# Patent Claim Processing for Readability

## - Structure Analysis and Term Explanation -

**Akihiro SHINMORI**
Department of Computational Intelligence and Systems Sciences, Tokyo Institute of Technology, and INTEC Web and Genome Informatics Co.
shinmori@isl.intec.co.jp

**Manabu OKUMURA**
Precision and Intelligence Laboratory, Tokyo Institute of Technology
oku@pi.titech.ac.jp

**Yuzo MARUKAWA**
Japan Science and Technology Corp., and National Institute of Informatics
maru@nii.ac.jp

**Makoto IWAYAMA**
Precision and Intelligence Laboratory, Tokyo Institute of Technology, and Hitachi, Ltd.
iwayama@pi.titech.ac.jp

### Abstract

Patent corpus processing should be centered around patent claim processing because claims are the most important part in patent specifications. It is common that claims written in Japanese are described in one sentence with peculiar style and wording and are difficult to understand for ordinary people. The peculiarity is caused by structural complexity of the sentences and many difficult terms used in the description. We have already proposed a framework to represent the structure of patent claims and a method to automatically analyze it. We are currently investigating a method to clarify terms in patent claims and to find the explanatory portions from the detailed description part of the patent specifications. Through both approaches, we believe we can improve readability of patent claims.

## 1 Introduction

The importance of intellectual property, specifically patent, is being recognized more than ever. In the academia, patent is being considered as the core component for technology transfer to industry. With the upsurge of business method patents and software patents, more and more business persons are concerned about patent.

Patent is described in patent specification which is a kind of legal documents. The most important part of patent specification is where the claims are written, because "the claims specify the boundaries of the legal monopoly created by the patent" (Burgunder, 1995). Therefore, we believe that patent corpus processing should be centered around patent claim processing.

It is common that Japanese patent claims are described in one sentence with peculiar style and wording and that they are difficult to read and understand for ordinary people. After surveying related literature and investigating NTCIR3 patent collection (Iwayama et al., 2003), we found the difficulty has two aspects: structural difficulty and term difficulty.

In this paper, we first present the characteristics of patent claims. Next, we present our work on the structure analysis of patent claims. Third, we introduce our on-going research on term explanation for patent claims.

## 2 Characteristics of Patent Claim

Typical Japanese patent claims taken from two patents are shown in Figure 1 and 2.

In general, Japanese sentences are inserted with the touten "、" or ", " (comma) and end with the kuten "。" or ". " (period) . The touten plays a role of segmenting the sentence for disambiguating the meaning and for improving readability. According to the literature (Maekawa, 1995), the average length of Japanese sentences is 55.85 characters in newspaper articles on politics and 75.37 characters on social affairs articles.

The claims of Figure 1 and 2 are both written in one sentence. Though they are appropriately in-

操作手段によりアクチュエータを駆動して所望の作業を行なう作業機において、前記作業機の作業機構に作用する負荷を検出する 負荷検出手段 と、この 負荷検出手段 の検出値に応じた周波数の信号を出力する 第１の周波数変換器 と、当該負荷検出手段の検出値に応じた周波数のパルスを出力する 第２の周波数変換器 と、前記 第１の周波数変換器 から出力される信号を前記 第２の周波数変換器 からのパルスの出力期間だけ間欠的に出力する変調手段と、この変調手段の出力信号に応じて振動を発生する振動発生手段とを設けたことを特徴とする作業機の操作用仮想振動生成装置。

Figure 1: A sample Japanese patent claim (Publication Number=10-011111)

印刷ヘッドがヘッドキャップに密着した状態の経過時間を計測する時間計測手段と、<nl>
前記時間計測手段の計測結果に基づいて、前記印刷ヘッドの増粘インク除去処理を、定期的に実行する増粘インク除去手段と、を備えたことを特徴とするインクジェットプリンタ。

Figure 2: A sample Japanese patent claim containing a newline (Publication Number=10-146993) (Note: <nl> means a newline.)

serted with the touten “、”, they are unusually long with the length of 295 characters and 119 characters. It is definitely true that most Japanese who are not accustomed to reading patent claims have difficulty in reading them. In fact, according to (Kasuya, 1999), Japanese patent attorneys themselves recognize that Japanese patent claims are difficult to read.

