手がかり句を用いた特許請求項の構造解析

新森昭宏[†1] 奥村学[†2]
丸川雄三[†3] 岩山真[†4]

特許の内容を記述する「特許明細書」において最も重要な箇所は,「特許請求項」(クレーム)の部分である.日本語の特許請求項の多くは,1文で発明内容を記述するという制約と,その独特の記述形式により,専門家以外の人にとってはきわめて読みにくいものになっている.本稿では,特許請求項の可読性を向上させることを目的とした,構造解析手法を提案する.まず,文献調査と予備調査の結果に基づいて,特許請求項の構造を表現するための枠組みを提案する.次に,この枠組みに基づいて,特許請求項の構造を自動解析する手法を提案する.自動解析においては,実際の特許請求項記述で多用される,いくつかの定型的表現を手がかりとして利用する.大規模テキストコレクションであるNTCIR3特許データコレクションのデータを用いて提案手法の評価を行い,間接評価と直接評価により本手法の有効性を示した.

Structure Analysis of Japanese Patent Claims Using Cue Phrases

Akihiro Shinmori,[†1] Manabu Okumura,[†2] Yuzo Marukawa[†3]
and Makoto Iwayama[†4]

The most important part of patent specification documents is where the claims are written. It is common that claims written in Japanese are described in one sentence with peculiar style and are difficult to understand for ordinary people. In this paper, we propose a method to analyze the structure of patent claims for the purpose of improving readability. First, based on the result of literature survey and preliminary study, we propose a framework to represent the structure of patent claims. Next, based on this framework we propose a method to automatically analyze the structure of patent claims. This method utilizes several cue phrases which are used in patent claim descriptions. By using the NTCIR3 patent data collection, we show the effectiveness of our proposal through indirect and direct evaluation.

1. はじめに

かつて特許は,機械・製薬・化学など特定分野の企業において,知的財産権担当者や研究者が主に関わるものであった.しかし,ビジネスやサービスの方法を権利の対象とする「ビジネスモデル特許」の出現や,コンピュータプログラムを対象とした「ソフトウエア特許」の認知により,広い範囲の企業関係者が特許に関わらざるをえない状況が生まれている.大学においても,研究成果や技術移転のコアとしての特許の重要性が認識されるようになった.

特許行政は,いち早く電子化が推進された分野であり,過去に出願された特許明細書の電子データは膨大に蓄積されている.さらに,年間40万件以上といわれる新規出願にともない,データ量は日々増加している.こうした膨大な特許データを対象とした研究は従来,検索に関するものがほとんどであった.

特許は一種の法的文書であり,その内容を記述する「特許明細書」は,独特の記述形式を持っている.最も重要な箇所である「特許請求項」(クレーム)は,1文で発明内容を記述するという制約と,独特の記述スタイルにより,専門家以外の人にとってはきわめて読みにくいものになっているのが通例である[1),☆].

以上を背景として,本稿では,特許請求項の可読性

†1 インテック・ウェブ・アンド・ゲノム・インフォマティクス
INTEC Web and Genome Informatics Corporation
†2 東京工業大学精密工学研究所
Precision and Intelligence Laboratory, Tokyo Institute of Technology
†3 科学技術振興機構戦略的創造研究推進事業(CREST)
Core Research for Evolutional Science and Technology, Japan Science and Technology Agency
†4 東京工業大学精密工学研究所/日立製作所
Precision and Intelligence Laboratory, Tokyo Institute of Technology/Hitachi, Ltd.

☆ 文献5)では,専門家である弁理士自身が特許請求項の読みにくさを認めている.

を向上させることを目的とした，構造解析手法を提案する．まず，文献調査と予備調査により，特許請求項の記述特性を明らかにする．次に，修辞構造理論を基にして，特許請求項の構造を表現するための枠組みを提案する．次に，この枠組みに基づいて，特許請求項の構造を自動解析する手法を提案する．最後に，大規模テキストコレクションである NTCIR3 「特許データコレクション」[2] のデータを用いて提案手法の評価を行い，間接評価と直接評価により本手法の有効性を示す．

2. 特許請求項の記述特性

日本語の特許明細書における特許請求項の典型例として，公開番号「特開平 10-011111」の第 1 請求項と，公開番号「特開平 10-11007」の第 1 請求項をそれぞれ，**図 1** と**図 2** に示す．

日本語の特許請求項は通常，記述末尾に名詞または記号が存在し☆，その直前に名詞または記号または助詞「の」が連続的に出現して「名詞まとまり」を形成し，それに対して修飾部が前置する形で記述されている．記述末尾の「名詞まとまり」は，その発明の名称を表していることが多い．図 1 の場合，記述末尾に「作業機の操作用仮想振動生成装置」という「名詞まとまり」が存在し，それに対して「を設けたことを特徴とする」で終わる修飾部が前置している．また，「において、」や「であって、」などの文字列を用いて記述を前半部と後半部に分割し，前半部分にも「名詞まとまり」とそれに対する修飾部を前置して記述したものも多い．図 2 は，このようなスタイルで記述されており，記述末尾と，「において、」の直前に「大学構内掲示板サービスシステム」という「名詞まとまり」が存在している．

日本語の特許請求項は一般に，以下のような特徴を持っている．

(1) 文長がきわめて長い．
(2) 記述スタイルが独特である．
(3) 構文が複雑である．

第 1 の点を実証するために，NTCIR3 の特許データコレクションのうち，公開日が 1998 年 1 月から 3 月までのものからなる「サンプルデータ」(59,968 件) の第 1 請求項の長さを調査したところ，平均の文長は 241.97 文字であった．各種の日本語文章の平均文長を調べた文献 3) によれば，新聞記事の平均文長は政治面で 55.85 文字，社会面で 75.37 文字，小説の平均文長は夏目漱石の「坊ちゃん」で 30.9 文字，例外的に長文作家である谷崎潤一郎の「細雪」で 170.1 文字と報告されている．このことから，特許請求項は日本語の文として，異例の長さであるといえる．

第 2 の点に関して，特許出願者の立場から「どのように特許請求項を記述すべきか」を解説・考察した文献[4),5)]，および特許翻訳者の立場から日米の特許における用語と構文について調査・考察した文献[6)] を調査した．その結果，日本語の特許請求項には，確立した記述スタイルと制約が存在することが分かった．ま

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。**

図 1 日本語の特許請求項の例（特開平 10-011111 より引用）（259 文字）
Fig. 1 A sample Japanese patent (Publication Number=10-01111) (259 characters).

各種情報を蓄積し掲示用情報として出力するホストコンピュータと、このホストコンピュータと大学構内に付設されたデータ回線網を介して接続し、前記掲示用情報を入力あるいは受信して表示し、前記掲示用情報に含まれる各種サービスの要求と任意の情報の入力および出力とを行う複数の端末とからなる大学構内掲示板サービスシステム**において、**前記端末として各種掲示用情報を入力し利用者の要求を受付ける事務端末および図書端末と、前記利用者が使用し前記掲示用情報の取得と前記掲示用情報に含まれる各種サービスの要求と登録と予約とこれらの入力に対応する回答とを表示し出力する利用者端末とを備えることを特徴とする**大学構内掲示板サービスシステム。**

図 2 日本語の特許請求項の例 2（特開平 10-11007 より引用）（304 文字）
Fig. 2 A sample Japanese patent (Publication Number=10-11007) (304 characters).

