U.D.C. 519.768.4:〔802.0:809.56〕

機械翻訳システムの原理と実際

Principle and Practice of Machine Translation System

岡島　惇*　*Atsushi Okajima*
新田義彦*　*Yoshihiko Nitta*
山野文行*　*Fumiyuki Yamano*

近年，海外との情報交換の量が飛躍的に増大しつつあり，言語の障壁を打破するための機械翻訳システム実現の要請が高まっている。また，最も人間的な所産といわれる自然言語を，計算機がどのように処理するのかは，多くの人々の関心を集めるものであろう。

本論文では，自然言語処理の一例として，機械翻訳システムを取り上げ，その動作原理，技術の要点及び実用化の課題について，実例に即して述べる。

実例として取り上げるものは，日立製作所で開発した英和機械翻訳実験システムであり，拡張された句構造を中間語とする言語モデルと，人間の翻訳過程に近いヒューリスティックな翻訳方式に特徴がある。

1　緒　　言

情報化社会の進展に伴い，我々人間が処理しなければならない情報(そのほとんどは自然言語で書かれている。)は，日増しに増加している。更に，政治・経済・商業などの分野での国際化の進展に伴い，外国語で書かれた文書を読む作業も膨大なものになりつつある。このような海外との情報交換での言葉の障壁を打破する決め手として，機械翻訳実現の要請が強まっている。

一方，計算機の発展は，会計処理といった数値処理から，ワード プロセッサに代表されるような非数値処理へとその利用可能な範囲を拡大しつつあり，その最先端技術として，自然言語理解や推論といった人間特有の能力にまで迫ろうという研究が，第5世代コンピュータ開発といった国家的規模で行なわれようとしている。

以下，本論文では，自然言語理解のうち特に機械翻訳システムについて，その動作原理と技術の要点を実例にそって述べるとともに，真の実用化のために残されている課題について報告する。

2　機械翻訳の原理

翻訳とは，ある文法に従って作られている1次元の文字の並びを解析し，その結果から再度異なる文法に従った1次元の文字の並びを作り出すことである。特に，言語間に隔りの大きい欧米語と日本語では，同じ意味の文を全く異なる表現で表わすことも多く，次の四つのことが大きな問題となる[1),2)]。

(1)　構文解析法

定められた文法に従って入力された文を，どのように解析するか。

(2)　中間語表現法

構文解析した結果を，どのように表現・格納しておくか。

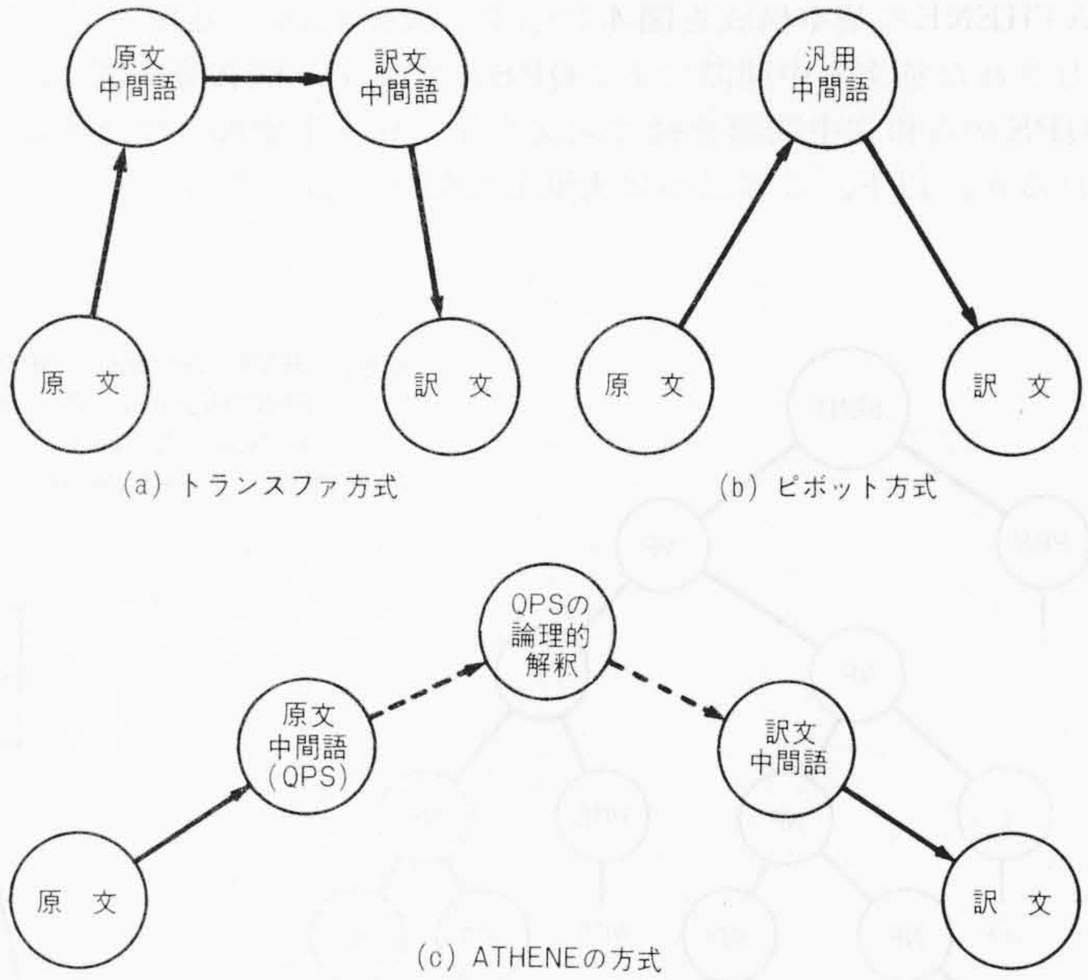
(3)　訳文の生成法

中間語から，どのようにして1次元的な出力文(訳文)を得るか。

(4)　辞書の形態

構文解析や訳文の生成のためにどのような辞書をもつか。

機械翻訳で一般的にとられる方式を**図1**に示す。同図(a)の方式は，二つの言語間の相違を重視して，各々の言語に適した文法で構文解析と訳文生成を行ない，その間のギャップは，二つの言語の中間語の間の変換(これを文型変換と呼ぶ。)により吸収する方式で，トランスファ方式と呼ばれる。(b)は，汎用的な「文の意味表現」を設けて，これを中間語とするピボット方式であり，これら2種類が代表的な機械翻訳の方式である。(a)，(b)を比較すると，(a)は言語特有の細かな規則の記述に適しているが，「文の意味」を論理的にとらえる点で弱く，(b)はこの逆に論理的に文をとらえられるため，推論などの高度な知的処理に適しているが，翻訳としては粗い訳文しか得られない危険性がある。



注：略語説明
QPS(Quasi Phrase Structure)
ATHENE(Automatic Translator of
Hitachi for English into
Nihongo with Editing Support)

図1　機械翻訳の方式　機械翻訳の方式を大別すると，(a)のトランスファ方式と(b)のピボット方式がある。ATHENEは二つを融合した方式と考えられる。

* 日立製作所システム開発研究所

3 ATHENEでの機械翻訳の方式

3.1 全体方式[3),4)]

