A new electronic dictionary with meaning description of case frame

Kouhei Shimizu and Masafumi Hagiwara
Department of Information and Computer Science, Keio University
3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan
{shimizu,hagiwara}@soft.ics.keio.ac.jp

***Abstract*— In this paper, we propose a new electronic dictionary which helps computers to understand languages. In the field of natural language processing, meaning analysis is generally a difficult problem. The research has been quite few. On the other hand, meaning comprehension is quite interesting and expected to contribute to many fields. For meaning analysis, knowledge about words is necessary. Thesauruses provide such knowledge by classifying words by their concepts. They can be used as electronic dictionaries. However, the contents are merely the clusters of words. They have few knowledge about the relation of multiple clusters, resulting in no capacity for grasping relative assertions of words. For this reason, thesauruses are insufficient for meaning analysis. A novel approach is presented in this paper for improving the capability of electronic dictionaries. Our method enables computers to comprehend concretely the relation between multiple concepts of words. The contents of the new electronic dictionary are set of nodes and links, the same style as thesauruses. Each node comprises a verb and some nouns. Thus, a node-itself can insist meaning as a form of sentence, which differs from thesauruses. Meaning comprehension is achieved by routing nouns through nodes connected with links. Experimental results indicate the good ability to analyze meaning of a whole sentence. It is concluded that such ability of the dictionary is quite useful even in the field of composition support system.**

I. Introduction

Languages are deeply concerned with our life. For example, languages are necessary to communicate with other people. In addition, we need to say to ourselves for logical thinking. Languages take an active part not only in communicating with each other, but also in working for oneself.

Since the attempt to use computers for understanding languages is very challenging, it started as machine translation in 1947[1]. We have no idea about the algorithm of the brain for handling languages. The history of natural language processing can be regarded as a challenge to analyze brain's process.

Languages are generally composed of letters. For computers' meaning comprehension, letters as a sentence to be analyzed are given to the computer. At first, words are extracted from the letters. In addition, part of speech of each word should be resolved. This is morphological analysis. Next, grammatical relation of the words are resolved. This is syntactic analysis.

For morphological analysis and syntactic analysis, we are not concerned about meaning of the words and that of the sentence. However we should not ignore meaning, since meaning is an important element of languages. Therefore, research of meaning analysis becomes active in recent years instead of morphological analysis and syntactic analysis.

Computers have no knowledge about meaning of words. Dictionaries to teach meaning are required. When one finds a strange word, he/she looks up the word in a dictionary to know its meaning. It is clear that dictionaries are worthless to ones who cannot understand explanation about the strange words. Even if computers have the ability of morphological analysis and syntactic analysis, they cannot deal with our general dictionaries. Thus, for meaning analysis, electronic dictionaries for ignorant computers are necessary.

EDR is famous electronic dictionaries for Japanese. EDR is composed of Tango, Kyouki, and Gainen dictionary. Tango dictionary provides grammatical and statistical characters of words. Kyouki dictionary provides information about cooccurrence relations of two words emerged in sentences. Gainen dictionary provides the clusters of the words in terms of concepts. This is described as a tree structure.

Nakayama[3] used EDR for calculating distance between two words in terms of words' similarity. They used number of steps extracted from the tree structure of EDR Gainen Dictionary. Their results are better than that of similar approach using conventional Bunrui Goishuu[4].

On the other hand, there is an another approach which does word-association experiments to collect clusters of words, of which the results reflect well humans' sentiments. Mochihashi[5] obtained numerical values of distance between two words by calculating probability of word-association. Their idea is that word-association makes another one, like chain reactions. By the effect, distance between two words, not associated directly, can be obtained also.

Distance values between words extracted from EDR Dictionary or word-association experiments are practical in analysis of two words' similarity. However, it should be pointed out that we cannot obtain anything but cluster of words from such values. They have few information about meaning. They are useless when it comes to meaning of a whole sentence, because their contents are just the similarity between two words. They cannot explain verbs' meaning against a sentence. For this reason, these dictionaries seem to be poor when we consider some tasks like whole sentence composition, whole sentence analysis, and so on.

