Development of the Concept Dictionary - Implementation of Lexical Knowledge

Tomoyoshi MATSUKAWA, Eiji YOKOTA
Japan Electronic Dictionary Research Institute, Ltd. (EDR)
Mita-Kokusai-Bldg. 4-28, Mita 1-chome,Minato-ku, Tokyo, 108, JAPAN
e-mail : matsu@edr5r.edr.co.jp
yoko@edr7r.edr.co.jp
tel: 81-3-3798-5521,
fax: 81-3-3798-5335

Summary

The methodology of development of the Concept Dictionary being compiled by EDR, which is to be a neutral dictionary for semantic processing of natural languages available for various application systems, is described. The Concept Dictionary is based on several linguistic semantic representation theories and consists of : a) concept descriptions, and b) the concept taxonomy. Moreover, preference knowledge is being collected from output data of various testing systems.

1. Introduction

A dictionary in which dependencies among 400,000 word senses of the English and Japanese languages are described in detail (Concept Dictionary) is being developed by EDR. The goal of the development is to build a neutral dictionary for natural language semantic processing that is available for various application systems. The implementation of the dictionary is based on several linguistic semantic representation theories.

For a long time, a series of trials for describing dependencies among words or word senses by bundling verbs, adjectives, etc., has been conducted. Establishing a deep case level and using a formalism independent of each language, Fillmore developed a theory of representation of dependency among words (Fillmore 1968). On the other hand, Fodor and Katz explained a mechanism of selecting interpretations of constituents in a sentence by using a formalism composed of a semantic marker, distinguisher and selection restriction (Katz and Fodor 1963). In contrast to these theories, Wilks proposed a point of view to consider word dependency not as a constraint but as a preference (Wilks 1975). In addition, Schank proposed to abstract connotations not only from senses of nouns but also those of verbs, and he named them "primitive actions" (Schank 1975). These semantic representation theories have been reviewed and used in developing practical natural language processing systems (Nagao 1985, etc.).

As such development of practical natural language processing systems progressed, the importance of accumulating lexical descriptions became recognized by developers of such systems. That is, a dictionary large enough in terms of both the granularity of semantic markers and the number of words or word senses became necessary to build. Against the background of the situation, the development of the Concept Dictionary began (Kakizaki 1987, Yokoi et. al. 1989, Uchida, 1990, Miike et. al. 1990a). The methodology of development of

the Concept Dictionary, which consists of a) concept descriptions, which represent dependencies among concepts and categories, and b) the concept taxonomy, which represent super-sub relations among concepts, is described in sections 2 and 3. Preference knowledge, which represents preference order of concept descriptions, is explained in section 4. In section 5, spheres of applications and limitations of the dictionary are discussed.

2. Development of Concept Descriptions

Concept relations are described at the following three levels:

a) concept-concept relation descriptions
b) concept-category relation descriptions
c) category-category relation descriptions

2.1 Concept-concept Relation Descriptions

We are building an on-line corpus which includes 1,000,000 practical example sentences that are analyzed lexically, syntactically and semantically for the most part manually (EDR corpus). Figure 1 is an example of an entry of the corpus.

Firstly, (a) is the word sense selection (lexical analysis) section, where a word sense (concept) has been selected for each word in the sentence. Secondly, (b) is the syntactic analysis section, where all binding relations among words has been analyzed. Finally, (c) is the semantic analysis section, where the semantic network representing the meaning of the sentence is decomposed into a set of triplets. These triplets correspond to the following concept-concept relation descriptions:

(1)

c#enlarge —<and>→ c#new	c#membership ←<object>— c#enlarge
c#membership ←<object>— c#new	c#membership —<modify>→ c#its
c#bear —<location>→ c#it	c#bring —<goal>→ c#bear
c#bring —<cause>→ c#membership	c#pressure ←<object>— c#bring
c#vulnerable —<goal>→ c#pressure	c#vulnerable —<and>→ c#fluid
c#fluid —<modify>→ c#still	c#vulnerable —<modify>→ c#still
c#fluid —<object>→ c#the_U.N	c#vulnerable —<object>→ c#the_U.N
c#fluid —<modify>→ c#structurally	c#vulnerable —<modify>→ c#structurally

As shown above, concept-concept relation descriptions are extracted directly from the semantic analysis section (and word sense selection section) in the EDR Corpus. A method of collecting and selecting source sentences for the EDR corpus is described in (Nakao 1990a) and a method of extracting concept descriptions from the EDR corpus is explained in detail in (Nakao 1990b).

Source texts of the EDR corpus are selected so as to diversify as much as possible the concepts in them. However, it is impossible to collect all concepts or concept relations from the corpus even if the amount of texts is very large. To compensate for the shortage of examples, we also create example sentences and analyze them lexically and semantically. Concept-concept relation descriptions are also extracted from the sentences.

<<Text No : 00040000187d : 6/13/90 from e0327#03.ya1/4>>

=>> Structurally, the U.N. is still fluid and vulnerable to the pressures that its new and enlarged memberships are bringing to bear upon it.

```
$$(LEX_Start)$$
  1  structurally (structurally) <ADV>
        c#(0da914)in_a_structurel_manner)
  4.3 U.N. (U.N.) <NOUN>
        c#(ZZZZZZZ)the_organization_named_U.N.
  8  still (still) <ADV>
        c#(0da0f3)even_up_to_now_or_then_and_at_this_or_that_moment
  10 fluid (fluid) <ADJ>
        c#(0bd848)unsettled;_not_fixed
  13 vulnerable (vulnerable) <ADJ>
        c#(0a16e3)(of_a_place_or_thing)weak;_not_well_protected;_easily_attacked
  19 pressure (pressure) <NOUN>
        c#(0d05d5)trouble_that_causes_anxiety_and_difficulty
  26 new (new) <ADJ>
        c#(0ca953)having_begun_or_been_made_only_a_short_time_ago_or_before
  30 enlarg (enlarge) <VT>
        c#(0bad63)to_cause_to_grow_larger_or_wider
  33 membership (membership) <NOUN>
        c#(0c8477)the_state_of_being_or_status_as_a_member
  37 bring (bring) <VT>
        c#(0b0f68)to_cause_to_reach_a_certain_state
  42 bear (bear) <VI>
        c#(0af3df)to_exert_pressure
$$(LEX_End)$$
```

