

## 結合価フレームを利用した 文解析システム

三浦 貢\* 板橋 秀一\* 西野 博二\*\*

\* 筑波大学

\*\* 東京工科大学

我々は、日本語の文の統語ならびに意味の解析をおこなうシステムMOMOを開発した。統語構造の分析やひらがなべたがき文を漢字かな混じり文に変換するには、自然言語の意味表現の枠組みと意味辞書とが必要である。MOMOにおいて、意味表現は、結合価フレームによって表され、意味辞書もまた結合価フレームによって作られる。MOMOは、意味辞書を利用して日本語を解析し、その結果として、インスタンスフレームを生成する。また、得られたインスタンスフレームを使って意味辞書を強化することが可能である。

### JAPANESE TEXT ANALYSIS SYSTEM WITH VALENCY FRAME

Mitsugu MIURA \* Shuichi ITAHASHI \*  
Hiroji NISHINO \*\*

\* University of Tsukuba,

1-1-1, Tennodai, Sakuramura, Niiharigun, Ibarakiken, 305 Japan

\*\* Tokyo Engineering University,

1401-1, Katakurachou, Hachioji, Tokyo, 192 Japan

We have developed a system MOMO which analyzes syntactic and semantic structure of Japanese sentences. In order to analyze syntactic structure and translate kana-texts into kanji-texts, a framework for semantic expression of natural language and a semantic dictionary are required. In the MOMO, semantic expression is represented by valency frame; the dictionary consists of valency frames. This system analyzes Japanese sentences with the semantic dictionary; it generates instance frames from the analyzed results. The system can revise the dictionary with the derived instance frames.

## 1. はじめに

品詞情報や文法規則のみで自然言語を処理した場合、一意に解析木が決定できることは、まれである。数個の単語の文から数十通りもの統語構造が得られることもよくある。同様に、平仮名で書かれた文を漢字かな混じりの文へ変換する場合、同音異義語の弁別を行わなければならない。弁別を行わない場合、仮名漢字変換後、得られる文の数は、同音異義語の数の積になるため、100のオーダーを越えることもしばしばある。複数の解析木から正しいものを選択したり、同音異義語の弁別を行うには、意味的な情報あるいは知識を解析システム中に導入する必要がある。<sup>1)</sup>

意味を統語構造の決定に用いる手法は、いくつかあるが、現在行われている標準的な方法の一つに述語と述語要素との間の意味的整合性を利用するものがある。この場合、述語要素に付加する意味情報をどのように与えるか、という問題が常に残る。また、その様な意味的な情報を持った辞書「意味辞書」をいかにして作成するかという問題もある。意味辞書は、解析システムとは独立に開発される事が多いが、システムと対応が取れていない場合、解析結果が不安定になりやすい。

このような問題に対して、システムの内部表現と意味辞書を同じ表現形式で記述する手法が有効であると考えられる。内部表現と意味辞書の記述形式が同じである場合、解析結果を利用し、意味辞書を強化することが簡単に行える。また、そのようにして強化された意味辞書をシステムが利用することにより、より正確な解析を行うことが可能となると考えられる。本稿では、以上の操作を行う解析システム(MOMO)について述べる。

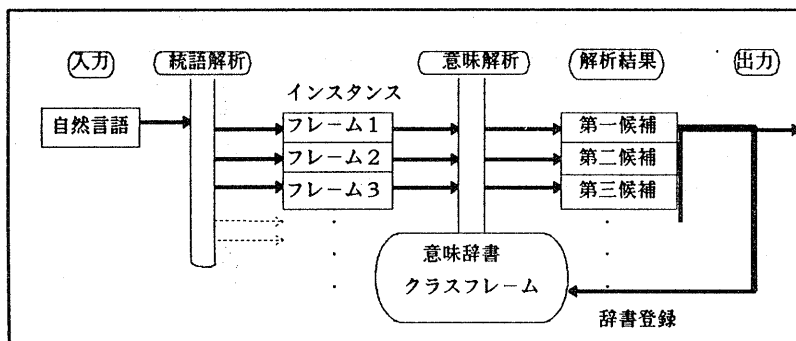
## 2. システムの概要

MOMOの特徴は、以下の通りである。

- 1) 結合価文法を採用した。
  - 2) 意味素性を用いない。
  - 3) 解析結果を登録することにより意味辞書を強化出来る。
- また、MOMOは、以下のように日本語を解析し意味辞書を強化する。(図1. 参照)
- 1) 日本語で書かれた文を読み込み、フレーム形式の意味表現に変換する。
  - 2) 結合価フレームによる意味辞書を用いて、同音異義語の弁別を行い、優先順位を付ける。
  - 3) 構文解析時に複数の解析結果が得られた場合、意味辞書を用いて、解析結果に優先順位を付ける。
  - 4) 第一候補が正しい場合は、その文の処理を終了し、誤った場合は正しい解析結果を辞書に登録する。
  - 5) 1~4を繰り返し意味辞書を強化する。
  - 6) 辞書を整理する。