We propose a novel style of electronic dictionary. Our concern is meaning of a whole sentence, which differs from

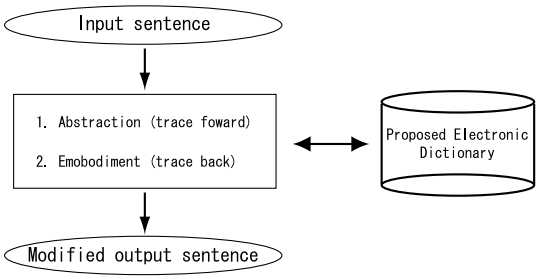

Fig. 1. Overview of the system

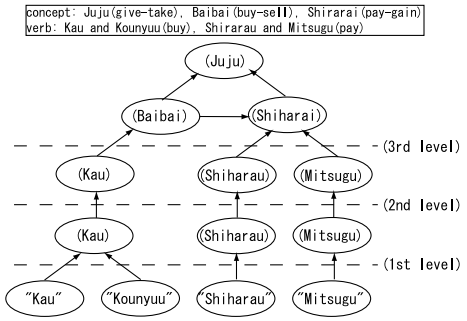

Fig. 2. Part of the proposed dictionary

conventional approach. The proposed dictionary is mainly described using case frames of Fillmore's[6]. This paper is organized as follows. The structure of the proposed dictionary is presented in Section 2. Section 3 shows good efficiency through an experiment of synonymous sentence composition. Finally, section 4 concludes the paper.

II. The Proposed System

A. Outline

We propose a new style of electronic dictionary for improving the ability of computers' meaning analysis. It is described with sentence based form of case-frames. This style enables us to explain words' meaning against a sentence. Furthermore, it means to describe our common knowledge.

In this paper, we consider that one has ability of comprehending meaning if he/she can compose synonymous sentences when an input sentence is given. The general overview of the sentence composition system is presented in Fig. 1. By analyzing a sentence with the new electronic dictionary, multiple synonymous sentences of various styles can be composed.

An example for the structure of the proposed dictionary is presented in Fig. 2. The dictionary comprises nodes and their links, the same style as conventional thesauruses. Each node makes an case-frame. Links have embraceable relation between two case-frames. By tracing links and detecting nodes, new interpretations described with case-frames are generated. We can find multiple synonymous expressions by doing these operations. Each node is classified to one of the four levels to keep the dictionary's consistency. The number of the nodes is determined so that the results of word acquisition

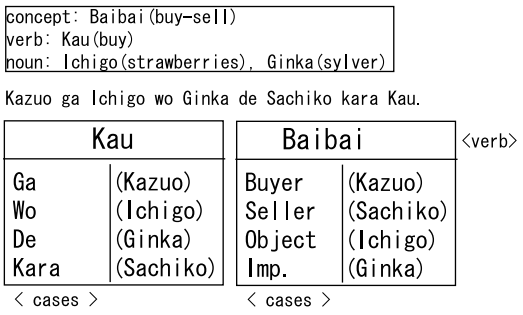

Fig. 3. Two example of node

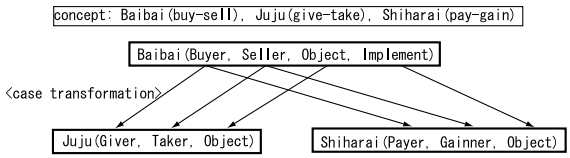

Fig. 4. Embraceable relation

using conventional dictionaries are easily applied. Therefore, the number has relation with the method of word acquisitoin.

B. Node

Nodes produce case-frames. Each node comprises a verb and cases. The verb is specified by node-itself. The cases are slots to stock indefinite nouns. A specific example is presented in Fig. 3. Each node has a particular verb and receives nouns to be effected on by the verb. The full expression of case-frame is generated when nouns are given to the node. Then the meaning description as a sentence is appeared. Both surface cases and deep cases are available in the system. The proposed system adopts surface cases in the 1st to the 3rd levels and deep cases in the 4th level.