(a) Word Sense Selection Section

```
$$(SYN_Start)$$
  ; 1 structurally ----------------------------------------------------------------------------------------------------------------------------
  ; 4 the_U.N.  ----------------------------------------------------------------------------------------------------------------------------
  ; 6 is        ------------------------------------------------------ ; 13-3,S
  ; 8 still     ----------------------------;13-2,M                :
  ; 10 fluid    ; 13-1,S                  :                        :
  ; 13 vulnerable  fluid,_vulnarable -->fluid,_vulnarable -->is_fluid,_vulnarable ----------------------------------------------------------
  ; 15 to       ----------------------------------------------------------------------------------------------------; 19-3,S
  ; 17 the      ; 19-1,S                                                                                                        :
  ; 19 preasure   pressure_s -->the_pressure ----------------------------------------------->the_pressure -->to_the_pressure ;
  ; 22 that     ----------------------------------------------------------------------------; 37-5,S     :
  ; 24 its      ----------------------------------------------------; 33-2,M                 :          :
  ; 26 new      ----------------; 30-2,S                           :                       :          :
  ; 28 and      ----------------;                                  :                       :          :
  ; 30 enlarg   enlarged -->new_and_enlarged ; 33-1,M             :                       :          :
  ; 33 membership  ----------------------->membership -->membership ; 37-4,M             :          :
  ; 35 are       ----------; 37-2,S                                :                       :          :
  ; 37 bring    bringing -->are_bringing -->are_bringing ---------->are_bringing -->are_bringing ; 19-2,M
  ; 40 to       ; 42-1,S                  :
  ; 42 bear     to_bear -->to_bear----------; 37-3,M
  ; 44 upon     ; 46-1,S       :
  ; 46 it       upon_it ------; 42-2,M
$$(SYN_End)$$
```

(b) Syntactic Analysis Section

```
$$(SEM_Start)$$
  30 3 enlarged_          26 new_                 S and       > *
  33 2 membership_        30 new_and_enlarged_    M object    <
  33 3 membership_        24 its                  M modify    >
  42 2 bear_              46 upon_it              M location  >
  37 4 are_bringing_      42 to_bear_             M goal      >
  37 5 are_bringing_      33 membership_          M cause     >
  19 4 the_pressures_     37 are_bringing_        M object    <
  13 2 vulnerable_        19 to_the_pressures_    M goal      >
  13 3 vulnerable_        10 fluid                S and       > *
  13 4 fluid,_vulnerable_  8 still_               M modify    >
  13 6 is_fluid,_vulnerable_  4.3 the_U.N._       M object    >
  13 7 is_fluid,_vulnerable_  1 structurally,_    M modify    >
$$(SEM_End)$$
```

(c) Semantic Analysis Section

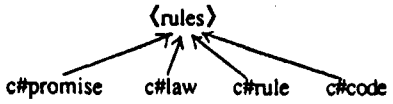
Figure 1. An Entry of the EDR Corpus

2.2 Concept-category Relation Descriptions

If some concept-concept relations share a concept, it is possible to bundle them into a representation. For example, concept-concept relation descriptions (2) can be bundled into a concept-category relation description (4), if a super-sub relation (3) is also described simultaneously within the concept taxonomy. This level of description corresponds to Fodor and Katz's representation using semantic markers and selection restrictions.

(2) c#break —<object>→ c#promise (to break a promise)
c#break —<object>→ c#law (to break a law)
c#break —<object>→ c#rule (to break a rule)
c#break —<object>→ c#code (to break a code)

(3)



(4) c#break —<object>→ ⟨rules⟩

2.3 Category-category Relation Descriptions

In the previous section, we discussed the cases in which filler concepts of deep case patterns can be bundled into categories, namely concept-category relations. Moreover, frame concepts in deep case patterns can also be bundled and represented by categories (frame categories) (Ogino et. al. 1989). For example, the three categories ○1.2.6, ○6.8.2 and ○9.14.4 defined in Figure 2 (Hereafter, the notation "○" also means a category,) bundle concepts and are linked with other categories to describe category-category relations (5), (6) and (7) :

(5) ○1.2.6 —<agent>→ ⟨Animals⟩
○1.2.6 —<object>→ ⟨Physical_Objects⟩
○1.2.6 —<implement>→ ⟨Parts_of_Animals⟩

(6) ○6.8.2 —<agent>→ ⟨Human_Beings⟩
○6.8.2 —<object>→ ⟨Information⟩ ; ⟨Things_With_Information⟩
○6.8.2 —<goal>→ ⟨Human_Beings⟩ ; ⟨Information_Accepters⟩

(7) ○9.14.4 —<object>→ ⟨Animals⟩

This level of descriptions corresponds to Schank's representation using primitive actions, in a sense. For example, ○6.8.2 includes a connotation that can be represented by MTRANS, which is one of the primitive actions, although other frame categories do not always correspond to a primitive action. In addition, relations between verbs and adverbs, for example those mentioned in (Lakoff 1966), are also described at this level.

○1.2.6 ⟨(For_an_Animal_to_Touch_a_Physical_Object)(With_a_Part_of_the_Body)⟩

[<agent> : ⟨Animals⟩ ,
<object> : ⟨Physical_Objects⟩ ,
<implement> : ⟨Parts_of_Animals⟩]

[push (<agent>: "a person", <object>: "a button", <implement>: with "a finger")]
[press (<agent>: "a person", <object>: against "a door", <implement>: with "one's hands")]
[kick (<agent>: "a person", <object>: "a ball", <implement>: with "a foot")]
[step (<agent>: "a person", <object>: on "a can", <implement>: with "a foot")]
[grasp (<agent>: "a person", <object>: "a ball", <implement>: with "a hand")]
[lift (<agent>: "a person", <object>: "a box", <implement>: on "one's shoulder")]

○6.8.2 ⟨(For_Human_Beings)_To_Send_(Information)(To_Information_Accepter)⟩

[<agent> : ⟨Human_Beings⟩ ,
<object> : ⟨Information⟩ | ⟨Things_with_Information⟩ ,
<goal> : ⟨Human_Beings⟩ | ⟨Information_Accepters_other_than_Human_Beings⟩

[speak (<agent>: "he", <object>: about "the_story", <goal>: to "her")]
[tell (<agent>: "he", <object>: "the_way", <goal>: "the_traveler")]
[describe (<agent>: "he", <object>: "the_situation", <goal>: in "the_book")]
[explain (<agent>: "he", <object>: "the_plan", <goal>: to "his_boss")]
[write (<agent>: "he", <object>: "his_name", <goal>: on "the_sheet")]
[input (<agent>: "he", <object>: "the_data", <goal>: into "the_file")]
[copy (<agent>: "he", <object>: "the_document", <goal>: into "his_notebook")]

○9.14.4 ⟨For_Functions_(Of_Human_Beings)_to_Become_Lower⟩

[<object> : ⟨Animals⟩]

c#beaten, c#go_down, c#ill, c#collapse, c#dispirited
c#pyrosis, c#sinophobia, c#malnutrition, etc.