MOMOは、UTILISP で記述されており、FACOM-α上で作動する。現在、システムに登録されている語は9000語程である。動詞に関しては1003語が登録されており、その大部分は情報処理振興事業協会によるIPAL「計算機用日本語基本動詞辞書」<sup>2)3)</sup>に基づいている。そのサブフレームの総数は現在、3500個をこえている。形容詞に関しては約460語、形容動詞は、約340語が登録されている。名詞に関しては見出し語数は4197語が登録されているが、結合価フレームが付けられたものはまだ800程度である。現在、名詞の辞書を拡張中である。また、システムに与える例文は、2000文近くを三省堂の「新明解国語辞典」ならびにIPALから採取し、辞書の強化に役立てている。

図1 MOMOの処理のあらまし



### 3. 結合価文法

従来、機械翻訳あるいは質問応答システム等を構築する場合、意味情報を解析システム中に導入する枠組みとして、よく使われてきたものに格文法がある。

しかし、格文法には、

- 1) 格の種類不安定性
- 2) 格の概念不安定性

などのいわゆる「格不安定性」と呼ばれる問題が存在している。このため格文法を辞書記述の枠組みとして採用すると、小規模な辞書は作製できるが、実用的なレベルの大規模な辞書開発が難しくなる。そこで、MOMOには、結合価文法を採用した。結合価文法は、文の表層に近い、浅いレベルの意味を対象とするため、その格は安定していると言われている。<sup>4)</sup>

結合価文法は格文法によく似た述語を中心とする文法である。

その基本的な概念は、

- 1) 結合価 (Valenz)
- 2) 分布 (Distribution)

である。

1) の結合価は、格文法における表層の格にほぼ等しく、述語が取る述語要素を記述したものである。

2) の分布とは、述語に結び付く名詞に制限を加えることである。その選択制限は、述語要素の意味的特徴を示すことによりなされる。従来は、意味素性を名詞に付加することにより選択制限を実現させていた。しかし、意味素性には、問題が多い。

MOMOにおいては、意味素性の代わりに結合価フレームを語の意味的特徴を示すための枠組みとして採用した。

### 4. 意味情報

意味素性あるいは、意味マーカと呼ばれる述語要素の意味情報は、結合価文法や格文法を利用した解析システムによく用いられている。意味素性は語の意味を表す方法として簡便であるため、数千語程度の中規模の辞書も開発されている。

しかし、そのような辞書においては、単語に対する素性の割付がアドホックになされることが多く、解析システムとその辞書の対応がとれていない場合が多い。そのような辞書は、語と語の結びつきよりも単語の上位一下位概念を軸に形成されている。実際にはその作成者が、

ある単語の上位概念に対応すると思われる意味素性を選びだし、それを単語に割り当てて行く、という手順で辞書が作られる。この場合、意味素性は表1のように名詞に付加されよう。

表1. 意味素性の例

意味素性	例
動物	犬 鳥 猿 コアラ
自然物	山 川 石 空
生産物	紙 バン 鉄 車
現象	風 雨 涙 音 光
言語	名前 ニュース 説教

このように付加された意味素性は、統語構造の決定や同音異義語の弁別にはあまり役立たない。システムは、述語と述語要素との意味的関係の深さから統語構造を決定するが、表1に見られるように意味素性は、そのような意味的な関係を表してはいない。

例えば、統語構造の決定には、「風」と「吹く」、「涙」と「拭く」のような語と語の結びつきが必要であるが、「風」と同じ意味素性を付加された「涙」は、同じ意味的深さで「吹く」と結びついてしまう。このような状態でシステムに同音異義語の処理などを行わせると以下の例のような現象が多発してしまい、初期の目的を達せられなくなる。

「ふいた」の弁別

入力文：昨日 ふいた 風は・・・

A：昨日 (吹いた) 風は・・・

B：昨日 (拭いた) 風は・・・

AもBも同じ程度の確からしきで出力される。

また、素性間のマッチングミスを減らすために意味素性間に階層関係を導入し、処理を行うシステムもあるがこのような機構は、前述の例を減少させられず、むしろ増やす結果をもたらしてしまう。

