Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in Theoretical Computer Science 225 (2009) 329–340

Electronic Notes in Theoretical Computer Science

www.elsevier.com/locate/entcs

Automatic Extraction of Super-Function From Bilingual Corpus

Manabu Sasayama[a,1,2] Fuji Ren[a,b,3] Shingo Kuroiwa[a,4]

[a] *Department of Information Science & Intelligent Systems Faculty of Engineering*
Tokushima University
Tokushima, Japan

[b] *School of Information Engineering*
Beijing University of Posts and Telecommunications
Beijing 100876, China

Abstract

Extraction of a large Super-Function (SF) is one of the most important factor in realizing SF based machine translation. This paper presents a method to automatically extract SF from a Japanese-English bilingual corpus. The extraction process matches Japanese noun and English noun in each bilingual sentence in a bilingual corpus using a bilingual dictionary. The experimental results show that this method performs very well in automatically extracting SF for machine translation. Then, we discuss a problem of SF based machine translation from the result of the evaluation experiment using extracted SF.

Keywords: Super-Function, translation, automatic extraction

1 Introduction

Increasingly, when information is searched for on the internet, the infomation desired is written in a foreign language. When internet users get a foreign document from a search engine, they take a longer time to understand the document than a native document. Therefore, machine translations is indispensable for reading a foreign document. However, current machine translation systems have many problems. Almost no users are satisfied with current machine translation systems, Because the accuracy and the quality of translations have not reached a level which fills the demand of a user.

[1] This research has been partially supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan under Grant-in-Aid for Scientific Research (B), 19300029, 17300065, Exploratory Research, 17656128.

[2] Email: sasayama@is.tokushima-u.ac.jp

[3] Email: ren@is.tokushima-u.ac.jp

[4] Email: kuroiwa@is.tokushima-u.ac.jp

doi:10.1016/j.entcs.2008.12.084

If a sentence which users want to translate exists in an exemplary bilingual corpus, they can get a good translation by only performing search. But coverage of a user's input sentence is limited by only using a bilingual corpus as it is. Therefore, we are continuing a trial, in which the coverage is extended by functionising a bilingual corpus ([3]∼[7]). The functionisation is done by introducing the Super-Function (SF) concept for each bilingual sentence in a corpus, and the functionised corpus is used for translation. Translation using SF has the advantage that coverage is wide and flexibility is higher than using the bilingual corpus as it is. However, there was a problem that the process was manual, we could only create a limited quantity of SF, and we had not yet clarified the problem of SF for creating a pratical system.

To solve this problem, this paper presents a method to automatically extract SF from a Japanese-English bilingual corpus. In particular, while examining the structure of a sentence we pay attention to the nouns contained in each bilingual sentence in the bilingual corpus, so that SF are automatically extracted by matching a common noun to source and target language using a dictionaly. We conduct a translation experiment using SF extracted from a large bilingual corpus, and clarify the validity and problems of translation using SF.

In the next section we describe the purpose of this research. In Section 3 we explain a definition and a structure of SF about SF based machine translation. In Section 4 we describe the method of extracting SF from a bilingual corpus automatically. In Section 5 we show the validity of this method by conducting the translation experiment, and consider the experimental result. Finally in Section 6 we give remarks about the conclusion and future works.

2 Purpose

Currently many commercial machine translation systems perform syntactic analysis, and the grammatical and conversion rules between structures are described manually ([1],[2]). Because input sentences are various, detailed grammatical and conversion rules are needed for performing flexible translation in large quantities. It is very difficult to describe manually all the rules that cover many sentences. Moreover, even though an output is a grammatical sentence, it may be unnatural as a translation. To solve this problem, a method of translation using a large bilingual corpora is proposed ([3]∼[16]). These methods are classified into two categories, example based machine translation ([3]∼[7],[9]∼[16]) and statistical based machine translation ([8]). The advantage of using the bilingual corpus itself is to generate a natural translation, because the language phenomena is made into the rule. However, becauese the language phenomena are used directly, coverage is limited, and flexibility is low. Therefore, a large corpus is needed in order to realize accuracy of high translation and quality. SF based machine translation, which is a type of an example based machine translation, extends coverage and raises flexibility by using a functionised corpus. There is Template-based machine translation ([15],[16]), which is a similar method to SF based machine translation. This method uses syntactic

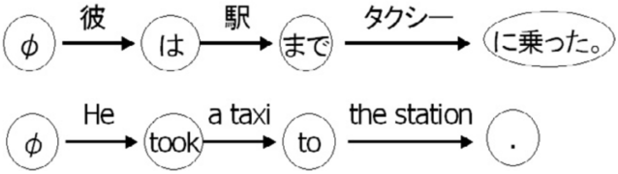


Fig. 1. Directional Graph of Japanese sentence(top) and English(bottom)

analysis to translate where SF based machine translation does not.

