

# Analysis of Parallel Structures in Patent Sentences, Focusing on the Head Words

Shoichi Yokoyama

Graduate School of Science and Engineering (Informatics), Yamagata University  
4-3-16 Jonan, Yonezawa, Yamagata 992-8510, Japan  
yokoyama@yz.yamagata-u.ac.jp

## Abstract

One of the characteristics of patent sentences is long, complicated modifications. A modification is identified by the presence of a head word in the modifier. We extracted head words with a high occurrence frequency from about 1 million patent sentences. Based on the results, we constructed a modifier correcting system using these head words. About 60% of the errors could be modified with our system.

## 1 Introduction

In patent sentences, the “problem to be solved” and “solution” parts have complicated and long modificational phrases.

In order to analyze this complicated modification structure, we have investigated parallel conjunctions and parallel particles (Yokoyama 2005, 2007, 2009, 2011).

Here, we first considered the center of a modified noun as the head. Based on this, we constructed a system which corrects errors of modification. About 60% of the errors could be modified with our system. This paper is mainly based on the work of (Yokoyama 2012) and (Sakamoto 2012).

## 2 Parallel Structures in Patent Sentences

The majority of the “problem to be solved” and/or “solution” sentences in patents are often very long and complicated. These sentences sometimes have parallel structures with long modificational phrases.

We investigated parallel conjunctions and parallel particles to clarify the structure of the modifiers, and constructed a system which corrects errors of modification.

Here, we focused on the head words. First, we will describe the characteristics of each part of speech.

### (a) Parallel conjunctions

Parallel conjunctions in Japanese include “mata wa” (or), “mosikuwa” (or), “oyobi” (and), and “narabini” (and).

Legal language places these conjunctions in a hierarchical structure to reduce and remove the ambiguity of law (Tajima 2006). However, following our investigations, no hierarchical structures were found in patent sentences (Yokoyama 2011).

### (b) Parallel particles

In Japanese, parallel particles include “to” (and), “ya” (and), and “ka” (or). Patent sentences are often written in the form “A to B to no...” (the...of A and B). Our system can effectively correct relatively simple structures utilizing this form, but not more complicated ones (Yokoyama 2007).

### (c) Head words

Head words are defined as the central nouns in modificational phrases such as “bu” (section) in the Japanese phrase “A-bu to, B-bu to...” (section A and section B ...), and “sou” (layer) in the phrase “C-sou oyobi D-sou de wa ...” (in layer C and layer D).

Fig. 1 shows an example abstract of a Japanese patent translated into English. In Fig. 1, the underlined and bold “section” words create parallel phrases in Japanese, but the underlined “section” words do not. In the Figure, these phrases are clearly separated by semicolons, but in Japanese they are connected by the coordinating conjunction “to” (and).

Here, we demonstrate that these head words clarify parallel structures.

(57)Abstract:  
 SOLUTION: The power line communication apparatus includes: an initial connection section for transmitting/receiving the device information between the other devices connected to the power line; a communicable party information creating section for creating the communicable party information of own, device based on the device information transmitted and received by the initial connection section; a communication-available party information transmitting/receiving section for transmitting and receiving the communication-available party information of own device, created by the communicable party information creating section between the other devices; a registration information creating section for creating the registration information where the other devices capable of communicating with own device are mapped; and a registration information transmitting/receiving section for transmitting and receiving the registration information created by the registration information creating section among other devices.

Fig.1 Patent language example (J2010-021954)

### 3 Materials and Methods

#### 3.1 Materials

We used a patent information database made by AAMT (Asia-Pacific Association for Machine Translation)/Japio (Japan Patent Information Organization) Special Interest Group on patent translation (AAMT/Japio 2004). This database includes all patent applications filed in 2004, which consists of 339,716 patents, and 1,013,582 sentences.

#### 3.2 Extraction and Classification of Head Words

All sentences are input into the common-use modification analyzer Cabocha (Cabocha 2012), and the analyzed results are output.

Fig. 2 shows an example of analysis by Cabocha (some parts omitted). In Fig. 2, the part of speech (for example, “noun, general”) is shown translated into English, and the English translation of each word is in parentheses. Numbers like “0 2D” show the modification. Here, it shows that the phrase “tahoubu ni” (in the other section) modifies “aze-keisei souti” (ridge-forming device).