C. Link

Links are meant to describe embraceable relation between two case-frames produced by nodes. Concerning about embraceable relation, there seem to be many varieties of them. By specifying such relations, we can teach our general knowledge to computers. Fig. 4 presents an example of a link connecting a node to the other. This link demonstrates our common recognition that deeds about buy-sell necessarily lead to those about give-take, and those about pay-gain. Furthermore, case transformations are contained in the links for specifying who get the goods, what are the goods, and who pay. By relating a node to another one, general knowledge can be taught to computers.

D. Level

The proposed system deals with natural languages for users' interface. Natural languages have commonly complicated peculiarities and their trends vary widely. It is desirable that the main description for diverse knowledge is independent of languages. Therefore, surface cases are used for resolving

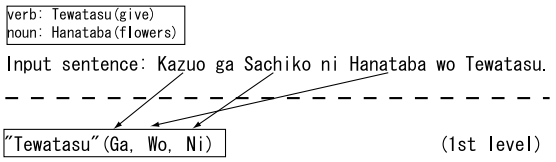

Fig. 5. 1st level node detection

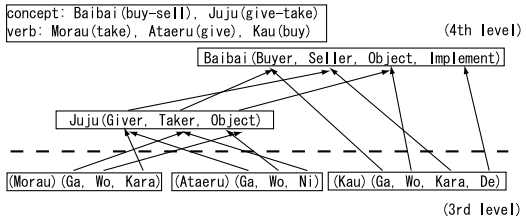

Fig. 6. 3rd level and 4th level nodes

languages' peculiarities. Deep cases are used in the 4th level for describing our various knowledge. To make this clear, all nodes are classified to one of the fours levels, as can be seen in Fig. 2. This way seems good when considering about view and unity of the task of dictionary designing. Each level's role is explained below.

The role of the nodes in the 1st level is to receive the string of verb in the input sentence. Therefore, a corresponding node in the 1st level is at first detected according to the input sentence. To handle the natural languages' properties, surface cases are adopted. Detection of the node in the 1st level and generation of case-frame is presented in Fig. 5, for instance.

The role of the nodes in 2nd level is to integrate multiple synonymous verbs. It seems more efficient to treat the similar verbs all together in various point of view. Surface cases are adopted, the same as the 1st level.

The role of the nodes in the 3rd level is to provide a bridge between surface cases in the 2nd level and deep cases in the 4th level. The correspondence is described as case transformation. By simplifying the interface between the 2nd level and the 4th level in this way, designers' task is reduced. In addition, the amount of the dictionary decreases.

The role of the nodes in 4th level is to describe our common knowledge and recognition. This is achieved by specifying the embraceable relations between nodes in the 4th level with the links and their case-frame transformation. On the whole, linguistic properties are resolved in the 1st to the 3rd levels, and meaning comprehension is begun first in the 4th level. An example for the relation of the 3rd level and the 4th level is presented in Fig. 6. This figure shows the relation among Morau(take), Ataeru(give), and Kau(buy). In general thesauruses, these verbs are close together. They have the same abstract concept. However, we can also consider Morau and Ataeru to be the exact opposite. This complicated relation can be easily described when using case-frame transformation, which is not done by conventional electronic dictionaries.

- Kazuo ga Sachiko kara Ichigo wo Morau.