SF in machine translation is defined as showing correspondence of a source and a target language ([3],[4]). SF based machine translation does not use syntactic analysis to translate as most conventional MT systems do. Therefore, the process is simple. Because SF is corpus based, a fluent translation is generated. However, previously SF was extracted from a corpus manually. As a result, we could extract only SF of a limited quantity. Also, a SF's problem for creating a practical system has not been clarified. In this paper, the method of extracting SF from a bilingual corpus automatically by the following methods is proposed:

(i) The noun of each bilingual sentence is determined using a noun classification rule, and a sentence is expressed using a direction graph.

(ii) Based on the above-mentioned results, a noun of a source and a target language sentence is matched in a bilingual word dictionary.

(iii) SF is obtained from the consequence of matching.

Moreover, we conduct a translation experiment using extracted SF, and show that suitable translation is possible.

3 Super-Function based machine translation

SF is a function that shows some defined relations between a source and a target language sentence pair. We assume that a sentence consists of some constant and variable parts. By replacing the variable parts, one can obtain a multitude of different sentences. In this paper, we only focus on nouns as variables for translation.

SF consists of two architectures. One is Directional Graph. Another one is Transformation Table. A Directional Graph is composed of nodes and edges. In Directional Graph in SF, nodes represent constant parts and edges represent variable parts.

Fig. 1 shows the example of Directional Graph. The top in the figure is Directional Graph of Japanese sentence and the bottom is Directional Graph of English sentence. ϕ in value of the first circle is a null character. A null character is used when there is no word which corresponds to a first character.

Transformation table consists of Node Table and Edge Table. Node Table is a correspondence table of source and target languages. Edge Table represents an order of a variable of a source language in a target language. Edge Table describes a condition of a noun to distinguish whether a noun is a pronoun or not. Table 1 shows the example of Node Table and Edge Table. On the left is Node Table and

Table 1
Node Table(left) and Edge Table(right)

J	E	J	E	condition
ϕ	ϕ	1	1	1p
ha	took	2	3	a
made	to	3	2	the
ninotta.	.	-	-	-

Table 2
Unite table

Japanese node	Z:ha:made:ninotta
order of noun	1:3:2
condition of noun	1p:a:the
English node	Z:took:to:.

on the right is Edge Table. The condition in Table 1 represents that the condition '1p' represents a pronoun code we defined.

We use the unification of the two tables. The unified table is called Unite Table. Table 2 shows the Unite Table. We obtain the English node, the location and the condition using Table 2 when we obtain the Japanese node. 'Z' denotes a null character.

3.1 Translation process

Here is the flow of machine translation using SF.

(i) A Japanese sentence is separated into nouns(edges) and constants(nodes) by morphological analysis using ChaSen [17].

(ii) The constants(nodes of the Japanese sentence) are matched with a NTB in the SF database and an English node, location and condition are obtained. However, in certain cases, a SF has multiple English nodes and edges. Section 3.2 describes these cases.

(iii) All nouns are translated using a bilingual dictionary.

(iv) Based on the order of the nouns in the ETB a rearrangement of the nouns within the English nodes takes place. Finally a translation sentence is outputted.

3.2 Multiple target language nodes

An example of a Japanese node having multiple English nodes and edges is shown in Table 3. In this instance, we select one of the English nodes and edges by matching

Table 3
Multiple target language nodes

Japanese node	Z:ha:made:ninotta
order of noun1	1:3:2
condition of noun1	1p:a:the
English node1	Z:took:to:.
order of noun2	1:3:2
condition of noun2	1p:a:the
English node2	Z:rode:to:.

the condition of the noun in the English edge with a Japanese noun. Cases in which the English nodes and edges cannot be disambiguated using the condition are a problem that must be solved. However, the frequency of this problem is not known, because a large quantity of SFs was not obtained manually. In this paper, we calculate the frequency of SFs which have two or more English nodes and edges using a large quantity of SFs, which were created using the proposed method.

