(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号
特開2003－196274
(P2003－196274A)
(43)公開日　平成15年７月11日(2003.7.11)

(51)Int.Cl.[7]	識別記号	ＦＩ	ﾃｰﾏｺｰﾄﾞ(参考)
Ｇ０６Ｆ　17/27		Ｇ０６Ｆ　17/27	Ｊ　５Ｂ０９１

審査請求　有　　請求項の数６　　ＯＬ　（全　６　頁）

(21)出願番号　特願2001－395617(P2001－395617)

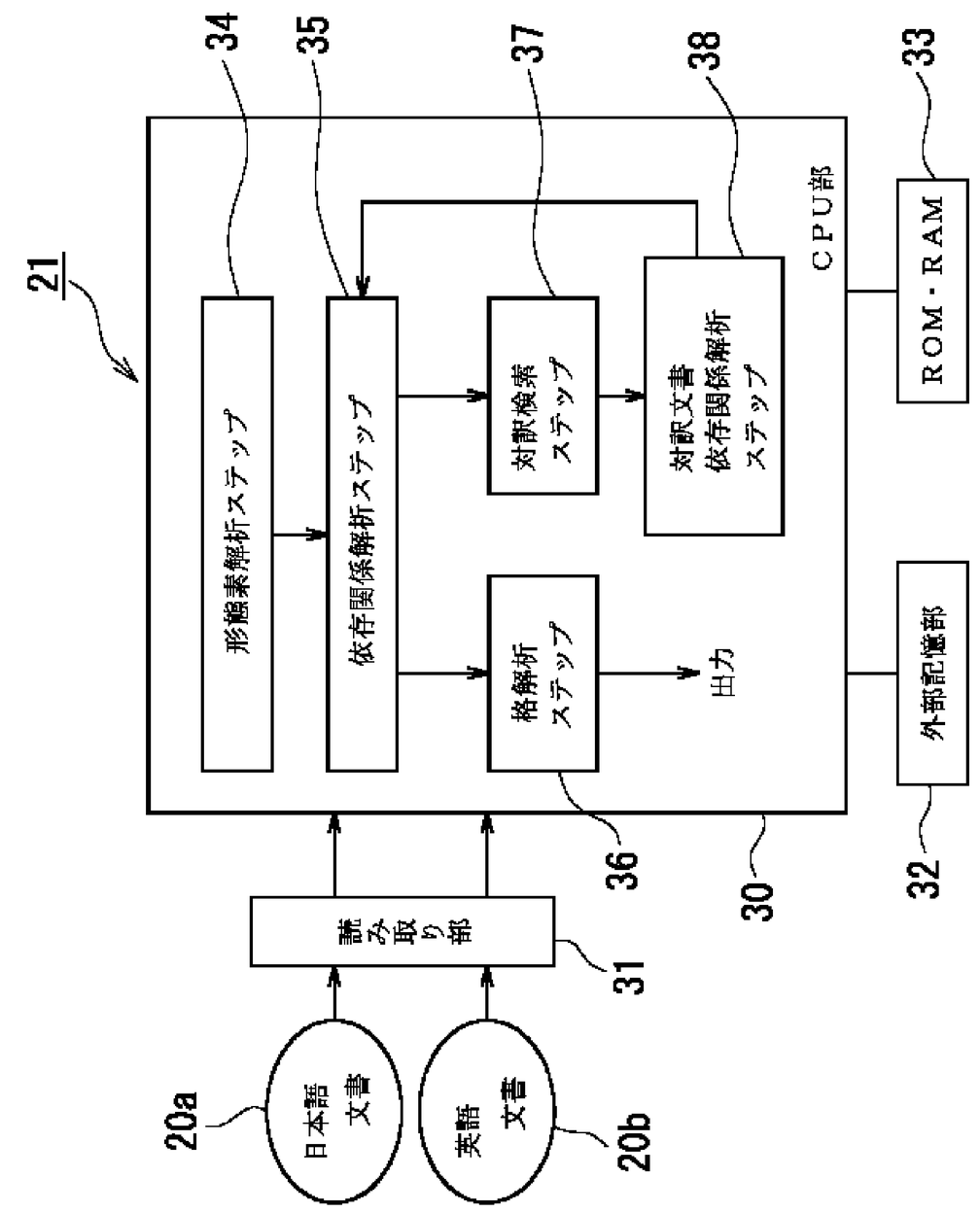
(22)出願日　平成13年12月27日(2001.12.27)