ATHENE (Automatic Translator of Hitachi for English into Nihongo with Editing Support)は，英和翻訳システムであり，その基本方式は，**図1**(c)のように，トランスファ方式の2国語間の文型変換を基本とし，更にピボット方式の論理性を目指したものである。中間語は，トランスファ方式の基本である句構造表現(例えば，冠詞＋名詞で名詞句となるといった，句生成規則を木の形で表現する。)を拡張したQPS(Quasi Phrase Structure：擬似的句構造)を用いている[5)]。

QPSは，次の二つの要素から成っている。

(1) 木構造：主語，主動詞，目的語などの必須格的要素

(2) リンク構造：副詞，前置詞句などの任意格的要素

図2に，両方式による表現例を示す。

QPSの特徴は，解析木の段数の少ないこと，論理表現に素直に変換しやすいこと，解釈の多義性(あいまいさ)を見やすく表現できることにあり，ATHENEは，QPSを論理的な中間語とする擬似的なピボット方式と考えることもできる。一方，ピボット方式では対処が難しい言語固有の表現間の違いに対しては，従来，イディオムと呼ばれているものを拡張した「拡張イディオム表現」を設け，これをユーザーが辞書に記述できるようにすることによって，できるだけ汎用的な処理方式で，言語固有の記述も可能であるという，アルゴリズム(プログラム)と辞書記述とのバランスの良いシステム作りをねらっている。**図3**に拡張イディオムの記述形式と幾つかの例を示す。この例で分かるように，ユーザーは単なる単語のつながり(連語的イディオム)の記述だけでなく，英語の構文そのものに近い部分までも記述できることが分かる。ATHENEの基本構成を**図4**に示す。機械翻訳の処理は，入力された英文を中間語であるQPSに変換する構文解析部と，QPSから和文中間語を経て訳文を作る和文生成部とに分けられるが，以下，この二つに大別して処理を説明する。

```
〈イディオムパターン〉 ::= S α S
α ::= ＊ | ＋ | ＊P | ＊PP | ＊SG | ＊ING
S ::= 〈品詞記号列〉 | 〈単語〉 | S S
〈品詞記号列〉 ::= ！〈品詞記号〉〔〈細分類〉〔,〈属性1〉
〔,〈属性2〉〔,〈属性3〉〔,〈属性4〉〕〕〕〕
```

(a) 拡張イディオムの記述形式

(I) 連接型レベル(主に名詞句などで，連接しているもの)
the United Nations, New York, make sense など

(II) 分離型レベル(主に動詞句などで，制限語句の介在を許すもの)
bring～to an end, half past～ など

(III) 節構造レベル(節間にまたがり構文に関与するもの)
so～ that…, too～ to……など

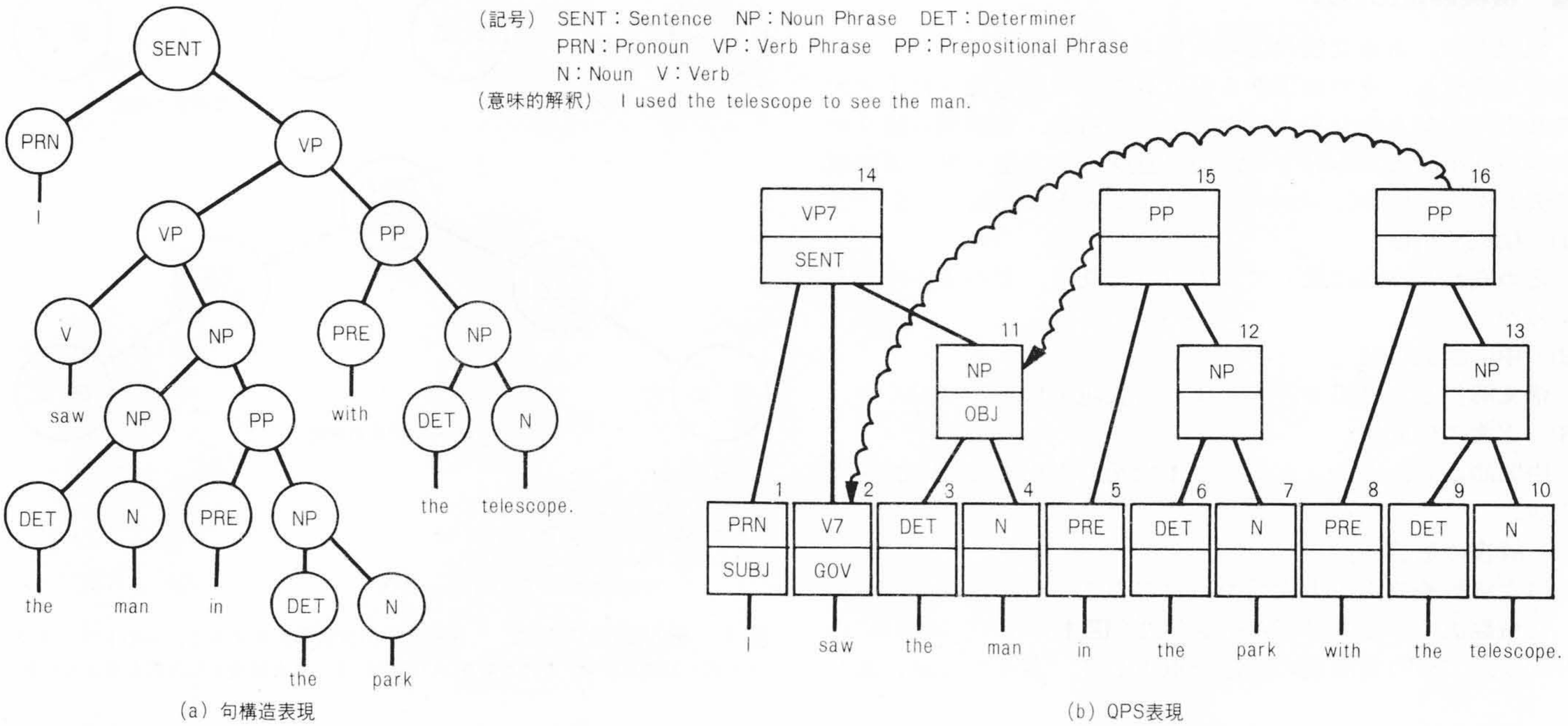
(b) 拡張イディオムの例

図3 ATHENEでのイディオムの記述方式 ATHENEでは，(a)のような記述形式により，(b)(III)のような節構造までもイディオムとして記述できる。

3.2 構文解析方式

構文解析部は，対象となる英文を受けとると，これを単語に切り辞書検索を行なう。また，動詞や名詞，形容詞といった活用をするものについては，形態素解析を行なって標準表現(例えばstudiesならstudy＋3単現)に帰着させる。次に多品詞解消を行ない，それ以降の解析であまり多くの解析木を作り出さないように，可能性のない品詞を削除する。例えば，studyには動詞と名詞の二つの可能性があるが，周りの品詞やその他の情報(これらを品詞細分や属性の形でもっておく。)から，どちらか一方は可能性なしと判定することができる。次に，品詞や属性の情報に従って，句を認識する。ここで句とは，関係代名詞節などを含まない，いわゆる単純句と呼ばれるものであり，先の**図2**の例文でいえば，NPとPPまでを作り出すものである。

次に節構造の解析を行なう。ここで節とは，述部(その中



(a) 句構造表現　　(b) QPS表現

図2 従来の句構造表現とQPS表現の違い QPSは従来方式に比べ，解析木の段数の少ないこと，論理表現への変換がしやすいこと，多義性(あいまいさ)を見やすく表現できる点に特徴がある。