(The notation " [...] " is a deep case pattern to distinguish the category from the other categories at the same fine semantic cluster; called "distinctive pattern")

Figure 2. Categories ○1.2.6, ○6.8.2, ○9.14.4 and Examples of Their Sub-Concepts

3 Development of the Concept Taxonomy

The two kinds of concept descriptions including categories, namely concept-category relation descriptions and category-category relation descriptions, mentioned in the previous sections, must have their descendant concept-concept relations in order to become useful. That is, concepts must be able to be actually classified into categories included in such descriptions.

A concept can generally be classified into more than one categories (multiple classification). However it is difficult to make exhaustive multiple classification from the beginning, because in the case of multiple classification, we must compare concepts with categories mn times when there are m concepts and n categories. In the case of exclusive classification using a distinctive tree whose leaves mean categories, on the other hand, we must only compare concepts with nodes on the tree $O(m(\log n))$ times. Additionally, the

number of categories which share same sub-concepts with a category (cross categories) is generally much less than the number of all categories. Moreover, it is not so difficult to make a list of cross categories for each category (cross category list) in advance.

Considering the points mentioned above, we use the following method for concept classification : 1) exclusive classification : selecting categories which hardly share same sub-concepts (exclusive categories), and making the first classification using a distinctive tree locating the exclusive categories at its leaf level. 2) cross classification : making the second classification into categories other than exclusive categories, based on cross category lists, and building a concept taxonomy from the results of the second classification. 3) improvement of the Concept Dictionary : modifying the Concept Dictionary based on the results from tests using various testing systems and automatic concept clustering from concept-concept relation descriptions. In the following sections 3.1 and 3.2, we explain the first exclusive classification and the second cross classification respectively. In section 3.3, we describe a method for modification of the Concept Dictionary.

3.1 Exclusive Classification of Concepts

3.1.1 Classification into MONO-Categories

The first classification into categories for nominal concepts (MONO-concepts) is made by using the MONO-concept taxonomy as shown in Figure 3 as a distinctive tree. That is, the classification starts from the top node, decends along branches of the tree, and when reaching a node, compares the node's children nodes with the input concept. This process is repeated, and if one of the leaves of the tree is reached, the MONO-category corresponding to the leave should be selected.

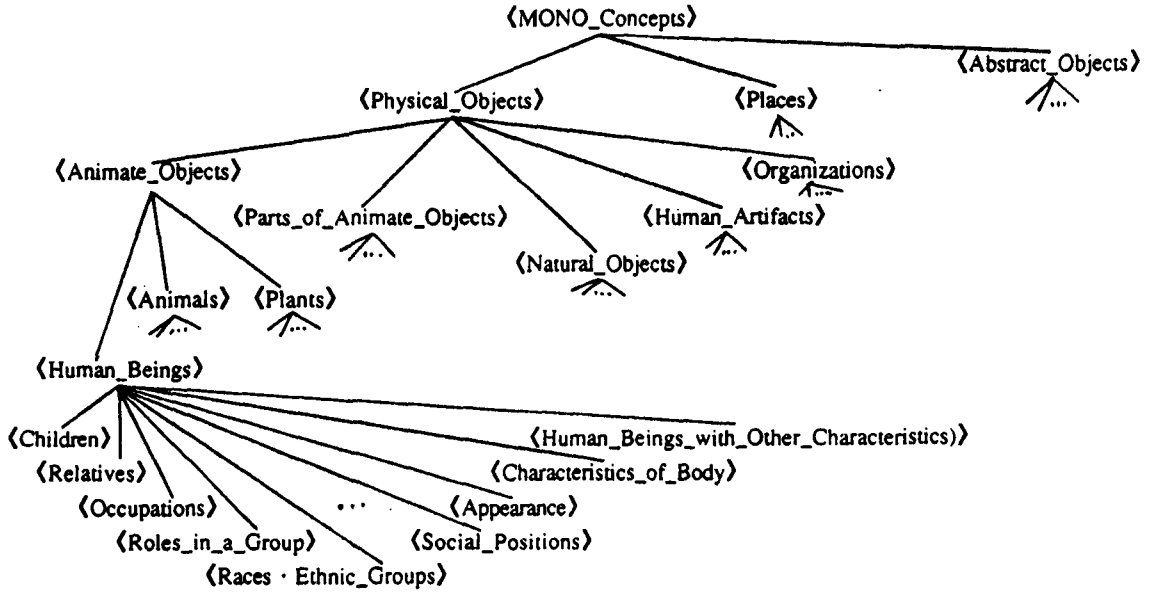


Figure 3. the MONO-Concept Taxonomy

For example, in the case of the concept "c#police_man", when we start with the question, "Is that a physical object, a place or an abstract object?" (answer: a physical object), and pass through the questions, "Is that an animate object, a part of the body of an animate object, a natural object, a human artifact or an organization?" (answer: an animate object), "Is that a human being, an animal or a plant?" (answer: a human being), and "Is that a child, a relative, an occupation ...?" (answer: an occupation), then we can classify the concept into the category 〈Occupations〉.

3.1.2 Classification into KOTO-Categories

As mentioned above, the first classification of the MONO-concepts is made by using the MONO-concept taxonomy's hierarchy as a distinctive tree. On the other hand, the method for the first classification of verbal concepts (KOTO-concepts) is not made by using the hierarchy as a distinctive tree but by semantic association from the meanings of the concepts and examples of deap case patterns of the concepts.