(71)出願人　301022471
独立行政法人通信総合研究所
東京都小金井市貫井北町４－２－１
(72)発明者　内元　清貴
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内
(72)発明者　井佐原　均
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内
(74)代理人　100090893
弁理士　渡邊　敏

Ｆターム(参考)　5B091 CA05

(54)【発明の名称】　構文解析方法及び装置

(57)【要約】
【課題】　精度の高い構文解析方法を提供することによって、より正確な言語処理技術に寄与すること。
【解決手段】　単言語文書２０ａと共に、対訳関係にある対訳文書２０ｂを入力し、単言語文書２０ａの構文解析、例えば依存関係解析３５で複数の解析結果が生じて特定が困難な場合に、対訳文書２０ｂの依存関係３８を調べ、その情報に基づいて最適な依存関係解析３５を行う。



【特許請求の範囲】
【請求項１】言語処理における構文解析方法であって、
構文解析を行う対象テキストと、該対象テキストと少なくともその一部が対訳関係にある１つ以上の対訳テキストとを入力し、
該対象テキスト及び対訳テキストの構文解析を行い、
該対象テキストについて、少なくとも２通り以上の構文解析情報が得られる場合に、
いずれかの対訳テキストの構文解析情報を用い、該対訳テキストの構文解析情報から最も適当な対象テキストの構文解析情報を特定し、
該対象テキストの構文解析結果として出力することを特徴とする構文解析方法。
【請求項２】前記構文解析方法が、
前記対象テキストについて、少なくとも２通り以上の構文解析情報が得られる場合に、
いずれかの対訳テキストにおける語順に係る情報又は、文法的情報又は、省略の有無情報又は、語義情報の少なくともいずれかに基づいて構文解析情報を得、該対訳テキストの構文解析情報から最も適当な対象テキストの構文解析情報を特定することを特徴とする請求項１に記載の構文解析方法。
【請求項３】前記構文解析方法が、
複数の言語テキストを用いて新たな第３の言語テキストを生成する過程に用いられる請求項１又は２に記載の構文解析方法。
【請求項４】言語処理における構文解析を行う装置であって、
構文解析を行う対象テキストを入力する対象テキスト入力手段、
該対象テキストと、少なくともその一部が対訳関係にある対訳テキストを、対訳関係を関連づけながら入力する対訳テキスト入力手段、
該入力された対象テキスト及び対訳テキストに係る形態素解析を行う形態素解析手段、
該形態素解析の結果に対して構文解析を行う構文解析手段、
該対象テキストの構文解析結果において、複数の構文解析結果が得られた場合、又は複数の構文解析結果のうち１つの構文解析結果が所定の確からしさを超えない場合に、対訳テキストの構文解析結果を参照し、
最も適当な対象テキストの構文解析結果を特定する最適結果特定手段、
該最適な結果を出力する構文解析結果出力手段を備えることを特徴とする構文解析装置。
【請求項５】前記構文解析装置の構文解析手段において、
前記対象テキストについて、少なくとも２通り以上の構文解析情報が得られる場合に、
いずれかの対訳テキストにおける語順に係る情報又は、文法的情報又は、省略の有無情報又は、語義情報の少なくともいずれかに基づいて構文解析情報を得、
前記最適結果特定手段が、
該対訳テキストの構文解析情報から最も適当な対象テキストの構文解析情報を特定することを特徴とする請求項４に記載の構文解析装置。
【請求項６】前記構文解析装置が、
複数の言語テキストを用いて新たな第３の言語テキストを生成する装置に具備される請求項４又は５に記載の構文解析装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は言語処理における構文解析の精度を向上させる技術に関するものであり、より詳しくは、複数の言語を入力して構文解析の精度向上を図る技術に関わる。
【０００２】
【従来の技術】近年、コンピュータによって言語のテキストを解析する技術、或いは生成する技術の開発が進み、特にそれらの技術を用いた機械翻訳や、要約システムの提供が図られている。その中で、文章の係り受け関係などを解析する構文解析技術は正確な文脈の把握に極めて重要であり、従来から高精度な構文解析技術の研究が進められている。
【０００３】特に、日本語のように、係り受け関係が曖昧であり、省略される語句も多い言語を解析する際には、解析時に複数の解析可能性が存在し、解析結果が不確定になる場合が少なくない。また、単語には複数の語義があることが多いが、１つの言語を解析しただけでは、いずれの語義で用いられているのかが不明な場合も多い。従来の構文解析技術では、当該解析対象の言語についてより多くの文法情報を与え、それによって解析精度の向上を図るものが多い。しかし、このような手法では、確率的により最適なものが選択できるようになるだけで、必ずしも正しい解析結果を得ることができなかった。
【０００４】
【発明が解決しようとする課題】本発明は、上記従来技術の有する問題点に鑑みて創出されたものであり、その目的は、精度の高い構文解析方法を提供することによって、より正確な言語処理技術に寄与することである。
【０００５】
【課題を解決するための手段】本発明は、上記の課題を解決するために、次のような構文解析方法を創出する。すなわち、本発明では構文解析方法に、従来では解析対象であった１言語のテキストだけでなく、該対象のテキストとは異なる言語による対訳テキストを入力することによって、より高精度な構文解析を実現する。具体的には次の手法による。まず、構文解析を行う対象テキストと、該対象テキストと少なくともその一部が対訳関係に