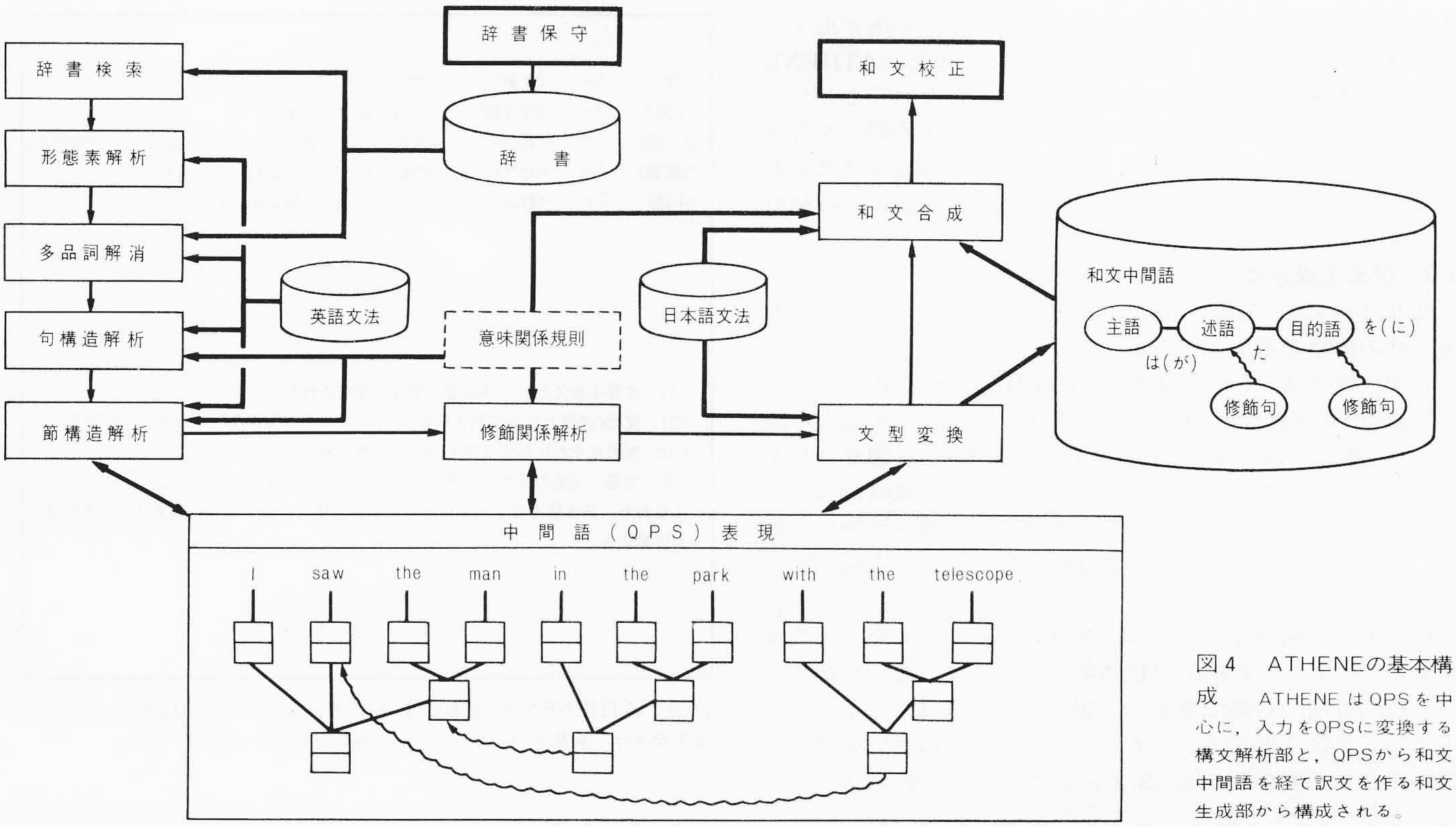


図4　ATHENEの基本構成　ATHENEはQPSを中心に，入力をQPSに変換する構文解析部と，QPSから和文中間語を経て訳文を作る和文生成部から構成される。

心は動詞)を含むものであり，関係代名詞節や従属文の分詞(ing節)などである。節解析は動詞を軸として行なわれ[6]，この際，主語，目的語，補語といった各要素の役割(これを役割子と呼ぶ。)も認識される。先の**図2**でいえば，PRN，V7，NPの三つからVP7(SENT)が作られる過程である。この際，節が目的語になるといった再帰的構造も認識・処理される。また，ある動詞が自動詞でも他動詞でもあるといったことによる解析木の多義性も解消する。

次に，修飾関係の解析を行なう。修飾関係としては，副詞や前置詞句のほか，関係代名詞節やing節などがあり，この場合も前置詞句の中に節があるという，再帰的構造を含めた解析が必要となる。

句・節の関係は，QPSにより木で，また，修飾関係はリンク(先の**図2**では波線)で各々表示される。

修飾関係の解析が終わると，文全体の解析が終了する。

修飾関係の解析は，非常に多くの多義性の中から正解を見つける操作となる。例えば，先の**図2**ではwith the telescopeという前置詞は，1，2，11，12，14のどれにも係り得る。

