Handling Japanese Homophone Errors in Revision Support System for Japanese Texts; REVISE

Masahiro Oku
NTT Information and Communication Systems Laboratories
1-2356 Take, Yokosuka-shi, Kanagawa, 238-03 Japan
E-mail oku@nttnly.ntt.jp

Abstract

Japanese texts frequently suffer from the homophone errors caused by the KANA-KANJI conversion needed to input the text. It is critical, therefore, for Japanese revision support systems to detect and to correct homophone errors. This paper proposes a method for detecting and correcting Japanese homophone errors in compound nouns. This method can not only detect Japanese homophone errors in compound nouns, but also can find the correct candidates for the detected errors automatically. Finding the correct candidates is one superiority of this method over existing methods. The basic idea of this method is that a compound noun component places some restrictions on the semantic categories of the adjoining words. The method accurately determines that a homophone is misused in a compound noun if one or both of its neighbors is not a member of the semantic set defined by the homophone. Also, the method successfully indicates the correct candidates for the detected homophone errors.

1 Introduction

We have been using morphological analysis to develop REVISE, a revision support system that corrects Japanese input errors (Ikehara, Yasuda, Shimazaki, and Takagi, 1987; Ohara, Takagi, Hayashi, and Takeishi, 1991). REVISE can detect and correct various types of errors, such as character deletion, character insertion and some grammatical errors, using knowledge bases that describe the characteristics of each error type (see figure 1). Homophone errors are one of the error types that can be detected and corrected in REVISE.

Most Japanese texts are made with Japanese word processors. As Japanese texts consist of phonograms, KANA, and ideograms, KANJI, Japanese word processors always use KANA-KANJI conversion in which KANA sequences (i.e. readings) input through the key board are converted into KANA-KANJI sequences. Therefore, Japanese texts suffer from homophone errors caused by erroneous KANA-KANJI conversion. A homophone error occurs when a KANA sequence is converted into the wrong word which has the same KANA sequence (i.e. the same reading). Therefore, detecting and correcting homophone errors is an important topic.

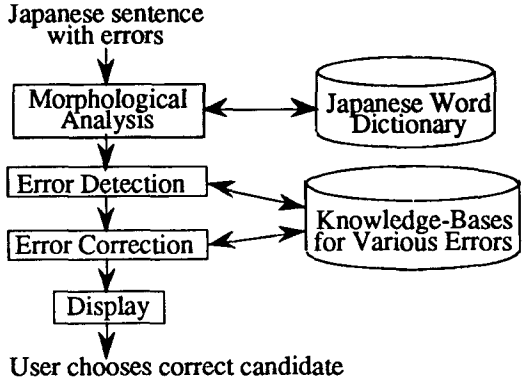

Figure 1: Processing flow of REVISE.

Previous research into detecting homophone errors with revision support systems used two approaches;
(a) using correct-wrong word pairs (Kuga, 1986),
(b) using KWIC (Key Word In Context) lists (Fukushima, Ohtake, Ohyama, and Shutoh, 1986; Suzuki and Takeda, 1989).

Previous research into correct homophone selection in KANA-KANJI conversion used the following two methods;
(c) using collocation of words (Nakano, 1982; Tanaka, Mizutani, and Yoshida, 1984; Makino and Kizawa, 1981).
(d) using case frame grammar (Oshima, Abe, Yuura, and Takeichi, 1986).

Method (a) has a drawback in that only pre-defined wrong words in correct-wrong word pairs are detected. Method (b) only indicates which words are in the KWIC list. Therefore, method (b) cannot automatically detect if the word is misused. Method (c) demands the creation of a huge dictionary which must describe all possible word collocations. Method (d) can select the correct homophone by using the semantic restriction between a verb and its cases based on case frame grammar. It is difficult, however, to use method (d) for detecting the homophone

errors in compound nouns because it mainly depends on JOSHI (i.e. Japanese postpositions) which are absent in compound nouns. Furthermore, it is difficult, if not impossible, for existing methods, (a)~(d), to correct homophone errors.

