The Concept Dictionary consists of three components: a set of concepts, a set of the Concept Descriptions and the Concept Classification, as shown in Figure 4.

The Concept Descriptions and the Concept Classification in the Concept Dictionary enable the following processing:

(1) Generating the appropriate semantic representations for sentences

(2) Determining the similarity of semantic contents

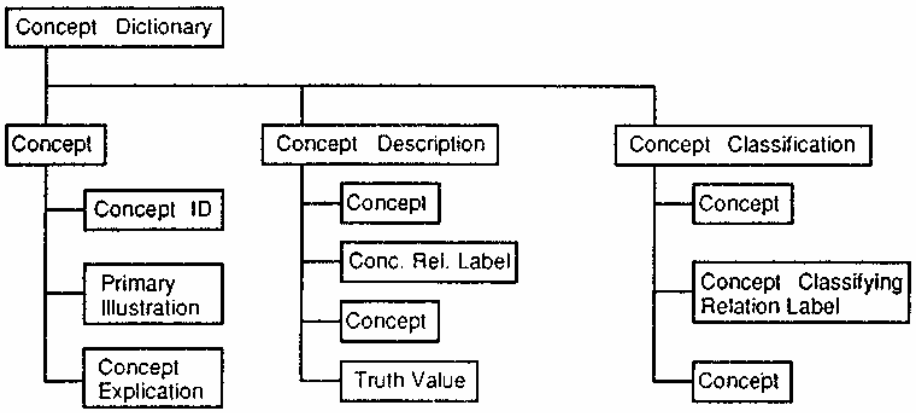


Figure 4. Structure of the Concept Dictionary

(3) Converting semantic contents into similar contents

## 4.1 Concept

A concept is an abstract entity which is used to represent a meaning of sentences. Most concepts are introduced as an entity corresponding to a meaning of a word. They are reduced by 'unification' [10]; when two or more concepts are judged to correspond to the same meaning, they are unified. Most of the concepts are linked to either Japanese words or English words, or both. Other concepts are introduced as categorical concepts under which similar concepts are classified.

Each concept has a concept identifier and a concept illustration, as in Figure 4. A concept identifier is a number which is uniquely assigned to a concept. A concept illustration contains data which gives an idea to human users as to what that concept really is. A concept explication is a description of the concept in natural language [10, 11].

## 4.2 Concept Description

A Concept Description is a tuple of a concept relation label and two concepts. The Concept Descriptions indicate the appropriateness of links between concepts, as shown in Figure 5. There are two kinds of Concept Descriptions. One is a description by a lexicographer based on his intuition, or is an abstraction from a set of descriptions in the EDR Corpus [13]. The other is descriptions found in the EDR Corpus.

## 4.3 Concept Classification

Concept Classification shows the hierarchical relationships between concepts, as in Figure 5. The Concept Classification classifies concepts under the same concept when they are judged to behave alike with respect to the Concept Descriptions. Classification is implemented by a set of tuples of "kind-of(concept,concept)." For example, two concepts located close in the Concept Classification indicate that they are similar or equivalent in meaning. When no concept relation exists between particular concepts in the Concept Descriptions, the Concept Classification can be used to refer to

the Concept Description between synonymous concepts.

With regard to meaning, no real common ground has yet been established on its theory, the form of representation or the mode of processing. Given this situation, we have taken the following approaches to developing a large-scale Concept Dictionary:

(1) Extensive and uniform, although not deep, meanings are covered.

(2) A general representation format, which is applicable to various languages, has been adopted.

(3) A flexible representation format, which enables transfer and expansion to various semantic theories and representation forms, has been adopted.

For further details, see the technical guide [9].