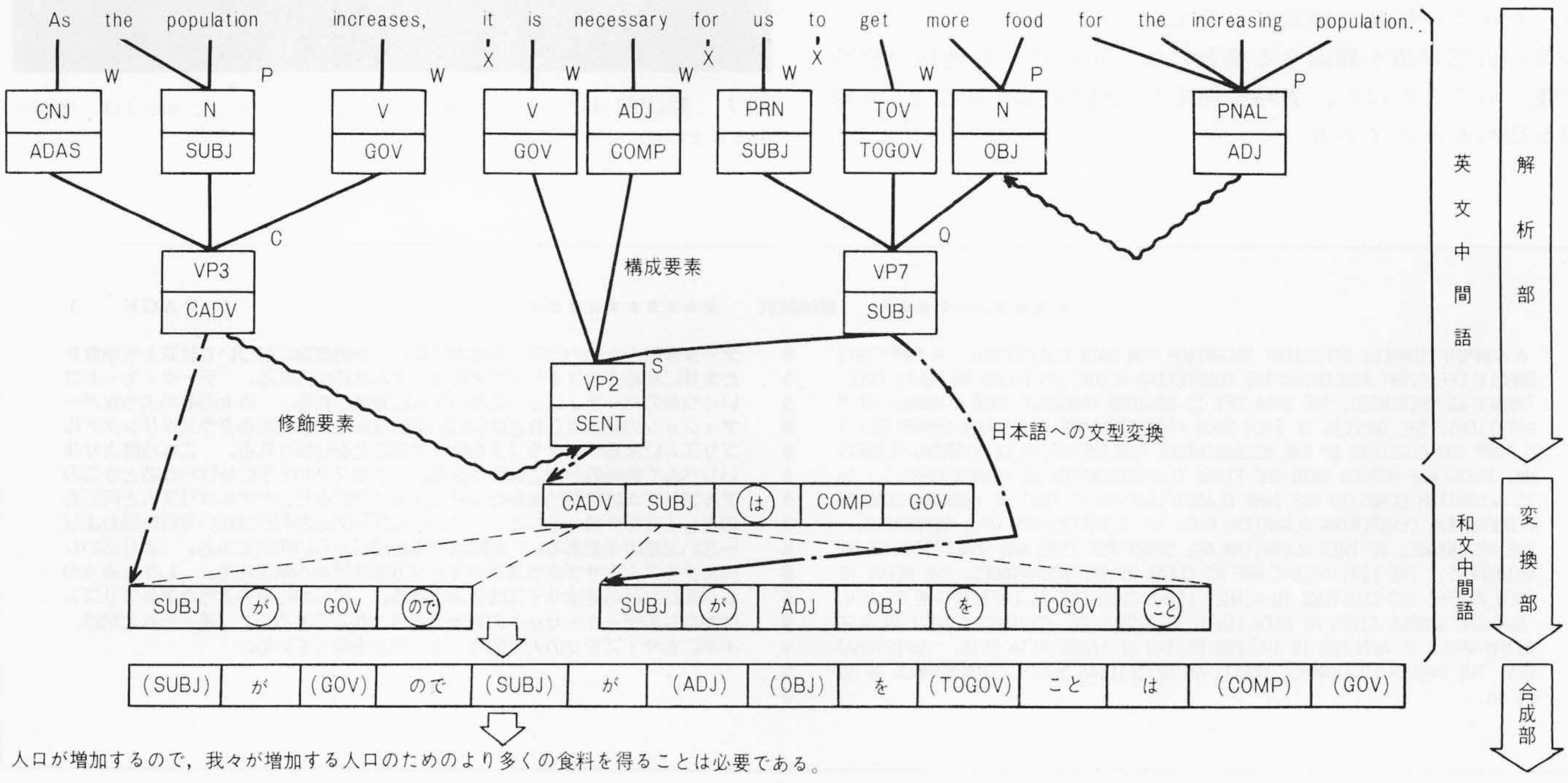


図5　ATHENEの和文生成処理の概略　和文生成はQPSから和文中間語を経て，再び一次元的な訳文を作る過程である。

この多義性の中から文の意味として最適のものを選び出すのは，構文的にはほとんど不可能である。このためATHENEでは，意味関係規則によって多義性解消を行なおうとしている。ただし，現在のものは文法的な関係に意味的制約を付与した規則群によるものであり，いわゆる人工知能などでいわれる常識ベースといった，文法以外の知識によるものの利用にはまだ至っていない。

3.3 訳文生成方式

図5に複文で，itの特殊構文を含むものについて，和文生成処理の概略を示す[7]。

和文生成では，まず中間語であるQPSを和文訳出のための中間語(和文中間語と呼ぶ。)に変換する。次にこの和文中間語から日本語文法に従って自然な語順を生成する(**図6**に和文生成のための日本語文法のモデルと制約規則の概略を示す)。

次に，役割子に従って助詞などの付属語情報を付加し，更に，動詞の活用形などに従った様態・時制などの表現に対し適切な和文を合成する。

和文生成で問題となるのは，英日両言語間の差をどう吸収するかということである。(1)語順については，文型変換という一般的方法で対処できるし，(2)単語対応のイディオム的表現に対しては，拡張イディオムで対処する。和文生成部で，特に注意が必要なものは，(3)2言語間のずれと(4)訳語のずれ及び多義の二つである。このうち(3)は，和文の言いまわし(ごろ)といった，構文的に解決してもかなり高度な処理であり，(4)は，意味処理の中で解決すべき問題であり，共に今後の大きな課題である。

4 翻訳の例

ここでは，ATHENEの実際の出力例を**図7**に示す。また，**図8**には科学技術文に対する翻訳例を示す。to不定詞の多義性など難しいものもあるが，全体の大意はとれると思われる。**図9**には，翻訳システムが入力された英文をどのように解析したかを表示した例を示す。表示は，CRT(Cathode Ray Tube)画面上にも，またプリンタにも出力することができる。更に**図10**には，**図9**の例文を句単位の表示を含めて，プリンタに出力した例を示す。上記の例のような比較的簡単な単文ならば，かなりの精度で翻訳が可能である。しかし，入力に制限のない自然言語を翻訳する場合には，句読点や並列句の解析が難しいことのほか，省略や照応(itが何を指すかなど)の難問が待ちかまえている。

```
〈文〉      :=  〈文節〉  |  〈文〉〈文節〉
〈文節〉    :=  〈自立語〉  |  〈自立語〉〈付属語〉
〈自立語〉  :=  〈名詞〉  |  〈用言〉  |  〈副詞〉  |  〈連体詞〉  |  〈接続詞〉
〈付属語〉  :=  〈助詞〉  |  〈助動詞〉  |  〈付属語〉〈付属語〉
〈用語〉    :=  〈動詞〉  |  〈形容詞〉  |  〈形容動詞〉
```

(a) 日本語文法モデル

(1) 文節を単位とした係り受け関係で構成される。
(2) 文節(述語となる文節を除く)は，それ以外の文節のいずれか一つに係る。
(3) 文節はそれ以外の文節を0個以上受ける。
(4) 文節と文節の係り受け関係は，他の係り受け関係と交差しない。