The salient characteristics of Japanese patent claims from the viewpoint of readability are as follows:

1. The length of sentence is long.
2. The structure of description is complex.
3. There are several terms which are difficult to understand or requires explanation for understanding.

To examine the first point, we extracted all of the first claims of the sample data (59,968 patents) in the NTCIR3 patent collection, and calculated the average sentence length. We found that it is 242 characters and confirmed that Japanese patent claims are unusually long.

With regard to the second point, we surveyed several books and articles written for patent applicants to explain how to draft patent claims(Kasai, 1999; Kasuya, 1999) and how to translate patent claims(Lise, 2002).

Based on the survey, we classify the description style into the following three. [Note: In the following explanation, Japanese phrases are followed by their literal expression in [] and their English translation in (). ]

**Process sequence style** As in “... し [shi](does), ... し [shi](does), ... した [shita] (and does)...”, the sequence of processes is described. Mainly used in method inventions.

**Element enumeration style** As in “... と [to](and), ... と [to](and), ... とからなる [to kara naru](comprising), ...”, the set of element is described. Mainly used in product inventions.

**Jepson-like style** As in “... において [ni oite](in), ... を特徴とする [wo tokuchou to suru](be characterized by), ...”, the description consists of the first half part and the last half part. In the first half part, either the known or the precondition part is described. In the last half part, either the new or the main part is described [1].

These patterns are not mutually exclusive. For example, the first half part of the Jepson-like style may be written in the process sequence style or in the element enumeration style.

With regard to the third point, Figure 1 contains the term “アクチュエータ”(an actuator) and Figure 2 contains the term “増粘インク”(sticky ink) which require explanation for understanding.

Because of these characteristics, the well-known Japanese parser KNP (Kurohashi, 2000) incorrectly analyze or cannot process most of the Japanese patent claims.

KNP’s dependency analysis works by detecting parallel structure utilizing thesaurus and dynamic programming, but it does not work well for patent

[1]Note that the term “Jepson claim” is rigidly defined and used in Europe or in the USA to describe the kind of claims in which the known part and the new part are clearly separated. In Japan, that is not common and the separation is more vague(Lise, 2002). That’s why we name this as “Jepson-like style”.

Table 1: Relations for Japanese patent claims

| Type | Relation | Explanation | Example |
|---|---|---|---|
| Multi-Nuclear | PROCEDURE | Process Sequence Style | [～し、][～し、][～する]X<br>[Note: The above means "X which [does ～,] [does ～,] [and does ～]."] |
| Multi-Nuclear | COMPONENT | Element Enumeration Style | [～と、][～と、][～と] を<br>[Note: The above means "[～,] [～,] [and ～]."] |
| Mono-Nuclear | ELABORATION | S elaborates N. | [X を Y した][Z の A]<br>[Note: The above means "[A of Z] [which Y X]."] |
| Mono-Nuclear | FEATURE | Characterization | [X である Y][を特徴とする]<br>[Note: The above means "[characterized by] [Y which is X]."]. |
| Mono-Nuclear | PRECONDITION | Jepson-like Style | [X であって、][Y した Z]<br>[Note: The above means "[In X,] [Z which Y].". |
| Mono-Nuclear | COMPOSE | Composition | [～と、～と、～と][を備えた]X<br>[Note: The above means "X [composed of] [～, ～, and ～]."]. |

claims because they often include "chain expressions" in which one concept is first defined and next another concept is defined using the first. For the claim in Figure 1, although "負荷検出手段" (a load detection method), "第 1 の周波数変換器" (a frequency transfer device no.1), "第 2 の周波数変換器" (a frequency transfer device no.2), "変調手段" (a modulation method), and "振動発生手段" (an oscillation generation method) need to be recognized as parallel, it cannot be recognized due to the existence of the expressions designated by the underline.

## 3 Structure Analysis of Patent Claims

### 3.1 Background

To improve readability of Japanese patent claims, we claim that the structure of description needs to be presented in a readable way. To do so, the structure needs to be analyzed first.

Japanese patent claims are described in such a way that multiple sentences are coerced into one sentence(Kasuya, 1999). In other words, a claim is composed of multiple sentences that have some kind of relationships with each other. Therefore, we decided to apply the RST (Rhetorical Structure Theory) (Mann, 1999) that was proposed to analyze discourse structure composed of multiple sentences.