☆ 後述するように，本研究では，形態素解析ツールとして「茶筌」[18)] を使用しているため，品詞名については「IPA 品詞体系」を用いている．記述末尾に「記号」が出現する例としては，「～燃焼支援喫煙具本体 (1)。」のような場合がある．

た，第 3 の点である構文の複雑性は，1 文で請求項を記述するという制約から生じているということも明確になった．

本稿では，こうした文献調査，特に文献 4) での特許請求項記述スタイル分類に基づいて，特許請求項の記述スタイルを以下の 3 つのスタイルに類型化する．ここで，各類型は排他的なものではない．たとえば，ジェプソン的形式の前半部または後半部が，順次列挙形式または構成要素列挙形式で記述されることは多い．

順次列挙形式 「〜し、〜し、〜した、〜」のように，処理を順序的に記述する形式である．方法の発明で用いられることが多いが，物の発明において用いられることもある．なお，文献 4) では,「書き流し形式」という用語が使われているが,「書き流す」とは本来,「あまり考えたり注意したりせずにさらさらと書く」（広辞苑第 5 版での語釈）を意味する言葉であり，本形式の特徴を正確に表現するものとはいい難い．このため，本研究では「順次列挙形式」という用語を採用した．

構成要素列挙形式 「〜と、〜と、〜とからなる、〜」のように，構成要素を列挙する形で記述する形式である．主に，物の発明で用いられる．なお，文献 4) では,「要件列挙形式」という用語が使われているが，この形式において列挙されているのは「要件」というよりも，発明を構成する「要素」であるため，本研究では「構成要素列挙形式」という用語を採用した．

ジェプソン（Jepson）的形式 「〜において、」や「〜であって、」などの文字列を用いて記述を前半部と後半部に分割し，前半部と後半部にはそれぞれ「名詞まとまり」とそれに対する修飾部が前置されている形式である．特に，後半部の修飾部は「〜を特徴とする〜」という形で記述されることが多い．前半部では，公知部分（すでに知られている内容）または前提条件を述べ，後半部では新規部分（この発明の特徴となる部分）または本論部分を記述する．なお，文献 4) では,「ジェプソン（Jepson）形式」という用語が使われているが，本来の意味での「ジェプソン（Jepson）形式」とは，公知部分と新規部分を明確に区別した形で表現されるものであり，文献 6)，7) によれば，日本語の特許請求項記述で多用される「において」形式や「であって」形式は，必ずしもこれと同義ではない☆．このため，本研究では,「ジェプソン的形式」という用語を採用した．

なお，特許の手続きや制度を解説した文献（たとえば，文献 8)，9)）には，上記以外に，以下のような特許請求項の記述スタイルを説明している．しかし，可読性向上を目的として本稿で提案する構造解析の枠組みにおいては，当面これらを考慮する必要がないと判断した．

Markush 形式 「1 つのグループを形成する相互に関連する化学的物質に一定の名称が付されていない場合に，その一群の物質を特定して限定する形式」[9] である.「または」などの用語により，択一的な表現を行う．本稿で提案する構造解析は，このレベルよりも高いレベルの構造を解析するものであるため，考慮しない．

product-by-process 形式 「製造法によって物質を特定する」[8] ものである．表現形式に着目した記述スタイルというよりも，意味内容に着目した記述スタイルといえる．また，多くの場合，順次列挙形式で記述されると考えられるため，考慮しない．

means-plus-function 形式 機能（function）を達成する手段（means）を記述する形のものである[9]．これも，表現形式に着目した記述スタイルというよりも，意味内容に着目した記述スタイルといえる．多くの場合，構成要素列挙形式で記述されると考えられるため，考慮しない．

日本語の特許請求項は上述したような特性を持つため，日本語の係り受け解析ツールとしてよく知られている KNP [10] を日本語の特許請求項に対して実行すると，多くの場合に解析に失敗する．KNP の解析失敗は，a) 結果を出力できない場合と，b) 間違った結果を出力する場合，とに分けられる．a) は，特許請求項の文長の長さと構文の複雑性に起因していると考えられる．b) は，特許請求項において，1 つの事項を説明した後でそれを用いて別の事項を説明するという，連鎖的な記述が多くみられることに起因していると考えられる．たとえば，図 1 の「特開平 10-011111」の場合,「〜負荷検出手段と」・「〜第一の周波数変換器と」・「〜第二の周波数変換器と」・「〜変調手段と」・「〜振動発生手段と」の 5 つの部分を並列構造としなければならないが，下線で示すような連鎖の存在により，この並列構造がうまく検出されない．

☆ 文献 6) は，“the Japanese claims drafter seldom intends a true Jepson claim, and the Jepson claims are indeed very rare in US patents” と指摘している．

表 1 特許請求項用の構造的関係
Table 1 Structural relations for patent claim.

区分	構造的関係	説明，および例
多核	PROCEDURE	順次列挙形式において，各処理を記述断片とし，それらの間の関係を表現する． 例：[〜し、][〜し、][〜した、]X
多核	COMPONENT	構成要素列挙形式において，各構成要素を記述断片とし，それらの間の関係を表現する． 例：[〜と、][〜と、][〜と] からなる X
単核	PRECONDITION	ジェプソン的形式において，前半部と後半部のそれぞれを記述断片または記述スパンとし，それらの間の関係を表現する． 例：[X であって、][〜した Y]
単核	COMPOSE	構成要素列挙形式において,「を備えた」や「からなる」など，要素構成を表す表現を 1 つの記述断片とし，それ以前の記述スパンとの関係を表現する． 例：[〜と、〜と、〜と][を備えた]X
単核	FEATURE	ジェプソン的形式の後半部に出現する「を特徴と (した | する)」という表現を 1 つの記述断片とし，それ以前の記述断片または記述スパンとの関係を表現する． 例：[〜した X][を特徴とする]
単核	ELABORATION	記述末尾の「名詞まとまり」を含む記述断片とそれに前置する記述断片または記述スパンとの関係，またはジェプソン的形式の「において」や「であって」の直前の「名詞まとまり」を含む記述断片とそれに前置する記述または記述スパンとの関係を表現する． 例：[X を〜した][Y の Z]

3. 修辞構造理論を基にした特許請求項の構造解析

前述したような特性を持つ日本語の特許請求項について，その記述内容を分かりやすく提示するためには，その構造を解析することが必要である．これにより，その特許請求項を構成する要素または処理が明確になるためである．

日本語の特許請求項は複数の要素や処理の説明を 1 文の中に詰め込んだ形で記述されている[5]．つまり，相互に関係を持つ複数の節から成り立っていると考えることができる．そこで我々は，複数の文・節から構成された談話の構造を解析するための理論である修辞構造理論（RST: Rhetorical Structure Theory）[11] のアイディアを借用することを考えた．

修辞構造理論は，1980 年代に提唱され，自動要約[12] や自動レイアウト[14] などに応用されている．修辞構造を対話型でグラフィカルに編集するためのツールとして，Tcl/Tk による RSTTool[15] も開発されている．

修辞構造理論においては，通常複数の文から構成されるテキストの構造を解明するために，修辞構造解析（rhetorical structure analysis）が行われる．修辞構造解析では，テキストを記述のまとまりごとに断片（segment）に分割し，断片間の関係付けを行い，さらに関係付けられた複数断片のまとまりであるスパン（span）間の関係付けを行いながら修辞構造木（rhetorical structure tree）を組み上げることで，その構造を解明する．断片とスパンの間を関係付ける際には，あらかじめ定義してある修辞関係（rhetorical relation）の 1 つが割り当てられる．修辞関係には，関係を構成する要素群が対等である関係と，重要な要素（nucleus：核=主要部）と補足的な要素（satellite：衛星=周辺部）とから構成される関係とがある．前者を多核（multi-nuclear）関係と呼び，後者を単核（mono-nuclear）関係と呼ぶ．