すなわち，日本語文は主語・述語関係の係り受け，及び依存修飾関係で表わせると仮定する。

(b) 日本語文法モデルの制約規則

図6 ATHENEでの日本語文法モデルとその制約規則 日本語文法モデルは，(a)基本的な文法モデルと(b)制約規則から成る。

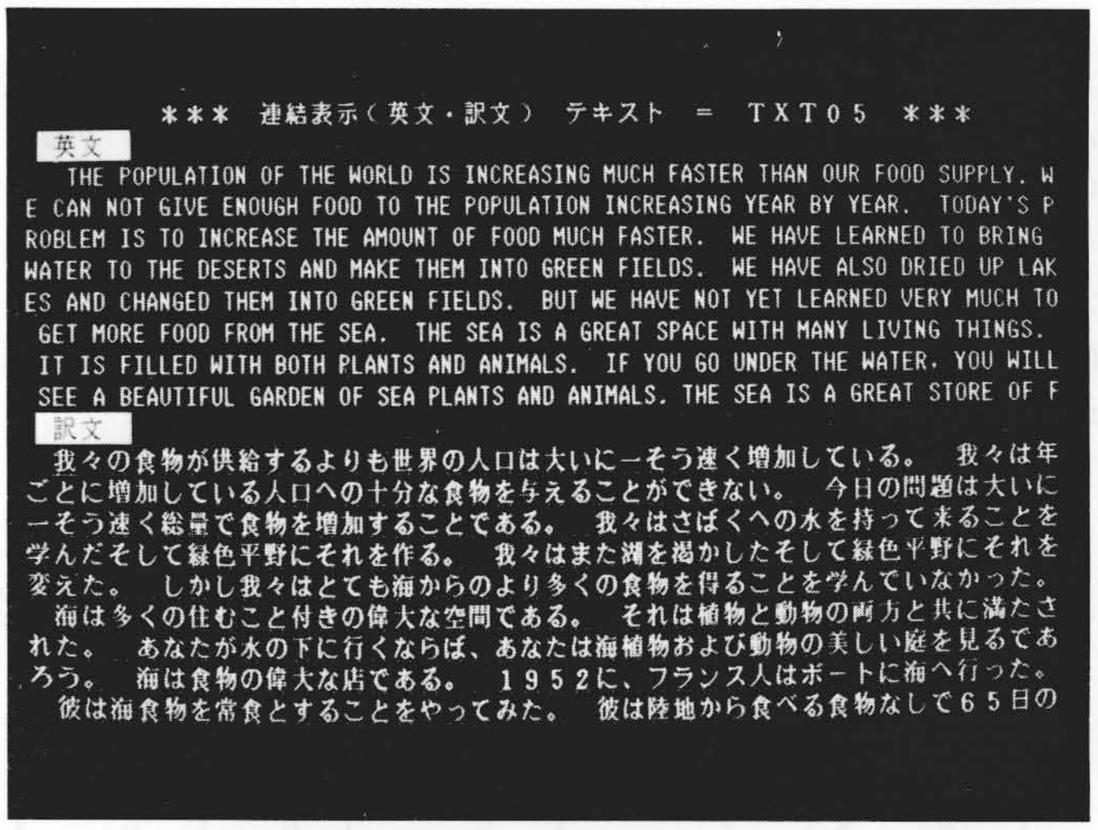


図7 翻訳例(1) 1952に「年」が入らないといったことを除けば，ほぼ訳出できている。

```
☆☆☆☆☆☆☆☆☆☆  機械翻訳  ☆☆☆☆☆☆☆☆☆☆  PAGE 1

A COMPUTATIONALLY EFFICIENT TECHNIQUE FOR DATA-CLUSTERING. A COMPUTATIONALLY EFFICIENT AGGLOMERATIVE CLUSTERING ALGORITHM BASED ON MULTILEVEL THEORY IS PRESENTED. THE DATA SET IS DIVIDED RANDOMLY INTO A NUMBER OF PARTITIONS. THE SAMPLES OF EACH SUCH PARTITION ARE CLUSTERED SEPARATELY TO FORM SUB-CLUSTERS BY THE HIERARCHICAL AGGLOMERATIVE CLUSTERING ALGORITHM. THESE ARE MERGED INTO THE FINAL CLASSIFICATION AT HIGHER LEVELS. THIS ALGORITHM LEADS TO THE SAME CLASSIFICATION AS THAT OF HIERARCHICAL AGGLOMERATIVE CLUSTERING ALGORITHM WHEN THE CLUSTERS ARE WELL SEPARATED. THE ADVANTAGES OF THIS ALGORITHM ARE SHORT RUN TIME AND SMALL STORAGE REQUIREMENT. THE LIMITATIONS ARE AS CLEAR AS THE ADVANTAGES. THE MERGE METHOD OF THE SUB-CLUSTERS IN HIGHER LEVEL CLUSTERS IS LEFT TO THE READER. BUT THE SAMPLE SIZES AT EACH LEVEL IS DISCUSSED. OBVIOUSLY SUCH AN ALGORITHM WOULD BE APPLIED IN THE PROCESSING OF LARGE DATA SETS. UNFORTUNATELY, THE PRESENTED EXAMPLE TREATS AN ARTIFICIAL 2-DIMENSION SAMPLE OF SIZE 50.

データ分類のための計算上有効な技術。 多値理論に基づいて計算上有効なかたまりになるクラスタリングアルゴリズムは述べられる。 データ・セットはいくつかのパーティションにランダムに分割される。 各々のそのようなパーティションの見本はこれとは別に階層的なかたまりになるクラスタリングアルゴリズムによるサブクラスタを形成することを分類される。 これらはより高いレベルで最終の分類に合併される。 クラスタが上手に分けられるときこのアルゴリズムは階層的なかたまりになるクラスタリングアルゴリズムと同じものとして同じ分類へ通じる。 このアルゴリズムの利点は短い実行時間および小さい記憶容量である。 制限は利点と同じくらい明白である。 より高いレベルクラスタのサブクラスタのマージ方法は読者へ残される。 しかし各々のレベルにおける見本サイズは議論される。 明らかにそのようなアルゴリズムはかなりのデータ・セットの処理で適用されるであろう。 述べられた例は、不幸にもサイズ50の人工的な二次元見本を扱っている。
```

図8 翻訳例(2) to不定詞の難しいものなどあるが，文の大意は分かる程度の翻訳が可能である。

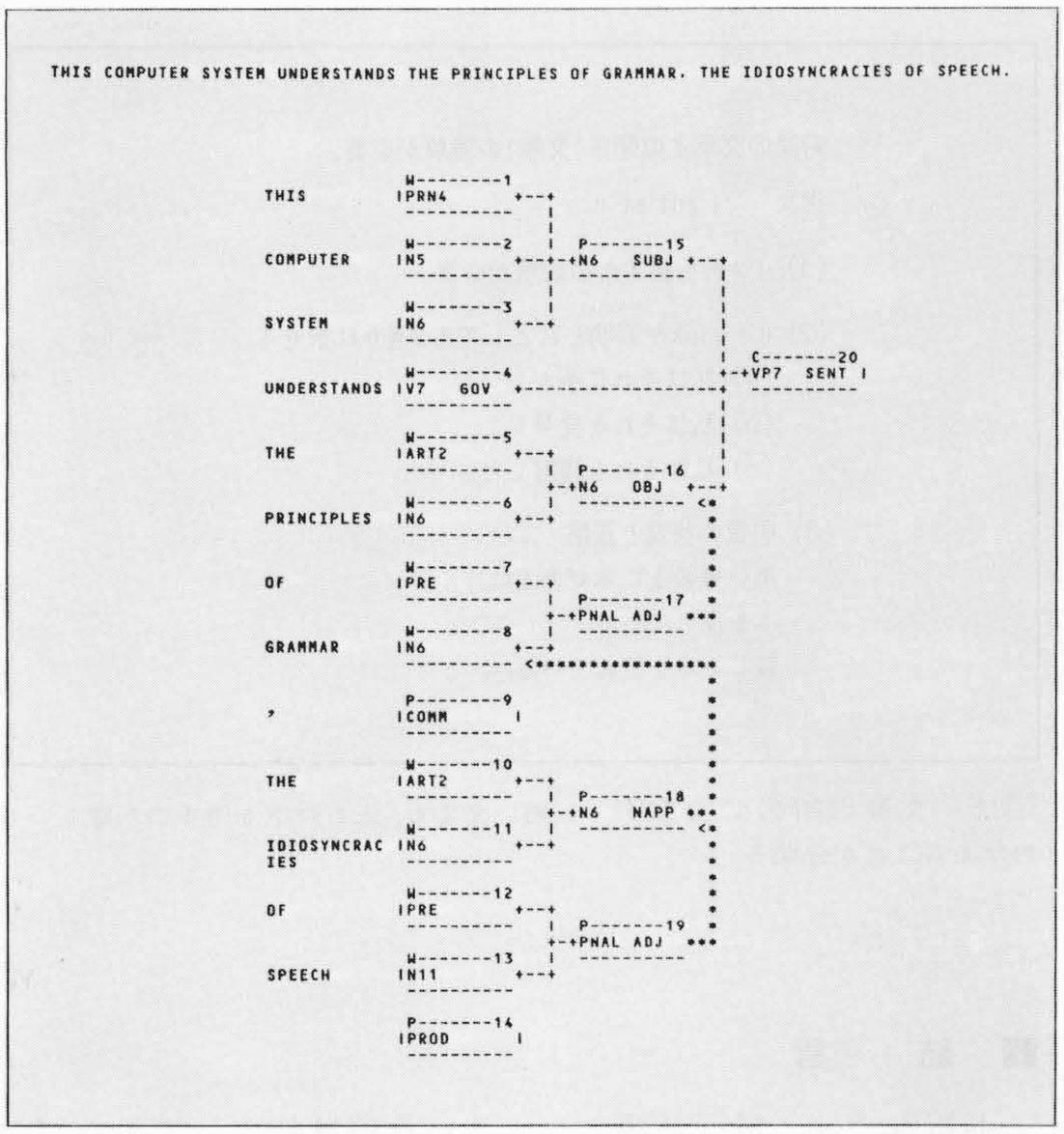


図 9　解析結果の表示例　図中のNAPPは名詞の同格を，ADJは修飾句(節)の形容詞的用法を示す。

1　入力英文

THIS COMPUTER SYSTEM UNDERSTANDS THE PRINCIPLES OF GRAMMAR, THE IDIOSYNCRACIES OF SPEECH.