The hierarchy has three levels. The highest level has been divided coarsely based on semantic association (coarse semantic clusters; all can be seen in Figure 4). The second level has been also divided based on semantic association (fine semantic clusters; all below coarse semantic cluster ☆1 can be seen in Figure 5). On the contrary, the third level has been divided based on the deep case pattern shared by concepts (KOTO-categories; all below fine semantic cluster ● 1.2 can be seen in Figure 6), where one category has only one deep case pattern which is specified with its distinctive pattern (expressed with the notation " 【...】 " as in Figure 2). We have now 14 coarse semantic clusters, 253 fine semantic clusters and 984 KOTO-categories in the hierarchy.

〈KOTO-Concepts〉

☆1<SPATIAL_RELATIONS> : relations among physical objects meaning states and changes in space
☆2<SPATIAL_ATTRIBUTES> : attributes of physical objects meaning spatial measures in space
☆3<SOCIAL_RELATIONS> : social relations among persons
☆4<CLASS_RELATIONS> : inclusive relations and comparative relations among objects
☆5<POSSESSION> : relations among possessors and possessions
☆6<INFORMATION> : relations among information and information processers
☆7<ESTIMATION> : relations and attributes of objects meaning states and changes of their estimations
☆8<POSSIBILITY> : relations and attributes of events meaning states and changes of possibilities
☆9<FUNCTION> : relations and attributes of objects meaning states and changes of their functions
☆10<PROGRESS> : relations and attributes of events meaning degrees of actualization
☆11<TIME> : relations and attributes of events meaning temporal order or distance
☆12<QUANTITY> : relations and attributes of objects meaning quantity or degree
☆13<OTHER_ATTRIBUTES> : attributes other than those above
☆14<EXISTENCE> : relations meaning appearance, continuance and disappearance of existences

Figure 4. Coarse Semantic Clusters and Their Definitions

☆1<SPATIAL_RELATIONS>

- ●1.1<For_a_Physical_Object_itself_to_Change_in_Space>
- ●1.2<To_Touch_a_Place_or_Physical_Objects_in_Space>
- ●1.3<To_Separate_from_a_Thing_Touching_it_in_Space>
- ●1.4<For_Physical_Objects_to_Unite_in_Space>
- ●1.5<For_United_Physical_Objects_to_Separate_in_Space>
- ●1.6<To_Move_Some_Distance_in_Space>
- ●1.7<To_Move_Some_Distance_to_a_Direction_in_Space>
- ●1.8<To_Move_Inside_in_Space>
- ●1.9<To_Move_Outside_in_Space>
- ●1.10<To_Approach_a_Goal_Some_Distance_in_Space>
- ●1.11<To_Leave_a_Source_Some_Distance_in_Space>
- ●1.12<For_Physical_Objects_to_Gather_at_Some_Distance_in_Space>
- ●1.13<For_Physical_Objects_to_Disperse_from_Some_Distance_in_Space>
- ●1.14<For_Physical_Objects_to_Fill_in_Space>
- ●1.15<For_an_Angle_to_decrease_in_Space>
- ●1.16<For_an_Angle_to_increase_in_Space>
- ●1.17<For_Order_of_Physical_Objects_to_Change_in_Space>
- ●1.18<For_a_Wearable_Object_to_Touch_a_Body>
- ●1.19<For_a_Wearable_Object_to_Separate_from_a_Body>
- ●1.20<To_Move_into_a_Body_Physiologically>
- ●1.21<To_Move_out_of_a_Body_Physiologically>

Figure 5. Fine Semantic Clusters Below ☆1<Spatial_Relations>

●1.2<To_Touch_a_Place_or_Physical_Object_in_Space>

- ○1.2.1 ⟨(For_a_Physical_Object)_To_Touch_(Another_Physical_Object)⟩
- ○1.2.2 ⟨(For_a_Physical_Object)_To_Touch_another_Physical_Object⟩
- ○1.2.3 ⟨(For_an_Animal)_To_Touch_(A_Physical_Object)_Intentionally⟩
- ○1.2.4 ⟨(For_an_Animal)_To_Touch_a_Physical_Object_Intentionally⟩
- ○1.2.5 ⟨(For_an_Intentional_Object)_To_Touch_a_Physical_Object_Intentionally⟩
- ○1.2.6 ⟨(For_an_Animal)_to_Touch_(A_Physical_Object)(With_a_Part_of_the_Body)⟩
- ○1.2.7 ⟨(For_an_Animal)_to_Touch_a_Physical_Object_(With_a_Part_of_the_Body)⟩
- ○1.2.8 ⟨(For_an_Animal)_to_Touch_(A_Physical_Object)_with_a_Part_of_the_Body⟩
- ○1.2.9 ⟨(For_an_Animal)_to_Touch_(A_Physical_Object)(With_an_Implement)⟩
- ○1.2.10 ⟨(For_an_Animal)_to_Touch_(A_Physical_Object)_with_an_Implement⟩
- ○1.2.11 ⟨(For_an_Intentional_Object)_to_Send_(A_Physical_Object)(To_Some_Place)⟩
- ○1.2.12 ⟨(For_an_Animal)_to_Go_(to_Some_Place)_Intentionally⟩
- ○1.2.13 ⟨(For_a_Person)_to_Meet_with_(another_Person)_Intentionally⟩
- ○1.2.14 ⟨(For_an_Animal)_to_Grasp_(A_Physical_Object)(With_a_Part_of_the_Body)⟩
- ○1.2.15 ⟨(For_an_Animal)_to_Grasp_(A_Physical_Object)(With_an_Implement)⟩
- ○1.2.16 ⟨(For_an_Person)_to_Put_(A_Physical_Object)(On_Some_Place)⟩
- ○1.2.17 ⟨(For_an_Person)_to_Cause_(An_Animal_and_another_Animal)_to_Touch_Each_Other⟩
- ○1.2.18 ⟨(For_an_Animal_and_another_Animal)_to_Touch_Each_Other⟩
- ○1.2.19 ⟨(For_an_Animal)_to_Touch_(An_Physical_Object)⟩
- ○1.2.20 ⟨(For_an_Intentional_Object)_to_Cause_a_Physical_Object_to_Touch_(a_Physical_Object⟩

Figure 6. KOTO-Categories Below ●1.2

The first classification of a concept into the KOTO-categories is made based on semantic association with the concept and deap case patterns created with the concept. The procedures are as follows: 1) assigning basic concepts into KOTO-categories: classifying about 4,000 basic concepts into fine semantic clusters, describing deep case patterns underlying example sentences created with the concepts and dividing the clusters into KOTO-categories to make each of them have only one deap case pattern. 2) Establishing two indexes : making a) a word index for retrieving categories by a word, and b) a case frame index for retrieving categories by a deep case set. 3) Searching category candidates: a) searching category candidates by associating basic concepts which seem to share a deap case pattern with the concept and retrieving the word index by words meaning the basic concepts. b) In a case in that it is impossible to associate any basic concepts with the concept, finding category candidates by creating example sentences, making deap case frames from the sentences, and retrieving the case frame index by the frames to find category candidates. 4) Selecting a category from the category candidates: classifying concepts into the most appropriate category by considering from the following three points of view: a) the names of the categories and their upper clusters, b) the distinctive patterns of the categories, and c) the basic concepts assigned to the categories.