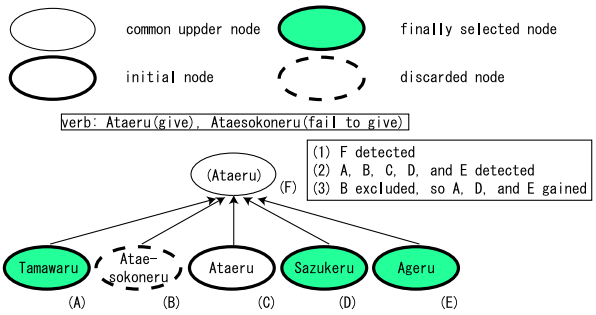

Fig. 7. Verb association

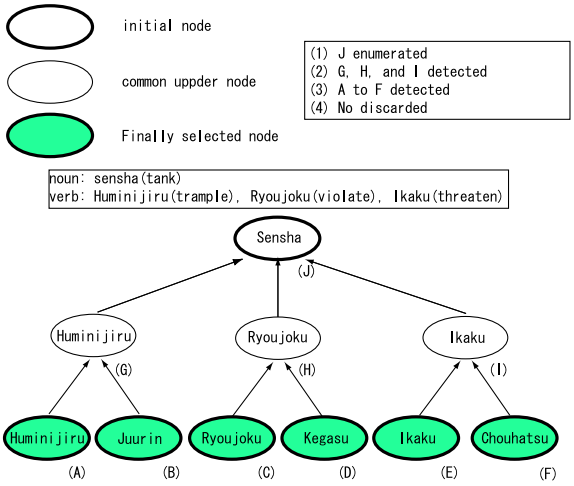

Fig. 8. Noun association

- Sachiko ga Kazuo ni Ichigo wo Ataeru.

The both sentences above are synonymous in terms of possession. The conventional approach concerning only about relating two words has no capability of treating this knowledge. By describing relation between two sentences of case-frames, this goal can be achieved. The main process of the proposed system is search of synonymous relation like this. Meaning interpretation of a whole sentence is constructed with the new dictionary. As a result, the system can make wide synonymous expressions.

E. Word acquisition

One of the main problems in natural language processing like machine translation is word acquisition. Generally humans learn quite many words in childhood. However, we do not know how it is done. For the proposed system, word acquisition is necessary. The most simple way for word acquisition is hand work by editors. Regrettably, it requires enormous time and great labor, which is impractical. It seems better to adopt appropriate methods to reduce editors' burden. Two methods using EDR for word acquisition are presented below.

In the proposed dictionary, antonyms like Ataeru and Morau are integrated in some nodes in the 4th level. In this point of view, the proposed dictionary is similar to conventional

thesauruses. However, it differs from thesauruses in that the discrete relation of the antonyms should be signified by case transformation. The proposed dictionary includes thesauruses. So the EDR Gainen Taikei dictionary is useful for supporting word acquisition of the proposed dictionary. EDR Gainen Taikei dictionary classifies concepts to one of the nodes in the tree structure. Lower nodes inherited from common upper nodes probably resemble together, since they have the same concepts. The process of collecting verbs which are similar to Ataeru(give) and different from Morau(take) is presented below.

- (1) Search upper nodes from which Ataeru is inherited.
- (2) Collect all lower nodes of the detected upper nodes above.
- (3) Choose appropriate nodes from the detected nodes above.

This framework is illustrated in Fig. 7. By (1), the upper node (F) is detected. By (2), many lower concepts inherited from (F) are collected. By (3), inappropriate nodes like Ataesokoneru(fail to give) are removed. The task of (3) depends on designer's work, since it is impossible for computers. The number of the nodes detected by (2) is usually less than 20. Therefore, the efficiency of the word acquisition is not deteriorated. As a whole, many desirable verbs are obtained from only one verb, Ataeru. This way is based on the thought that nodes inherited from common nodes resemble together, even if a few inappropriate nodes are sometimes found. But search scope is limited to very narrow range.

Broader search is required for improvement. Another method using EDR Gainen Kijutsu dictionary for satisfying the demand is presented below. The Gainen Kijutsu dictionary provides cooccurrence relation of two concepts. In addition, their grammatical status is signified with deep cases like Agent, Object, Source, Goal, and Implement. Examples are presented below.

- Yomu(read), object, Hon(book)
- Shinu(die), cause, Byouki(ill)

Our various knowledge about the world can be obtained from cooccurrence relation. The process of collecting verbs which are similar to Kougeki(attack) and different from Mamoru(guard) is presented below.

- (4) Enumerate nouns associated with Kougeki.
- (5) Collect all verbal nodes related as Implement with the enumerated nouns above.
- (6) Collect all lower nodes of the detected upper nodes above.
- (7) Choose appropriate nodes from the detected nodes above.