2　和文生成

英字句	構文子	訳語	助詞
THIS COMPUTER SYSTEM	主語	この計算機システム	は
UNDERSTANDS	主動詞	理解する	
THE PRINCIPLES	目的語	原理	を
OF GRAMMAR	形容	文法の	
, THE IDIOSYNCRACIES	名同格	特質	、すなわち
OF SPEECH	形容	言語の	

3　訳文

この計算機システムは言語の特質、すなわち文法の原理を理解する。

図10　翻訳結果のプリンタ出力例　構文子の欄は，図9のノード番号(各箱の右肩に表示)の15，4，16，17，18，19のものに対応している。これを日本語の語順に従って並べ替えると訳文が得られる。

5　校正及び辞書

計算機が人間に近い翻訳能力をもつことは当面不可能であり，実用性を重んじるシステムにとっては，不十分な訳文を校正する作業の合理化が重要な課題である。機械翻訳の校正が，一般的なワードプロセッサの機能とは若干異なることから，ATHENEでは翻訳特有の校正機能を実験的に実現している。従来のワードプロセッシングのほかに特徴的なことを以下に列挙する。

(1)　訳の多義表示と選択
(2)　語・句単位の移動，並べ替え
(3)　句単位の訳文の校正

図11に校正のための句単位の表示を，また**図12**に訳の多義表示と選択機能の一例を示す。**図12**では，カーソルのある「ので」に対する「AS」が点滅しており，画面右上の2を選択すると「ので」が「ように」に変化する。

また，翻訳の質を向上するには辞書の充実が不可欠であり，基本語のほかに分野ごとに大量の専門語辞書を用意しなければならない。また，これらの辞書は絶えず増補改訂が必要であり，そのための辞書保守用プログラムも充実する必要がある[8]。

6　今後の課題

機械翻訳システムの実用化のためには，まだ多くの課題が残されている。その一つは翻訳能力の向上であり，もう一つは周辺機能とでも言うべき問題である。

翻訳能力の向上に関しては，膨大な構文規則を計算機の中に入れることのほかにも，意味解析，推論機能，学習機能といった将来的に取り組んでいかなければならないテーマが山積している。ここでは，意味解析の必要な例の幾つかを**図13**に示す。(1)はand処理，(2)は訳の多義性，(3)は修飾の多義性を示す例である。このうち幾つかは，既に解決可能になっている。例えば，(3)については，**図14**に示すような階層形シソーラスに従った意味的制約規則(意味文法)を利用すれば，下