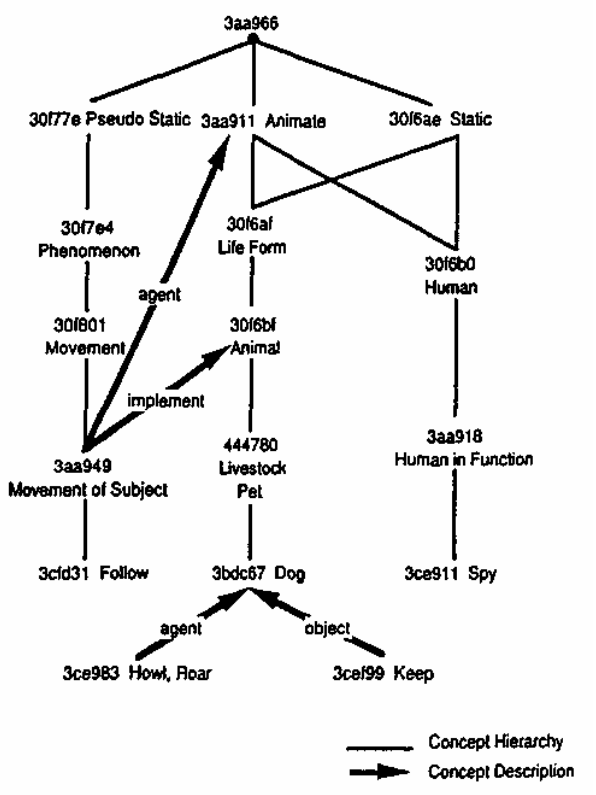


Figure 5. A Part of the Concept Dictionary

## 5. Cooccurrence and Bilingual Dictionaries

### 5.1 Cooccurrence Dictionary

The Cooccurrence Dictionary provides information on word usage. This information consists of the types of word combinations used to construct a sentence representing events or states [8]. The information on language use, contained in the Cooccurrence Dictionary, covers the relationships showing how given words are used with other words, and the types of words with which they can be used. This information constitutes the cooccurrence relation.

The Cooccurrence Dictionary extends ordinary cooccurrence relations to include syntactic relationships, by defining cooccurrence relations in terms of syntactic relationships in which words occur. The extended cooccurrence relations are listed in the Cooccurrence Dictionary. Even when a word cannot be identified solely by its surface information, detailed cooccurrence relations can be defined by specifying its grammatical characteristics and the concepts expressed by the word.

The Cooccurrence Dictionary is used to select the optimum word to express a given concept in relation to other words. This type of information is useful in selecting corresponding words in the target language in machine translation.

### 5.2 Bilingual Dictionary

While existing machine-readable dictionaries provide only meaning retrieval and display functions, the EDR Bilingual Dictionary provides useful information for natural language processing by computers [8].

The Bilingual Dictionary is composed of Japanese-English and English-Japanese Dictionaries. The Bilingual Dictionary defines the corresponding words (i.e. translation equivalents) for each headword in both the Japanese Word Dictionary and the English Word Dictionary. The entries in the Bilingual Dictionary contain corresponding words, supplementary information, as well as the correspondence relations between headwords and their corresponding words.

In order to facilitate natural language processing, the information in an entry is described following the guidelines below:

(1) Compact representation of corresponding words have priority (no explanatory or sample sentences are used in the corresponding word description).

(2) Correspondence relations between a headword and its corresponding word are clearly indicated by four inter-lingual correspondence labels: equivalent relation, synonymous relation, subset relation and superset relation.

(3) Supplementary explanation for synonymous, subset and superset corresponding relations is given separately from the corresponding word. For further details, see the technical guide [9].

## 6. EDR Corpus

### 6.1 EDR Corpus

The EDR Electronic Dictionary consists of a number of interrelated large-scale dictionaries such as the Word Dictionary and the Concept Dictionary. In order to develop these Dictionaries, a new methodology is required. The development of the EDR Corpus is central to the progress of the development of individual dictionaries [13].

The EDR Corpus is created in the following processes:

(1) Text collection
(2) Creation of KWIC data
(3) Sentence selection
(4) Sentence analysis

The collection of a text base includes the materials of the EDR Corpus, and its eventual scale will be 20 million sentences in each of the English Corpus and the Japanese Corpus. The principal sources of text are newspapers, encyclopedias, educational and reference texts.

KWIC (Key Word in Context) data is created through morphological analysis of all the sentences in the text base. KWIC data is used to detect words that are not yet registered in the Word Dictionary, to calculate the frequency of words and to verify grammatical information of words.

The process of sentence selection selects the sentences which must be analyzed from the sentences used to create KWIC data.