This framework is illustrated in Fig. 8. By (4), a noun Sensha(tank) is enumerated. By (5), verbal concepts like Huminijiru(trample) and Ikaku(threaten) are detected. (6) and (7) are the same as (2) and (3). This way depends on nouns association. It has the virtue of achieving broader search scope than the former one. Furthermore, many words which were not expected by designers can be also collected.

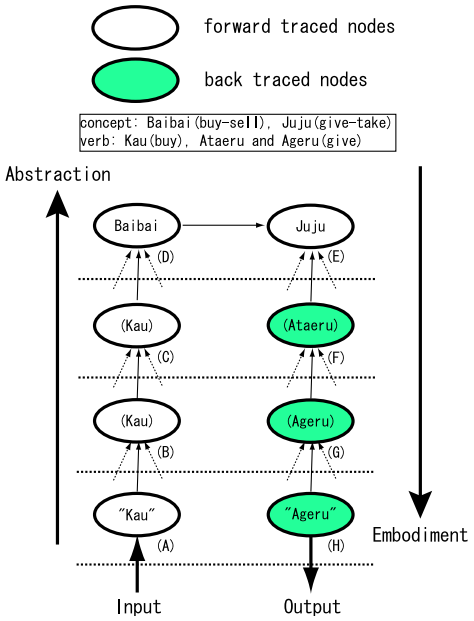

Fig. 9. Abstraction and embodiment

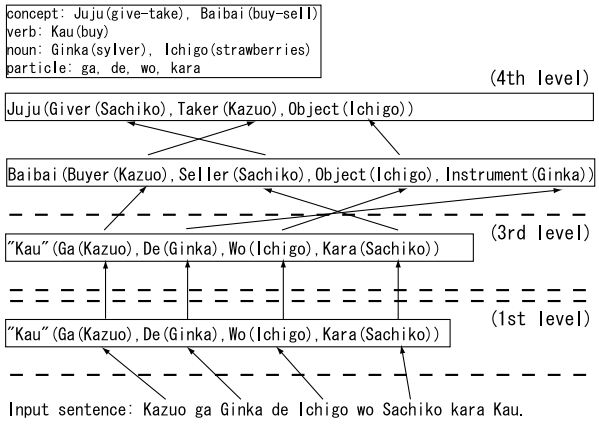

Fig. 10. Abstraction

Two methods were presented above. Empirical knowledge indicates that 50 to 200 verbs are obtained by one operation. So it is clear that they are much more efficient than the method depending fully on editors' hand work.

F. Synonymous sentence composition

With the proposed dictionary, a novel system for synonymous sentence composition is achieved. The process is expressed as

- (8) Detect the node corresponding to the verb in the given input sentence.
- (9) Step on nodes with tracing forward the links. (abstraction)
- (10) Step on nodes with tracing back the links. (embodiment)

The whole procedure of composition system is illustrated in Fig. 9. As can be seen, the main two steps, abstraction and

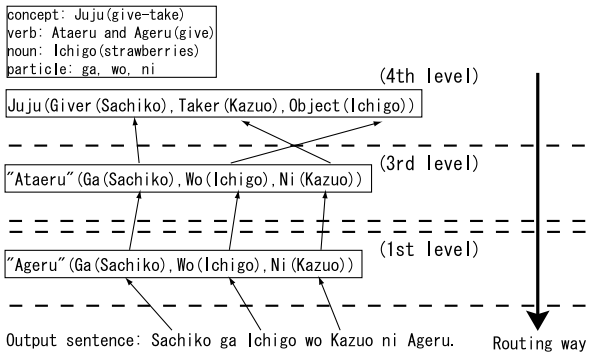

Fig. 11. Embodiment

embodiment are done by referring to the proposed dictionary. Details are explained below with a sample input sentence.

For the input, the following sentence is given.

- Kazuo ga Ginka de Ichigo wo Sachiko kara Kau.

Fig. 10 shows the overview of the first step, abstraction. The node in the 1st level corresponding to the verb Kau(buy) is detected. Then, Kazuo, Ginka(silver), Ichigo(strawberries), and Sachiko are passed respectively to surface cases, ga, de, wo, and kara. The case-frame of the node in the 1st level is formed.