ある１つ以上の対訳テキストとを入力する。そして、該対象テキスト及び対訳テキストの構文解析を行う。このとき、必ずしもすべての文章について構文解析を行う必要はなく、対象テキストの構文解析を行いながら、必要に応じて対訳テキストの構文解析を行ってもよい。

【０００６】さらに、該対象テキストについて、少なくとも２通り以上の構文解析情報が得られる場合、つまり対象テキストの構文解析の結果、複数の解析情報が得られていずれが最も適当な解析情報であるかの判別が困難な場合には、上記対訳テキストの構文解析結果を用いる。対訳テキストが複数ある場合には、例えば最も確からしい解析情報が得られた対訳テキストの情報を用い、それを複数の対象テキストの構文解析情報のうちから最も適当な対象テキストの特定に利用する。特定された結果を対象テキストの適当な構文解析結果として出力することによって、従来の１言語のみによる解析では困難な構文解析も、高精度な解析結果を得ることができるようになる。

【０００７】さらに、本発明における前記構文解析方法は、複数の言語を入力して新たな第３の言語を生成する過程に導入することもできる。従来ある言語から第３の言語を生成するとき、単一の言語を入力するよりも、複数の言語を用いて生成するとより精密な生成が可能であることが分かっている。すなわち、本件出願人らによる特願２００１－２４３１１８号公報に開示される方法などにおいては、すでに複数の対訳テキストが入力として用いられることから、本発明の実施には極めて好適であり、構文解析方法の高精度化により、第３言語生成にも有効である。

【０００８】また、本発明では言語処理における構文解析装置を提供することもできる。本装置には、まず構文解析を行う対象テキストを入力する対象テキスト入力手段と、該対象テキストと、少なくともその一部が対訳関係にある対訳テキストを、対訳関係を関連づけながら入力する対訳テキスト入力手段を備える。入力された対象テキスト及び対訳テキストは、形態素解析手段によって形態素解析を行う。該形態素解析の結果に対しては、構文解析手段が、対象テキストの全形態素及び、対訳テキストの少なくとも必要な形態素の構文解析を行う。該対象テキストの構文解析結果において、複数の構文解析結果が得られた場合、又は複数の構文解析結果のうち１つの構文解析結果が所定の確からしさを超えない場合に、対訳テキストの構文解析結果を参照し、最も適当な対象テキストの構文解析結果を特定する最適結果特定手段を有する。本装置は、構文解析結果出力手段を用いて最適な結果を出力する。

