Phrase-Level Pattern-Based Machine Translation Based on Analogical Mapping Method

Jun Sakata, Masato Tokuhisa, and Jin'ichi Murakami

Information and Knowledge Engineering, Tottori University
4-101 Koyama-Minami, Tottori 680-8550, Japan
{d112004,tokuhisa,murakami}@ike.tottori-u.ac.jp

Abstract. To overcome the conventional method based on Compositional Semantics, the Analogical Mapping Method was developed. Implementing this method requires a translation method based on sentence patterns. Although a word-level pattern-based translation system already exists in Japanese-English machine translation, this paper describes the new phrase-level pattern-based translation system. The results of translation experiments show that the quality of phrase translation is still low. However, these problems are to be resolved in our future work.

Keywords: Pattern-Based Machine Translation, Sentence and Phrase Pattern, Phrase Translation.

1 Introduction

The conventional machine translation method based on compositional semantics has a problem in that it cannot generate sentence meaning when it generates the target sentence. To resolve this problem, Ikehara et al. proposed a transfer method named the "Analogical Mapping Method" [1]. This method uses sentence patterns that have "linear" and "non-linear" parts. In translation, we perform local translation of variables that are equivalent to the "linear part" and insert these results into the target sentence pattern.

Machine translation based on the Analogical Mapping Method requires many sentence patterns, a pattern matching system, and the translation system with matching sentence patterns. To implement Japanese-English MT based on the Analogical Mapping Method, we developed the "Japanese-English compound and complex sentence pattern dictionary", a structural pattern matching system named "SPM"[2], and a generating system named "ITM"[1]. The dictionary has 226,817 sentence pattern pairs from Japanese/English compound/complex sentences pairs. It also has three levels of sentence pattern: word-level (121,904 pattern pairs), phrase-level (79,438 pattern pairs), and clause-level (25,475 pattern pairs) [1]. The sentences were collected from various Japanese-to-English parallel corpora.

[1] "ITM" has never been described in a paper in detail.

H. Isahara and K. Kanzaki (Eds.): JapTAL 2012, LNAI 7614, pp. 28–39, 2012.

Currently, only the word-level pattern-based translation is implemented in ITM. However, this pattern matching rate (SPM + ITM) is still low. To increase the pattern matching rate, we try to use phrase-level patterns [3]. For this, we need phrase translation. The phrase pattern dictionary has been developed from the Japanese-English compound and complex sentence pattern dictionary. Moreover, we implement the phrase-level pattern-based translation that performs phrase translation using this phrase pattern dictionary.

The rest of this paper is organized as follows. Section 2 describes an example of our sentence pattern dictionary and phrase pattern dictionary. Section 3 explains the ordinary word-level pattern-based translation method. Section 4 presents our proposed method. Section 5 discusses results of the experiments and assesses the proposed method. Finally, section 6 offers our conclusions.

2 Sentence Pattern Dictionary and Phrase Pattern Dictionary

2.1 Sentence Pattern

Fig. 1 shows an example of our sentence patterns. Sentence patterns have letters, variables, functions, and markers. Japanese/English word/phrase/clause alignment are replaced with variables. Word-level patterns have word variables. "*N*2^poss" means it N2 is the possessive case in English. "*V*5^past" means *V*5 is the past tense. In Japanese, a subject is often omitted, so "<*N*1 は >" means whether the subject is omitted or not in pattern. In English patterns, <I|*N*1> is "*N*1" if Japanese matches *N*1, or "I" if not. ".hitei" and ".kako" are tense and modality function, respectively [4]. In total, 226,817 sentence patterns have already been created [1].

<table>
<tr><td>AC000004-00</td><td></td></tr>
<tr><td>Japanese Sentence:</td><td>彼のお母さんがああ若いとは思わなかった。</td></tr>
<tr><td></td><td>[Kare no okaasan ga aa wakai towa omowa naka tta。]</td></tr>
<tr><td>English Sentence:</td><td>I never expected his mother to be so young.</td></tr>
<tr><td>Word-Level JP.Pattern:</td><td><N1 は >N2 の N3 がああ AJ4 とは V5.hitei.kako。</td></tr>
<tr><td>Word-Level EN.Pattern:</td><td><I|N1> never V5 ^past N2^poss N3 to be so AJ4.</td></tr>
<tr><td>Phrase-Level JP.Pattern:</td><td><N1 は >NP2 がああ AJ3 とは V4.hitei.kako。</td></tr>
<tr><td>Phrase-Level EN.Pattern:</td><td><I|N1> never V4^past NP2 to be so AJ3.</td></tr>
</table>

Fig. 1. Description of Japanese-English compound and complex sentence pattern dictionary

2.2 Phrase Pattern

Our phrase pattern dictionary is automatically extracted from the Japanese-English compound and complex sentence pattern dictionary. Phrase patterns are extracted from the word-level sentence patterns.