4 Automatic extraction of SF from corpus

The method of extracting SF from a bilingual corpus automatically is shown in Figure 2.

First, Japanese and English sentences are separated into nouns and constants by using morphological analysis (Figure 2(1)). Second, We obtain the location relationship between Japanese and English nouns by matching using a bilingual dictionary(Figure 2(2)). This process is called noun matching and will be described in section 4.2. Third, If there is not a corresponding noun then it is an isolated noun. In this case, we perform Isolated noun processing, which is detailed in section 4.3.

A node table and an edge table are created from these processes. Finally, we merge extracted SFs with Japanese nodes.

4.1 Morphological analysis

This step classifies morphemes as either a noun or an other constant. This step requires a noun classification rule in order to identify nouns. There are two sets of noun classification rules. One is for Japanese noun classification and the other is for English classification.

Both store rules to classify whether a morpheme is a noun or not(Table 4). To do this the next morpheme is also checked. For example, if the next morpheme of Sahen is Noun, Sahen is classified as a noun. The set of Japanese noun classification rules has 32 rules and the set of English noun classification rules has 40 rules.

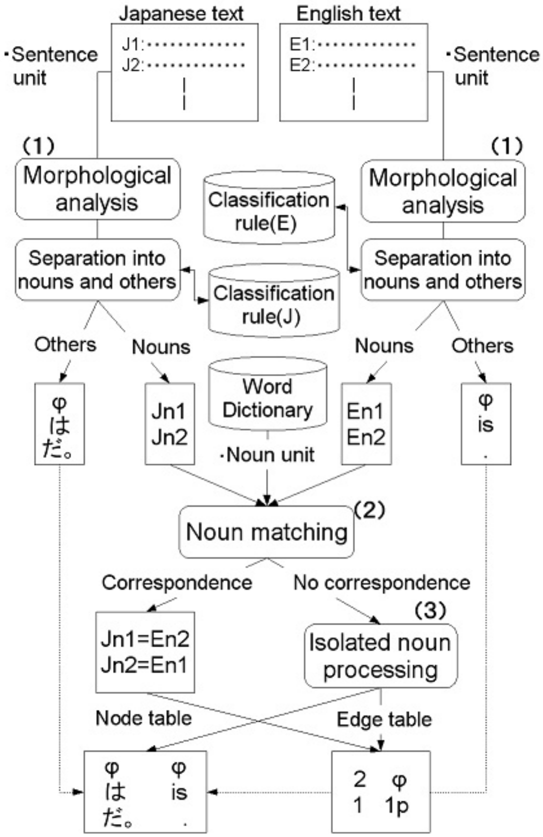


Fig. 2. Outline of extraction process

When an unexpected rule appears in a sentence pair, we do not extract a SF from the sentence pair, to prevent extraction of bad SF.

4.2 Noun matching

This step matches a Japanese noun with an English noun to obtain the location of the Japanese noun in the English sentence. The flow of noun matching is as follows.

(i) We obtain some Japanese nouns by translating an English noun into Japanese using a bilingual dictionary.

(ii) We search for a Japanese noun corresponding to one of translations.

(iii) We obtain the location and the condition of the English noun(which was translated into Japanese) from the matching in Step ii. Articles are always stored as a condition an pronouns are stored as a condition when the matched noun was a pronoun.

(iv) An English noun becomes an isolated noun when none of its translation are matched with a Japanese noun in Step ii. When this happens, Isolated noun processing is performed(section 4.3).

(v) We repeat steps i) to iv) for all english nouns.

Table 4
A part of classification rule

Rule	Example
sahen varb + varb → varb	benkyo suru
sahen varb + except → varb noun	benkyo ga
sahen varb + noun → noun	benkyo kikan
noun + nominal adjective → noun	sekai heiwa
nominal adjective + noun → noun	zyuyou anken

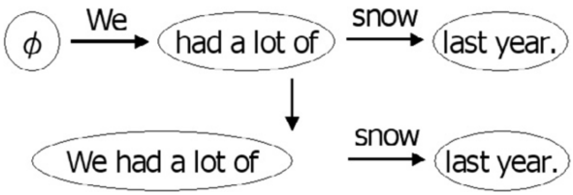


Fig. 3. Insert processing to a node

(vi) Finally, a Japanese noun which does not correspond with an English nouns becomes an isolated noun.