文献 5) によれば，特許請求項はまず,「対象の発明を分析し認識」したうえで「各要件を単文で定義」し，次に「重文化」を行うというプロセスによって作文される．また，文献 4) によれば，その記述スタイルは，2 章に示した 3 つに類型化される．そこで，特許請求項の構造解析を行うにあたり，修辞構造理論の考え方を参考にして，以下のように「記述断片」と「記述スパン」を定義する．そして，**表 1** に示すような形で，記述断片または記述スパン間の「構造的関係」を定義する☆．

☆ 表 1 の例の欄において，“[” と “]” で囲まれた部分が記述断片または記述スパンである．X，Y，Z は名詞または記号を表す．また，単核の関係の場合，下線が引かれている部分が核である．

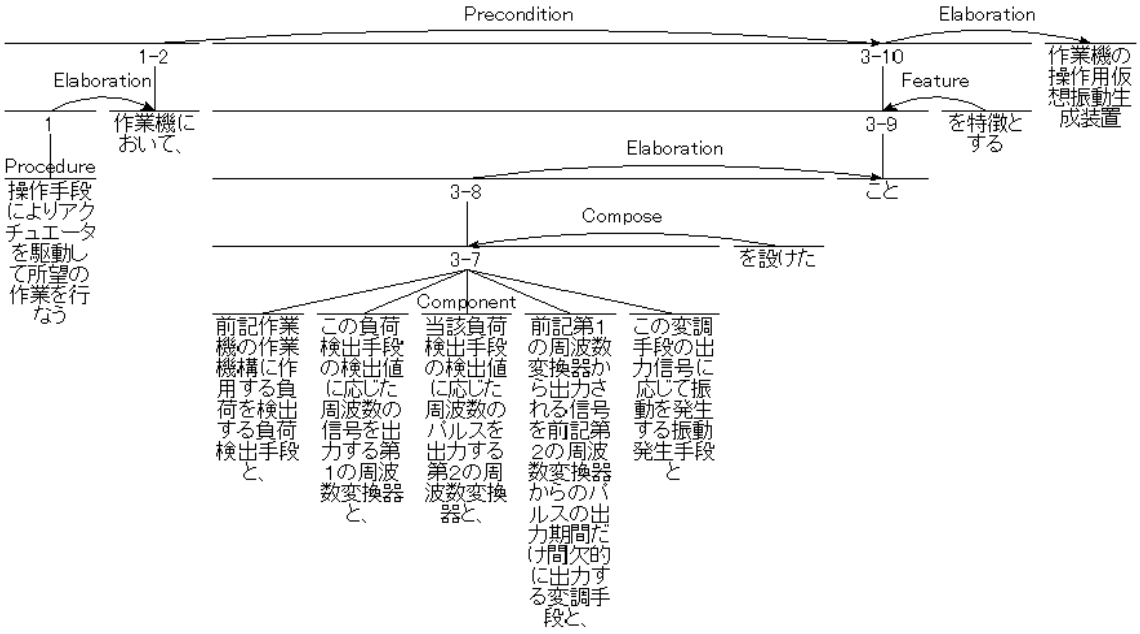

図 3 特許請求項の構造解析例（RSTTool v2.7 による表示）
Fig. 3 A result of structure analysis of patent claim.

- 順次列挙形式の場合は，各処理を記述断片とする．
- 構成要素列挙形式の場合は，各構成要素を記述断片とする．
- ジェプソン的形式の場合は，前半部と後半部のそれぞれを記述断片または記述スパンとする（前半部または後半部が，順次列挙形式または構成要素列挙形式になっている場合は記述スパンになり，そうでない場合は記述断片になる）．
- 構成要素列挙形式において,「を備えた」や「からなる」など，要素を構成していることを表す表現を 1 つの記述断片とする．
- ジェプソン的形式の後半部に典型的に出現する「を特徴とする」または「を特徴とした」という表現を 1 つの記述断片とする．
- ジェプソン的形式の場合,「において」や「であって」など，ジェプソン的形式を象徴的に表す文字列とその直前の「名詞まとまり」，およびこれに前置する修飾部のうち，上記処理の対象とならなかった部分（直前修飾部）をまとめて 1 つの記述断片とする．
- 記述末尾の「名詞まとまり」と，それに前置する修飾部のうち，上記処理の対象とならなかった部分（直前修飾部）をまとめて 1 つの記述断片とする．

ここで，断片・スパン・修辞関係など修辞構造理論の用語ではなく，記述断片・記述スパン・構造的関係という独自の用語を用いている理由は，本稿で提案している構造解析は，修辞構造理論からその着想を得てはいるものの，厳密には意味合いが異なるものであるためである．たとえば，表 1 の ELABORATION は，修辞構造理論で使われている ELABORATION とは意味合いが異なり，記述末尾の「名詞まとまり」を含む記述断片とそれに前置する記述断片または記述スパンとの関係，またはジェプソン的形式の「において」や「であって」直前の「名詞まとまり」を含む記述断片とそれに前置する記述断片または記述スパンとの関係を表現するために導入しているものである．

上記のような枠組みのもとで構造解析を行うことは，文献 5) に記述されているような，弁理士の実際の特許請求項作文プロセスを，言語処理によって解明するものであるといえる．

図 3 に，図 1 の特許請求項を表 1 で定義した構造的関係を用いて構造解析し，RSTTool によりグラフィカル表示した様子を示す．また，**図 4** に，構造解析の結果に基づいて，改行と箇条書き記号挿入およびインデンテーションを行った形式で表示した様子を示す☆．

いずれの表示形式にせよ，特許請求項の構造を視覚

☆ 図 4 では，順次列挙形式部分を箇条書きするために「－」を，構成要素列挙形式部分を箇条書きするために「＋」をそれぞれ使用している．

－ 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。

図 4 改行挿入とインデンテーションを行った例
Fig. 4 An example of newline insertion and indentation.

化することで，特許請求項に記述されている要素や処理を明確にすることができるため，特許請求項の読みやすさは大幅に向上する．

4. 手がかり句を用いた構造解析

4.1 手がかり句を用いたアプローチ

日本語の特許請求項の記述特性を調査する過程で，我々は，一部の特許請求項において，テキスト中に改行が明示的に挿入されていること，その挿入箇所が，長い特許請求項における記述断片または記述スパンの境界（以下では,「断片境界」と略称する）に一致していることが多いことを発見した．テキスト中に改行が挿入されている特許請求項の例を**図 5** と**図 6** に示す☆．

ここで，図 6 では，ジェプソン的形式を明示する手がかり句「であって、」は，1 行目末尾と 2 行目途中の 2 カ所に出現している．この例の場合，明らかに，改行が後続する 1 行目末尾のものが，この特許請求項の構造における断片境界を示すものである．

現在主流の出願方式である電子出願では，電子的な出願書類のフォーマットとして，HTML の一部のタグを使用したものが採用されており[16]，改行を特許明細書作成者が特許請求項記述の途中に明示的に挿入することができるようになっている．実際，弁理士によって作成された特許明細書の原稿を調査した結果，特許請求項の記述に改行が挿入される場合，断片境界やそれ以外の記述まとまり単位で行われるものであるということが分かった．すなわち，特許請求項に明示的に挿入された改行は，断片境界やそれ以外の記述まとまりを示すものとして，特許明細書作成者（弁理士など）によって挿入されたものであるといえる．実際，特許請求項の書き方について解説した文献 4)，5) の記述も，そのことを裏付けている．

NTCIR3 の特許データコレクションの「サンプルデータ」を調査した結果，第 1 請求項の記述中に改行が挿入されているものの割合は，48.5%であった．ま

☆ 図 5 と図 6 において，<nl> と明記した箇所に，改行コードが挿入されている．なお，表示レイアウトのためのマークアップ（
）は削除している．

原稿が載置される原稿台と、<nl>
この原稿台に対して主走査方向に移動する走査光学手段と、<nl>
この走査光学手段上に配置され原稿を副走査方向に照明する照明手段と、を備えた画像読取装置において、<nl>
前記照明手段は、前記走査光学手段に対して走査移動平面に略平行に回動自在に取付けられることを特徴とする画像読取装置。

図 5 明示的に改行が挿入されている特許請求項例（特開平 10-010848 より引用）
Fig. 5 An example of claim which newlines are explicitly inserted.