Phrase patterns are categorized as noun phrases (*NP*), verb phrases (*VP*), adjective phrases (*AJP)*, adjective-verb phrases (*AJVP*), and adverbial phrases (*ADVP*).

Fig. 2 shows examples of each phrase pattern. In the *VP* patterns, verbs are letters and not changed variables, because verbs have nonlinearity.

<NP>

Japanese Pattern:	$N1$ の $N2$ [$N1$ no $N2$]
English Pattern:	$N1$ˆposs $N2$
Original Japanese:	彼 の お母さん [kare no okaasan]
Original English:	his mother

<VP>

Japanese Pattern:	ああいう $N1$ と’ 付き合う’ [aayuu $N1$ to ’tukiau’]
English Pattern:	’associate’ with that kind of $N1$
Original Japanese:	ああいう 人 と 付き合う [aayuu hito to tukiau]
Original English:	associate with that kind of person

<AJP>

Japanese Pattern:	実に $AJ1$ˆrentai [jituni $AJ1$ˆrentai]
English Pattern:	very $AJ1$
Original Japanese:	実に 痛ましい [jituni itamashii]
Original English:	very painful

<AJVP>

Japanese Pattern:	$ADV1$ $AJV2$ˆrentai
English Pattern:	$ADV1$ $AJ2$
Original Japanese:	とても 静かな [totemo shizukana]
Original English:	very quiet

<ADVP>

Japanese Pattern:	$N1$ の あと [$N1$ no ato]
English Pattern:	after $N1$
Original Japanese:	手術 の あと [shujutu no ato]
Original English:	after the operation

Fig. 2. Description of phrase patterns

3 Pattern-Based Machine Translation

3.1 Japanese Sentence Pattern Matching (SPM)

The Japanese pattern matching system named SPM has already been developed. The SPM [2] implements the Augmented Transition Network (ATN) algorithm

[5] with breadth-first search and uses sentence patterns. The input sentence for SPM is already morphological and has semantic codes added. SPM performs pattern matching between the input sentence and sentence patterns. Moreover, SPM outputs the pattern matching results. Fig. 3 shows an example of an input sentence.

```
1. /彼 (1710,{NI:23,NI:48})
2. +の (7410)
3. /お母さん (1100,{NI:80,NI:49})
4. +が (7410)
5. /ああ (1110)
6. /若い (3106,{NY:5})
7. +と (7420)
8. +は (7530)
9. /思わ (2392, 思う, 思わ,{NY:32,NY:31})
10. +なかっ(7184, ない, なかっ)
11. +た (7216)
12. +。 (0110)
13. /nil
```

Fig. 3. Example of an input sentence

In the first line in Fig. 3, “彼” is a Japanese morpheme, “1710” is tagging code, and “NI:23,NI:48” are indeclinable semantic codes [6]. In the sixth line, “NY:5” is a declinable semantic code. Each line shows a Japanese morpheme and semantic information.

```
PATTERN=PJAC000004-00
          =[NP2, が, ああ,AJ3, とは, V4,.hitei,.kako,。]
          =[1,2,3,4,5,6,7,8,9,10,11,12]=12
NP2=[1,2,3]=3
AJ3=[6]=1
V4=[9]=1
```

Fig. 4. Results of sentence pattern matching

Fig. 4 shows the sentence pattern matching result. “AC000004-00” is the Japanese pattern ID. “*NP*2=[1,2,3]” shows that the morpheme numbers 1, 2, and 3 are matched as the phrase variable *NP*2. *NP*2 is “彼 の お母さん [kare no okaasan]”. “*AJ*3=[6]” shows “若い [wakai]” is matched as *AJ*3. “*V*4=[9]” shows “思わ [omowa]” is matched as *V*4.

If several sentence patterns are matched, we select only one sentence pattern in accordance with the following steps.