4.3 Isolated noun processing

In this section, we explain about Isolated noun processing. This process has two cases: 1) the isolated noun is a Japanese noun, 2) the isolated noun is an English noun.

When the isolated noun is an English noun(procedure iv), the English noun becomes a node instead of an edge, and performs admission processing to a node (Figure 3). Using this process, we can effectively deal with elliptical nouns.

Next, when an isolated noun is a Japanese noun, the Japanese noun is registered into SF as a condition for the noun as an isolated noun. It is because the Japanese noun is not necessarily an English noun. An example is shown in Figure 4. The translation to 'netukikyu' is usually 'a hot-air balloon'. However, in this example, because it is 'a hot-air', 'a hot-air' does not exist in a word dictionary but instead becomes an isolated noun.

5 Automatic extraction of SF and experimental evaluation

We extracted SF automatically from a Japanese-English bilingual corpus, and we compared automatic and manual extraction to verify that an automatically extracted SF and a manually extracted SF were the same. Moreover, in order to

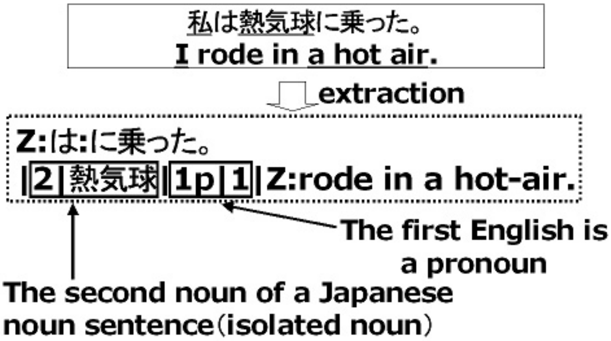

私は熱気球に乗った。
I rode in a hot air.
extraction
Z:は:に乗った。
|2|熱気球||1p|1||Z:rode in a hot-air.
The first English is a pronoun
The second noun of a Japanese noun sentence(isolated noun)

Fig. 4. Structure of SF

Table 5
Extracted SF

number of nodes	number of SF	total(%)
1node	7,430	4.6
2node	32,286	20.4
3node	57,726	36.4
4node	41,782	26.3
5node	14,550	9.1
6node	3,488	2.2
above 7node	1,372	0.8
total	158,634	100

check whether suitable translation is possible using the extracted SF, the translation experiment (hereinafter closedtest) using the bilingual corpus which the SF were extracted from was conducted.

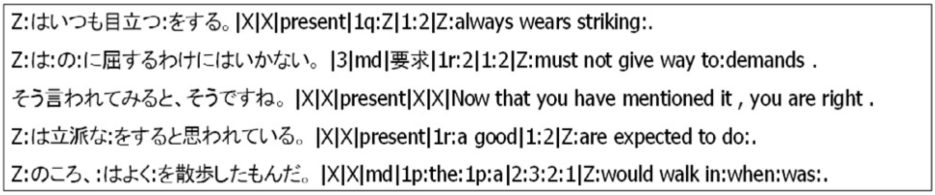
5.1 *Automatic extraction of SF*

5.1.1 *Experimental conditions*

We used 202,597 bilingual sentences in the Japanese-English bilingual corpus [19]. The word dictionary was created from the English-Japanese bilingual dictionary of EDR[20]. We used ChaSen [17] for Japanese morphological analysis and Brill tagger [18] for English.

5.1.2 *Experimental results*

The proposed method was able to extract SF from 194,689 sentences. 158,634 SF has been extracted except for overlapping SF. There were 7,908 failed sentences. Table 5 shows the extraction result classified by node. According to Table 5, 99%



```
Z:はいつも目立つ:をする。|X|X|present|1q:Z|1:2|Z:always wears striking:.
Z:は:の:に屈するわけにはいかない。|3|md|要求|1r:2|1:2|Z:must not give way to:demands .
そう言われてみると、そうですね。|X|X|present|X|X|Now that you have mentioned it , you are right .
Z:は立派な:をすると思われている。|X|X|present|1r:a good|1:2|Z:are expected to do:.
Z:のころ、:はよく:を散歩したもんだ。|X|X|md|1p:the:1p:a|2:3:2:1|Z:would walk in:when:was:.
```

Fig. 5. Super-Function

of the SF were formed by less than six nodes. In contrast, there were extremely few SF with seven or more nodes. An example of an extracted SF is shown in Figure 5.