【０００９】

【発明の実施の形態】以下、本発明の実施方法を図面に示した実施例に基づいて説明する。なお、本発明の実施形態は以下に限定されず、適宜変更可能である。本発明は、従来の構文解析技術では困難であった的確な構文解析を実現する技術であり、人手により作成された高精度の複数の対訳文書、例えば日英２つの言語を用いて、極めて高精度な構文解析技術を提供するものである。

【００１０】以下、本発明の利用形態の１つとして、構文解析を行う対象言語文書と共に、その対訳関係にある言語の文書を入力し、最終的に目標言語を生成して出力する翻訳システムに実装した場合を例として説明する。図１に従来から行われている単言語文書を目標言語に変換、生成するフローチャートを、図２に本発明に係る日米対訳文書から目標言語に変換、生成するフローチャートを示す。

【００１１】従来の方法において、単言語文書（１０）を目標言語文書（１４）に翻訳するプロセスは、大きく分類して、構文解析装置（１１）、変換装置（１２）、生成装置（１３）を経て行うのが一般的であった。それら各装置（１１）（１２）（１３）の開発に当たっては、人手による規則の作成（１５）が不可欠であって、高精度なシステム開発には大規模な文書の解析作業が必要であった。たとえば、学習に用いる大規模なテキストコーパスは莫大なコストと、研究が必要であり、現状では主要言語のみにようやく整備されつつあるものの、非主要言語において用意される望みは極めて薄い。

【００１２】そこで、図２に示すように、主要言語等のコーパスが整備された単言語文書（２０ａ）と共に、該単言語文書と対訳関係にある対訳文書（２０ｂ）を用い、目標言語の的確な翻訳を実現する翻訳システムが提供される。本システムでは、２つ以上の対訳テキストを入力する入力手段（図示しない）によって文書の入力を行い、各対訳テキストにつき、各言語毎に、又は各言語を任意に２つ以上組み合わせて、言語情報の解析を行う解析手段として本発明による構文解析装置（２１）に至る。

【００１３】さらに、構文解析装置（２１）における解析結果に基づき、第３言語への言語変換を行う変換手段として変換装置（２２）、該変換ステップにおける変換結果に基づき第３言語によるテキストを生成する生成手段として生成装置（２３）を備える。変換装置（２２）、生成装置（２３）にはそれぞれ変換・生成に必要な情報が変換知識（２５）、生成知識（２６）として備えられている。最終的には、別に配設する出力手段（図示しない）によって目標言語文書（２４）が出力可能である。

【００１４】入力する言語は、例えば日本語と英語の対訳関係にある文書である。この際、その全部が完全な対訳関係にある場合だけでなく、一部が対訳関係にある文書でもよい。また、入力する言語は２つ以上であればよく、例えば３言語によってより高精度な構文解析を実現することもできる。

【００１５】従って、本発明における対訳文書の言語の

組み合わせとしては、日本語と英語や、日本語と中国語、或いはその３言語を用いるなど、言語体系が異なる言語を用いると特に好適である。逆に、英語とフランス語のみ等では本発明による効果は必ずしも大きくないが、英語・フランス語・日本語のように組み合わせると、英語・日本語のみの場合よりも高精度な解析が行える可能性が高く、そのような構成でもよい。

【００１６】次に、本発明に係る構文解析装置（２１）につき詳述する。本システムは、日英二言語の対訳文書（２０ａ）（２０ｂ）の入力を前提に、語と語(あるいは日本語の文節のようにもう少し大きい単位)の間の依存関係(係り受け関係）を解析する。依存関係はすでに本件出願人らが提案している日本語の係り受けモデル（内元清貴、村田真樹、関根聡、井佐原均、「後方文脈を考慮した係り受けモデル」、自然言語処理，Vol.7，No.5，pp.3-17（2000)、に記載）を他言語にも適用することによって決定する。

【００１７】このモデルは、二つの語(あるいは文節)が依存関係にあるかないかを学習するもので、機械学習モデルを用いて実現される。依存関係は学習されたモデルによって計算される確率の積が一文全体で最も高くなるように決定する。さらに、この依存関係構造から格解析(意味解析)を行う。依存関係の処理においては、二言語対訳入力の有効性は、依存構造における係り受けの正解率の向上で計量可能である。