This paper describes a method for detecting and correcting homophone errors in compound nouns used in REVISE. The idea underlying this method is that a compound noun component semantically restricts the semantic categories of adjoining words. Using semantic categories reduces dictionary size; moreover, this method needs no syntactic information such as case frames. Also described are the experimental results made to certify the validity of this method.

2 Definition of key terms

Key terms used in this paper are defined as follows:

- Japanese compound noun;
  A noun that consists of several nouns, none of which have JOSHI (i.e. Japanese postpositions).
- Homophone;
  A word that sounds the same as another but has different spelling (i.e. KANJI sequence) and meaning.
- Homophone error;
  An error that occurs when a KANA sequence is converted into the wrong word which has the same KANA sequence (i.e. the same reading) as the correct one.
- Semantic category;
  A class for dividing nouns themselves into concepts according to their meaning. For example, both "自然" and "天然" belong to the same semantic category [nature].

3 A variety of homophone errors

It is necessary to use semantic information, such as the semantic restriction between words in a sentence, to handle homophone errors. We note that it is difficult, if may not impossible, to handle all homophone errors uniformly. For example, within a compound noun, the semantic restriction is mainly seen between adjacent words. The case frame semantic restriction encompasses the whole sentence. Therefore, the discussion of this paper focuses on the detection and correction of homophone errors in compound nouns.

4 A method for handling homophone errors

Tanaka and Yoshida (1987) pointed out that the collocation of words in compound nouns is restricted semantically. This means that the existence of compound noun component "X" semantically restricts the set of words that can appear next to "X". In order to describe this set, we use semantic categories instead of the words themselves to significantly reduce dictionary size. Namely, if a word is to be accepted as an immediate neighbor of "X", its semantic category must be within the set defined by "X".

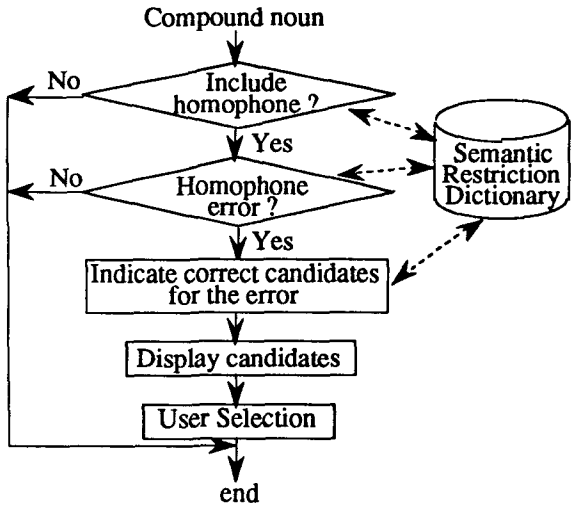

Figure 2: Flow diagram of handling homophone errors.

Figure 2 shows the flow diagram of handling homophone errors. Handling consists of two processes: error detection and error correction. In the error correction process, the correct candidates for detected homophone errors can be indicated to the user automatically. The user is responsible for the final selection of the correct homophone from among the indicated candidates. Semantic restrictions, which are used in both processes, are described in a semantic restriction dictionary using semantic categories.

4.1 Detecting homophone errors in compound nouns

The compound noun that includes only one homophone, h_i, is represented as;

$$w_p\ h_i\ w_n,$$

where w_p, w_n are words that have no homophones. The set of words with the same reading as h_i is

$$H = \{ h_1, h_2, \cdots, h_i, \cdots, h_m \}.$$

PS_i is the set of semantic categories that can appear immediately before homophone h_i. NS_i is the set of semantic categories that can appear immediately after h_i.