RST was proposed in the 1980's and has been successfully applied to automatic summarization (Marcu, 2000), automatic layout (John Bateman, 2000), and so on. A Tcl/Tk-based interactive tool (OD'onnel, 1997) was developed to support to manually edit and to visually show the structure.

### 3.2 Framework

For the structure analysis of Japanese patent claims, we defined six relations as in Table 1. Two of them are multi-nuclear where composing elements are equally important. Four of them are mono-nuclear where one element is nucleus, the other is satellite, and the nucleus is more important than the satellite. In the "Example" column of Table 1, the regions enclosed with "[" and "]" are segments or spans and the underlined ones are nuclei.

Given the patent claims in Figure 1 and Figure 2, we can analyze their structure and present them visually by using RSTTool (OD'onnel, 1997) as in Figure 3 and Figure 4 [2].

### 3.3 Cue-phrase-based Approach

In designing the algorithm, we took a similar approach to (Marcu, 2000). We collected cue phrases that can be used for segmenting long claims and establishing relations among segments or spans.

[2]Because RSTTool is written in Tcl/Tk and Tcl/Tk is an internationalized language, we did not have to localize it to display Japanese characters.

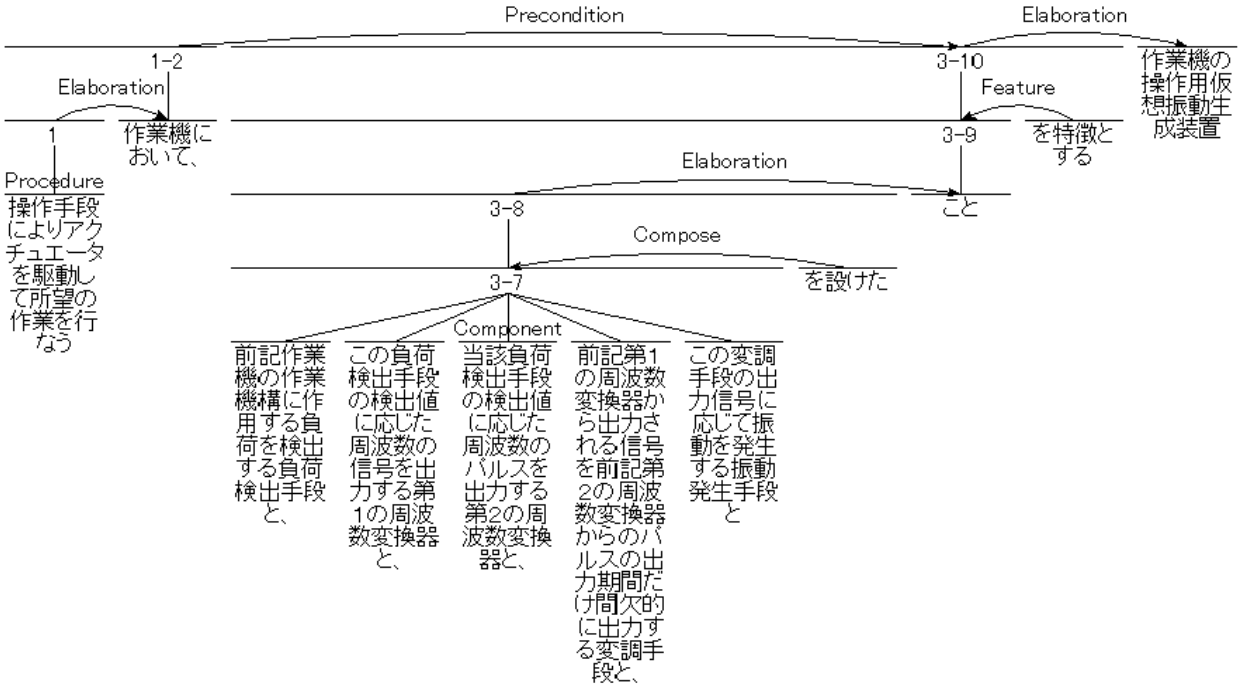


Figure 3: A result of structure analysis of patent claim in Figure 1 (using RSTTool v2.7)