【００１８】図３に本発明による構文解析装置の構成図を示す。本装置（２１）は、ＣＰＵ部（３０）、読み取り部（３１）、外部記憶部（３２）、ＲＯＭ・ＲＡＭ部（３３）から構成され、ＣＰＵ部（３０）における以下の処理を、ＲＯＭ・ＲＡＭ部（３３）が適宜記録しながら行う。構文解析の結果は、外部記憶部（３３）に出力されて蓄積し、変換装置（２２）における処理に向かう。

【００１９】ＣＰＵ部（３０）においては、まず形態素解析ステップ（３４）において入力される単言語文書（ここでは日本語文書）（２０ａ）と、対訳文書（ここでは英語文書）（２０ｂ）の形態素解析を行う。そして、形態素解析の結果に基づいて日本語文書（２０ａ）における語と語の依存関係を解析する。（依存関係解析ステップ（３５））依存関係解析ステップ（３５）における解析結果が、１個の解析結果を生じる場合、又はその解析結果が上記機械学習によって一定の確からしさを示している場合には、格解析ステップ（３６）において格解析を行い、その結果を外部記憶部（３２）に蓄積する。

【００２０】しかし、一般に単言語文書の入力だけで、確実な依存関係を決定することは難しい。依存関係解析ステップ（３５）において、特に重要な情報となるのが語順であり、例えば「私は少女と犬を見た。」という日本語を入力する場合には、「私」が「少女と犬を見た」とも解釈できるし、「私」が「少女」と共に「犬を見た」とも解釈できる。そこで、本発明では英語文書中の該当対訳部分を解析し、いずれの解釈が適当であるかを決定する。

【００２１】上記依存関係解析ステップ（３５）で、複数の解析結果が生じ、いずれの解析が適当であるかが判別出来ない場合、日本語文書（２０ａ）の当該文章に該当する個所を英語文書（２０ｂ）から検索する対訳検索ステップ（３７）に進む。そして、対訳文章が検索された場合、該文章についての依存関係を解析する。（対訳文書依存関係解析ステップ（３８））

【００２２】上記例文の場合、対訳として検索された文章「I saw a girl and a dog.」によれば、「私」が「少女と犬を見た」という前者の解釈が妥当であることが極めて容易に判明する。それは、後者の「私」が「少女」と共に見た場合には、対訳の文章が「I and a girl saw a dog」という語順でなければならず、検索された文章では取りえない解釈だからである。このように対訳文書における依存関係の情報を、依存関係解析ステップ（３５）にフィードバックすることで、従来では困難であった依存関係の正確な解析が可能となる。

【００２３】日本語文と英語文は語順がかなり異なる上、英語では語順に対する文法的な制約が厳しいため、日本語文側では曖昧な係り先が英語文側では明らかに決まる場合やその逆の場合が多くある。上記の例の通り、対訳文が「I saw a girl and a dog.／私は少女と犬を見た。」の場合、英語文では「and a dog」の係り先は明らかに「saw」であるが、日本語文では「少女と」が「見た」に係るのか、並列句として「犬を」に係るのかが曖昧である。

【００２４】また、対訳文が「I saw a girl with a telescope.／私は望遠鏡で少女を見た。」の場合、逆に、英語文では「with a telescope」の係り先が「saw」なのか「a girl」なのかが曖昧であるが、日本語文では「望遠鏡で」が「見た」に係るということは容易に解析できる。後者の例では、単言語文書として英語を入力した場合に、日本語の対訳文書の入力が有効に作用することを示している。

【００２５】さらに、語順以外にも文法的な情報、例えば、英語では冠詞や単複形、動名詞や不定詞など動詞の活用形の情報、日本語では助詞の情報も有効に用いることができる。例えば、日本語文で「彼は本を書き、出版している人を尊敬している。」では、「本を書」いているのが「彼」なのか「出版している人」なのか曖昧である。しかし、対訳文「He respects people who write books and publish them.」を入力することによって、文法的にwho以降の動詞はいずれもpeopleに係る(三単現のsがないため)ことが分かり、「本を書」いているのは「出版している人」であることが正確に解析できる。