3.2 Cross Classification of Concepts

Cross classification of concepts is made in the following way:

1) Making cross category lists for each exclusive categories. Types of cross relations are assorted into the following three types: a) a cross category which implies an exclusive category, b) a cross category which intersects an exclusive category, and c) a cross category which includes an exclusive category.

2) Contrasting each concept classified into an exclusive categoriy and each cross category listed in the cross category list of the exclusive category and judging whether or not the concept can be classified into the cross category. Here in the above case c), all concepts in the exclusive category can be automatically classified into the cross category.

3.3 Improvement of the Concept Dictionary

Through the following procedures, categories which should be modified are found and improvement of the Concept Dictionary is made:

1) Collecting negative examples:

When an answer other than correct answers is output from a testing system, an inappropriate concept-concept relation must be found deduced from the Concept Dictionary by viewing a debugging trace of the process of the system. Such concept-concept relations are collected as *negative examples*.

2) Collecting positive examples:

When a correct answer is not output from a testing system, a concept-concept relation must be found to be added to the Concept Dictionary. Moreover, all correct answers output from all testing systems

must have their corresponding concept-concept relations deduced from the Concept Dictionary. These concept-concept relations are collected as *positive examples*.

3) Estimation:

At a stage when negative and positive examples have been collected to some extent, the *divisibility* of each concept-category relation description and category-category relation description is estimated by using the following formula:

$$(8)\quad D(l, m, n) = \frac{\sum_{\frac{i}{m} > k} P(i)}{\sum_{i=0}^{l} P(i)} ,$$

$$\text{where } P(i) = \frac{\binom{i}{l}\binom{n-i}{m-l}}{\binom{n}{m}} ,$$

$0 \leqq k \leqq 1.0$ (a parameter),

n : the number of concepts under the category,

i : the number of incorrect classifications into the category,

m : the number of examples,

l : the number of negative examples.

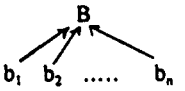
The formula (8) is derived as follows:

We suppose that the concept-category (or category-category) relation description (9) is an object of our estimation:

(9) a — rel → B

that the number of the concepts classified under the category B is n:

(10)



and that the number of (both negative and positive) examples is m and the number of negative examples in the examples is l:

(11) m examples

a — rel → b_1

* a — rel → b_2

a — rel → b_3

a — rel → b_4

* a — rel → b_5

⋮

a — rel → b_m

($(m-l)$ positive examples) (l negative examples)

If the number of concepts which are under the category B but not appropriate for the concept description (9) is i,

the probability that l negative examples are found out of m examples is given by the formula (12)[1].

$$(12) \quad P(i) = \frac{\binom{i}{l}\binom{n-i}{m-l}}{\binom{n}{m}} .$$

Therefore, the probability that the ratio of the concepts not appropriate for the description (9) to the concepts located under the category B is more than k is given by the formula (8). Here we use Bayse's Theorem because sellections of the number i are events independent from each other.

4) Deletion of the concept descriptions:

In cases in that the value of the formula (8) is more than 0.9 when k = 0.9, the concept description (9) is deleted from the Concept Dictionary and remaining positive examples are asserted as concept-concept relation descriptions into the Concept Dictionary because most of the examples for the description are negative.

5) Division of categories

In cases in that the value of the formula (8) is more than 0.9 when k = 0.1, the category B is divided in two in order to represent both a category satisfying the relation (9) and a category not satisfying the relation (9) and all concepts under the category B is reclassified into the two categories because we recognize that a) the number of examples are large enough for the estimation, and that b) the number of negative examples is too large to neglect. Here if the divided two categories exist as sub-categories of the category B in the concept taxonomy, the classification is not necessary.

6) Accumulation of preference knowledge

In cases in that the value of the formula (8) is not more than 0.9 when k = 0.1, the collected negative examples are translated to preference knowledge (for data structures and usages of preference knowledge, see Section 4). From a debugging trace of a testing system, together with a negative example, a concept, a word or a pronumciation corresponding to the negative example and a concept description more appropriate than the negative example must be also gained. This information is represented by preterence knowdedge with the following format (13) and accumulated:

```
(13) on <concept> | <word> | <pronunciation>
       give preference to
         <a-more-appropriate-concept-description>
       over
         <a-negative-example-of-concept-description>
```

---

[1] We may use Poisson distribution as an approximation to (12) if n is large. However, since $n \lesssim 3{,}000$, it is realistic to calculate the formula (12).

7) Clustering of concept-concept relation descriptions

Concept-concept relation descriptions remaining after all the above procedures are clustered by using an optimal scaling or using DM-decomposition and a probability-based estimation and the gained clusters are asserted as concept-category relation descriptions into the Concept Dictionary. The clustering algorithms are explained in detail in (Nakao 1988, Matsukawa 1989).

8) Reconstruction of the concept taxonomy

Category-category relation descriptions are clustered and hierachized by using DM-decomposition and set-relation calculations in order to bundle the descriptions into higher level categories. The hierachization algorithm is explained in detail in (Matsukawa 1990a, 1990b, Yokota 1990).