Here, we assume that each semantic restriction for each word in set H is exclusive. That is, for every i, j,

$$PS_i \cap PS_j = \phi\ ,$$
$$NS_i \cap NS_j = \phi\ , \qquad \cdots (1)$$
$$i \neq j,\ i, j = 1, 2, \cdots, m.$$

In the compound noun $w_p\ h_i\ w_n$, when h_i is the correct homophone, the semantic categories of w_p and w_n satisfy the semantic restrictions of h_i, i.e.,

the semantic category of $w_p \in PS_i$ and
the semantic category of $w_n \in NS_i$ ··· (2)

On the other hand, when h_i is the wrong homophone, semantic categories of w_p and w_n do not satisfy the semantic restriction for h_i, i.e., from (1) and (2);

the semantic category of $w_p \notin PS_i$ and/or
the semantic category of $w_n \notin NS_i$ ··· (3)

Therefore, we can detect homophone errors in compound nouns based on (2) and (3).

4.2 Insufficient semantic discrimination

It is possible that set H contains two or more words whose PSs and/or NSs overlap, such that the semantic sets do not yield sufficient discrimination performance. Namely, several semantic restrictions for words in set H do not satisfy formula (1), i.e., for the semantic categories of several words in set H,

$$PS_i \cap PS_j \neq \phi,$$
$$NS_i \cap NS_j \neq \phi, \quad \cdots (4)$$
$$i \neq j.$$

In this case, semantic categories which do not belong to $PS_i \cap PS_j$ or $NS_i \cap NS_j$ can also be used for detecting homophone errors based on (2) and/or (3). The words with semantic categories belonging to $PS_i \cap PS_j$ or $NS_i \cap NS_j$, however, fail to distinguish h_i and h_j because such categories satisfy both semantic restrictions in terms of h_i and h_j.

It is very difficult to construct a semantic category system that would satisfy formula (1) for all words. Therefore, in REVISE, when a word whose semantic categories belong to $PS_i \cap PS_j$ or $NS_i \cap NS_j$ adjoin h_i or h_j in compound nouns, h_i or h_j is detected as a homophone error. This may wrongly indicate correct homophones as errors but no error will be missed. This is a basic requirement of any text revision support system and/or any text proofreading system.

4.3 Correcting homophone errors in compound nouns

The correct homophone in a compound noun should satisfy the semantic restrictions established by its adjoining words. The semantic category for the adjoining word of the homophone error should be included in the sets of semantic categories that can appear immediately before/after the correct homophone. Namely, it is the correct candidates for the detected homophone error that satisfy formula (2) and that have the same KANA sequence (i.e. the same reading) as the error. When the semantic category sets of homophones partially overlap and the category of the adjoining word falls into the overlap region, the homophone is detected as erroneous even if it is correct, as described above in 4.2. In this case, the detected homophone itself is also indicated as one of the correct candidates if it satisfies formula (2). To indicate only candidates which satisfy formula (2) leads us to a shortened correction process because the correct homophone will be included in the candidates.

4.4 Semantic restriction dictionary

The semantic restriction dictionary describes which semantic categories can adjoin, either before or after, each homophone. Figure 3 shows the format of the semantic restriction dictionary. A record consists of the following four items;

- homophone reading: the semantic restriction dictionary is retrieved by the homophone reading in the error correction process, to find the correct candidates for the detected homophone error.
- KANJI homophone spelling: the dictionary is retrieved by the KANJI homophone spelling in the error detection process, to determine whether the homophone is misused in the compound noun or not.
- information whether semantic restrictions in this record apply to the preceding or following word.
- semantic restrictions: this is the set of semantic categories that can adjoin the homophone. Semantic categories which are included in two or more sets of the homophones are marked as to show insufficient semantic discrimination.

Ways of using the semantic restriction dictionary in both processes, error detection and error correction, will be described using examples in the next section.

reading	spelling	preceding or following	semantic restrictions

Figure 3: The format of the semantic restriction dictionary.