【００２６】さらに、省略の有無の情報を用いることも

できる。日本語文では主語が省略されることが多い(ゼロ代名詞がよく使われる)が、英語文では主語が必須の場合が多いため、省略のため曖昧になっている部分を英語側の情報で補うことも可能になる。これは特に格解析で主語を特定する必要がある場合に有効に用いることができる。

【００２７】例えば、日本語文「友達とレストランに行きました。有名人に会えてラッキーでした。」の場合、ラッキーだったのは話者の私なのか、友達なのか、両方なのか、また、有名人は一人だったのか複数いたのか曖昧である。この日本語文の対訳が「I went to the restaurant with my friend. We were lucky because wemet a celebrity.」であれば、二人ともラッキーで、会えたのは一人の有名人ということが分かる。

【００２８】語義の曖昧性が相手側の言語で解消されて、構文的な係り受けの曖昧性の解消につながる場合も考えられる。例えば、英語文を対象言語とし、日本語を対訳で入力する場合を考える。「He saw a girl laughing at the second story.」という英語文の場合、彼女が二番目の話を聞いて笑っているのか、彼が二階で見たのか、二階で笑っている少女を見たのか、つまり、「at the second story」が「laughing」に係るのか「saw」に係るのか、曖昧である。ここで対訳が「彼は二番目の話を聞いて笑っている少女を見た。」であれば、「story」は「階」を意味するのではなく、「話」を意味することから、「laughing」に係っていることが正確に解析できる。

【００２９】上記例とも関連するが、対訳の情報は構文解析だけでなく、語義の曖昧性解消にも寄与する。単純な例として、英語の「bank」の曖昧性に着目する。英語の「bank」の語義には「銀行」と「土手」という意味の曖昧性がある一方、日本語の「銀行」と「土手」という単語は同じ意味として使われることはない。そこで、「bank」が対訳の日本語文側にどちらの単語として現れているかを調べれば容易に曖昧性を解消することができる。そして、語義の曖昧性が相手側の言語によって解消されることによって係り先が容易に決まることによって、正確な構文解析に寄与することもできる。

【００３０】本発明では、以上に示したように従来の単言語文書を構文解析する手法に、対訳文書の入力を行うことで、極めて高精度な構文解析が行えることを編み出し、新しい構文解析装置の創出を行った。特に、語順の制約が緩やかな言語と厳しい言語では、厳しい言語による語順を解析することによって、緩やかな言語で複数の解析結果が得られた場合に、該解析結果中、制約の厳しい言語で認められ得る解析結果を採用すればよく、簡便に、かつ高精度に構文解析を行うことができる。

【００３１】

【発明の効果】本発明は、以上の構成を備えるので、次の効果を奏する。請求項１ないし３に記載の構文解析方法では、従来複数の解析結果が得られた場合に困難であった最適な構文解析結果の特定が可能となり、高精度な構文解析方法に寄与する。特に、日本語などで語順の制約の緩さから、複数の解釈が可能な文章について、従来の手法では膨大な知識を備えて、より確からしい解釈を行うほかなかったが、本発明によると語順の制約が厳しい言語を対訳文書として入力することにより、適当な解釈の採用が可能となる。

【００３２】さらに、本発明では語順以外にも文法的な情報も有効に用いることができるので、例えば日本語文では主語が曖昧な場合にも、英語の単複形から的確な特定が可能となるなど、解析精度の向上に寄与することができる。他に、省略の有無の情報を用いることもできる。日本語文では主語が省略され格解析で主語を特定する必要がある場合に、従来の単言語のみの解析ではその予測が困難であったが、本発明によれば、主語を英語文を参照することによって正確に特定できるため、解析精度が向上する。単言語のみの入力では、１つの単語に複数の語義がある場合は少なくなく、従来の構文解析方法では、しばしば誤った語義の認識から、誤った解析が行われる例が見られた。この点についても、本発明では対訳から的確な語義の特定が可能となり、構文解析精度の向上を図ることができる。