Tahoubu ni tutimori-souti to aze-keisei souti to (filling device and ridge-forming device in other part)  
 \*0 2D 1/2 0.644114  
 tahou (other) noun, general,\*,\*,\*,\*,  
bu (part) noun, suffix, general,\*,\*,\*,  
 ni particle, case particle, general, \*,\*,\*,  
 \*1 2D 2/3 0.000000  
 tuti (mud) noun, general, \*,\*,\*,\*,  
 mori (filling) noun, proper noun, general, \*,\*,  
souti (device) noun, suru verb, \*,\*,\*,\*,  
 to (and) particle, parallel particle, \*,\*,\*,\*,  
 \*2-1D 2/3 0.000000  
 aze (ridge) noun, general, \*,\*,\*,\*,  
 keisei (forming) noun, suru verb, \*,\*,\*,\*,  
souti (device) noun, suru verb, \*,\*,\*,\*,  
 to (and) particle, parallel particle, \*,\*,\*,\*,  
 EOS

Fig. 2 Example analyzed by Cabocha

In Fig. 2, underlined “bu” (part) is a noun and suffix, and functions as a head word. The two underlined “souti” (device) words occurring in the second and third phrases can also be considered head nouns.

#### 3.3 Investigation of Head Words with High Occurrence Frequency

We investigated 1 million sentences. To identify parallel phrases (Yamamoto 1996, Iwamoto 1993), we used coordinate particles and parallel conjunctions such as “to”, “ya”, “ka”, “,”, “katu”, “oyobi”, “mata”, “narabini” (these translate into English as “and”), “aruiwa”, “mosikuwa” (or), and “dake de (wa) naku” (not only...but also...). We ignored numbers and words written in original text using as the number of devices.

Table 1 Examples of head nouns

Word (Jap.)	Eng.	POS	Occur. Freq.
syudan	means	n. gr.	23,523
souti	device	n. v.	19,906
koutei	process	n. gr.	12,305
houhou	method	n. gr.	10,683
zyouhou	information	n. gr.	7,229
ki	radical	n. gr.	6,579
de-ta	data	n. gr.	4,671
buzai	component	n. gr.	4,534
suteppu	step	n. gr.	3,967
iti	location	n. v.	3,674

Table 1 gives some common head words occurring frequently in the text we searched. The columns show the head words, their English translation, part of speech, and occurrence frequency. We used the top 100 words with an occurrence frequency higher than 412 for our system. In Table 1, “n. gr.” means “noun, general”, and “n. v.” means “noun, suru verb”.

### 3.4 Investigation of Occurrence Frequency in a Specific Field

International patents are categorized by technical content, that is, using IPC (International Patent Classification). They are classified within hierarchies such as section, subsection, class, subclass, main group, and subgroup.

Sections are divided into 8 fields: A (human necessities), B (performing operations; transporting), C (chemistry; metallurgy), D (textiles; paper), E (fixed constructions), F (mechanical engineering; lighting; heating; weapons; blasting), G (physics), and H (electricity) (WIPO 2013).

Table 2 shows the high frequency words in Section C (chemistry; metallurgy) (27,969 patents, 76,517 sentences). There, nouns such as “atom” and “acid” (which do not occur very frequently in search results from all fields) have relatively high frequency. In Table 2, “n. adv.” means nouns that can be adverbs.

Table 2 Examples of head nouns in Section C

Word (Jap.)	Eng.	POS	Occur. Freq.
ki	radical	n. gr.	3,729
ika	less than	n. adv.	1,941
koutei	process	n. gr.	1,844
zyusi	resin	n. gr.	1,654
houhou	method	n. gr.	1,613
genshi	atom	n. gr.	633
san	acid	n. gr.	580

## 4 Modification Correction System and Evaluation

### 4.1 Modification Correction System

Use of head nouns makes possible to deal with complicated telescopic modificational structures. We constructed a system to modify erroneous modification.

0 28D		mata, (and,)
1 2D		sono (its)
2 3D		tame no (for)
3 28D		seigyō reikyaku soutei wa (control cooling device)
4 5D		atuen tyokugo no (just after rolling)
5 6D		kouhan no (steel plate)
6 7D		men ondo bunpu wo (surface temperature distribution)
7 8D		sokutei suru (measure)
8 15D (8 17D)		ondosokutei [ <u>soutei</u> ] wo (temperature measuring device)
9 11D		reikyaku sui hedda- to (cool water header)
10 11D		kore ni (this)
11 12D		setuzoku sareta (connected)
12 13D		ramina-zyou no (laminar-formed)
13 14D		reikyaku sui wo (cool water)
14 15D		kyoukyu suru (supply)
15 16D		nozuru to wo (nozzle)
16 17D		hukumu (including)
17 23D (17 27D)		reikyaku <[ <u>soutei</u> ]> to (cooling device)
18 19D		syotei no (designated)
19 20D		keisan puroguramu ni (computer program)
20 23D		sitagatte (following)
21 22D		kouhan no (steel plate)
22 23D		men ondo bunpu wo (surface temperature distribution)
23 25D		kin'ituka suru you ni (in order to standardize)
24 26D		reikyaku suiryō wo (cooling water volume)
25 27D		seigyō suru (control)
26 27D		reikyaku suiryō no (cooling water volume)
27 28D		seigyō <[ <u>soutei</u> ]> to wo (controlled device)
28-1D		sonaeru (have)