4 Preference Knowledge

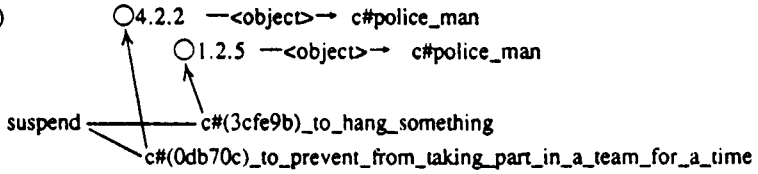
All concepts, categories and concept descriptions have an ID number called concept ID. Knowledge for ordering input sentence interpretations given by using the Concept Dictionary are represented by data individually expressing the order of the concept IDs that co-occur with each word pronunciation, word and concept, respectively (preference knowledge). We use the following three methods for ordering concept IDs:

a) Linear lists of concept IDs
b) Association lists of concept IDs and the concept IDs' preference value
c) Directed graphs including arcs meaning preference relations between concept IDs

As mentioned in section 3.3, modification of the Concept Dictionary is made based on feedback information from tests performed by various kinds of processes in application systems (testing systems). Word sense selection and translation word candidates selection are ones of these processes. When an output answer given by such a testing system is different from correct answers, the reason for the difference is analyzed by viewing traces of processes of the system, and the Concept Dictionary and/or the preference knowledge are/is modified. After such modifications, the correct answers become able to be selected by using the Concept Dictionary and the preference knowledge. For example, the word "suspend" has five senses, as shown in Figure 7. If the concept-concept relation shown in (14) is input, only two out of the five senses match the relation. The two senses are shown in (15):

(14) c#suspend —<object>→ c#police_man
(to suspend the policeman)

(15)



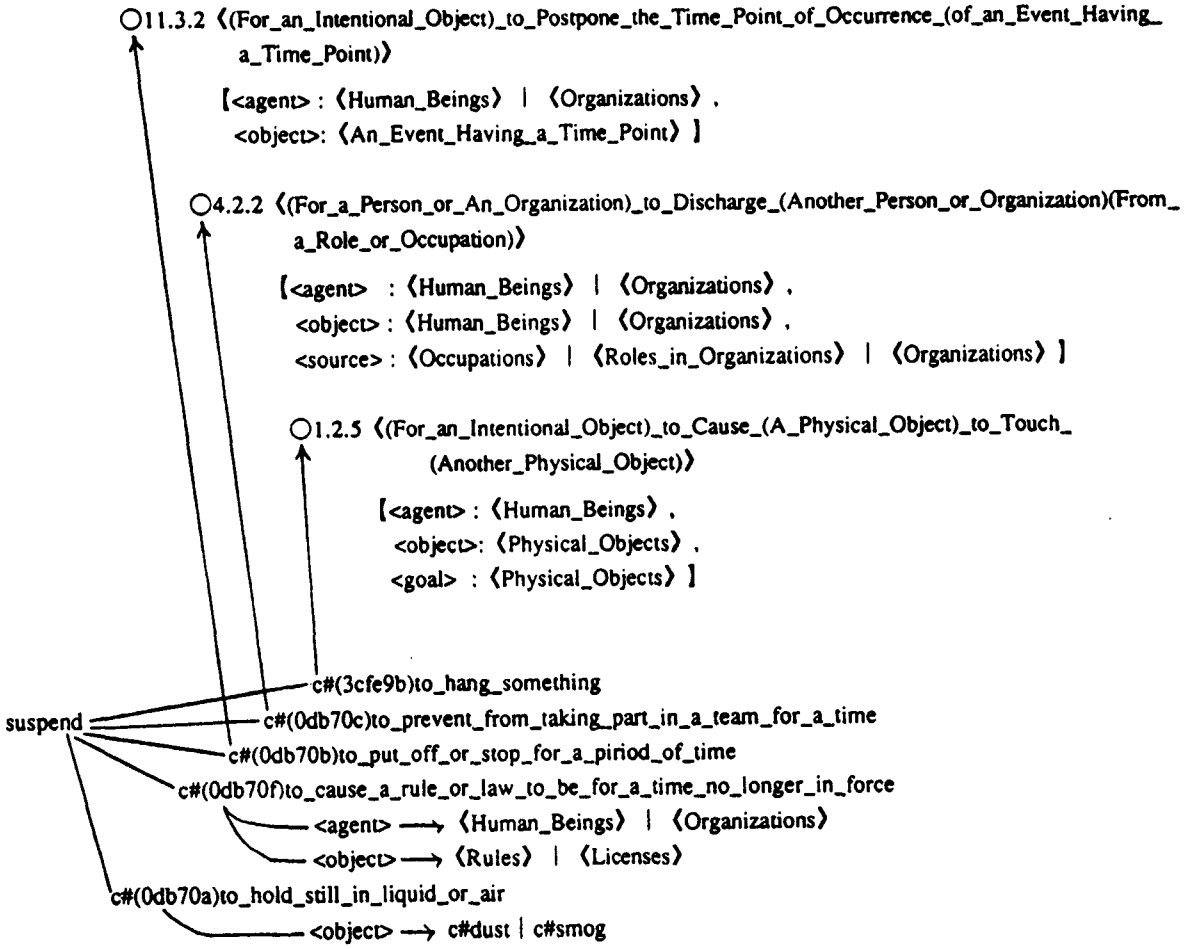


Figure 7. The Five Senses of Word "suspend" and Their Concept Descriptions

If we have preference knowledge on "c#police_man" as shown in (16), we can select only one sense out of the two senses, namely "c#(0db70c)to_prevent_from_taking_part_in_a_team_for_a_time."

```
(16) on c#police_man
        give preference to
           ○4.2.2 —<object>→ ⟨Human_Beings⟩ ;
        over
           ○1.2.5 —<object>→ ⟨Physical_Objects⟩ ;
```

Moreover, such preference knowledge is given to each concept, word or pronunciation. For example, "c#bird" and "c#rabbit" have different preference knowledge as follows:

(17) **on** c#bird
give preference to
c#fly —<agent>→ ⟨Intentional-Objects⟩ ;
over
c#hop —<agent>→ ⟨Animals⟩ ;

on c#rabbit
give preference to
c#hop —<agent>→ ⟨Animals⟩ ;
over
c#fly —<agent>→ ⟨Intentional-Objects⟩ ;

By using this knowledge with the concept descriptions shown in Figure 8, Japanese sentences can be properly translated, for example as shown in (18) (for a method for unification of concepts expressed by different words or in different languages, see (Miike 1990b, Tominaga 1991)):

(18) a) TORI - GA (a bird) TOBU. ⟶ A bird flies.

b) USAGI - GA (a rabbit) TOBU. ⟶ A rabbit hops.