湾曲状に形成された空気取入口を有する筐体に対して着脱自在になされるフィルタの着脱機構 であって、<nl>
前記空気取入口の上下部の筐体内面 であって、該空気取入口の湾曲面に沿って前記筐体に設けられた溝部と、<nl>
前記空気取入口の開口形状よりも大きな形状で可逆性の枠体を有したフィルタとを備え、<nl>
前記枠体の上下部が前記空気取入口の上下部の溝部に入り込んだときには、その枠体の両端が該空気取入口の両側に引っかかるようにして前記フィルタが前記筐体に係止されると共に、前記フィルタが前記筐体の内側に向けて押されたときには、前記溝部からフィルタが外れるようになされたことを特徴とするフィルタの着脱機構。

図 6 明示的に改行が挿入されている特許請求項例 2（特開平 11-19437 より引用）
Fig. 6 An example of claim which newlines are explicitly inserted.

表 2 明示的に改行が挿入された請求項における，改行直前のパターン
Table 2 Expression pattern just before newline of the claims in which newlines are explicitly inserted.

No.	パターン	割合	累積割合
1	(名詞 | 記号) と (、|,)	46.1%	46.1%
2	(動詞連用形 | 助動詞連用形)(、|,)	17.5%	63.6%
3	(名詞 | 記号)(において | に於 (い)?て)(、|,)	16.5%	80.1%
4	(名詞 | 記号) であって (、|,)	7.2%	87.3%

た，改行が挿入されている特許請求項について，改行直前の形態素を 3 つ分抽出して集計したところ，**表 2** のような結果が得られた．第 1 位から第 4 位までのパターンで，87%以上のカバー率となっている．

上記の分析と 2 章の調査結果に基づき，我々は，日本語の特許請求項の構造を解析するために，手がかり句を用いたアプローチ[13] を採用した．すなわち，手がかり句（cue phrase）を収集し，それを用いて，長い請求項を断片化し，記述断片または記述スパン間の構造的関係を設定することにした．

表 2 の結果と 2 章の調査結果から収集した手がかり句を**表 3** にまとめる．なお，表 3 では，手がかり句の表記に，Perl の正規表現記法を使用している．また，手がかり句を類型化し，各類型に対して，トークンを割り当てている．トークンの割当ては，手がかり句の文字列に対して 1 つのトークンを割り当てる場合（表中の上から 3 つ）と，複数のトークンを割り当てる場合（表中の下 2 つ）とがある．

4.2 構造解析アルゴリズム

日本語で記述され，記述が 1 つの請求項で完結するもの（独立形式請求項）☆を対象として，手がかり句を用いて構造解析を行うアルゴリズムを設計した．

特許請求項は自然言語で記述されるとはいえ，前述したように，その記述スタイルが確立しパターン化している．このため，その解析にあたっては，字句解析と構文解析から構成されるコンパイラの手法をベースとしたものを採用した．ただし，字句解析において，文脈を判定し，文脈に依存した形でトークンを出力させることにした．

また，図 6 のように，特許請求項記述に改行が挿入されている場合は，手がかり句の判定に改行を利用する形で解析を行うことにした．

以下に，アルゴリズムの概要を示す．

(1) 茶筌による形態素解析
字句解析の前処理として，単語の切り出しを行い，かつその品詞情報を得るために，特許請求項テキストを茶筌[18] で形態素解析する．もともと挿入されている改行コードは，そのままの状態で入力する．
- 茶筌には，-j オプションを使用し，区切り文字を「。::」のいずれかとする．

(2) 字句解析
茶筌の出力結果を入力し，文脈を判定しながら，トークンと文字列のペアを出力する．トークンは，表 3 中に示すもの（JEPSON_CUE, FEATURE_CUE, COMPOSE_CUE, NOUN, POSTP_TO, PUNCT_TOUTEN, VERB_RENYOU, VERB_KIHON）に，以下の 2 つを加えた 10 種類を考える．

POSTP_NO 「名詞まとまり」中の助詞「の」に対して割り当てるトークンである．

WORD 他の 9 種類のトークンが割り当てられなかった単語に対して割り当てるトークンである．

(3) 構文解析（=構造解析）
Bison[19] 互換のパーサジェネレータを用いて生成したパーサを使用して構文解析を行う．この場合，この処理はそのまま構造解析となっている．なお，今回，特許請求項の構造解析用に記述した文脈自由文法（LALR(1) 文法）の規模は，ルール数が 57，トークン数が 10，非終端記号数が 19 である．この文法は，shift/reduce conflict や reduce/reduce conflict が生じない形で記述されたものである．構造的関係の認識の概要は以下のとおりである．
- JEPSON_CUE，COMPOSE_CUE，VERB_KIHON，FEATURE_CUE の各トークンの存在により対応するルールが起動され，PRECONDITION，COMPONENT，PROCEDURE，FEATURE の各構造的関係が認識される．なお，COMPOSE_CUE については同時に，COMPOSE 関係も認識される．

☆ このほか，請求項には，他の請求項を引用する形式のもの（引用形式請求項）が存在する．

表 3 特許請求項解析のための手がかり句
Table 3 Cue phrases which can be used to analyze patent claims.

手がかり句	割り当てるトークン
において (、|,) に於 (い)?て (、|,) であって (、|,) にあたり (、|,) に当 (た)?り (、|,)	JEPSON_CUE
を特徴と (した｜する)(、|,)?	FEATURE_CUE
を搭載して構成され (た｜る｜ている) (、|,)? を (、|,)?(具｜備｜そな) え (た｜る｜ている｜てなる) (、|,)? を (、|,)?具備 (した｜する｜している｜してなる) (、|,)? (で｜から) 構成され (た｜ている) (、|,)? を (、|,)?有 (する｜した) (、|,)? を (、|,)?包含 (する｜した) (、|,)? を (、|,)?含 (む｜んだ) (、|,)? (から｜より)(、|,)?(なる｜なった｜なっている) (、|,)? (から｜より)(、|,)?(成る｜成った｜成っている) (、|,)? を (、|,)?設け (た｜る｜ている) (、|,)? を (、|,)?装備 (する｜した｜している) (、|,)?	COMPOSE_CUE
（COMPOSE_CUE に対応する手がかり句の前に存在する） (名詞｜記号) と (、|,)	（「(名詞｜記号)」に対して）NOUN （「と」に対して）POSTP_TO （「(、|,)」に対して）PUNCT_TOUTEN
（「(動詞基本形｜助動詞基本形) ＜名詞まとまり＞」の前に存在する） (動詞連用形｜助動詞連用形)(、|,) (動詞基本形｜助動詞基本形) と (共｜とも) に	（「(動詞連用形｜助動詞連用形)」，または 「(動詞基本形｜助動詞基本形)」に対して） VERB_RENYOU （「(、|,)」，または「と (共｜とも) に」に対して） PUNCT_TOUTEN （＜名詞まとまり＞に前置する「(動詞基本形｜助動詞基本形)」に対して） VERB_KIHON

- 記述末尾または JEPSON_CUE に対応する文字列直前に存在する「名詞まとまり」を認識するルールにより，当該「名詞まとまり」が認識され，それを含む記述断片とその直前の記述断片または記述スパンとの間での ELABORATION 関係が認識される．