4.5 Examples of handling homophone errors

An example of detecting homophone errors in the compound noun "自然化学", which includes the homophone "化学(chemistry)" is shown in figure 4. "化学", whose reading is "かがく(kagaku)", has the homophonic word "科学(science)" while "自然" has no homophonic word. The word preceding homophone "化学" in the compound noun "自然化学" is "自然" and it has the semantic category [nature]. As shown in figure 4, semantic category [nature] is not included in the set that represents the semantic restriction on the possible prior neighbors of "化学". Therefore, "化学" is detected as a homophone error in the compound noun "自然化学" based on formula (3). Next, the error correction process is invoked after detecting homophone error "化学". In order to indicate the correct candidates for the error, the semantic restriction dictionary is accessed using the reading "かがく". The semantic set of possible prior

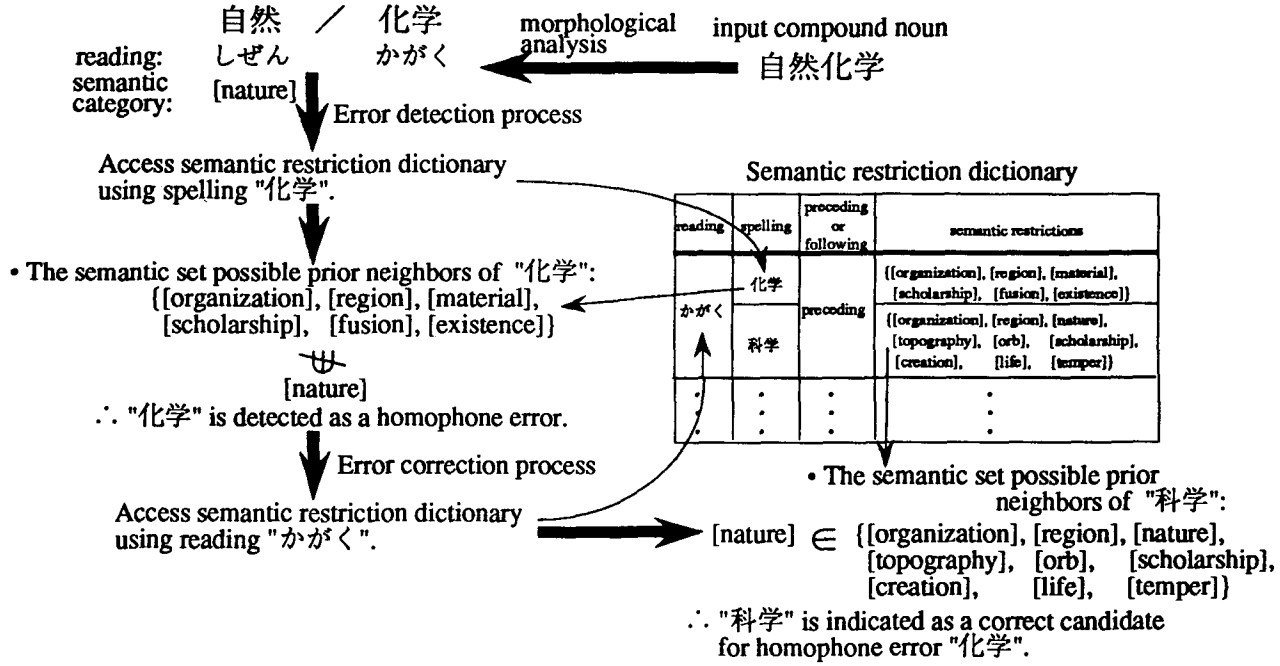

Figure 4: An example of handling a homophone error.

neighbors of homophonic word "科学" is then obtained. Because the semantic category [nature] for "自然" is included in this set, "科学" is indicated as a correct candidate for homophone error "化学" in the compound noun "自然化学" based on formula (2).