Connected nodes in the 2nd and the 3rd levels are detected by tracing forward the links. Since both levels' function is purely synonyms integrating, nouns like Kazuo and Ginka are passed to ga and de as it is.

The node in the 4th level are detected by tracing forward the links. In the 4th level, deep cases are used and independent of languages' properties, which differs from the 1st, the 2nd, and the 3rd levels. For this reason, the links between the 3rd and the 4th levels include case transformation for specifying the appropriate correspondence of surface cases and deep cases. Nouns Kazuo, Ginka, Ichigo, and Sachiko are passed respectively to deep cases, Buyer, Seller, Object, and Implement.

By the same way, the connected nodes in the 4th level are detected, as can be seen in Fig. 10. The links mean that concept of Juju(give-take) is necessarily led to by concept of Baibai(buy-sell). However, the opposite inference is not correct: concept of Baibai is not necessarily led to by concept of Juju. The direction of the links means this fact.

The above explains (8) and (9), abstraction. The next step, embodiment is the pure opposite of abstraction. From the node detected finally in abstraction, the connected nodes are detected tracing back the links. This situation is shown in Fig. 11. Since the direction of tracing is the opposite, the direction of case transformation is also opposite. By this operation, the meaning comprehension keeps consistent. Continuing embodiment, one of the node in the 1st level is finally detected. Using the routed nouns, the following output sentence is produced.

- Sachiko ga Ichigo wo Kazuo ni Ageru.

If the above sentence is given to the system as an input sentence, the node Juju is surely detected in abstraction. The original input sentence leads also to the same node in abstraction. In this point of view, the both sentences have the common meaning. So the output sentence is considered to be synonymous to the input sentence in terms of the common node. In addition, the common node is the finally detected node in the former abstraction.

Abstraction is thought to be extending range of meaning comprehension. This does not modify the original meaning of the input sentence. Embodiment is thought to be narrowing range of meaning comprehension according to the reversed links. Contrary to abstraction, this may modify the meaning, resulting in the gap between the input and the output. For example, the output sentence above loses completely the information about the noun Ginka.

The core of the proposed system is tracing links with case-frame transformation. The difference between abstraction and embodiment is simply the direction of tracing. In addition, there are many choices of the nodes to be detected in abstraction. For this reason, plentiful synonymous sentences can be composed from only one input sentence.

III. Evaluation of the Dictionary

We evaluated the new dictionary through an experiment in synonymous sentence composition as an application of creative thinking support system. We prepared the dictionary's contents for the experiment. For increasing efficiency of word acquisition, we used EDR Gainen Taikei dictionary and EDR Kijutsu dictionary.

With the proposed system, we tried to get various kinds of synonymous sentence through abstraction and embodiment. Consideration about obtained synonymous sentences is explained below with examples of input and output.

- Input A: Sachiko Ga Ichigo Wo Uru.
  (Sachiko sells strawberries.)
- Output A: Sachiko Ga Ichigo Wo Urisabaku(synonymous of Uru).

The rewriting above is done by transposing the verb with its synonym. Practical synonym dictionaries are not available yet. Conventional thesauruses partially work as synonym dictionary. However, opposite words are sometimes located close together. For this reason, thesauruses seem to be inadequate for synonym dictionary.

- Input B: Kazuo Ga Gabyou Wo Kabe Ni Sasu.
  (Kazuo puts a thumbtack into the wall.)
- Output B: Kazuo Ga Gabyou Wo Kabe Ni Uchikomu.

The above is also achieved by transposing the verb. However, Uchikomu is generally a bit far from Sasu. The conventional methods require complicated calculation to treat this knowledge. The calculation depends on hop counts extracted from thesauruses' tree structure as well as cooccurrence relations, which seem to be ad-hoc. Thesauruses have no explicit flags about physical concepts like in-out. On the other hand, the proposed system has the knowledge owing to deep cases and explicit case-frame transformation. Therefore, the proposed system can analyze more easily the correct difference of words including common concepts than conventional systems.

- Input C: Kazuo Ga Gabyou Wo Kabe Ni Sasu.
  (Kazuo puts a thumbtack into the wall.)
- Output C: Gabyou Ga Kabe Ni Korogarikomu(get in).