1. Sentence patterns are selected by using semantic codes in the input sentence.

2. Pattern matching test is done with all Japanese sentences in the Japanese-English compound and complex sentence pattern dictionary.
   The most matched pattern is selected.
3. If several patterns are left, only one pattern is randomly selected.

3.2 Word-Level Pattern-Based Machine Translation (ITM-w)

An English sentence is translated with matched sentence pattern pairs and the translation system "ITM". ITM-w (word-level translation) performs word translation for Japanese linear parts of SPM results. Word translation is performed by a word dictionary. Several results of word translation are inserted into the English pattern, and the maximum likelihood sentence is selected by the English language model. The translation steps of ITM-w are as follows.

1. The sentence pattern pair is selected (section 3.1).
2. Word translations of Japanese linear parts are performed.
3. Word translation results are changed to the assigned form.
4. Candidates of word translation are inserted into the English sentence pattern.
5. The maximum likelihood sentence is selected by the English language model.

4 Phrase-Level Pattern-Based Machine Translation

We implemented phrase-level pattern-based translation. The proposed method is constructed of SPM-s (sentence pattern matching), SPM-p (phrase pattern matching), ITM-p (phrase-level translation), and ITM-w. SPM-s is already described section 3.1, and ITM-w is described section 3.2. SPM-p is the same program as SPM-s but uses phrase patterns. ITM-p performs phrase translation with phrase patterns and word translation with word dictionary. In ITM-p, SPM-p and ITM-w are activated for phrase translation.

Phrase patterns have word variables and are translated by the word dictionary. Word translation results are inserted into the phrase pattern, and the maximum likelihood phrase is selected by the English language model. If several phrase patterns are matched, all phrase patterns are used.

The steps in our proposed method are as follows.

Process 1 Sentence pattern matching is performed by SPM-s.
Process 2 The sentence pattern pair is selected (section 3.1).
Process 3 Phrase translation is performed in ITM-p.
Process 4 Phrase pattern matching is performed by SPM-p.
Process 5 Phrase translation is performed by ITM-w.
Process 6 Word translation is performed in ITM-p.
Process 7 Candidates for all local translation results are inserted into the English sentence pattern.
Process 8 The maximum likelihood sentence is selected by the English language model.

Fig. 5 sketches the whole configuration of phrase-level pattern-based translation.

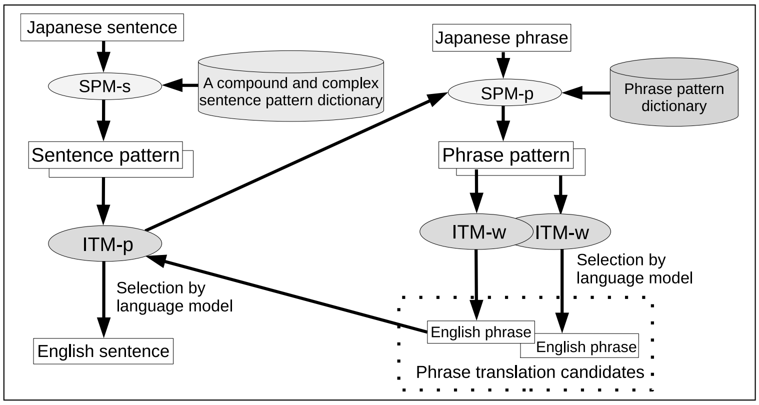
Japanese sentence
SPM-s
A compound and complex sentence pattern dictionary
Sentence pattern
ITM-p
Selection by language model
English sentence
Japanese phrase
SPM-p
Phrase pattern dictionary
Phrase pattern
ITM-w
ITM-w
Selection by language model
English phrase
English phrase
Phrase translation candidates

Fig. 5. Phrase-level ITM

4.1 Example of Phrase-level Translation (ITM-p)

Fig. 6 shows an example of translation.

Input Sentence:	彼のお母さんがああ若いとは思わなかった。 [kare no okaasan ga aa wakai towa omowa naka tta。]
Reference Sentence:	I never expected his mother to be so young.
English Pattern:	<I|$N1$> never $V4$ˆpast $NP2$ to be so $AJ3$.
Output Sentence:	I never expected his mother to be so young .

Fig. 6. Example of translation results

Processes 1 and 2 are already described in section 3.1.