4.3 字句解析の文脈依存処理の詳細

上記のアルゴリズムのうち，字句解析の文脈依存処理の詳細について，以下に説明する．

(1) 表 3 の JEPSON_CUE に対応する手がかり句を探索し，それが存在した場合，その文字列に対して JEPSON_CUE トークンを割り当てる．改行コードを含む特許請求項の場合，改行コードが後続する場合のみ認識させる．該当するものが 2 個以上存在する場合，後方に出現するものに対してだけ割り当てる．

(2) 表 3 の FEATURE_CUE に対応する手がかり句を探索し，それが存在した場合，その文字列に対して FEATURE_CUE トークンを割り当てる．

(3) 記述末尾から前方向に探索し，名詞または記号と助詞「の」が連続的に出現する「名詞まとまり」を検出し，名詞と記号に対して NOUN トークン，助詞「の」に対して POSTP_NO トークンを割り当てる．

(4) JEPSON_CUE，FEATURE_CUE の直前から前方向に探索し，上記と同様に「名詞まとまり」を検出し，NOUN トークンと POSTP_NO トークンを割り当てる．

(5) 非ジェプソン的形式の場合は全体に対して 1 回，ジェプソン的形式の場合は前半部と後半部のそれぞれに対して前方向に探索し，以下のいずれのパターンが後に出現するかを調べる．
  (a) (動詞基本形 ｜ 助動詞基本形)(、|,)? NOUN
  (b) COMPOSE_CUE に対応する手がかり句

(6) (a) の場合，動詞基本形または助動詞基本形に対して VERB_KIHON トークン，読点に対して PUNCT_TOUTEN トークンを割り当てる．さらに前方向に探索し，他の手がかり句トークンが存在するまでの範囲において，「(動詞連

用形 | 助動詞連用形)(、|,)」パターンまたは「(動詞基本形 | 助動詞基本形) と (共 | とも) に」を探索し，それらが存在した場合，前者の場合は動詞連用形または助動詞連用形に対して VERB_RENYOU トークンを，読点に対して PUNCT_TOUTEN トークンを割り当てる．後者の場合は「(動詞基本形 | 助動詞基本形)」に対して VERB_RENYOU トークンを,「と (共 | とも) に」に対して PUNCT_TOUTEN トークンを割り当てる．

(7) (b) の場合，COMPOSE_CUE に対応する手がかり句に対して，COMPOSE_CUE トークンを割り当てる．その直前に「と (、|,)?」が存在するときは，さらに前方向に探索し，他の手がかり句トークンが存在するまでの範囲において,「(名詞 | 記号) と (、|,)」パターンを探索し，それが存在した場合，名詞または記号に対して NOUN トークン,「と」に対して POSTP_TO トークン，読点に対して PUNCT_TOUTEN トークンを割り当てる．それが存在しない場合，他の手がかり句トークンが存在するまでの範囲において，「(動詞連用形 | 助動詞連用形)(、|,)」パターンを探索し，それが存在した場合，動詞連用形または助動詞連用形に対して VERB_RENYOU トークン，読点に対して PUNCT_TOUTEN トークンを割り当てる．

(8) 上記の処理によって NOUN トークンを割り当てる対象となった単語の直前を探索し，これに連続して出現する名詞・記号，または助詞「の」が存在した場合，それらに対してそれぞれ NOUN トークン，POSTP_NO トークンを割り当てる．

(9) 上記の処理によってトークンが割り当てられなかった単語に対して WORD トークンを割り当てる．

図 5 の特許請求項テキストを字句解析に入力したときの出力の一部を，付録 A.1 に示す．

4.4 実　　装

上記のアルゴリズムに基づき，構造解析を行うプログラム（rst_claim）を実装した．プログラムの入力は，特許明細書から特許請求項（独立形式請求項）の部分を切り出し，表示レイアウトのためのマークアップ（BR〉）を削除したテキストとした．このとき，もともと挿入されていた改行コードは保存するようにしている．プログラムの出力としては，コマンドオプションにより，以下の形式のいずれかを選択できるようにした．

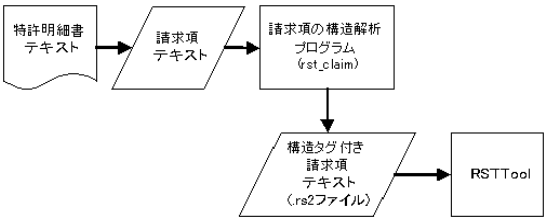

図 7 構造解析プログラム（rst_claim）の位置付け
Fig. 7 Positioning of structure analysis program.

- RSTTool で表示するためのファイルフォーマットである「. rs2 ファイル」の形式（XML ライクな形式で，構造をマークアップしたファイル）
- 図 4 に示されているように，構造解析の結果に基づいて，改行と箇条書き記号の挿入およびインデンテーションを行ったテキスト形式

図 7 に，構造解析プログラム（rst_claim）と RSTTool の関係を示す．

5. 評　　価

実装した構造解析プログラムについて，NTCIR3 の特許データコレクションを使用して評価を行った．

NTCIR3 の特許データコレクションは，1998 年の公開特許公報（約 34 万件の特許明細書）と 1999 年の公開特許公報（約 35 万件の特許明細書）を含んでいる．それぞれの公開特許公報は，公開日順に整列され，1,000 件単位にディレクトリとしてまとめられた形で提供されている．手がかり句の抽出とルール作成において参照した「サンプルデータ」は，1998 年の公開特許公報から抽出されたものであり，公開日が 1998 年 1 月から 3 月までの 59,968 件であった．評価を行うにあたり，1999 年の公開特許公報から，公開日が 1999 年 1 月から 3 月までの 59,956 件の第 1 請求項☆を抽出して使用した．評価に用いた 59,956 件について，分野の偏りがないかを調査するために，割り振られている国際特許分類（IPC）コードの先頭の記号（A から H まで）を抽出してその分布を確認したところ，特許庁が公開している 1999 年の全公開特許のそれと類似したものであった．すなわち，評価に使用したデータは，手がかり句の抽出とルール作成において参照したデータとは異なっており，かつ，分野について特定の偏りを持つものではない．

評価は，以下の観点で行った．

受理率　文脈自由文法から生成されたパーサが受理（accept）した特許請求項の割合

☆ 第 1 請求項は通常，独立請求項である．

表 4 解析精度に関する間接評価結果
Table 4 Indirect evaluation result on the analysis precision.

指標	ベースライン 1	ベースライン 2	構造解析結果を利用した改行挿入処理	推定上限値
再現率（R）	0.4779	0.8394	0.6741	0.8736
適合率（P）	0.3739	0.5151	0.6632	N/A
F-measure（F）	0.4195	0.6384	0.6686	N/A

処理速度 1 件の特許請求項を解析するために要する時間

解析精度 構造解析結果の精度

- 間接評価
- 直接評価

受理率は，99.77%であった．処理速度は，1 件あたり 0.30 秒であり，即時的に実行された．なお，処理速度評価には，Pentium III 1 GHz・メモリ 512 MB で Linux OS を搭載した PC を使用した．

以下では，解析精度に関する評価について述べる．

5.1 解析精度に関する間接評価

我々が実装した構造解析プログラムは，コマンドラインオプションの指定により，特許請求項記述に挿入されている改行コードをいっさい利用しないで動作することができるようになっている．これにより，特許請求項記述に改行コードが挿入されていた場合であっても，それを利用せずに，手がかり句だけを用いて構造解析を行い，その結果から，以下のようなヒューリスティックスを用いて改行挿入位置を推定することができる．