従来、格文法や結合価文法を利用した解析システムで用いられた素性の数は数十～数百のオーダーである。従って、意味素性によって表される名詞の意味は、名詞それぞれが持つ意味に比べるとかなり粗い。

このため、語彙数が少ない場合、つまり言語処理の対象とする分野が極めて狭い場合は、システムは、ある程度安定した解析を行えるが、対象分野を広げ、辞書を拡張すると実用上の効率が低下してしまい、解析木を減少させることができなくなる。<sup>9)</sup>

このような問題に対処する方法として次のようなものが考えられてきた。

- ・意味素性の種類を増やし、語をより細かく分類する。
- ・名詞に付加する意味素性を複数化し、その意味素性間にAND, OR, NOTなどを導入して、意味素性の種数を増やさずに意味の細分化を行う。

しかし、実際には上記の方法では対処できない場合が多い。

彼は**ピストル**で歌っている女を**撃**った。

彼は**カラオケ**で歌っている女を**撃**った。

このような場合、「ピストル」や「カラオケ」に与えられる意味素性は（道具）程度である。そのため、「ピストル」と「撃つ」、「カラオケ」と「歌う」などの係り受け関係を決定することができない。「ピストル」と「カラオケ」を弁別可能なレベルまで意味素性を細分化ことは難しいと考えられる。素性の種類が少ないと、語に付加される素性が単数になることが多く、素性間にANDやORを導入することが出来ない。

また、素性の種類を増やしても統語構造の決定には、あまり寄与しない。意味素性を細分化し、素性の種類を増やすと、述語につけられた意味素性と述語要素につけられた意味素性とのマッチングのミスが多発する。その

ため、誤った解析結果を選択することが多くなってしまふ。

以上のような現象が起こるのは、意味素性の表す意味的關係が、ISA やPARTOFに限定されているためと思われる。この2つの関係だけでは統語構造を決定するには、不十分である。自然言語の場合、ある語が他の語と強く結びつく条件は慣習的な共起性に依存することも多い。意味素性では、語の共起性を表すことができない。

#### 4. 結合価フレーム

フレーム形式は、意味素性にくらべ意味の記述をより厳密に行うことができる。フレームは、語の意味を述語に依存した形で記述する事が可能であるためISA やPART OF 以上の関係表現することができる。<sup>9)</sup> 前の例においても、「カラオケ」のフレーム中に「歌う」を記述すれば、解決がつく。

MOMOは、結合価中の分布に意味素性ではなく、フレーム（結合価フレーム）を用いている。

結合価フレームにおいて、語の意味は、その語を含む文により表される。このことにより、語の意味をより細かく記述することが可能となる。

図2に結合価フレームの概略を示す。結合価フレームにおいて、語を表す文は結合価スロットの形で記述される。また、スロット中には、OR, NOT, →（言い替え可能の記号）などの付加情報を書くことができる。それらがない場合はスロットはANDで結ばれている。結合価スロットが集まったものがサブフレームである。サブフレームの集まったものが結合価フレームとなる。サブフレーム同士はORでつながれている。スロットは、異なるサブフレームに共用されてもよい。サブフレームとスロット