Process 3. In SPM-p, "彼 の お母さん [kare no okaasan]" is matched as the phrase variable $NP2$ and is translated by ITM-w. Phrase translation results ("his mother", "he's mother", ……) are obtained with all phrase patterns.

Process 6. The Japanese morphemes "若い [wakai]" and "思わ [omowa]" are matched as the word variables $AJ3$ and $V4$, respectively. They are translated by the word dictionary, and several candidates are obtained. "$V4$ˆpast" means the past tense is selected from the translation candidates. For example, only past tense words ("thought", "expected", ……) are selected from translation candidates ("think", "thought", "expect", "expected", ……).

Process 7. All local translation results are inserted into the English pattern.

Process 8. The maximum likelihood sentence is selected by tri-gram. Fig. 7 shows an example of words selection by tri-gram.

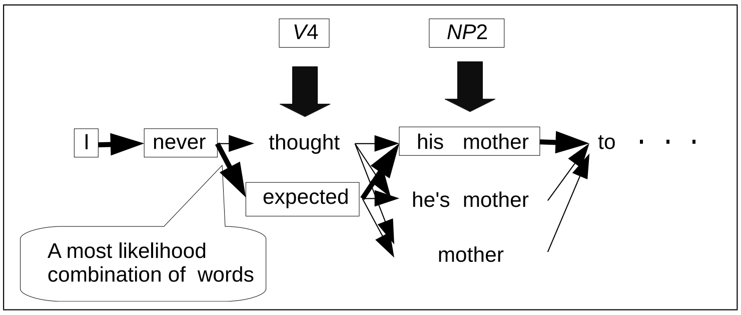

Fig. 7. Selection from translation candidates

5 Experiments

First, we carried out the closed-test to check the implementation of ITM-p. Next, we carried out the open-test to investigate effectiveness of the proposed method.

5.1 Closed-Test

Experimental Method. To survey our system, the closed-test was carried out with input sentences from the Japanese-English compound and complex sentence pattern dictionary. We used 500 input sentences for pattern matching and selected 100 sentences to evaluate translation accuracy. For pattern matching, we used 100 sentences that included at least one *NP* in the self-pattern. *VP*, *AJP*, *AJVP*, and *ADVP* were tested in the same way. We calculated the pattern matching rate and assessed the translation accuracy by human evaluation.

Pattern Matching Results. Table 1 shows the number of sentences that matched sentence patterns. In table 1, "Pattern mismatch" means that no pattern is matched to the input sentence. "Self-pattern mismatch" means that the self-pattern is not matched but other patterns are. "Self-pattern match" means that the self-pattern is matched.

Table 1. Results of self-pattern matching

	Pattern mismatch	Self-pattern mismatch	Self-pattern match
♯Sentence NP	4	7	89
♯Sentence VP	4	2	94
♯Sentence AJP	7	4	89
♯Sentence $AJVP$	11	16	73
♯Sentence $ADVP$	20	27	53

Example of Pattern Mismatch. Fig. 8 shows examples of pattern mismatch and self-pattern mismatch.

<Pattern mismatch>

Input St.	早くも子供のころからたくさん本を読むという傾向を示した。 [Hayakumo kodomo no koro kara takusan hon wo yomu to yuu keikou wo shimeshi ta。]
Self-Pt.JP.	<*N*1 は >*ADVP*2!*VP*3(という | と言う)*N*4 を *V*5.kako。 [<*N*1wa>*ADVP*2!*VP*3(toyuu|toyuu)*N*4 wo *V*5.kako。]

<Self-pattern mismatch>

Input St.	その品は品質がいいので高価なのももっともである。 [Sono shina wa hinshitu ga ii node koukana no mo mottomo dearu。]
Self-Pt.JP.	*NP*1 は *N*2 が *AJ*3^rentai ので *AJV*4^rentai!のも!*VP*5.tearu。 [*NP*1 wa *N*2 ga *AJ*3^rentai node *AJV*4^rentai! nomo !*VP*5.tearu。]

Fig. 8. Examples of pattern mismatch and self-pattern mismatch

The following three reasons cause sentence pattern mismatching and self-pattern mismatching.

Cause 1 Sub-networks are short for SPM-s.
Cause 2 The sentence pattern description is wrong.
Cause 3 The result of morphological analysis is wrong.