Fig. 3 Example of correction by the system

Fig. 3 shows the correction result of the output of the system. The Japanese sentence we are using is the combination of every word in Fig. 3, and it is too complicated to translate it into English. Here, translation is only shown word by word.

In Fig. 3, the numbers show the phrase number, and numbers such as 2D, 3D show the phrase number modified. The correction shows parentheses such as (8 17D) and (17 27D). Head words relative to a modifier are shown by [ ], and heads relative to a modificand are shown by < >.

## 4.2 Evaluation

500 sentences randomly selected from patents in 2004, including parallel structures, are analyzed.

Table 3 shows the results. In Table 3, “C>C” shows that the analysis of modification is correct in the original Cabocha system, and the analysis by our system is also correct. “C>E” shows that the original analysis is correct, but our analysis is wrong. Conversely, “E>C” shows that the analysis of the original system is wrong, but our analysis can modify the result. “E>E” shows that the modification does not work.

We were able to modify 97 (58.4%) of 166 (97 + 69) erroneous sentences.

Table 3 Results of correction

	C>C	C>E	E>C	E>E	Total
Sent.	318	16	97	69	500
%	63.6	3.2	19.4	13.8	100

## 5. Concluding Remarks

In this paper, we described our experiment to correct erroneous analysis of modification. However, the correction was not as successful as expected; one reason is that the number of head words was restricted to words with a high occurrence frequency.

Next we plan to increase the number of words. We are also planning to use a thesaurus, and focus on numbers and symbols just after head words.

## Acknowledgements

We thank Japio and the committee members for supporting this research and supplying the patent database.

## References

AAMT/Japio, 2004. AAMT (Asia-Pacific Association for Machine Translation)/Japio (Japan Patent Information Organization) Special Interest Group on Patent Translation Patent Database.

Cabocha, 2012. Nara Advanced Institute of Science and Technology <http://code.google.com/p/cabocha/>

Hideaki Iwamoto, Kaoru Nagano, Hidetoshi Nagai, Teigo Nakamura, and Hirosato Nomur. 2004a, 1993: An Analysis of Coordinate Structures and its Application to A Controlled Linguistic Model for Law Sentences, *Special Interest Group of Natural Language Processing in Information Processing Society Japan* (in Japanese) NL98-3

Nobutake Tajima, 2006. Basic Knowledge about Law Words (Third Edition) (in Japanese), Gyosei, Tokyo.

Kazuma Sakamoto, 2012. Analysis of Parallel Structures of Patent Sentences Focusing on the Head Words, *Bachelor Thesis of Department of Informatics, Faculty of Engineering, Yamagata University* (in Japanese).

WIPO (World Intellectual Patent Organization), 2013. IPC (international Patent Classification) <http://web2.wipo.int/ipcpub/#refresh=page>

Hiroomi Yamamura, Akira Suganuma, Kazuo Ushijima, 1996. A Simple Method to Analyze Coordinate Structures in a Japanese Document, and its Application to a Writing Tool, *Proceedings of 52<sup>nd</sup> Annual Meetings of Information Processing Society Japan* (in Japanese) 2J-1, pp.3-277-278.

Shoichi Yokoyama and Yuya Kaneda, 2005. Classification of Modified Relationships in Japanese Patent Sentences, *Proceedings of Workshop on Patent Translation*, pp.16-20.

Shoichi Yokoyama and Shigehiro Kennendai, 2007. Error Correcting System for Analysis of Japanese Patent Sentences, *Proceedings of Second Workshop on Patent Translation*, pp.24-27.

Shoichi Yokoyama and Masumi Okuyama, 2009. Translation Disambiguation of Patent Sentence using Case Frames, *Proceedings of Third Workshop on Patent Translation*.

Shoichi Yokoyama and Yuichi Takano, 2011. Investigation for Translation Disambiguation of Verbs in Patent Sentences using Word Grouping, *Proceedings of Fourth Workshop on Patent Translation*, pp.60-63.

Shoichi Yokoyama, 2012. Analysis of Parallel Structures of Patent Sentences Focusing on the Head Words, *Japio Year Book 2012* (in Japanese) pp.250-253.