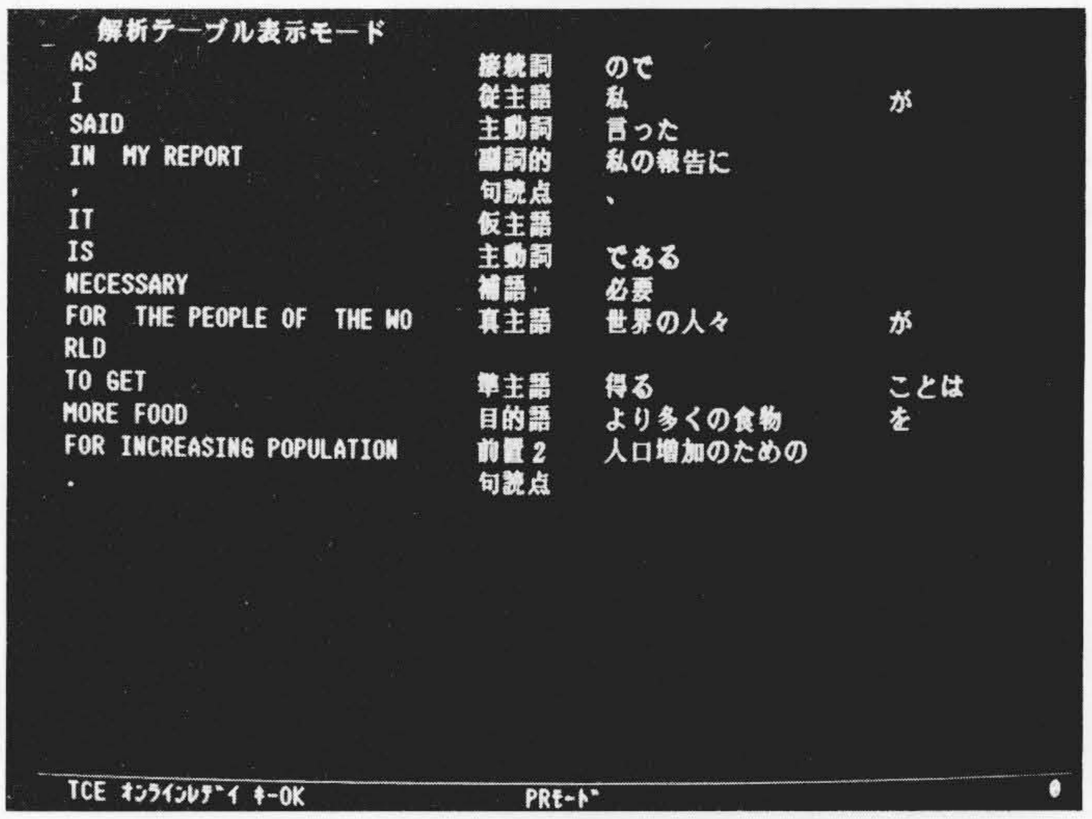


図11　校正のための句単位の解析結果表示　このように，訳文生成の際には不要となるit(仮主語)も校正のため表示される。

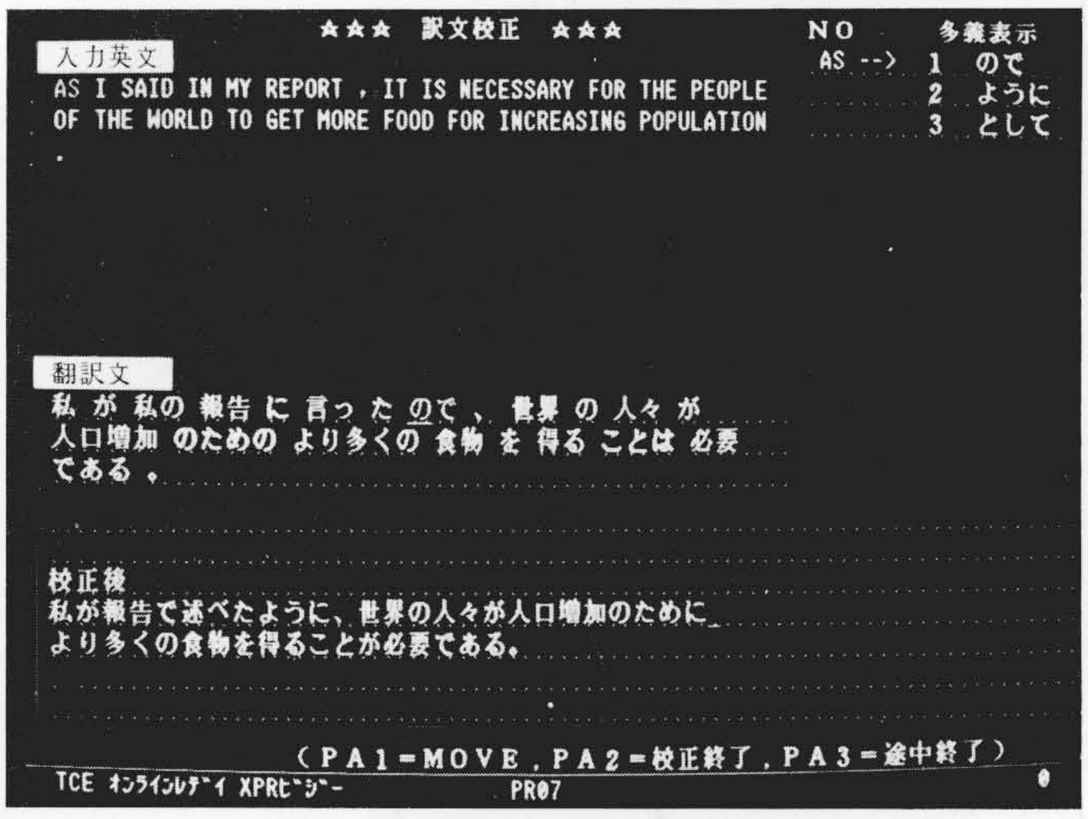


図12　訳の多義表示と選択機能　カーソルは「AS」の訳「ので」にあり，右上の訳のうち2を選択すると，「校正後」に示すように訳文が変更されて作られる。

の解釈が正しいと判定できる。しかし，実際の実用に到るまでには，**図14**に示したような組合せや分類の手数，更には意味コードの追加・変更に伴う保守の難しさなど，予測の難しい問題が待ちかまえている。

更に実際の翻訳では，人工知能研究で話題となっている文章(談話)理解まで必要な場合さえある。**図15**にその一例を示す。この例のように，文が短くても翻訳が容易ではないことが，自然言語の奥深さを我々に教えてくれる。

さて，実用化のもう一つの課題である周辺機能については，最近は明るい面が見られる。第5世代コンピュータ計画のような高速の推論マシンや，文法チェックを含むエディタ(校正システム)などが実現されつつある状況を考えると，知的な校正や辞書保守のためのサポート ツールの実現は，大いに期待がもてるし，それらが機械翻訳の実用化時期を早めてくれる可能性もある。