図2. 結合価フレーム

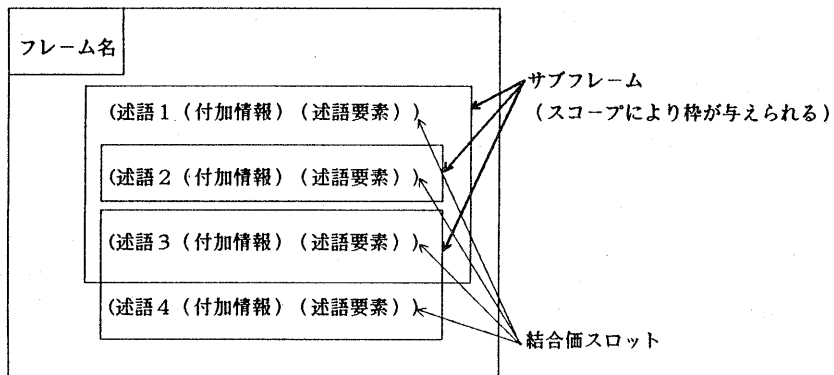


図3. 結合価フレームの例

動詞 「あがる」のフレーム

(あがる (NIL (が おとこ 花子) (へ 二階 上 屋根) (で 階段 梯子)))
(あがる (OR (が おとこ 花子) (に 二階 上 屋根) (で 階段 梯子)))
(あがる (NIL (が おとこ 花子) (を 坂 階段 梯子)))
(登る (→ (が おとこ 花子) (を 坂 階段 梯子)))
(あがる (NIL (が おとこ 花子) (から 風呂 プール)))
(出る (→ (が おとこ 花子) (から 風呂 プール)))
(あがる (NIL (が 値段 税金) (から 10円 2%) (へ 50円 6%)))
(する (→ (が 値段 税金) (Φ 値上がり)))
(あがる (NIL (が 仕事 稽古) (に 10時)))
(終わる (→ (が 仕事 稽古) (に 10時)))

名詞 「花」のフレーム

(ISA (NIL (は 桜 チューリップ) (である 花)))
(咲く (NIL (が 花) (に 木 枝 庭 山)))
(散る (NIL (が 花)))
(NO (NIL (の 桜 チューリップ) (Φ 花)))
(ISA (NIL (は 花) (である 植物)))

形容詞 「明るい」のフレーム

(明るい (NIL (は 太郎 おとこ) (に 地理 歴史)))
(明るい (NIL (は 太郎 おとこ) (が 性格)))
(明るい (NIL (は 空 外)))
(明るい (NIL (# 笑う 話す)))

形容動詞 「健康」のフレーム

(健康 (NIL (は 太郎 おとこ)))
(大切 (NIL (は 健康)))
(する (NIL (は 太郎 おとこ) (に 健康) (Φ 注意)))
(する (NIL (は 太郎 おとこ) (を 健康) (に 大切)))

の関係に登録するためのテーブル（スコープ）がシステムには準備されている。

同じ言葉でも意味的内容が異なる場合、あるいは、異なった結合価が与えられた場合はそれらは独立したサブフレームとなる。以下の言葉は異なったサブフレームが与えられている。

1) (風呂からあがる。 成績があがる。)

2) (息を飲んだ。 水を飲んだ。)

2)の例に見られるように慣用句の場合も独立したサブフレームが与えられる。慣用句は、語と語の結びつき方が極めて強い。このような場合、意味素性だけでは処理できない。意味辞書中に素性以外の語自体を記述出来るようにするか、あるいは、意味辞書とは、独立に慣用句専用の辞書をもたせなければならなくなる。結合価フレームを用いた場合、スロット中に書かれている語とのマッチングの分布を限定することにより、慣用句を記述することが簡単に行える。

2)の意味的違いを弁別するために用意されるサブフレームは、次のようなものである。

息を飲んだ。

(飲む (NIL (が 人) (を (息) )))
(驚く (→ (が 人)))

水を飲んだ。

(飲む (NIL (が 動物) (を 水 茶 酒)))
-----------------------------

「息を飲む」の場合(を)のスロットとマッチするものは「息」のみである。

「水を飲む」の場合(を)のスロットとマッチするものは「水」「酒」「牛乳」などが可能である。

結合価フレーム中では、動詞、形容詞、形容動詞、名詞、文などは、区別されず同じように表現される。これらの品詞の違いは述語要素中に書かれる格助詞あるいはスロット中で自分が置かれる位置のみに現れる。