5.1.3 Comparision with manual extraction

We compared automatic and manual extraction to verify that an automatically extracted SF and a manually extracted SF were the same.

First, we randomly selected 200 bilingual sentences from the bilingual sentence pairs which could have SF extracted. Next, we extracted SF automatically and manually from the 200. Finally, we compared the automatically extracted SF and the manually extracted SF. As a result, 184 SFs(92%) corresponded. The cause of the differences in the 16 sentences which did not correspond were morphological analysis errors and a shortage of noun classification rules. Here are examples of mismatches.

- Failure of morphological analysis

  SF of automatically extraction corresponding to 'Kanojo ha Kanazuti douzen oyogenai.' and 'She can no more swim than a hammer can.' were 'Z:ha:zuti douzen oyogenai.|2|kana|present|1|1q|Z:can no more swim than a hammer can.' SF of manually extraction was 'Z:ha:douzen oyogenai.|Z|Z|present|1q:Z|1:2|Z:can no more swim than a hammer can.' The result of automatic and manually extraction were different because the morphological analysis of 'Kanazuti' failed in automatically extraction.

- A shortage of noun classification rules

  SF of automatically extraction corresponding to 'Inu wo kautameno kihontekina rule ha nandesuka.' and 'What is the basic rule for keeping a dog?' were 'Z:wo kautameno:na:ha:desuka.|2:4|kihonteki:nani|past|the basic:a|3:1|What is:for keeping:?' SF of manually extraction was 'Z:wo kautameno kihontekina:ha:desuka.|4|nani|past|the basic:a|3:1|Z:can no more swim than a hammer can.' In this case, a classification rule of 'noun(kihon) nominal adjective(teki) auxiliary verb(na)' was not registered.

5.1.4 Discussions

Automatic extraction of SF was successful, because 92% of SF extracted manually were in agreement with the SF extracted automatically. Moreover, because an automatically extraction success rate is 96%, the SF extracted manually becomes 88% of the total.

Many sentences which were not able to be extracted stemmed from a shortage

Table 6
SF with many candidates

a number of candidates	japanese node of SF
284	Z:ha:desu.
229	Z:ha:da.
197	Z:ha:no:desu.
164	Z:ha:no:da.
131	Z:desu.

Table 7
Result of closedtest

successful	failure
165,145 sentences (84.1%)	29,199 sentences (15.9%)

of noun classification rules. This is solved by adding more noun classification rules. 11% of the SFs had multiple English nodes and edges. There were some SFs which had many English nodes and edges.(Table 6). This table shows that SF with two or more candidates has many sentences of fundamental structure. We discuss further in the next section(closed test) about this point.

5.2 Closed test

5.2.1 Experimental conditions

194,689 sentences which have extracted SF were used as the test set. All SF (158,634) extracted in Section 5.1 were used in the experiment. The success condition of this experiment is when the translation results in the original sentence, any other is considered a failure.

5.2.2 Experimental results

The experimental result is shown in Table 7. There were 165,145 (84.1%) successful sentences and there were 29,199 (15.9%) failed sentences. Because 84.1% of all the sentences were successful, it proves that SF extracted automatically has capability enough.

5.2.3 Discussions

We consider a failed sentence, a cause of failure and its solution. An example of a failed sentence is shown in Table 8. Because it is a closed test, the English node automatically extracted from the sentence is surely contained in SF as a candidate also about the failed sentence. However, there are two or more English nodes in

Table 8
Example of wrong sentence

<table>
<tr><td>Japanese sentence</td><td>KanojohaUmaninotta.</td></tr>
<tr><td>English sentence</td><td>She rode a horse.</td></tr>
<tr><td>corresponding SF</td><td>Z:ha:ninotta.
|Z|Z|past|1q:a|1:2|Z:rode:.
|Z|Z|past|1p:the|1:2|Z:got on:.
|Z|Z|past|1q:the|1:2|Z:got in:.
|2|Netukikyu|past|1p|1|Z:rode in a hot air .</td></tr>
</table>

this SF. Sometimes a candidate's narrowing down was not completed, because the first candidate was chosen. The example of Table 8 is the case that a verb changes depending on the noun. In English, 'ride' and 'take' are properly selected by nouns, but in Japanese, it is the same verb 'noru'. Because these are irregular, it is difficult for them to narrow down a candidate. One way of solving this problem is to add the concept of nouns to SF using a concept dictionary.