- JEPSON_CUE に対応する文字列の直後に改行コードが挿入されることが多い．
- ジェプソン的形式の後半部が順次列挙形式のときには，順次列挙の断片境界に改行コードが挿入されることが多い（非ジェプソン的形式の場合も同様）．
- ジェプソン的形式の後半部が構成要素列挙形式のときには，各構成要素の断片境界に改行コードが挿入されることが多い（非ジェプソン的形式の場合も同様）．

もともと改行が挿入されている特許請求項を対象とし，その改行挿入箇所を断片境界に関する正解データとして用いることで，改行コードを使わずに構造解析して推定した改行挿入位置の再現率（Recall）と適合率（Precision）を計算することができる．つまり，もともと挿入されている改行挿入箇所数を n，構造解析アルゴリズムが推定した改行挿入箇所の数を i，構造解析アルゴリズムが推定した改行挿入箇所の数のうちもともと挿入されている改行挿入箇所に一致した数を c として，再現率（R），適合率（P），さらに F-measure（F）を以下の式で算出することができる．

$$R = \frac{c}{n} \tag{1}$$

$$P = \frac{c}{i} \tag{2}$$

$$F = \frac{2 * R * P}{R + P} \tag{3}$$

なお，構造解析の結果を利用した改行挿入位置推定には，以下のような限界があることに留意する必要がある．

- 特許明細書作成者による改行挿入は，構造解析における断片境界以外の場所でも行われていることがあること（例：「〜は、」の後に改行が挿入される場合など）．このため，構造解析の結果を利用した改行挿入処理の再現率の上限値は，表 2 での調査結果から，0.8736 と推定されること☆．
- 特許明細書作成者による改行挿入は，必ずしも完全な一貫性を持ったものではないこと．たとえば，順次列挙形式でいくつかの断片境界には改行が挿入されていても，改行が挿入されていない断片境界が存在することもあること．

ベースラインとしては，以下の 2 つを設定した．

ベースライン 1 すべての「(名詞 | 記号) と (、|，)」と「(動詞連用形 | 助動詞連用形)(、|，)」の後に改行を機械的に挿入する．

- ベースラインをこのように設定した理由は，特許請求項の日英翻訳用にチューニングされていることをうたっている商用の機械翻訳ソフトウエア製品における「自動分割処理」が，このような処理を行っていると推定されるためである☆☆．

ベースライン 2 表 2 の 4 位までのパターンすべてについて，これが出現した後に改行を機械的に挿入する．

表 4 に，評価結果を示す．

☆ 表 2 に記載している 4 つに加えて,「に (あたり | 当たり | 当り)(、|，)」というパターンが 0.06%出現している．

☆☆ これは，商用の機械翻訳ソフトウエア製品 2 つに対する，2001 年 7 月と 2002 年 7 月時点でのテスト結果に基づく推定である．

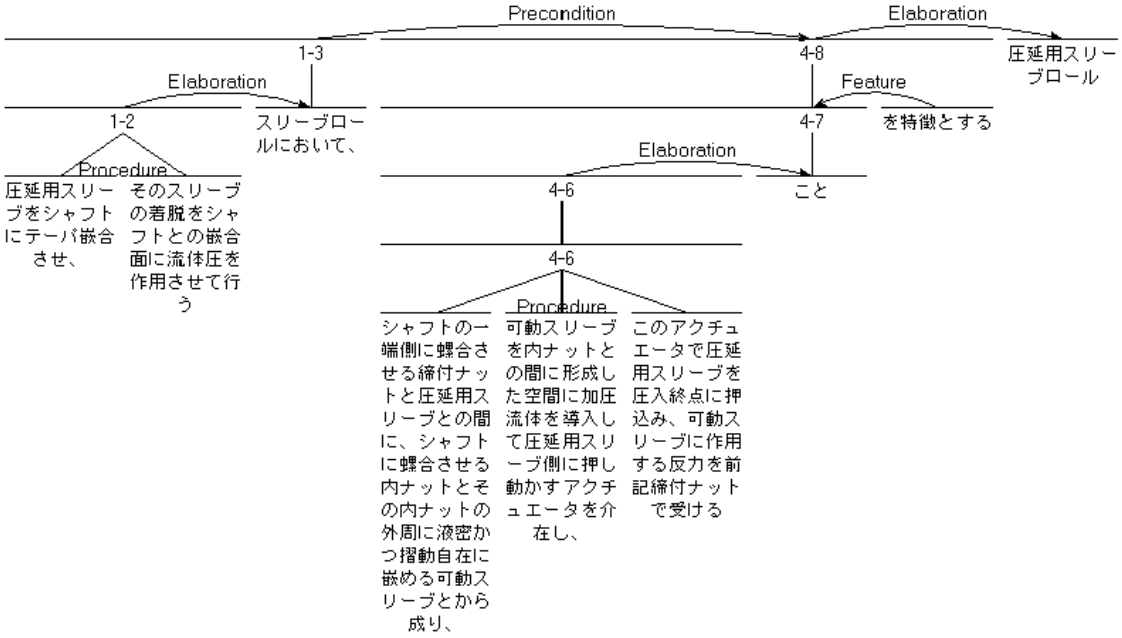

図 8 断片化不十分の例（特開平 11-33609 の第 1 請求項の構造解析結果）
Fig. 8 An example of under-segmentation.

表 5 解析精度に関する直接評価結果
Table 5 Direct evaluation result on the analysis precision.

区分	数	割合（「判定対象外」を除く）
完全正解	87	91.6%
部分的正解	2	2.1%
完全不正解	6	6.3%
判定対象外	5	-

5.2 解析精度に関する直接評価

解析精度に関する直接評価は，59,956 件の中からランダムに選択した 100 件の特許請求項の構造解析結果を 1 名の被験者が評価することで行った．被験者は，特許請求項テキストを読解してその大まかな内容を把握したうえで，構造解析結果を図 3 と同様のグラフィカル表示を目視してその解析結果を以下の評価基準に基づいて評価した．この被験者は，本研究とは別の観点で特許情報処理について研究していた大学研究者である．つまり，弁理士や知的財産権担当者などの専門家ではないが，特許明細書に日常的に接する経験を有していた．特許請求項の読解は，その内容把握のレベルを大まかなものにとどめたとはいえ，時間を要する作業となり，100 件の評価には約 10 時間を要した．

評価の基準は，以下のとおりである．

(1) 特許請求項がジェプソン的形式である場合，そのことが正しく認識され，かつ，前半部と後半部の分離が正しく行われているかどうか．
(2) 特許請求項がジェプソン的形式である場合，前半部の構造が正しく解析されているかどうか．
(3) 特許請求項がジェプソン的形式である場合，後半部の構造が正しく解析されているかどうか．
(4) 特許請求項がジェプソン的形式でない場合，全文の構造が正しく解析されているかどうか．

上記の 4 つの基準に照らし合わせて，完全に正解であった場合,「完全正解」とした．評価基準 (2), (3), (4) の解析において，記述断片の認識が部分的に誤っていた場合は,「部分的正解」とした．

表 5 に，評価結果を示す．

「部分的正解」の例を**図 8** に示す．これは，順次列挙形式の記述断片認識に関して，記述断片の認識に 1 カ所の間違いがあったものである．つまり，右から 4 番目の記述断片中の,「押込み、」の部分で断片化が行われるべきであったものが，認識されなかったため,「部分的正解」とした．

また，評価対象 100 件のうち 63 件を占めたジェプソン的形式に対して，本プログラムはそのすべてを正しく認識した．しかし，ジェプソン的形式ではない 1 件について，本プログラムはこれをジェプソン的形式として誤認識した．本プログラムが誤認識した特許請求項を**図 9** 示す．図 9 では,「ここにおいて、」という表現が存在し，この「において、」の部分をジェプソ