図3に結合価フレームの例を示す。これらのフレームは、MOMOに日本語のテキストを与えた結果、作り出されたものである。

結合価フレームを用いると、

- 1) 意味素性の細分類を行わずにすむ。
- 2) 動詞、名詞、形容詞、形容動詞、文、などを同じ形式で記述できる。

3) 語の定義が同音異義語の弁別や統語構造の決定、同音文の判定などに利用できる。

などが可能となる。

また、属性遺伝や暗黙値付加の機能を使えるため語の省略などにもある程度の対処が可能となっている。結合価フレームのリンクは、多重継承を許している。

## 5. 意味辞書

MOMOにおける意味辞書は、結合価フレームが集まったものである。MOMOは、解析結果を利用して意味辞書を自動的に強化することができる。

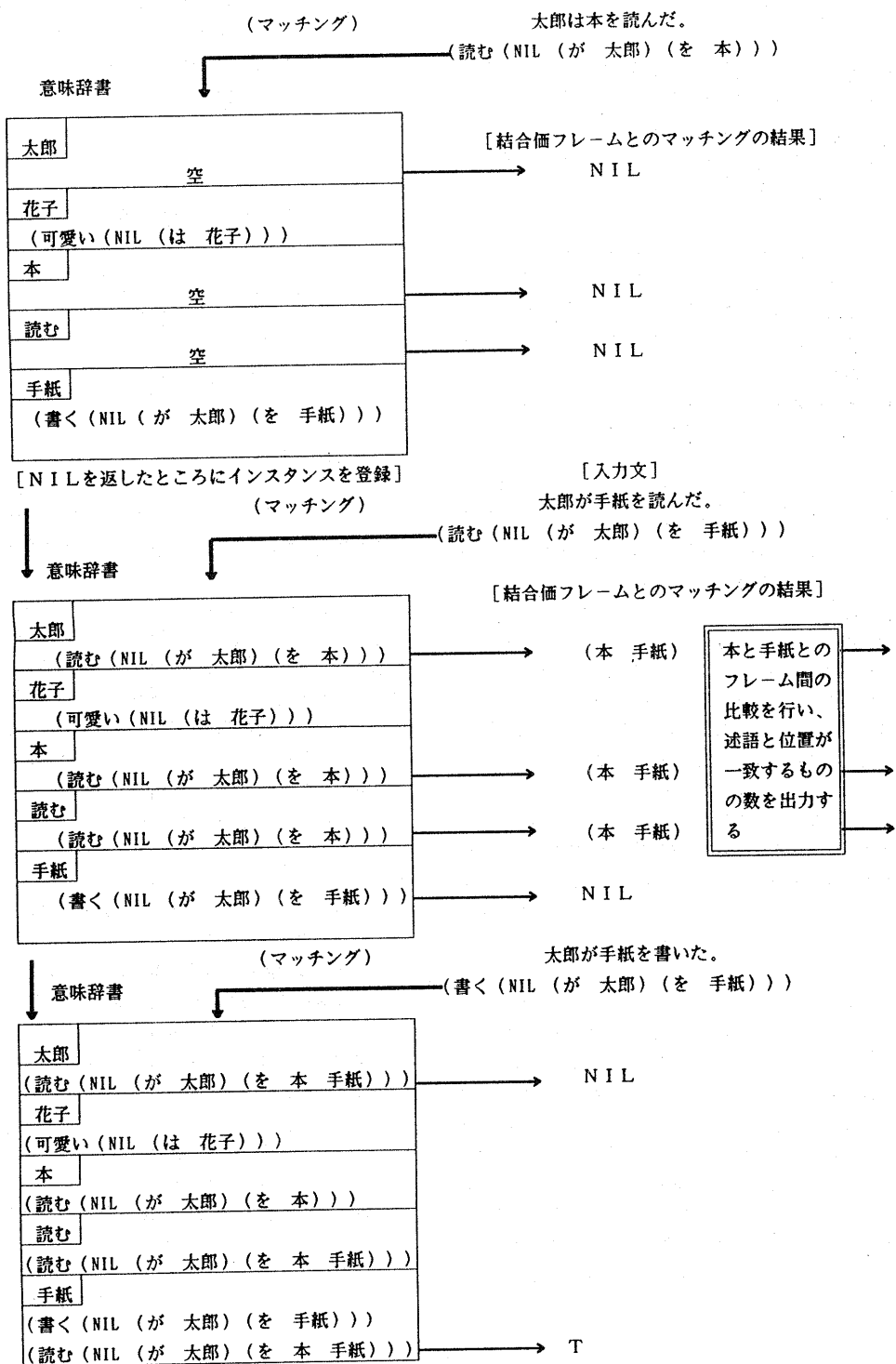
統語解析後、出力される複数のインスタンスフレームの中から正しいものを選び出すのが意味解析部の役割である。意味解析部は、インスタンスフレームと意味辞書中に書かれた結合価フレームとのマッチングを行い、マッチングの成功の度合いに応じて、解析結果に優先順位をつけて出力する。