【００３３】以上の方法は、実際にしばしば存在する対訳文書を用いるだけで正確な構文解析が可能となるので、構文解析中に人手を介在させて最適な解析結果を選択するよりも極めて簡便であり、構文解析ひいては言語処理の自動化の要請にも応えるものである。

【００３４】請求項４ないし６に記載の構文解析装置によると、対訳関係にある２つ以上の文書を入力することで、形態素解析、依存関係解析や格解析等の構文解析を自動的に行い、例えば依存関係で不明な場合には対訳関係の文書を解析し、その結果から適切な依存関係を決定することができるので、従来の構文解析装置を置換可能な高精度な構文解析装置を提供することが出来る。本発明は、すでに提供されている複数の対訳言語を入力することで第３言語を生成する翻訳システムに導入することで、特に効果的である。

【図面の簡単な説明】

【図１】従来の技術における、単言語文書を目標言語に変換、生成するフローチャートである。

【図２】本発明の構文解析装置を導入するのに好適な翻訳システムのフローチャートである。

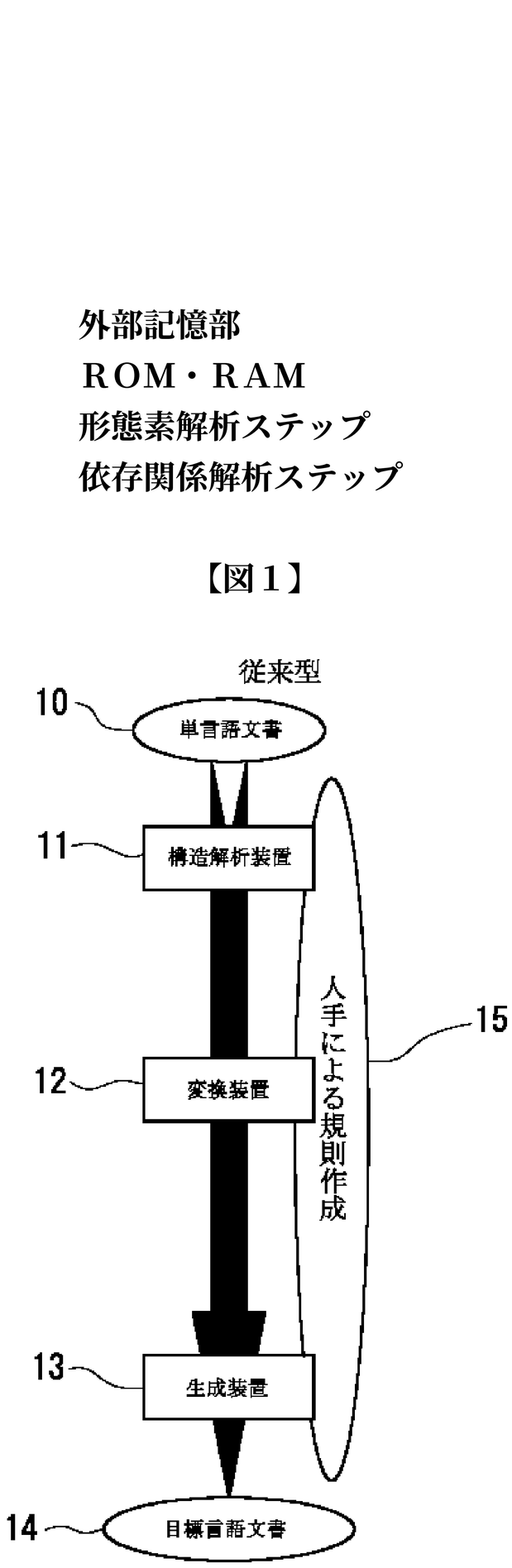
【図３】本発明における構文解析装置の構成図である。

【符号の説明】

２０ａ	単言語文書
２０ｂ	対訳文書
２１	本発明に係る構文解析装置
３０	ＣＰＵ部
３１	読み取り部

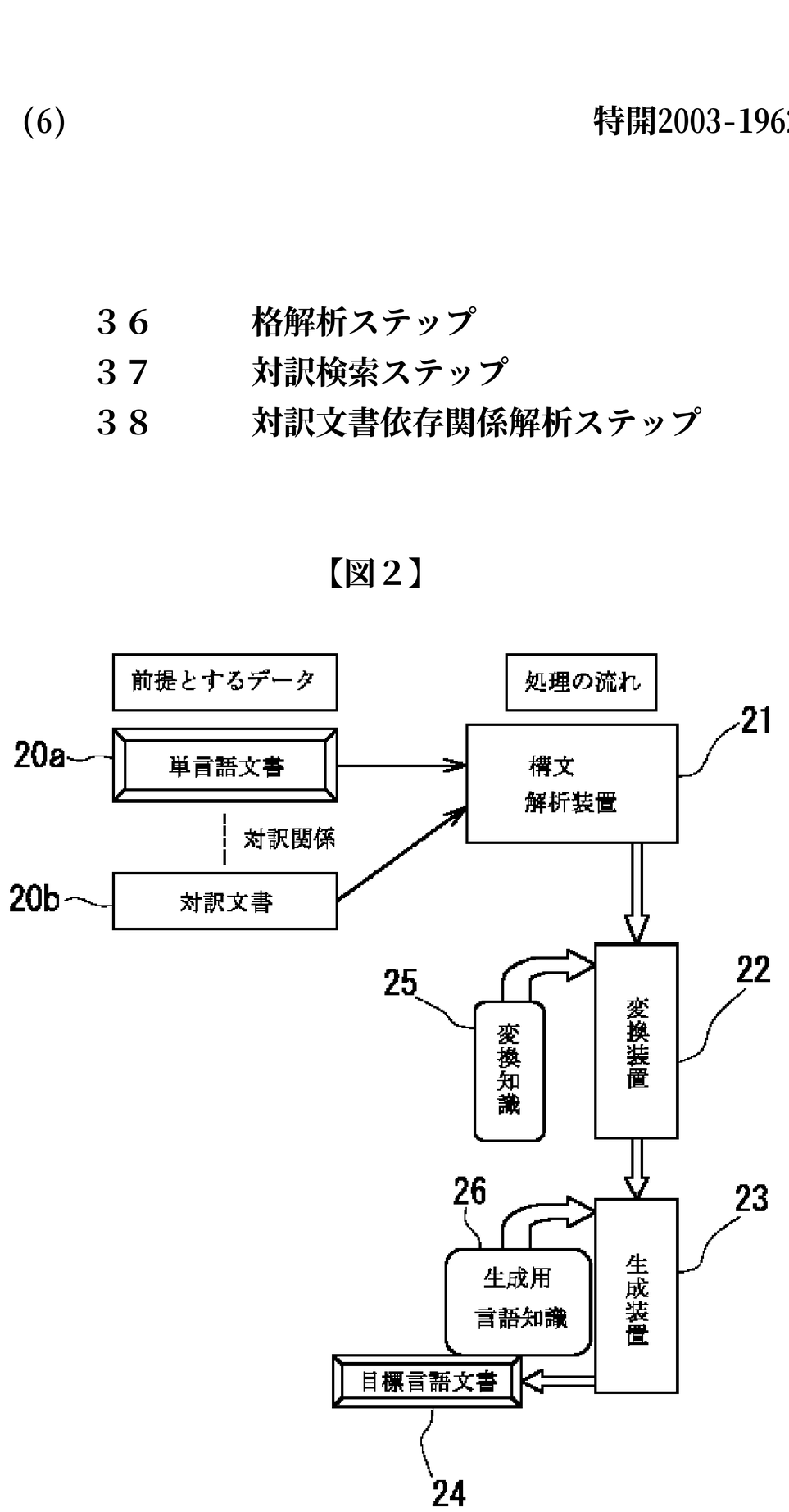
３２　　外部記憶部
３３　　ＲＯＭ・ＲＡＭ
３４　　形態素解析ステップ
３５　　依存関係解析ステップ

【図１】



３６　　格解析ステップ
３７　　対訳検索ステップ
３８　　対訳文書依存関係解析ステップ

【図２】



【図３】