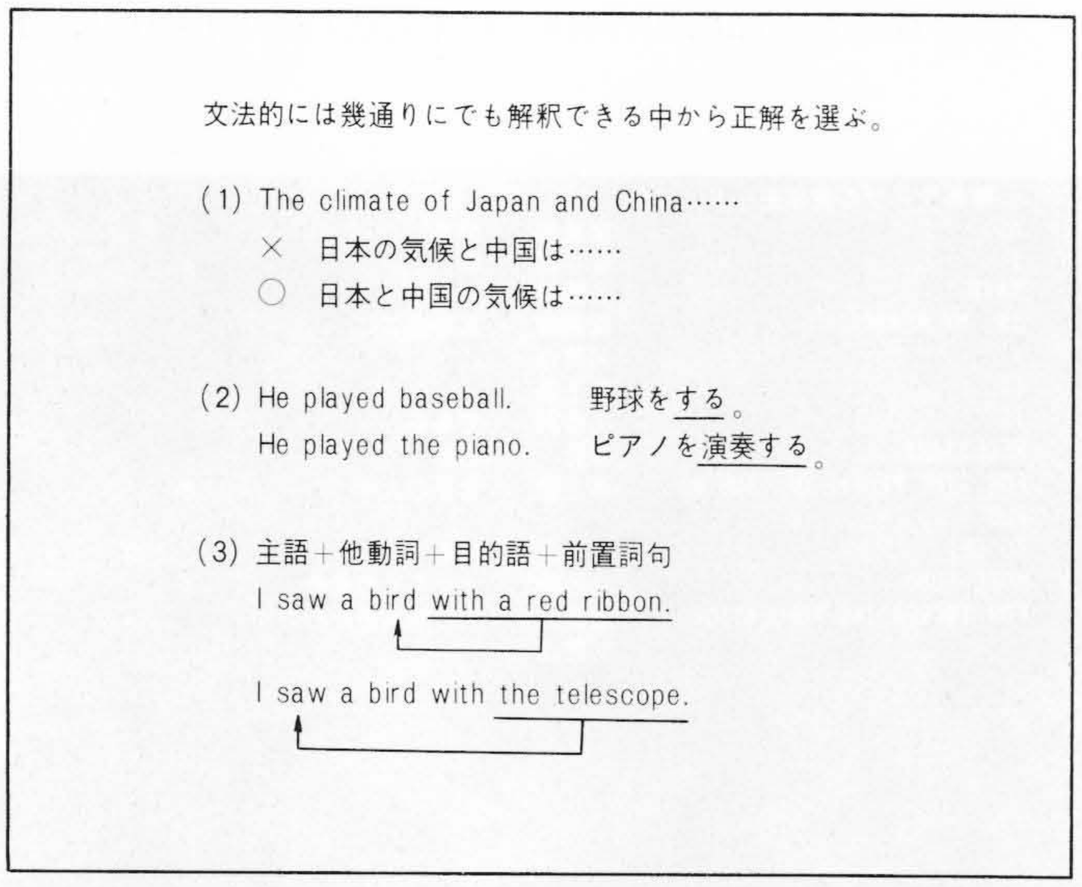


図13 意味解析の必要な例 (1)はand処理，(2)は訳の多義性，(3)は修飾の多義性に意味処理が必要である例である。

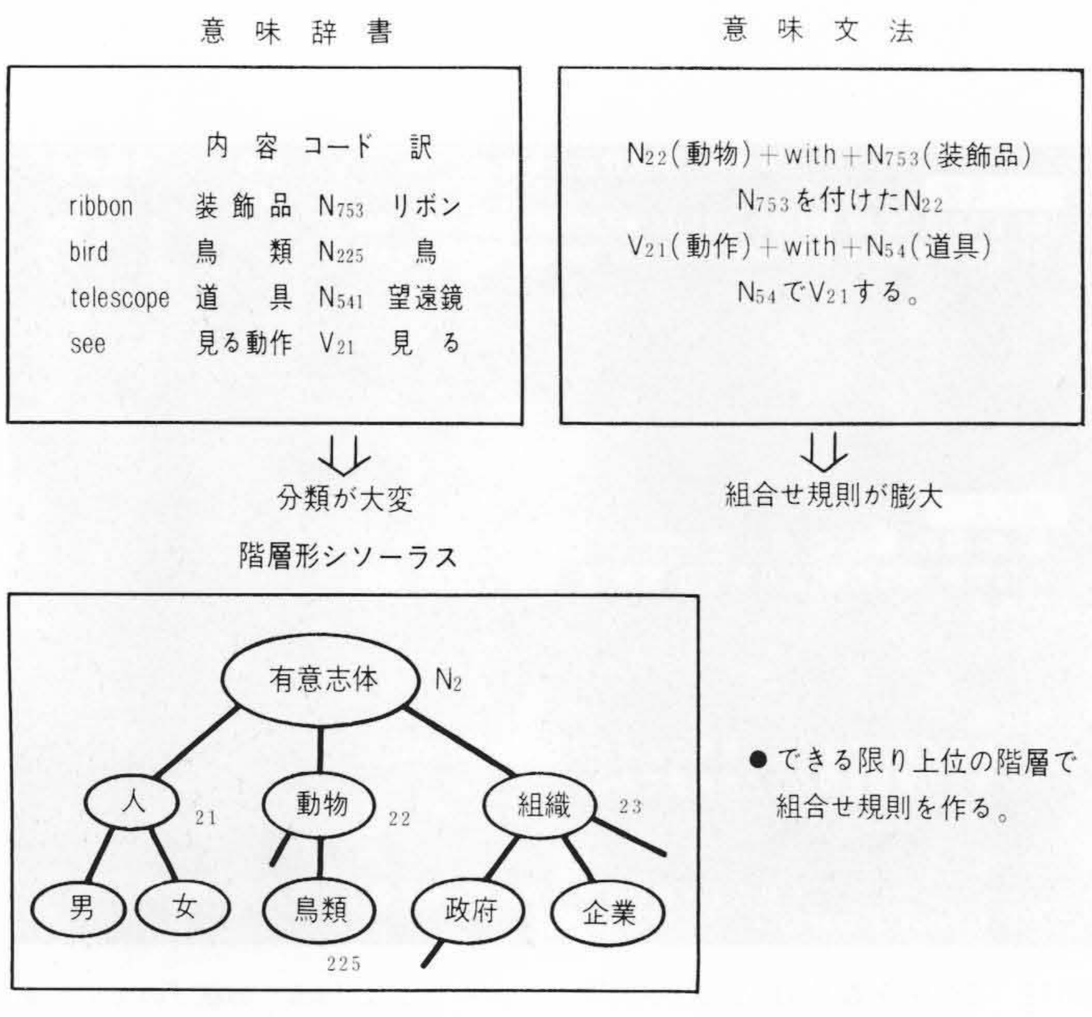


図14 意味解析と階層形シソーラス 右上の意味文法に従って図13(3)の二つの修飾関係が正しく解析できる。

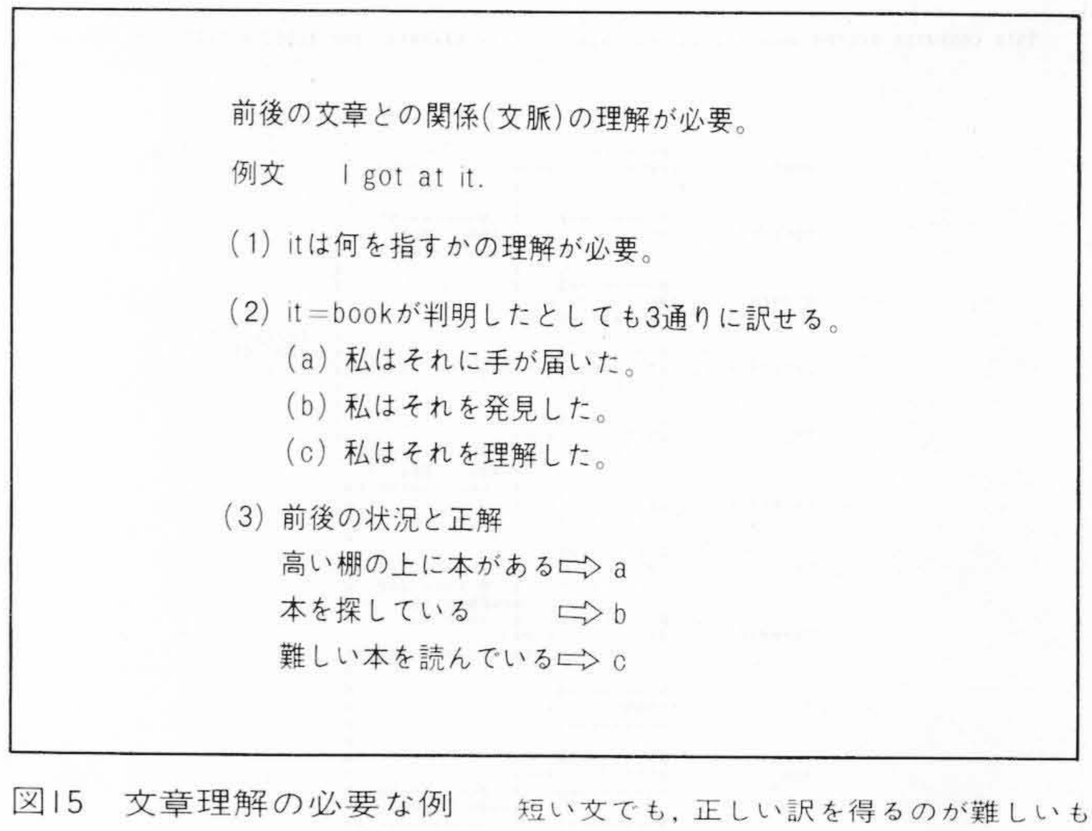


図15 文章理解の必要な例 短い文でも，正しい訳を得るのが難しいものがあることが分かる。

7 結 言

本報告では，最近話題となっている機械翻訳システムについて，その原理と実際の例を日立製作所が開発した英和機械翻訳実験システムについて述べるとともに，今後の実用化に向けての課題について検討した。

実験システムでは，従来の2国語間の文型変換を基本とし，汎用的な中間語による論理的解釈のしやすさをも重視した準論理的中間語を使い，更に，言語固有のイディオム的表現を構文的なものまで辞書に記述可能とした拡張イディオムを用意することによって柔軟な解析を可能としている。

翻訳システムを援助するものとしては，校正システムと辞書保守システムが重要であるが，今後の課題としては，これら周辺機能の拡充や規則の追加による翻訳能力向上のほか，意味処理も必要となる。実験システムで実現したその初歩的利用について一例を述べた。

真の意味での自然言語理解は，そうたやすいことではないが，用途を限定し，人手による援助を仮定すれば，機械翻訳システムも実用に向かって少しずつ前進しているといえよう。

参考文献

1) 長尾，外：自然言語処理プログラム，情報処理，Vol. 18, No. 1，63～75 (1977-1)
2) 辻井：人間と機械の自然な対話を目指す自然言語理解，日経エレクトロニクス，No. 300 (1982-9-27)
3) Y. Nitta, et al.：A Heuristic Approach to English-into-Japanese Machine Translation, Proc. COLING 82, North-Holland, (July 1982)
4) 新田，外：英和機械翻訳のための構文解析技法，情報処理自然言語処理研究会，28-4 (1981, 10)
5) 岡島，外：自然言語処理における中間語表現と多義解消のしやすさとの関係，情報処理「自然言語処理技術」シンポジウム (1983, 6)
6) 岡島：自然言語処理における述部駆動型構文解析と並列処理について，Proc. of The Logic Programming Conference '83, (Mar. 1983)
7) 山野，外：英和機械翻訳における和文生成について，昭和57年後期情報処理全国大会 (7K-4)
8) A. Okajima, et al.：Lexicon Structure for Machine Translation, Proc. ICTP '83, (Oct. 1983) to appear