Preference knowledge is collected not only through processings of word sense selection and translation word candidate selection, but also through those of structual disambiguation, paraphrasing and the like. Therefore, the knowledge includes descriptions corresponding to *lexical preference* proposed by Ford, Bresnan and Kaplan for structual disambiguation (Ford Bresnan and Kaplan 1982). Although such knowledge provides just a bias of interpretations of ambiguous structures, the knowledge is indispensable for deterministic sentence analysis without any knowledge about the discourse to be refered in order to use the *principle of parsimony* , the *principle of a priori plausibility* , etc. (Crain and Steedman 1984, Hirst 1984).

5 Discussion

Concept-category relations are similar to what are called selection restrictions. However, the goal of the development of the Concept Dictionary is not to express word sense with minimum semantic markers such as in (Katz and Fodor 1963). What is important is to actualize a methodology for weeding out incorrect, too coarse or useless concept descriptions by using them on various application systems. For the purpose, we may have redundant semantic markers (namely, categories) at the first stage and do have descriptions not including semantic markers (namely, concept-concept relation descriptions).

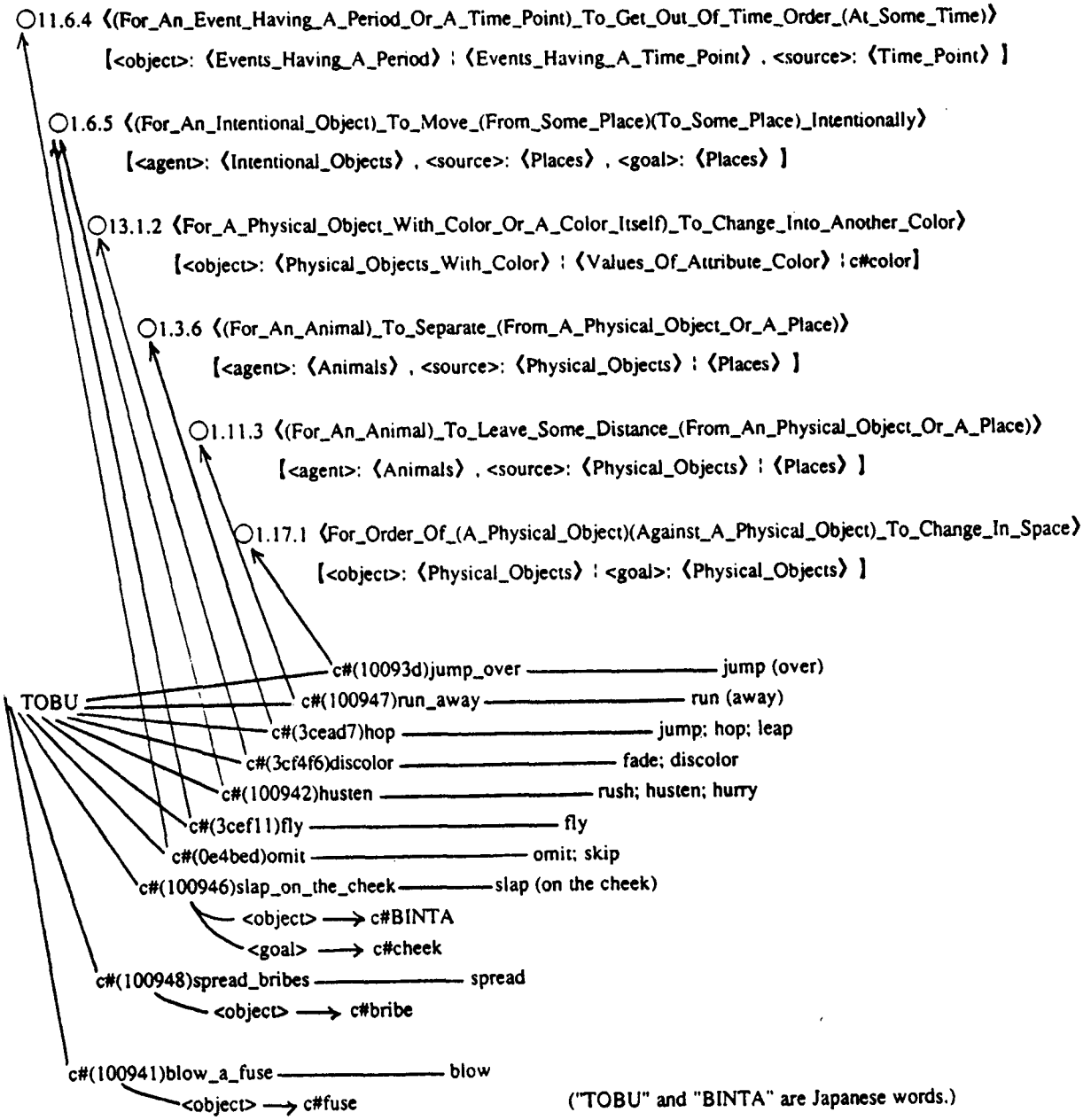


Figure 8. The Translation Word Candidates of Pronunciation "TOBU" and Their Concept Descriptions

We are compiling and improving the Concept Dictionary based on the results of many practical tests performed by various application systems. Actually, however, the number of the application systems we are using for testing dictionary data is only some dozen, and the functions we require the dictionary to fulfill are just at the level of practical necessity for the systems. In other words, compared with "the whole knowledge", the Concept Dictionary actually lacks various parts. For example:

1) Coverage of Concept Relations and Categories:

a) Coverage of concept relations:

The number of sentences in the EDR corpus is 1,000,000. As mentioned above, we also create example sentences for some concepts and we use categories to describe concept relations. However, the coverage of concept relations is incomplete. The situation is similar to that of ordinary lexicography in that even if we had a very large corpus and abundant lexicographers, they could never completely assign example sentences covering all kinds of word co-occurrences in all types of contexts

b) Coverage of categories:

Some categories are useful to describe concept relations, while other categories are not. Since the cost/performance of the implementation of useless categories is low, we do not implement such categories in the concept taxonomy.