Let's consider an example that exhibits insufficient semantic discrimination. The compound noun "工作機械" shown in figure 5 includes the homophone "機械 (machine)" whose reading is "きかい(kikai)" and the word "工作(operation)" whose semantic category is [act]. "機械" has homophonic words "器械(machine)" and "機会(chance)", while "工作" has no homophonic word. Although, as shown in figure 5, semantic category [act] is included in the semantic category set for the words preceding "機械", this category is also included in the other semantic category set (in figure 5, this fact is shown by outlining). As mentioned in section 4.2, such a case is flagged as a homophone error even though it is correct

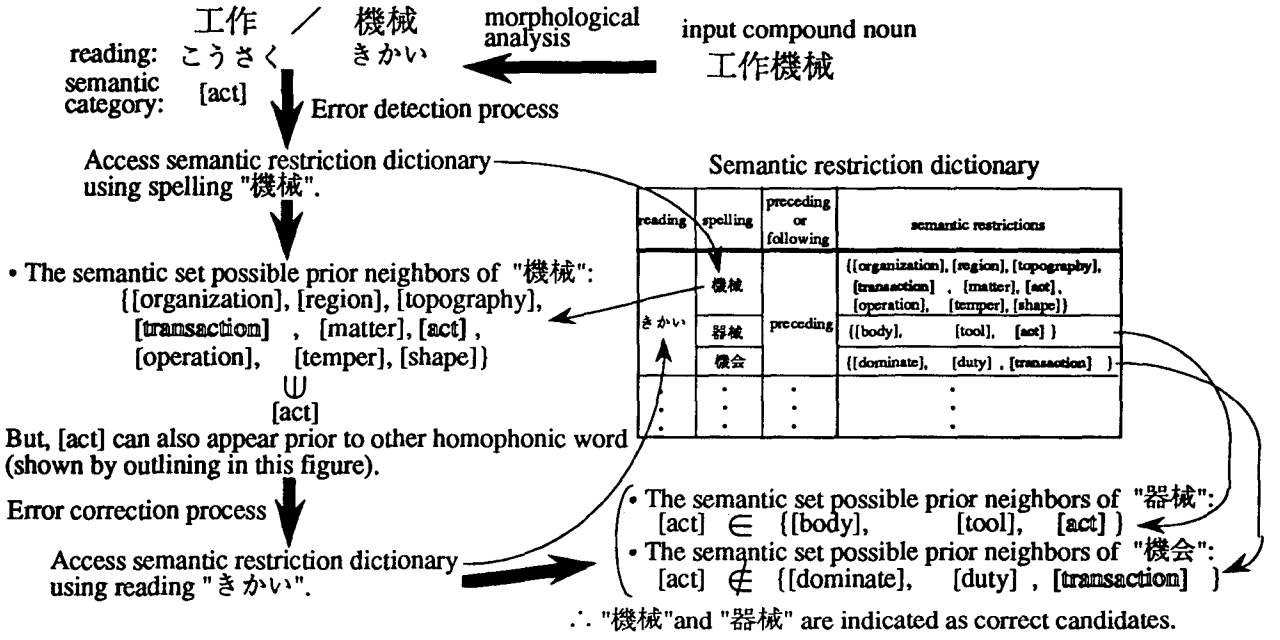

Figure 5: Another example of handling a homophone error.

(actually "機械" is correct in this example). Therefore, "機械" in the compound noun "工作機械" is detected as the error, and the correction process is invoked. The semantic restriction dictionary is accessed using the reading "きかい". The semantic set of possible prior neighbors of homophonic words "器械" and "機会" are then obtained. The semantic category [act] is an element of the set for "器械" but is not included the set for "機会". According to formulae (2) and (3), only "機械" and "器械" are indicated as correct candidates. Although the correct homophone is detected as the error, that the correct homophone (the original homophone) will be a candidate shortens the correction process.

5 Experiments

The validity of this method was confirmed with experiments in detecting and correcting homophone errors. We assumed that the input compound nouns were already segmented into component words and that their reading and semantic categories were already added.