The process of sentence analysis extracts syntactic and semantic structures of sentences. The semantic structure is represented by the concept relation representation used in the Concept Dictionary. As shown in Figure 6, each sentence in the EDR Corpus is attached with the following syntactic and semantic information:

(1) Constituent morphological information
(2) Syntactic tree of a sentence
(3) Concept relation representations

The data in the Concept Dictionary and Cooccurrence Dictionary is extracted from the concept relation representations and the syntactic trees in the EDR Corpus. Besides the development of the EDR Electronic Dictionary, the EDR Corpus is applicable to a variety of research tasks in the field of natural language processing.

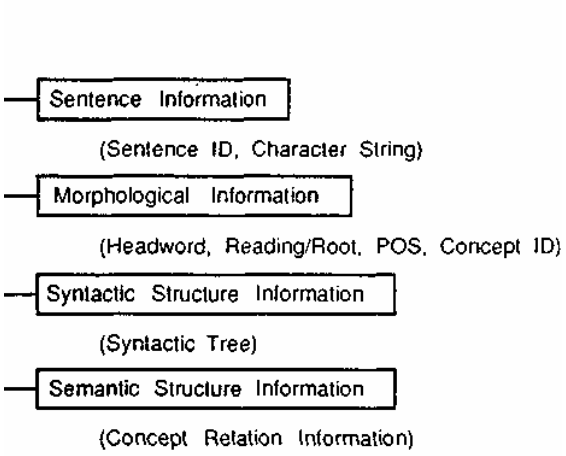


Figure 6. Structure of the EDR Corpus

## 6.2 Dictionary Development Based on the EDR Corpus

The development of the EDR Dictionary based on the EDR Corpus is performed as in Figure 7. In the corpus creation process, automatic analysis using the EDR Dictionary plays an important role. Every selected sentence is analyzed automatically and then revised manually. In order to list supplementary information on selected sentences in the EDR Corpus, the sentences are first morphologically analyzed, and then syntactically and semantically analyzed. This information is also revised manually.

The results of sentence analysis are stored in the EDR Corpus, which verifies and improves the EDR Dictionary as below:

(1) Concept relation representations in the EDR Corpus are used to verify and improve the Concept Dictionary.

(2) Syntactic trees in the EDR Corpus are used to create cooccurrence relations in the Cooccurrence Dictionary.

(3) Constituent information is used to obtain statistical data on word usage.

Frequency of a word and its usage supplies the data for assessing the importance of words and validity of its descriptions in the Word Dictionary.

Through the development of the interrelated EDR Dictionaries (the Word Dictionary, the Concept Dictionary, the Cooccurrence Dictionary and the Bilingual Dictionary), we have found the effectiveness of the large-scale EDR Corpus in improving these individual dictionaries. Besides the development of the EDR Electronic Dictionary, the EDR Corpus is useful in various tasks concerning natural language processing.

## 7. Conclusion

We have discussed our approach to creating the EDR Electronic Dictionary and presented an overview of its structure and features. Having developed the Word Dictionary, the Cooccurrence Dictionary and the Bilingual Dictionary, we are now completing the development of the Concept Dictionary. We are also evaluating and improving these dictionaries through a number of evaluation and application systems, including machine translation systems [14] and a voice-activated document preparation system [15]. We plan to continue to strive toward improving our EDR Electronic Dictionary, which we feel will play a significant role in many intelligent information processing and knowledge-based systems.