Cause 1. "Pattern mismatch" in Fig. 8 is caused by a shortage of sub-networks for SPM-s. In this example, "早くも子供のころから [hayakumo kodomo no koro kara]" is *ADVP*. Sub-networks for *ADVP* are short, and this sentence is not matched to this pattern. Therefore, this sentence is not matched to any sentence patterns. In table 1, the number of pattern mismatches in *ADVP* is the largest. In *ADVP*, many cases of pattern mismatching are caused by this problem. They seem to require many sub-networks for each different structure to respective phrase patterns. To resolve this problem, we have to add sub-networks to SPM-s.

Cause 2. "Self-pattern mismatch" in Fig. 8 is caused by the wrong pattern. " もっとも である [mottomo dearu]" is *AJVP*, but it is described as "*VP*5.tearu" in this sentence pattern. To resolve this problem, we need to examine and classify different kinds of wrong descriptions and correct them manually.

Cause 3. We omitted the case for failing morphological analysis.

Human Evaluation. Table 2 shows evaluation criteria.

Table 2. Evaluation criteria

Eval. 1	The sentence structure is correctly composed, and all local translation are not mistranslation.
Eval. 2	The sentence structure is correctly composed, but a local translation is mistranslation.
Eval. 3	The sentence structure is incorrectly composed.

Table 3 shows the evaluation results.

Table 3. Evaluation results

	Eval.1	Eval.2	Eval.3
♯Sentence	27	68	5

Fig. 9 shows examples of Eval. 1 and Eval. 2. All results in Eval. 3 are caused by wrong sentence pattern description. In Eval. 1, the sentence structure and all local translations are correct. In Eval. 2, sentence structure is correct, but local translations of *NP*1 and *N*3 are not correct.

<Eval.1>

Input Sentence:	これを読んで泣かざるを得ぬ。 [Kore wo yon de naka zaru wo e nu。]
Reference Sentence:	I can not read this without crying.
English Pattern:	<I|N1> can not VP2ˆbase without V3ˆing.
Matched Morphemes:	*VP*2 = kore wo yon, *V*3 = naka
Output Sentence:	I can not read this without crying.

<Eval.2>

Input Sentence:	その晩餐会は彼をたたえるために開かれた。 [Sono bansankai wa kare wo tatae ru tameni hiraka re ta。]
Reference Sentence:	The dinner party was a tribute paid to him.
English Pattern:	*NP*1 @beˆpast *N*3 paid to *N*2ˆobj .
Matched Morphemes:	*NP*1 = Sono bansankai, *N*2 = kare, *N*3 = tatae
Output Sentence:	That it was name paid to him .

Fig. 9. Examples of Eval.1 and Eval.2

In Table 3, the translation accuracy is low in spite of the self-sentence pattern being used. The most significant reason for this is the selection of the improper phrase translation. Such an example is shown in Eval. 2 in Fig. 9. "その 晩餐会 [sono bansankai]", which corresponds to "the dinner party", is matched as $NP1$, and the selected translation result is “that it”. In this case, selected phrase translation uses the phrase pattern “$AJ1$ $N2$ˆpron”. “その [sono]” is matched as $AJ1$, and the translation result of “その [sono]” is “that”. “晩餐会 [bansankai]” is matched as $N2$ˆpron. “$N2$ˆpron” means that $N2$ is the pronoun. Thus, the translation result of “晩餐会 [bansankai]” is changed to the pronoun “it”. On the other hand, the correct local translation “the dinner party ($AJ1$ $N2$)” is included in the candidates. This suggests failed selection from sentence candidates in ITM-p. It seems that the selection by the language-model is not good enough.

Functions in phrase patterns were added in sentence patterns. They are often improper in phrase patterns and often cause improper phrase translation results. If improper pronouns are generated by the functions in phrase translation, these phrase translation results are probably selected by tri-gram. We should remove improper functions from phrase patterns. Moreover, selecting the phrase pattern with semantic code will be required.

5.2 Open-Test

Experimental Method. Three-hundred compound or complex sentences are used for pattern matching. If sentence pattern matching and phrase pattern matching succeed, translation is performed. The translation accuracy is assessed with the evaluation criteria in section 5.1.

Pattern Matching Results. Sixty-one sentences are matched to sentence patterns. In these sentences, 14 sentences are matched to phrase patterns and 47 are not.