Table 1: Homophones used in experiments.

reading	spelling				
kagaku	化学	科学			
kakou	加工	下降	火口	河口	
kikai	機械	器械	機会		
kisei	帰省	寄生	規制	既製	既成
	規正	規整			
kyoukou	強硬	強行	恐慌	凶行	兇行
kyousou	競争	競走	強壮	狂騒	狂想
kyoudo	強度	郷土			
genka	原価	減価			
kougai	郊外	公害	口外	構外	梗概
kougyou	工業	鉱業	鋼業	興業	興行
koutai	交代	後退	交替	抗体	
kounai	構内	校内			
kouhyou	公表	好評	講評		
saitei	最低	裁定			
shiten	支店	視点	支点		
shougaku	小額	少額	奨学	商学	小学
shoukyaku	焼却	償却	消却		
shomei	署名	書名			
jiten	辞典	字典	事典		
senshin	先進	専心			
taikou	対抗	対校	対向		
chika	地下	地価			
teigaku	定額	低額	停学		
denki	電気	電機	電器	伝記	伝奇
toshi	都市	年			
naizou	内臓	内蔵			
nihon	日本	二本			
ninki	人気	任期			
hanmen	半面	反面			
fuyou	不要	扶養	不用	不溶	
bun	文	分			
hoken	保険	保健			

5.1 Experimental data

- Homophones used in experiments:
  Table 1 shows the 100 homophones (32 readings) that were used in the experiments.
- Compound nouns evaluated:
  We prepared two kinds of data: compound nouns that included correct homophones (correct homophone data sets) and compound nouns that included wrong homophones (wrong homophone data sets). Table 2 outlines the sets of experimental data used.

Table 2: Compound noun data set for experiments.

name	number	outline of data set
data set 1	461	compound nouns extracted from newspaper articles
data set 2	53	compound nouns extracted from text books in high schools
data set 3	1310	compound nouns formed by substituting a correct homophone in data set 1 with a wrong homophone
data set 4	170	compound nouns formed by substituting a correct homophone in data set 2 with a wrong homophone

5.2 Description of semantic restriction

- The semantic category system:
  The semantic category system used in the experiments was constructed by referring to BUNRUI-GOI-HYO edited by the National Language Research Institute (1964) and RUIGO-SHIN-JITEN written by Ono and Hamanishi (1981), which are the most famous semantic category systems for the Japanese language. The semantic system has about 200 nodes and covers about 35,000 words.
- The semantic restriction dictionary:
  Compound nouns including all homophones in table 1, were collected from newspaper articles over a 90 day period, and the semantic restriction dictionary was made based on the semantic restrictions between the homophones and the adjoining words in compound nouns.

5.3 Experimental results

Generally speaking, the performance of an error detection method can be measured by two indices: the detection rate indicates the percentage of errors correctly determined and the misdetection rate indicates the percentage of correct words that are erroneously detected as errors. The detection rate is defined as;

$$\text{Detection rate} = \frac{\text{the number of errors detected}}{\text{actual number of wrong compounds in the sample.}}$$

The misdetection rate is defined as;

$$\text{Misdetection rate} = \frac{\text{the number of homophones misdetected}}{\text{actual number of correct compounds in the sample.}}$$

The experimental results are shown in table 3. The detection rate is over 95%. This value is much higher than the 48.9% rate previously reported (Suzuki and Takeda, 1989). On the other hand, the misdetection rate is less than 30%. This value shows that the proposed method determined that over 70% of the correct homophones in compound nouns were correct. This means that the confirmation process can be significantly shortened because fewer correct compounds are presented for confirmation. Moreover, in the correction process, for more than 80% of detected errors, the correct homophone was a candidate. These results show that this method can detect and correct homophone errors in compound nouns successfully.