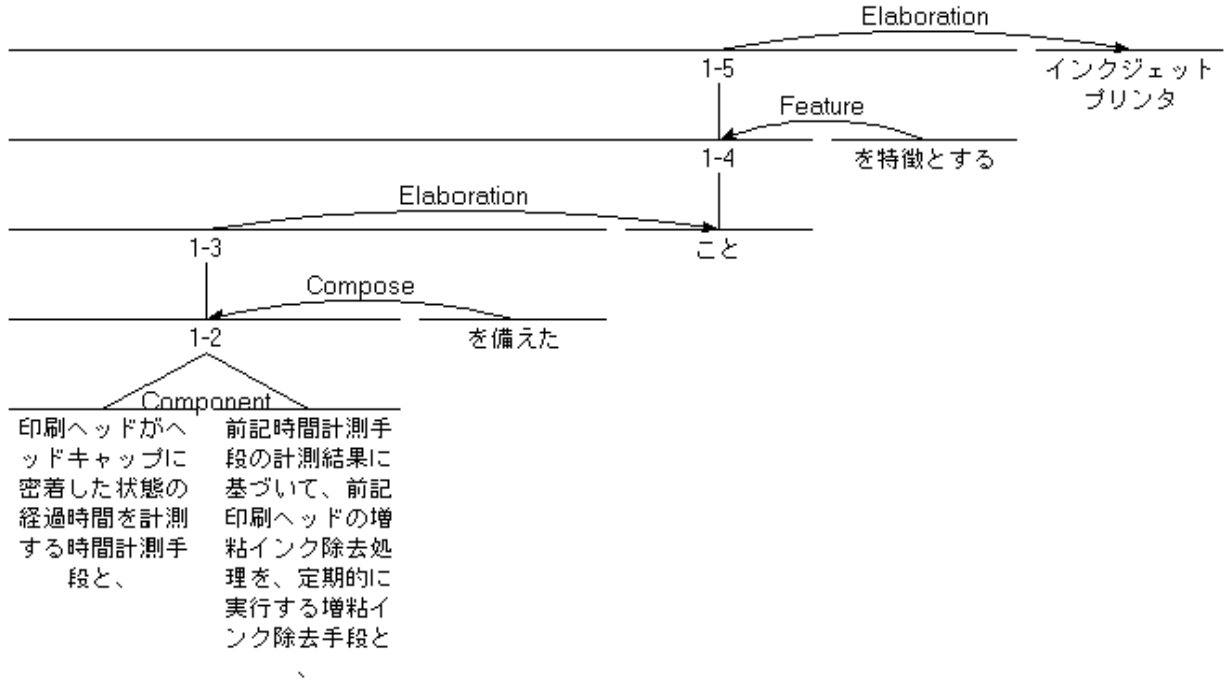


Figure 4: A result of structure analysis of patent claim in Figure 2 (using RSTTool v2.7)

Table 2: Description pattern just before the newlines in claims in which newline are explicitly inserted

| No | Pattern | Ratio |
|---|---|---|
| 1 | (Noun\|Symbol) と (、\|，) [Note: “と” means “and”.] | 46.1% |
| 2 | (Verb-Cont-Form\| AuxVerb-Cont-Form)(、\|，) | 17.5% |
| 3 | (Noun\|Symbol) において (、\|，) [Note: “において” means “in”.] | 16.4% |
| 4 | (Noun\|Symbol) であって (、\|，) [Note: “であって” means “in”.] | 7.2% |

Cue phrases were first collected manually by reading patent claims. Then we found that about half of the claims are inserted with newlines at seemingly segment boundaries as in Figure 2.

We investigated all of the extracted first claims of the sample data and 48.5% of them are newline-inserted claims. It seems that the drafters of patent claims explicitly inserted those newlines for readability for themselves. We checked the description pattern of the last three morphemes just before each newline of those claims. The result is shown in Table 2. In Table 2, “Verb-Cont-Form” means “動詞連用形” (verb in continuous form) and “AuxVerb-Cont-Form” means “助動詞連用形” (auxiliary verb in continuous form). Note that the description patterns are expressed in the regular expression notation of Perl.

Summarizing the above, we came up with the cue phrases in Table 3. In Table 3, “Verb-Basic-Form” means “動詞基本形” (verb in basic form) and “AuxVerb-Basic-Form” means “助動詞基本形” (auxiliary verb in basic form).

### 3.4 Algorithm and Implementation

We designed an algorithm for analyzing structure of independent claims[3]. Although patent claims are written in natural language, it’s not written in a free form and is restricted in a sense that there are description styles established in the community. So, we designed an algorithm composed of a lexical analyzer and a parser as in the formal language processors.