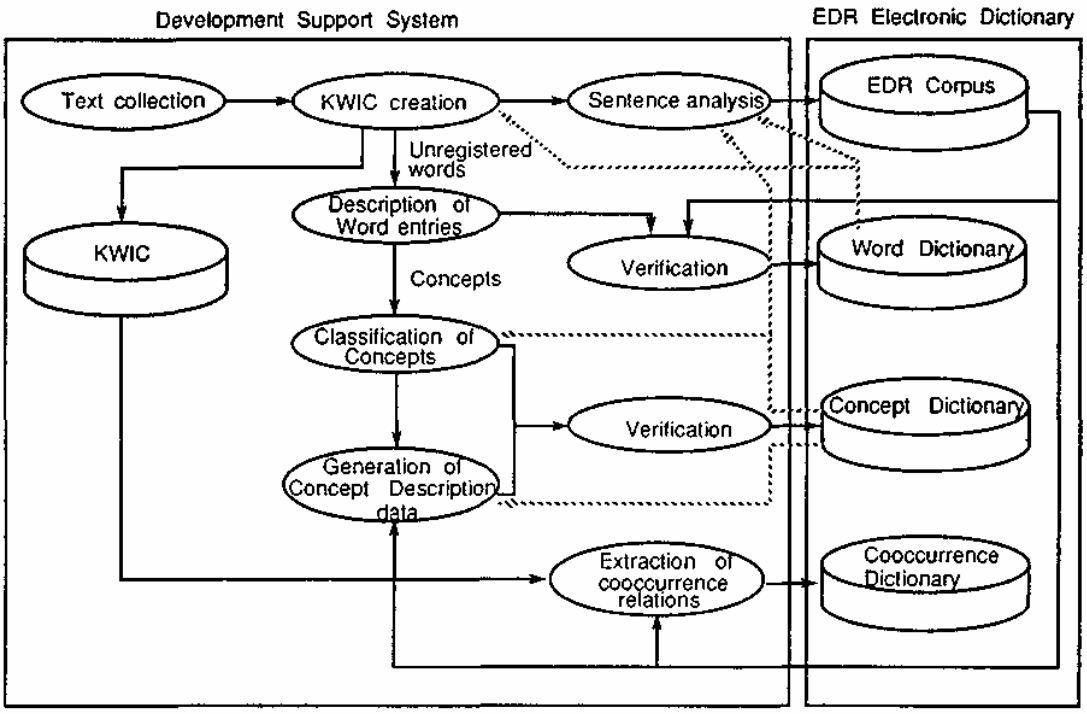


Figure 7. Procedures for Developing the EDR Dictionary


## Acknowledgments

The author is grateful to Chiaki Harada, Hiroyuki Suzuki, Yoshio Nakao, and Hiroshi Suematsu and his colleagues of EDR for their contribution to this paper, and to Toshio Yokoi and Seibi Chiba for valuable comments on drafts in the preparation of this paper.

## *References*


[ 1 ] Yokoi, T: "A Very Large-Scale Knowledge Base Based on the Amalgamation of Knowledge Processing and Natural Language Processing", Journal of JSAI, Vol.8, No.3, pp.286-296 (1993)

[ 2 ] Lenat, D. B and Guha, R.V. : Building Large Knowledge-Based Systems - Representation and Inference in the CYC Project, Addison-Wesley (1990)

[ 3 ] Lehnert, W. and Sundheim, B. : "A Performance Evaluation of Text-Analysis Technologies", AI Magazine, Vol.12, No.3, pp.81-94

[ 4 ] Minsky, M : Society of Mind, Simon and Schuster (1985)

[ 5 ] Sampson, G. : Note for Corpus Meeting, Wadham College. Oxford. (1990)

[ 6 ] Walker, D. : "The Ecology of Language", Proc. of International Workshop on Electronic Dictionaries, pp. 10-22 (1990)

[ 7 ] Neches, R., et al. : "Enabling Technology for Knowledge Sharing", AI Magazine, Vol.12, No.3, pp.36-56 (1991)

[ 8 ] Harada, C. : "EDR Word Dictionary, Cooccurrence Dictionary, and Bilingual Dictionary", Japan-U.S. Joint Workshop on Electronic Dictionaries and Language Technologies. (1993)

[ 9 ] EDR : "EDR Electronic Dictionary Technical Guide", TR-042 (1993)

[10] Miike, S.: "How to Define Concepts for Electronic Dictionaries", Proc. International Workshop on Electronic Dictionaries, EDR TR-031, pp.43-49. (1991)

[11] Suzuki, H.: "EDR Concept Dictionary", Japan-U.S.Joint Workshop on Electronic Dictionaries and Language Technologies. (1993)

[12] Nomura, N : "Functions of the Set of Concept Explications in an MTD and a Methodology for Developing Bilingual Concept Explications", Japan-U.S. Joint Workshop on Electronic Dictionaries and Language Technologies. (1993)

[13] Nakao, Y.: "EDR Corpus and Its Application to Concept Dictionary Development", Japan-U.S. Joint Workshop on Electronic Dictionaries and Language Technologies. (1993)

[14] Yasuhara, H : "An Example-Based Multilingual MT System in a Conceptual Language", Proc. of the MT SUMMIT (1993)

[15] Kanazawa, H., Takebayashi, Y : "A Text Generation Support System with EDR Electronic Dictionaries Using Speech Input", Information Processing Society of Japan, 2-243 (1993)