2) Granularity of Concepts

Since concepts, our representation primitives, are almost as fine as translation words' senses in bilingual dictionaries, it is impossible to implement knowledge including finer instances.

a) Real world instances :

From concept-concept relations, we can extract slots of a class in the real world. For example, from the concept-concept relation shown in (19), we can extract slot "haveALid?" of class "C#vessel" as shown in (20):

(19) c#vessel ←<part_of>— c#lid

(20) C#vessel
canHaveSlots: (haveALid?)

(Here we use the notation Cyc uses (Lenat and Guha 1989))

However, we cannot extract some attributes of real world instances from concept relation descriptions. For example the value "yes" of the slot "have ALid?" shown at (21) can never be decided with the Concept Dictionary:

(21) TheVesselOnTheDeskInFrontOfMeAt1:00amOnMarch3rdIn1991JST
instanceOf: C#vessel
haveALid?: yes

b) Distinction of pragmatic referents:

Finer referents are used for describing pragmatic ambiguities of a sentence. For example, in order to express pragmatically different interpretations, different referents are described into each mental space, according to Fauconnier's theory (Fauconnier 1984). For example, different interpretations of the sentence (22) require at least three different mental spaces and six different referents of the word "president".

(22) John believes that the president was a baby in 1929.

The Concept Dictionary can not give distinction of such referents although it can give predicate concepts used in annotations for mental spaces.

References

Crain, S. and Steedman, M. (1984) "On not Being Led Up the Garden Path: The Use of Context by the Psychological Parser", in: Dowty, D. R. ; Karttunen, L. J. and Zwicky, A. M. (editors). ***Syntactic Thecry and How People Parse Sentences***, Camgridge University Press, 1984.

Fauconnier, G. (1984) ***Espaces Mentaux***. Editions de Minuit.

Fillmore, C. J. (1968) "The Case for Case" in Bach E. and Harms R.T. (eds.), ***Universals in Linguistic Theory***, Holt, Rinehart and Winston, Chicago.

Ford, M., Bresnan, J.W. and Kaplan, R.M. (1982) "A Competence-based theory of syntactic closure", in: Bresnan, J.W. (editor) ***The Mental Representatin of Grammatical Relations***. Cambridge, Massashusetts: The MIT Press, 1982.

Hirst, G. J. (1984) ***Semantic Interpretation against Ambiguity***, University Microfilms International, pp. 196-200.

Kakizaki, N. (1987) "Research and Development of an Electronic Dictionary", ***Machine Translation Summit*** pp.61-64.

Katz, J.J. and Fodor, J.A. (1963) "The Structure of a Semantic Theory", ***Language*** 39, pp.170-210.

Lakoff, G. (1966) "Stative Adjectives and Verbs in English", ***Mathematical Linguistics and Automatic Translation*** 17, pp.1-16, Report to the National Science Foundation.

Lenat, D.B. and Guha, R.V. (1989) ***Building Large Knowledge-Based Systems***, Addison-Wesley Publishing Company, Inc., pp. 160-162.

Matsukawa, T., Nakamura, J. and Nagao, M. (1989) "An Algorithm of Word Clustering from Co-occurrence Data Using DM Decomposition and Statistical Estimation", ***Information Processing Society of Japan***, NL-72-9.

Matsukawa, T., Kishimoto, Y., Miike, S., Yokota, E., Takai, S. and Amano, S. (1990a) "Construction of a Hierarchical

Concept Classification Based On Compaction of Concept Descriptions", *Information Processing Society of Japan*, NL-78-6.

Matsukawa, T., Nakazawa, M., Adachi, H. and Amano, S. (1990b) "Basic Functions of the Environment for Binary Relation Categorization", *Proceedings of 41th Conference of Information Processing Society of Japan*, 7S-7.

Miike, S., Amano, S., Uchida, H. and Yokoi, T. (1990a) "The Structure and Function of the EDR Concept Dictionary", *TKE '90: Terminology and Knowledge Engineering, Frankfurt*, INDEKS VERLAG.

Miike, S. (1990b) "How to Define Concepts for Electronic Dictionaries", *Proceedings of International Workshop on Electronic Dictionaries*, pp. 43-49. TR-031, Japan Electronic Dictionary Research Institute, Ltd. Tokyo, Japan.

Nagao, M., Tsujii, J. and Nakamura, J. (1985) "The Japanese Government Project for Machine Translation", *Computational Linguistics*, Vol 11, Numbers 2-3, April-September.

Nakao, Y. and Momiyama, Y. (1988) "Word Clustering by Word Bindings," *Information Processing Society of Japan*, NL-65-1.

Nakao, Y. and Uchida, H. (1990a) "Corpus for Developing Dictionary," *Euralex 4th International Congress.*

Nakao, Y. (1990b) "How to Extract Dictionary Data from the EDR Copus", *Proceedings of International Workshop on Electronic Dictionaries*, pp. 58-62. TR-031, Japan Electronic Dictionary Research Institute, Ltd. Tokyo, Japan.

Ogino, T., Yamamoto, Y. Kiyono, M, Nawata, M. and Uchida, H. (1989) "Verb Classification Based On the Semantic Relation of Co-occurring Elements", *Information Processing Society of Japan*, NL-71-2.

Schank, R. C. (1975). *Conceptual Information Processing*, North-Holland.

Tominaga, M., Miike, S., Uchida, H. and Yokoi, T. (1991) "Development of the EDR Concept Dictionary", *Second Workshop of Japan-United Kingdom Bilateral Cooperative Research Programme on Computational Linguistics*, UMIST.

Uchida, H. (1990) "Electronic Dictionary", *Proceedings of International Workshop on Electronic Dictionaries*, pp. 23-42. TR-031, Japan Electronic Dictionary Research Institute, Ltd. Tokyo, Japan.

Wilks, Y. (1975). "Preference Semantics", in Keenan, Edward L. (ed.), *Formal Semantics of Natural Language*, Cambridge University Press, pp.329-348.

Yokoi, T., Uchida, H., Amano, S. and Kiyono, M. (1989) "Research and Development of Large-Scale Electronic Dictionaries - Current Status of the EDR Project", *Australian-Japanese Joint Symposium on Natural Language Processing.*

Yokota, E. (1990) "How to Organize a Concept Hierarchy", *Proceedings of International Workshop on Electronic Dictionaries*, pp. 50-57. TR-031, Japan Electronic Dictionary Research Institute, Ltd. Tokyo, Japan.