パーキングシフトレンジ位置から他の各シフトレンジ位置までシフト操作可能とされたシフトレバーと、
前記シフトレバーに装着されたノブボタンによって、移動操作されるグルーブドピンと、
前記グルーブドピンが挿通され、少なくともパーキングレンジ用凹部が形成されたディテント穴と、
前記シフトレバーが、パーキングシフトレンジ位置にあるときに、前記ノブボタンの操作により移動される前記グルーブドピンを溝部で受けることにより前記グルーブドピンの動作に連動して回動操作される第 2 操作部材と、
キーインターロック機構に連動するキーロックケーブルの一端が接続された第 1 操作部材と、
前記第 1 操作部材と、前記第 2 操作部材とを連結するとともに、所定以上の相対的回動力が働いた際、両者の連結を解除する連結手段とを有し、ここにおいて、
前記ノブボタンを途中まで押し操作した際、前記キーインターロック機構がかかったロック状態に移行し、前記キーインターロック機構からの反力で前記第 1 操作部材の移動が制止され、前記ノブボタンをさらに最後まで押し操作する際に、前記連結手段が解除され、前記第 2 操作部材が前記ノブボタンの比較的弱い押し操作力で回動操作され、前記グルーブドピンが前記パーキングレンジ用凹部から抜け出すようにしたことを特徴とするシフトレバー装置。

図 9 ジェプソン的形式として誤認識した特許請求項の例（特開平 11-42952 より引用）
Fig. 9 A sample Japanese patent misjudged as Jepson-type (Publication Number=11-42952).

ン的形式の手がかり句として誤認識したためである．

なお，表 5 で,「判定対象外」は，もともとの請求項記述において，たとえば「ことを特徴とする」と記述すべきところを「こと特徴とする」と記述されている場合などの記述エラーを含むものを指している☆．

6. 考　　察

6.1 アルゴリズムとプログラムの特性について

上記で説明したアルゴリズムとプログラムは，特許請求項の構造を末端レベルまで解析するものではない．このことを,「特開平 10-11007」の構造を解析し，改行挿入とインデンテーションを行ったテキスト出力結果である**図 10** を例にとって説明する．

＋ 各種情報を蓄積し掲示用情報として出力するホストコンピュータと、
＋ このホストコンピュータと大学構内に付設されたデータ回線網を介して接続し、前記掲示用情報を入力あるいは受信して表示し、前記掲示用情報に含まれる各種サービスの要求と任意の情報の入力および出力とを行う複数の端末と
からなる
　大学構内掲示板サービスシステム
　　において、
＋ 前記端末として各種掲示用情報を入力し利用者の要求を受付ける事務端末および図書端末と、
＋ 前記利用者が使用し前記掲示用情報の取得と前記掲示用情報に含まれる各種サービスの要求と登録と予約とこれらの入力に対応する回答とを表示し出力する利用者端末と
を備える
　こと
　　を特徴とする
　　大学構内掲示板サービスシステム

図 10 「特開平 10-11007」の請求項 1 に，改行挿入とインデンテーションを行った出力
Fig. 10 Output of newline-insertion and indentation for the claim of 10-11007.

図 10 において,「このホストコンピュータと大学構内に付設されたデータ回線網を介して接続し、前記掲示用情報を入力あるいは受信して表示し、前記掲示用情報に含まれる各種サービスの要求と任意の情報の入力および出力とを行う複数の端末と」という記述断片は，さらに以下の下部処理からなっている．

- このホストコンピュータと大学構内に付設されたデータ回線網を介して接続し，
- 前記掲示用情報を入力あるいは受信して表示し，
- 前記掲示用情報に含まれる各種サービスの要求と任意の情報の入力および出力とを行う

3 章で定義した構造的関係では，構成要素列挙形式の各構成要素を記述断片としている．つまり，本稿で提案しているアルゴリズムとプログラムではこのレベルより下部構造の解析は行っていないことになる．とはいえ，図 10 のレベルまでの断片化ができれば，後の処理は，もう 1 度同様の処理を繰り返す，または KNP [10] などの既存の言語解析ツールで処理することで対応可能であると考えられる．

☆ このような記述エラーの場合，解析プログラムが自動的にエラー回復すべきとの立場もありうるが，我々は，特許明細書は法的文書であり解析プログラムはその記述内容を忠実に処理すべきとの立場をとることにした．

6.2 受理率について

特許請求項は，弁理士や知的財産権担当者などの専門家が作成したものだけでなく，個人発明家が作成したものや海外特許を翻訳して作成されたものがある．このため，本稿で述べたような記述スタイルを逸脱したものや，区切り文字が「。:; 」以外のもの，そして記述エラーを含むものが，少数とはいえ確実に存在する．構造解析プログラムで受理されないものには，このような例外的な特許請求項が多い．

こうした状況にもかかわらず，今回実装したプログラムは，受理率が 99.77%と高い数字を示した．これは,「サンプルデータ」を使い，十分な時間と労力をかけて，字句解析プログラム（特に，その文脈依存処理部分の処理）と文脈自由文法をチューニングしたためである．

今後の課題としては，統計的なアプローチや各種の学習理論を用いて，このようなプログラム開発の労力を削減したり，半自動化を行ったりすることが考えられる．

6.3 解析精度に関する間接評価について

この間接評価で評価できるのは，構造解析の一部として実施される断片化処理の精度であり，構造的関係の認識精度も含めた構造解析の精度そのものを評価できるわけではない．しかし，本手法によって断片化処理が正しく行われているということは，構造的関係の認識が正しく行われていることを高い蓋然性をもって推測させるものであり，本評価法を用いて構造解析精度を間接的に評価できるものと考える．

解析精度に関する間接評価に関して，構造解析結果を利用した改行挿入処理には 5.1 節で述べたような限界性があるにもかかわらず，表 4 が示すように，そのF-measure の値は，ベースライン 1 のそれよりもはるかに高くなっている．これに対して，ベースライン 2 との違いはそれほど大きくない．しかし，ベースライン 2 はもともと，改行挿入箇所の特徴を直接反映したものであり，実際，その再現率（R）も推定上限値と近い値となっている．それにもかかわらず，構造解析結果を利用した改行挿入処理の F-measure はベースライン 2 よりも上回っており，本手法による構造解析の有効性を間接的に示すものであると考える．

6.4 解析精度に関する直接評価について

本稿で提案している構造解析手法では，手がかり句を認識することがそのまま，その手がかり句に関連付けられた構造的関係を認識することにつながっている．前述の図 9 は,「において、」を間違った形で認識したため，構造解析に失敗したものであった．この場合のように，手がかり句の認識に失敗した場合は，そのまま完全不正解となってしまうが，直接評価においてはそのようなケースは 6 件だけであった．

手がかり句を正しく認識することができれば，後は記述断片をいかに正しく認識するかが課題となる.「部分的正解」は，記述断片の認識に失敗したものであるが，これは，2.1%であった．これらは，局所的な部分において記述断片の認識に関して間違いを含むものであり，それ以外の部分には問題がないものばかりであった．

記述断片の認識に失敗した原因として多いのは，形態素解析プログラムが間違った解析結果を返したことによるものである．たとえば，前述の図 8 では,「押込み、」を「動詞連用形 + 読点」としてではなく「名詞 + 読点」として認識したものである．このほか,「おむつ」を名詞として認識しなかったため記述断片境界を間違えたものがあった．これらについては，形態素解析の精度向上に関する別の取り組みが必要となるだろう．

今回実装したプログラムは，もともと人手でテキストの構造をタグ付けするツール（RSTTool）用の形式で出力を行うため，プログラムの解析結果に問題があった場合は，人手でその修正を行うことができる．今回の評価結果における部分的正解の場合は，こうした修正作業はきわめて容易である．