Table 3: Experimental results.

	misdetection rate [%]	detection rate [%]
data set 1	27.1	—
data set 2	28.3	—
data set 3	—	96.3
data set 4	—	97.1

5.4 Discussion

We analyzed the experimental results and determined that misdetection is caused by two factors;
(a) imperfect semantic restriction dictionary,
(b) semantic categories that belong to sets that can adjoin words having the same reading.
The number of compound nouns used to make the semantic restriction dictionary was different for each word reading. When the number of compound nouns used to construct the dictionary is large enough, misdetection caused by factor (a) will be minimized. Factor (b) can be offset by optimizing the semantic category system to improve semantic discrimination. This problem will be researched in the future.

6 Conclusion

This paper has described a method for detecting and correcting Japanese homophone errors in compound nouns used in a revision support system for Japanese texts; REVISE. The underlying concept of this method is that a compound noun component can restrict the set of semantic categories of words that can adjoin the component. The method accurately determines that a homophone is misused in a compound noun if one or both of its neighbors is not a member of the set defined by the homophone. Also, the method successfully indicates the correct candidates for the detected homophone errors automatically. Experiments indicate that the detection rate is over 95% and that the misdetection rate is less than 30%. These results confirm the validity of this method in detecting and correcting Japanese homophone errors in compound nouns.

References:

Fukushima, Toshikazu; Ohtake, Akiko; Ohyama, Yutaka; and Shutoh, Tomoki (1986). "Computer Assisted Environment for Japanese Text Creation : COMET." *IEICE technical report*, OS86-21, 15-22 (in Japanese).

Ikehara, Satoru; Yasuda, Tsuneo; Shimazaki, Katsumi; and Takagi, Shin-ichiro (1987). "Revision Support System for Japanese Verbal Error (REVISE)." *NTT Electrical Communications Laboratories Technical Journal*, 36, 9, 1159-1167 (in Japanese).

Kuga, Shigeki (1986). "Japanese Text Writing and Proofreading System WISE." *IEICE technical report*, OS86-28 , 13-18 (in Japanese).

Makino, Hiroshi, and Kizawa, Makoto (1981). "An Automatic Translation System of Non-segmented Kana Sentences into Kanji-Kana Sentences and its Homonym Analysis." *Trans. IPS Japan*, 22, 1, 59-67 (in Japanese).

Nakano, Hiroshi (1982). "Distinction between Homophones." *Technical report on SIG-NL of IPS Japan*, 33-4 (in Japanese).

Ohara, Hisashi; Takagi, Shin-ichiro; Hayashi, Yoshihiko; and Takeishi, Eiji (1991). "Revision Support Techniques for Japanese Text." *NTT R&D*, 40, 7, 905-913 (in Japanese).

Ono, Susumu, and Hamanishi, Masando (1981). *RUIGO-SHIN-JITEN*. Kadokawa Shoten (in Japanese).

Oshima, Yoshimitsu; Abe, Masahiro; Yuura, Katsuhiko; and Takeichi, Nobuyuki (1986). "A Disambiguation Method in Kana-to-Kanji Conversion Using Case Frame Grammar." *Trans. IPS Japan*, 27, 7 , 679-687 (in Japanese).

Suzuki, Emiko, and Takeda, Koichi (1989). "Design and Evaluation of a Japanese Text Proofreading System." *Trans. IPS Japan*, 30, 11 , 1402-1412 (in Japanese).

Tanaka, Yasuhito; Mizutani, Shizuo; and Yoshida, Sho (1984). "Relationship between words." *Technical report on SIG-NL of IPS Japan*, 41-4 (in Japanese).

Tanaka, Yasuhito, and Yoshida, Sho (1987). "A Method for Appropriately Selecting the Multivocal Words by Utilizing Knowledge Data (Relationship between Words)." *Technical report on SIG-NL of IPS Japan*, 60-3 (in Japanese).

The National Language Research Institute (1964). *BUNRUI-GOI-HYO*. Shuuei Shuppan (in Japanese).