The Input C is the same sentence as the Input B. It is shown that many various synonymous sentences can be provided from only one input sentence. In this case, the number of tracing steps in abstraction for the 4th level is large. So the grammatical structure of the output sentence is different from that of the input sentence, while keeping the both meaning accorded together. Increasing the number of tracing steps leads to the metaphor emphasis. The degree of modification can be controlled.

- Input D: Kazuo Ga Ichigo Wo Ginka De Sachiko Kara Kau.
  (Kazuo buys strawberries from Sachiko with silver.)
- Output D: Kazuo Ga Ginka Wo Sachiko Ni Harai-watasu(synonymous of pay).

The above sentences demonstrate that the proposed dictionary has organized knowledge about the concept Baibai. Case transformation in the proposed dictionary explicates verbs' effect on the whole sentence. Conventional thesauruses cannot achieve this processing, since they connect only words.

- Input E: Sachiko Ga Tsuchi Ni Ichigo Wo Umeru.
  (Sachiko buries strawberries in the soil.)
- Output1 E: Tsuchi Ga Ichigo Wo Tsumamigu-isuru(synonymous of eat).
- Output2 E: Tsuchi Ga Ichigo Wo Nade-mawasu(synonymous of touch).
- Input F: Sachiko Ga Tsuchi Kara Ichigo Wo Horidasu.
  (Sachiko digs out strawberries from the soil.)
- Output1 F: Tsuchi Ga Ichigo Wo Hakidasu(synonymous of emit).
- Output2 F: Sachiko Ga Tsuchi Kara Ichigo Wo Hagi-toru(synonymous of strip).

Although the subject is changed in Output 1E, 2E, and 1F, the relations of Sachiko, Ichigo, and Tsuchi are the same as those of input 1E in terms of their movement. Therefore, these outputs can be regareded as figurative sentences. This is due to information loss of meaning caused by a large number of abstraction.

In addition to non-contradiction in meaning, various types of figurative sentences are provided.

Consistency in meaning is accomplished by the extended links in the 4th level. Word acquisition and natural language processing are done in the 1st, the 2nd, and the 3rd levels. The nodes in the proposed system are not mere symbols, which differs from conventional thesauruses. They can insist meaning comprehension as a form of case-frame. Therefore, synonymous sentence composition involving inference like E and F is easily achieved by the simple two steps, abstraction and embodiment.

We evaluated our dictionary and are satisfied with its capability. The unit of our dictionary is not word, but sentence. This approach has the merit of being easy to express general knowledge effectively. Knowledge described with sentences is considered to form our recognition about the world. And this can be used for meaning composition and for sentence composition including inference.

IV. Conclusion

A novel approach of electronic dictionary with a new style is presented. With case-frames and case-frame transformation, embraceable meaning structure in wide range can be effectively described. The proposed system has capability of extending meaning comprehension, which achieves sentence composition. We consider this to contribute to the field of creative thinking support system also.

In this paper, we employed EDR for the support of word acquisition. However, this is not sufficient. For the method using EDR Gainen dictionary, the designer have to find initial nouns. We are now investigating an automatic system for detecting initial words to reduce burden of creating the dictionary.

V. Acknowledgement

This research is partly supported by Keio University Special Grant-in-Aid for Innovative Collaborative Research Projects.

References

[1] Shin Nagao: Shizengengoshori, Iwanami-Bunko, 1996.
[2] Nihon Denshikajisho Kenkyuujo: EDR Setsumeisho, 1993.
[3] Satoshi Nakayama, Mine Akinori, Higashi Yuu, Taniguchi Rinnichirou, Amemiya Masato: EDR Corpus wo Riyou Shita Doushi no Gogibunrui, Shingakugihou, NLC95-43, 1995.
[4] Kokuritsu Kokugo Kenkyuujo: Bunruigoihyou, 1993.
[5] Mochihashi Daichi: Rensou to Shite no Imi, Shizengengoshori.
[6] Fillmore, C, J., Tanaka: Kakubunpou no Genri, Sanseidoi, 1975.