[3]Independent claims are claims which do not refer to any other claims.

First, the input claim is analyzed with the morphological analyzer “chasen” (Matsumoto et al., 2002). Because some patent claims explicitly contain newlines as in Figure 2, we use the “-j” option setting the sentence delimiter as “。:; ” in “.chasenrc”.

Next, the output from chasen is analyzed with the lexical analyzer. The main point of our algorithm is the context-dependent behavior of the lexical analyzer as follows:

- The lexical analyzer outputs two types of token: cue phrase token and morpheme token.
- Outputting morpheme tokens is done depending on some contextual conditions to avoid ambiguities in the parsing.
- For other morphemes whose context did not satisfy the above conditions, an anonymous morpheme token (WORD) is output.

Next, the output from the lexical analyzer is processed with the parser generated from a context-free grammar (CFG) by using Bison (Donnelly and Stallman, 1995)-compatible parser generator. The CFG we designed for Japanese patent claim consists of 57 rules, 11 terminals, and 19 non-terminals.

Finally, a structure tree is constructed in the form of “.rs2” file used in RSTTool v2.7. By using RSTTool, the output is visually displayed as in Figure 3 and Figure 4.

### 3.5 Evaluation

The evaluation was done by using the first claims [4] of 59,956 patents extracted from the NTCIR3 patent data collection.

The NTCIR3 patent data collection consists of 697,262 patents opened to public in 1998 and in 1999. For the analysis, the collection of cue phrases, and the creation of the CFG, we used patents in 1998. For the evaluation, we used patents in 1999.

We checked the IPC (International Patent Classification) code of 59,956 patents and confirmed that the distribution is similar to the one of all opened patents in 1999 disclosed by JPO (Japan Patent Office).

The evaluation was done in the following points:

[4]First claims are always independent claims.

Table 3: Cue phrases which can be used to analyze patent claims

| Token Name | Cue Phrase | Gloss |
|---|---|---|
| JEPSON_CUE | に (お \| 於) いて (、\|，)<br>であって (、\|，)<br>にあたり (、\|，)<br>に当 (た)?り (、\|，) | [ni oite] (in)<br>[de atte] (in)<br>[ni atari] (in)<br>[ni atari] (in) |
| FEATURE_CUE | を特徴と (した \| する)(、\|，)? | [wo tokuchou to (shita\|suru)]<br>(characterized by) |
| COMPOSE_CUE | を搭載して構成され (た \| る \| ている) (、\|，)? | [wo tousaishite kousei sare (ta\|ru\|teiru)]<br>(comprising) |
| | を (、\|，)?(具 \| 備 \| そな) え (た \| る \| ている) (、\|，)? | [wo sonae (ta\|ru\|teiru)]<br>(comprising) |
| | を (、\|，)?具備 (した \| する \| している \| してなる) (、\|，)? | [wo gubi (shita\|suru\|shiteiru\|shitenaru)]<br>(comprising) |
| | (で \| から) 構成され (た \| ている) (、\|，)? | [(de\|kara) kousei sare (ta\|teiru)]<br>(comprising) |
| | を (、\|，)?有 (する \| した) (、\|，)? | [wo yuu (suru\|shita)]<br>(comprising) |
| | を (、\|，)?包含 (する \| した) (、\|，)? | [wo hougan (suru\|shita)]<br>(comprising) |
| | を (、\|，)?含 (む \| んだ) (、\|，)? | [wo fuku (mu\|nda)]<br>(comprising) |
| | から (、\|，)?(なる \| なった \| なっている) (、\|，)? | [kara (naru\|natta\|natteiru)]<br>(comprising) |
| | から (、\|，)?(成る \| 成った \| 成っている) (、\|，)? | [kara (naru\|natta\|natteiru)]<br>(comprising) |
| | を (、\|，)?設け (た \| ている) (、\|，)? | [wo mouke (ta\|teiru)]<br>(comprising) |
| | を (、\|，)?装備 (する \| した \| している) (、\|，)? | [wo soubi (suru\|shita\|shiteiru)]<br>(comprising) |
| NOUN<br>POSTP_TO<br>PUNCT_TOUTEN | The sequence of “(Noun\|Symbol) と (、\|，)” | |
| VERB_RENYOU<br>PUNCT_TOUTEN | The sequence of<br>“(Verb-Cont-Form\|AuxVerb-Cont-Form)(、\|，)”<br>which exist before<br>“(Verb-Basic-Form\|AuxVerb-Basic-Form)<br>(Noun\|Symbol)” | |

**Accept Ratio** The ratio of claims accepted by the parser generated by the CFG.

**Processing Speed** The time required to process one claim.

**Accuracy** The accuracy of the analysis result evaluated indirectly and directly.

The accept ratio was more than 99.77%. The processing speed was 0.30 second per each claim (evaluated on a Linux PC using Pentium III 1GHz and 512MB memory). So, it is almost real-time.