結合価フレームとのマッチングは、以下のように行われる。

- (-) インスタンスフレーム中の単語ごとにその結合価フレームが意味辞書から呼び出される。
- (A) 呼び出された結合価フレームの中でインスタンスフレームと同じ述語が存在するサブフレームが呼び出される。もし、同じ述語が存在しなければ、NILが返されて終了する。同じ述語があれば、処理(B)を行う。
- (B) インスタンスフレームとサブフレームとで格助詞の比較を行い、一致したサブフレームを呼び出して、処理(C)を行う。もし、一致しない格助詞が存在する場合は、その格助詞の組みが返されて終了する。
- (C) サブフレーム中に書かれているスロットとインスタンスフレーム中のスロットを比較する。もし、同じ格助詞のスロット中にインスタンスフレームのスロットに書かれた語があるときはTが返される。そうでなければ、サブフレームの中に書かれている語とインスタンスフレームの中に書かれている語との組、並びにそれらの意味的關係の深さを表す数を返す。意味的關係の深さは、処理(D)を行うことにより得る。
- (D) 意味的關係の深さを調べる2つの語の結合価フレームを意味辞書から呼び出す。呼び出した2つのフレームの間で述語が一致し、かつ述語要素の位置が一致するスロットの数を調べ、その数を意味

図-3

[入力文]



的關係の深さとして返す。

上記の処理を行い、多くのTが返されたものほど正しい解析結果とされ、高い優先順位が付けられる。

この後、辞書の強化がなされる。意味辞書の強化は、インスタンスフレームと意味辞書とのマッチングの度合いに応じて、次のように行われる。

- (a) NILを返したフレームに正しい解析結果のインスタンスフレームをサブフレームとして登録する
- (b) 格助詞の組みを返したフレームに正しい解析結果のインスタンスフレームをサブフレームとして登録する。
- (c) 単語の組を返したフレームのスロットの中にインスタンスフレーム中の単語を付け加える。

(b)の場合、新たにインスタンスフレームを付け加えるのではなく、格助詞のスロットのみをフレームに付け加えた方がより適切である場合も考えられるが、現在のところそのような処理は行っていない。

例えば、  
[結合価フレーム中に登録されているスロット]が、  
(書く(NIL(が 太郎)(を 答え)))であり、  
[インスタンスフレーム中のスロット]が、  
(書く(NIL(が 太郎)(を 答え)(に 黒板)))  
で、あるような場合、(に 黒板)がスロットに付け加えられず、インスタンスフレーム自体が新たなサブフレームとして登録されてしまう。

これは、必須格、任意格の判定機構が曖昧であるため起きている。また、このような場合、上記2つのスロットは、同一サブフレーム中でORで結ばれるべきだが、それは現在のところ行っていない。

## 6. 実験結果

システムの有効性を調べるために、かな漢字変換をMOMOに行わせた。テストに用いた100文の内、意味辞書が空であった場合、正解が第一候補に選ばれたものは、8文しかなかったが、例文を600文与え、意味辞書を強化した後、正解のみが第一候補となったものは、47文あり、正解が第一候補の中にあつたものは61文あつた。テストに用いた文を入力し、意味辞書を強化してから、再度、同例文を入力すると、正解のみが第一候補となったものは77文あり、正解が第一候補の中にあつたものは94文あつた。

現在、システムが統語構造の決定も含めた解析結果を出力するまでに要する時間は、以下の通りである。

解析に用いられた文の総数 20 個  
一文あたりの平均自立語数 6.2 個

A: 意味辞書強化のために使われた文の数  
B: 一文当たりの平均解析時間(単位は分)

A (個)	B (分)
0	6
100	11
200	18

## 7. 今後の課題

当システムには時間構造を扱う枠組みは与えられていない。そのため、テイルが含まれる文や時間的な関係から文の統語構造が決定出来る場合などの処理はおこなえない。

システムに組み込まれている推論機構は強力なものではないため、単純な言い換え以外、文と文の関係を処理できない。文脈から同音異義語や統語構造が決定される場合が自然言語には多いが、そのような処理も現在のところ行えない。因果関係なども処理出来ない。従って、文脈処理などは扱っていない。また、解析にかなりの時間がかかっており、現在、問題となっている。今後、これらの問題を解決する必要がある。

## 参考文献

- 1) 長尾 真 「言語工学」 昭光堂(1983)
- 2) 情報処理振興事業協会  
「計算機用日本語基本動詞」(1985)
- 3) 村木新次郎 青山文啓 六条範俊 村田賢一  
「辞書における格情報の記述」  
情報処理学会自然言語処理研究会資料46-3,1984
- 4) 水谷 静夫編  
「朝倉日本語新講座3 文法と意味」(1983)
- 5) 三浦貢 板橋秀一 西野博二  
「素性問題と結合価フレーム」  
62年度人工知能学会全国大会論文集 pp.369-372
- 6) Marvin Minsky,  
"A Framework for Representing Knowledge",  
MIT AI Memo No.306,1974