In the results of pattern matching, the pattern matching rate is about 5%. The reason for this seems to be the difference between input sentence styles and sentence pattern styles. Pattern matching is great influenced by the expression at the end of a sentence.

Phrase Pattern Matching. The main reason for failed phrase pattern matching is that the matched phrase string is longer than we expected. Fig. 10 shows this example.

Input Sentence	大阪までの間のどこかで駅弁を買って食べよう。 [Oosaka made no aida no dokoka de ekiben wo ka tte tabe you。]
Japanese Pattern	<$N1$ は >!$VP2$(て | で)VP3.you。 [<$N1$wa>!$VP2$(te|de)VP3.you。]
Matched Morph.	VP2= Oosaka made no aida no dokoka de ekiben wo ka VP3 = tabe

Fig. 10. Examples of phrase pattern mismatch

In SPM-s, "大阪 まで の 間 の どこか で 駅弁 を 買っ [Oosaka made no aida no dokokade de ekiben wo ka]" is matched as VP. It is longer than our phrase patterns. If there were the Japanese pattern "$ADVP1$ の $VP2$ て $V3$.you [$ADVP1$ no $VP2$ te $V3$.you]", in the above sentence, "大阪 まで の 間 [Oosaka made no aida]" would be matched as $ADVP1$, "どこか で 駅弁 を 買っ [dokoka de ekiben wo katt]", would be matched as $VP2$, and "食べ [tabe]" would be matched as $V3$. Then, phrase pattern matching would succeed.

Sentence Pattern Matching. Similarly, if we add sentence patterns, the sentence pattern matching rate is improved. However, it is expensive to add sentence patterns. It is found out that the sentence pattern matching rate with clause-level patterns is 78% (234 sentences are matched to the same 300 sentences). To increase the sentence pattern matching rate, the clause-level pattern-based translation should be implemented.

Human Evaluation. Table 4 shows results.

Table 4. Evaluation results

	Eval.1	Eval.2	Eval.3
♯Sentence	1	9	4

The number in Eval. 2 is the largest, which is similar to the results of the closed-test. The main reason for the low translation accuracy is, similar to the closed-test, selection of the improper phrase translation.

6 Conclusion

We presented the phrase-level pattern-based translation system. Our method is based-on the Analogical Mapping Method. In the evaluation, the translation accuracy of phrase was still low, caused by improper phrase selection. Moreover, pattern matching rates of sentence and phrase are low.

For future work, we will resolve these problems of low translation accuracy. Also, we will implement the clause-level pattern-based translation to improve the low pattern matching rates.

References

1. Ikehara, S., Tokuhisa, M., Murakami, J.: Analogical Mapping Method and Semantic Categorization of Japanese Compound and Complex Sentence Patterns. In: Proceedings of the 10th Conference of the Pacific Association For Computational Linguistics, pp. 181–190 (2007)
2. Tokuhisa, M., Murakami, J., Ikehara, S.: Pattern Search by Structural Matching from Japanese Compound and Complex Sentence Pattern Dictionary. IPSJ SIG Technical Report, 2004-NL-176, pp. 9–16 (2006) (in Japanese)
3. Ikehara, S., Tokuhisa, M., Murakami, J., Saraki, M., Miyazaki, M., Ikeda, N.: Pattern Dictionary Development based on Non-Compositional Language Model for Japanese Compound and Complex Sentences. In: Matsumoto, Y., Sproat, R.W., Wong, K.-F., Zhang, M. (eds.) ICCPOL 2006. LNCS (LNAI), vol. 4285, pp. 509–519. Springer, Heidelberg (2006)
4. Tokuhisa, M., Endo, K., Kanazawa, Y., Murakami, J., Ikehara, S.: Evaluation of Pattern Generalization Effect under Development of Pattern Dictionary for Machine Translation. In: Pacific Association For Computational Linguistic, pp. 311–318 (2005)
5. Shapiro, S.C.: Generalized Augmented Transition Network Grammars For Generation From Semantic Networks. Computational Linguistics Archive 8(1), 12–25 (1982)
6. Ikehara, S., Miyazaki, M., Shirai, S., Yokoo, A., Nakaiwa, H., Ogura, K., Ooyama, Y., Hayashi, Y.: Goi-Taikei: A Japanese Lexicon. Iwanami Shoten (1997) (in Japanese)