ジェプソン的形式の認識に関して，本プログラムは高い精度を示した．これは，特許請求項の記述において,「において、」や「であって、」などのジェプソン的形式を示す手がかり句の「重さ」を意味していると考えられる．実際，特許の専門家である弁理士や特許翻訳者による文献 4)~6) では，これらの手がかり句を特別扱いして説明している．

以上の評価結果と考察により，本手法の妥当性が示されたといえる．

7. 関 連 研 究

一般的なテキスト（英文）を対象とした，手がかり句に基づく構造解析の手法と評価は，文献 12)，13) に述べられている．

特許データを対象とした研究としては，情報検索に関するものが多く，言語処理に関するものは，あまり多くない．

本研究と同様に，特許請求項の読解支援を目的とした言語処理の研究には，先行研究 20) がある．先行研究 20) においては，本研究よりもより細かな粒度での解析が行われているが，これはあくまでも係り受け

解析であり，記述断片に分割してその構造を解析する我々のアプローチとは大きく異なる．また，先行研究20) では，大量の特許データを対象としたときの解析についての評価結果については報告されていない．

このほか，特許請求項の言語的な特異性に着目し，知識工学と自然言語生成の手法により，その生成を支援するシステムについての研究が，文献 17) に報告されている．

8. むすび

専門家以外の人にとってはきわめて読みにくい特許請求項を対象として，その構造を表現するための枠組みを定義し，手がかり句を用いた構造解析手法を提案した．大規模テキストコレクションである NTCIR3 の特許データコレクションのデータを用いて提案手法の評価を行い，間接評価と直接評価により本手法の有効性を示した．

特許請求項の構造を解析することで，その特許請求項を構成する要素や処理を明確化することができる．構造解析結果を利用することで，特許請求項を自動的に分かりやすくいい換えることも可能になると思われる．

本研究の目的は特許請求項の可読性向上であるが，そこで得られた成果はより挑戦的なタスクである「特許マップ」[21)] の自動生成に応用できるものと考えている．

謝辞 本研究では，NTCIR3 ワークショップで配布された「特許データコレクション」を使用しています．

参考文献

1) 新森昭宏，齋藤 豪，奥村 学：特許請求項の可読性向上のための自動言い換えについての考察，言語処理学会第 7 回年次大会併設ワークショップ「言い換え/パラフレーズの自動化に向けて」，pp.65–70 (2001).
2) 岩山 真，藤井 敦，高野明彦，神門典子：特許コーパスを用いた検索タスクの提案，情報処理学会研究報告—情報学基礎，FI-63-007 (2001).
3) 前川 守：1000 万人のコンピュータ科学 3 文学編—文章を科学する，岩波書店 (1995).
4) 葛西泰二：特許明細書のクレーム作成マニュアル，工業調査会 (1999).
5) 糟谷洋治：「請求の範囲」の文体と作文技法の考察，パテント (1999).
6) Lise, W.: An Investigation of Terminology and Syntax in Japanese and US Patents and the Implications for the Patent Translator (2002). http://www.lise.jp/patsur.html
7) 田辺 徹：英文特許の常識，工業調査会 (1989).
8) 竹田和彦：特許の知識，第 6 版，ダイヤモンド社 (1999).
9) 飯田幸郷：特許用語の基礎知識，発明協会 (1999).
10) 黒橋禎夫：結構やるな，KNP, 情報処理，Vol.41, No.11, pp.1215–1220 (2000).
11) Mann, B.: An Introduction to Rhetorical Structure Theory (RST) (1999). http://www.sil.org/~mannb/rst/rintro99.htm
12) Marcu, D.: *The Theory and Practice of Discourse Parsing and Summarization*, MIT Press (2000).
13) Marcu, D.: The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach, *Computational Linguistics*, Vol.26, No.3, pp.395–448 (2000).
14) Bateman, J., Kamps, T., Kleinz, J. and Reichenberger, K.: Toward Constructive Text, Diagram, and Layout Generation for Information Presentation, *Computational Linguistics*, Vol.27, No.3, pp.409–449 (2000).
15) O'Donnell, M.: RST-Tool: An RST Analysis Tool, *Proc. 6th European Workshop on Natural Language Generation* (1997).
16) 特許庁：パソコン出願ソフト 2 操作マニュアル (1999).
17) Sheremetyeva, S. and Nirenburg, S.: Knowledge Elicitation for Authoring Patent Claims, *IEEE Computer*, pp.57–63 (July 1996).
18) 松本裕治，形態素解析システム「茶筌」，情報処理，Vol.41, No.11 (2000).
19) Donnelly, C. and Stallman, R.: Bison: The YACC-compatible Parser Generator, Version 1.25 (1995).
20) 亀田雅之：日本語文書読解支援系 QJR の検討，情報処理学会研究報告—自然言語処理，NL-110-9, pp.57–64 (1995).
21) パテントマップ研究会：パテントマップと情報戦略，発明協会 (1988).

付 録

A.1 字句解析の出力例

以下において，各行は，トークンと形態素文字列のペアからなっている．ここでたとえば，「原稿」という名詞に対するトークンとして，出現文脈に応じて，NOUN と WORD のいずれかが与えられている．

```
WORD 原稿
WORD が
...
WORD れる
NOUN 原稿
```

NOUN 台
POSTP_TO と
PUNCT_TOUTEN 、
WORD この
WORD 原稿
...
WORD する
NOUN 走査
NOUN 光学
NOUN 手段
POSTP_TO と
PUNCT_TOUTEN 、
WORD この
...
WORD する
NOUN 照明
NOUN 手段
POSTP_TO と
PUNCT_TOUTEN 、
COMPOSE_CUE を備えた
WORD 画像
WORD 読取
NOUN 装置
JEPSON_CUE において、
WORD 前記
...
WORD に
WORD 取付け
VERB_KIHON られる
NOUN こと
FEATURE_CUE を特徴とする
WORD 画像
WORD 読取
NOUN 装置

(平成 15 年 3 月 5 日受付)
(平成 15 年 12 月 2 日採録)

新森 昭宏（正会員）

1983 年京都大学理学部卒業．1990 年コロラド大学大学院コンピュータサイエンス学科修士課程修了．1983 年（株）インテック入社．2000 年，インテック・ウェブ・アンド・ゲノム・インフォマティクス（株）に移籍．現在，同社技術部アドバンスト・リサーチ・グループ・マネージャ．技術士（情報工学部門）．言語処理学会，人工知能学会，ACL，AAAI 各会員．

奥村 学（正会員）

1962 年生．1984 年東京工業大学工学部情報工学科卒業．1989 年同大学院博士課程修了．同年東京工業大学工学部情報工学科助手．1992 年北陸先端科学技術大学院大学情報科学研究科助教授，2000 年東京工業大学精密工学研究所助教授，現在に至る．工学博士．自然言語処理，知的情報提示技術，語学学習支援，テキストマイニングに関する研究に従事．人工知能学会，AAAI，言語処理学会，ACL，認知科学会，計量国語学会各会員．

丸川 雄三

1998 年東京工業大学大学院理工学研究科修士課程修了．2003 年 3 月東京工業大学情報理工学研究科博士課程修了．2001 年 4 月〜2003 年 3 月東京工業大学精密工学研究所助手．2003 年 4 月〜現在科学技術振興機構戦略的創造研究推進事業（CREST）研究員．

岩山 真（正会員）

1992 年東京工業大学大学院博士後期課程修了．同年（株）日立製作所入社．2000 年より東京工業大学精密工学研究所客員助教授を兼任．博士（工学）．言語処理学会，人工知能学会，ACM 各会員．