6 Conclusions

We presented a method to automatically extract SF from a corpus. In particular, while examining the structure of a sentence we paid attention to the nouns contained in each bilingual sentence in the bilingual corpus. SF was automatically extracted by matching a noun common to source and target languages using a dictionary. We performed a closed test using extracted SF, and showed that it can translate by SF extracted automatically. Because SF with two or more translation candidates existed, it proved that we need to narrow down candidate SF.

In the future, we will examine a method to narrow down translation candidates.

References

[1] H. Nomura, Gengo-shori to Kikai honyaku (Language processing and machine translation), Kodansha Scientific (in Japanese), 1991

[2] H. Tanaka, Shizen Gengo Shori (Natural language processing), The Institute of Electronics, Information and Communication Engineers, 1999

[3] F. Ren, Super-function based machine translation, Language Engineering, Proceedings of JSCL and TsingHua University Press, pp.305-312. 1997.

[4] F. Ren, Super-function based machine translation, Communications of COLIPS, pp.83-100. 1999.

[5] M. Sasayama, F. Ren, S. Kuroiwa, X. Zhao, Japanese-English Machine Translation Systems for Web Users Using Super-Function, International Conference on Information-2002, Series of Information and Management Sciences, Vol.3,pp.366-371,2002

[6] M. Sasayama,F. Ren,S. Kuroiwa, Super-Function Based Japanese-English Machine Translation, IEEE International Conference on Natural Language Processing and Knowledge Engineering, pp.555-560,2003

[7] M. Sasayama,F. Ren,S. Kuroiwa, Super-Function based Japanese English Machine Translation Experiment and Evaluation, Proceedings of the third International Conference on Information, pp.195-198,2004

[8] P.F. Brown,J. Cocke,S.A.D. Pietra,V.J.D. Pietra,F. Jelinek,J.D. Lafferty,R.L. Mercer,and P.S. Roosin A statistical approach to machine translation, Computational linguistics vol.16, no.2,June, 1990

[9] K. McTAIT, Memory-based translation using translation Patterns, UMIST, 2001

[10] T. IZUHA, Machine Translation Using Bilingual Term Entres, the Institute of Electronics Information and Communication Engineers, vol.101,No.189, pp.1-7, July, 2001,

[11] H. Echizenya, K. Araki, Y. Momouchi, K. Tochinai, Machine Translation Using Recursive Chain-Link-Type Learning Based on Translation Example, the Institute of Electronics Information and Communication Engineers, vol.J85-D-2 No.12, pp.1840-1852, July, 2002,

[12] M. Kitamura, Y. Matsumoto, Automatic Acquisition of Translation Rules from Paralell Corpora, Information Processing Society of Japan, vol.37 No.6, pp.1030-1040 1996,

[13] C. Brockett, T. Aikawa, A. Aue, A. Menezes, C. Quirk and H. Suzuki English-Japanese example-based machine translation using abstract linguistic representations, COLING, 2002

[14] E. Aramaki, Sadao Kurohashi, Hideki Kashioka and Naoto Kato, Probabilistic Model for Example-based Machine Translation, MT Summit X, 2005

[15] H. Watanabe, S. Kurohashi, and E. Aramaki, Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation, International Conference on Computational Linguistics(COLING), pp.906–912, 2000

[16] H. Kaji, Y. Kida, and Y. Morimoto, Learning translation templates from bilingual text, Proc. of COLING, pp.672–678, 1992

[17] ChaSen version 1.0 is officially released on 19 February 1997 by Computational Linguistics Laboratory, Graduate School of Information Science, Nara Institute of Science and Technology

[18] E. Brill, Brill Tagger, http://www.cs.jhu.edu/~brill/home.html

[19] The Tanaka Corpus, http://www.csse.monash.edu.au/~jwb/tanakacorpus.html

[20] EDR Electronic Dictionary Specifications Guide, Japan Electronic Dictionary Research Institute