#### 3.5.1 Indirect Evaluation on Accuracy

By specifying a command-line switch, our program can be run without utilizing the originally inserted newlines. The newline insertion positions can be predicted by the result of structure analysis and some heuristics. So, indirect evaluation was done by comparing the newline insertion positions between the originally newline-inserted claims and the automatically newline-inserted claims utilizing the result of structure analysis. The recall(R), the precision(P), and the F-measure(F) are calculated by the followings, where $c$ is the number of correctly-inserted newlines, $n$ is the number of newlines in the original claim, and $i$ is the number of inserted newlines.

$$R = \frac{c}{n} \quad (1)$$

$$P = \frac{c}{i} \quad (2)$$

$$F = \frac{2 * R * P}{R + P} \quad (3)$$

The baseline was set in that the newlines are inserted mechanically at the end of every sequence of "(NOUN|SYMBOL)(、|，)" and "(Verb-Cont-Form|AuxVerb-Cont-Form)(、|，)".

Note that newlines are sometimes inserted at the positions that are not segment boundaries in the meaning of RST. For example, it is often the case that at the end of "は、" (a postpositional particle representing the subject), newlines are inserted. So, our newline-insertion prediction algorithm has the inherent upper limit whose recall is 0.873.

The result is shown in Table 4.

Table 4: Evaluation result (Indirect)

| Index | Baseline | Newline Insertion utilizing RST | Upper Limit |
|---|---|---|---|
| Recall(R) | 0.478 | 0.674 | 0.8736 |
| Precision(P) | 0.374 | 0.663 | N/A |
| F-measure | 0.420 | 0.669 | N/A |

Table 5: Evaluation result (Direct)

| Category | Count | Percentage (Except "No judgment") |
|---|---|---|
| Correct | 76 | 80.85% |
| Partially Correct | 11 | 11.70% |
| Incorrect | 7 | 7.45% |
| No judgment | 6 | - |

#### 3.5.2 Direct Evaluation on Accuracy

The direct evaluation on accuracy was done by using randomly selected 100 claims extracted. All of these claims are the first claims. Again, we checked the distribution of IPC and confirmed it's similar to the one of all opened patents in 1999 disclosed by JPO.

The 100 claims were analyzed by our program and the visually-displayed outputs like Figure 3 and 4 were presented to a subject who had some experience in reading patent specifications. The subject evaluated the result by the following criteria:

- when the claim is in the Jepson-like style, whether that is correctly recognized.
- when the claim is in the Jepson-like style, whether the structure is correctly analyzed for the first half part and for the last half part.
- when the claim is not in the Jepson-like style, whether the structure is correctly analyzed for the whole.

The result is shown in Table 5.

### 3.6 Application to Patent Claim Paraphrase

Once the structure of patent claims are analyzed, we can apply the result to paraphrase patent claims.

To do so, the following actions are incorporated into the lexical analyzer and the parser.

- The lexical analyzer deletes the words “前記” (the), “該” (the), and “上記” (the).
- For the parser, new actions are added which relocates the “noun group” located at the end to the front. Same thing for the “noun group” located just before JEPSON_CUE for the Jepson-like style claims.
- For the process sequence style, the lexical analyzer conjugates verbs and adverbs from their continuous form to basic form and replaces the touten “(、|，)” with the kuten “。”.
- For the element enumeration style, the lexical analyzer converts those cue phrases such as “からなる”(consist of) and “を有する” (include) to their “テイル形”(“teiru” form) plus “。” and deletes “と (、|，)” (and) at the end of each element.
- The lexical analyzer converts “こと”(thing) just before “を特徴とする”(characterized by) to “以下”(the following).
- For the Jepson-like style, the parser separates the first-half part and the last-half part by inserting a newline.

By doing the above processing, long patent claim sentences are divided into multiple sentences. But as there are cases where some of the generated sentences are still too long, those sentences longer than the threshold length (75 characters) are recursively processed.

An example of paraphrase is shown in Figure 5.

We believe that paraphrasing can not only improve readability of patent claims but also can work effectively as a preprocessing for machine translation [5].

[5]In fact, there are several commercial machine translation software which does special preprocessing for patent claims before translating from Japanese to English.

操作手段によりアクチュエータを駆動して所望の作業を行なう作業機。

以下を設けていることを特徴とする作業機の操作用仮想振動生成装置:
　作業機の作業機構に作用する負荷を検出する負荷検出手段
　この負荷検出手段の検出値に応じた周波数の信号を出力する第１の周波数変換器
　当該負荷検出手段の検出値に応じた周波数のパルスを出力する第２の周波数変換器
　第１の周波数変換器から出力される信号を第２の周波数変換器からのパルスの出力期間だけ間欠的に出力する変調手段
　この変調手段の出力信号に応じて振動を発生する振動発生手段

Figure 5: A sample paraphrase for Figure 1

## 4 Term Explanation for Patent Claims

### 4.1 Background and Motivation

Once the structure of patent claims are analyzed and presented visually, next hurdle for readability is terms.

There are many novel terms used in patent claim description. They can be classified into the following categories:

**Terms specific to the invention** Patent drafters often assign unique names to the invention, its elements, and its processes for their identification.

**Terms specific to the domain** The patent law requires patents should be written so that those who have ordinary knowledge in the domain can understand and perform the invention. So, technical terms that are established in the domain are often used. Additionally, there exist “patent jargons” which are created by combining two kanji characters such as “嵌挿” (put and insert) and “枢着” (put into the hall)(Kasai, 1999). They are first created by some patent drafters for the sake of brevity and have been widely used in the community. So, they are terms specific to the inventions of the domain. Those who do not have enough knowledge in the domain or those who are not accustomed to reading patent specifications have difficulty in understanding them.

Giving appropriate explanations for these terms would help to improve readability of patent claims.

### 4.2 Approach

First of all, it is necessary to recognize terms to be explained. There are many research issues in term extraction in general, but for our purpose we use the following morphological pattern to extract terms from patent claims:

> (Prefix)*(Noun|Unknown-Words|Symbol |Verb-Cont-Form|Verb-Compound-With-Indeclinable-Word)+

By using the above pattern, we can extract such terms as “熱風吹き付け手段” (method to blow heat wind), “読取値” (read value), and “液滴” (liquid drop) which contain verbs.

Second, by using the result of structure analysis, we can infer the categories of the terms as follows:

- If the term appears at the end of the claim or just before the JEPSON_CUE in the Jepson-like style, or just before “と” (and) in the element enumeration style, it is a term specific to the invention. For example, “操作用仮想振動生成装置” (an operational virtual oscillation generating device) and “負荷検出手段”(a load detection method) in Figure 1 are terms specific to this invention.

- If the term appears in the middle of the first half in the Jepson-like style, it can be a term specific to the domain. For example, “アクチュエータ”(an actuator) in Figure 1 is a technical term in the domain.

- If the term is a two-kanji character and is not listed in the ordinary dictionaries, it can be a patent jargon.

Finally, by looking at the detailed description of the invention or related inventions, we can back up the above inference as follows:

- The terms specific to the invention should be described after the “means to solve the problem” section in the detailed description of the invention.

- The terms specific to the domain are widely used in the inventions of the domain. So, it is highly possible that they occur frequently in the related inventions. We can consider the collection of search result as the related inventions.

- Some of the technical terms specific to the domain are described in the “prior art” section of the detailed description of the invention or related inventions in the domain.

For those technical terms specific to the domain, explanatory portions such as the following can be found:

- “... これに対応する油圧シリンダ等（アクチュエータ）をその操作量に応じた速度で駆動し、...”
  (... driving the oil pressure cylinder (or the actuator) at the speed of ...)

- “... 吐出口（オリフィス）...”
  (... the spout (or the orifice) ...)

- “... インクの予備的噴出（つまりインクパージ）...”
  (... blowing out ink preliminarily (namely, purging ink) ...”

- “... ホットメルトタイプのインク（固形インク）...”
  (... ink of the hot-melt type (or solid ink) ...

As can be seen in the above, explanatory portions can be found by using cue phrases such as “（” and “）”, “以下” (“in the following”), and “つまり” (“or” or “namely”).

### 4.3 Sample Scenario

From the patent claim in Figure 2, we find many terms that are candidates for explanation such as “時間計測” (time measurement), “時間計測手段” (the method to measure time), “計測結果” (measurement result), “増粘インク” (sticky ink), “増粘インク除去” (removal of sticky ink), “増粘インク除去処理” (removal processing of sticky ink), “増粘インク除去手段” (the method to remove sticky ink).

Among the above terms, “時間計測手段” (the method to measure time) and “増粘インク除去手段” (the method to remove sticky ink) are terms specific to the invention because they are judged as the elements by structure analysis.

By searching the detailed description, we can find the explanatory portion for “増粘インク” (sticky ink) as follows.

- "... インクの粘度が増加すること（以下「増粘インク」という）..."
  (... the ink of increased stickiness (in the following, we call it as "sticky ink" ...)

### 4.4 Further Analysis and Experimentation

We continue to analyze the NTCIR3 patent data collection, specifically "Patolis Test Collection" which is a test collection for patent retrieval consisting of a set of query and search result. We use each search result as "related inventions" and analyze them to collect cue phrases for finding explanatory portions for technical terms specific to the domain.

## 5 Related Work

A NLP research for patent claim is already reported in (Kameda, 1995). It is directed toward dependency analysis of patent claims. Although it is proposed to support "analytic reading" of patent claims, the evaluation result for large-scale real patent data is not reported. Our approach is different from (Kameda, 1995) in that the top-level structure is analyzed.

In (Sheremetyeva and Nirenburg, 1996), a research on a system for authoring patent claims using NLP and knowledge engineering technique is reported.

## 6 Concluding Remarks

We have presented a framework to represent the structure of patent claims and a method to automatically analyze it. The evaluation result suggest that our approach is robust and practical.

We are currently investigating a method to clarify terms in patent claims and to find the explanatory portions from the detailed description part of the patent specifications.

It is not only a step toward improving readability, but it can also lead to more challenging task of automatic patent map generation(Study group on patent map, 1990).

## Acknowledgements

The NTCIR3 patent data collection was used in our research.

## References

Lee B. Burgunder. 1995. *Legal Aspects of Managing Technology*. South Western.

Charles Donnelly and Richard Stallman, 1995. *Bison: The YACC-compatible Parser Generator, Version 1.25*.

Makoto Iwayama, Atsushi Fujii, Akihiko Takano, and Noriko Kando. 2003. Overview of patent retrieval task at ntcir-3. In *The Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*. National Institute of Informatics.

Jorg Kleinz Klaus Reichenberger John Bateman, Thomas Kamps. 2000. Toward constructive text, diagram, and layout generation for information presentation. *Computational Linguistics*, 27(3):409–449.

Masayuki Kameda. 1995. Support functions for reading japanese text. In *IPSJ SIGNotes Natural Language*, number 110. Information Processing Society of Japan. (in Japanese).

Yasuji Kasai. 1999. *Manual for Drafting Patent Claims*. Kougyo Chosakai. (in Japanese).

Youji Kasuya. 1999. On the description style of patent claims and the techniques to draft them. *Patent*, 52(2). (in Japanese).

Sadao Kurohashi. 2000. KNP - japanese parsing for real. *IPSJ MAGAZINE*, 41(11). (in Japanese).

William Lise. 2002. An investigation of terminology and syntax in japanese and us patents and the implications for the patent translator. http://www.lise.jp/patsur.html.

Mamoru Maekawa. 1995. *Science of Sentences*. Iwanama. (in Japanese).

Bill Mann. 1999. An introduction to rhetorical structure theory (RST). http://www.sil.org/ mannb/rst/rintro99.htm.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka, and Masayuki Asahara, 2002. *Morphological Analysis System ChaSen version 2.2.9 Manual*. Nara Institute of Science and Technology.

Michael OD'onnel. 1997. RST-Tool: An RST analysis tool. In *The 6th European Workshop on Natural Language Generation*.

Svelana Sheremetyeva and Sergey Nirenburg. 1996. Knowledge elicitation for authoring patent claims. *IEEE Computer*, 57–63.

Study group on patent map, editor. 1990. *Patent Map and Information Strategy*. Japan Institute of Invention and Innovation. (in Japanese).