

表層的情報による日本語長文の骨格構造解析

兵藤安昭 池田尚志
岐阜大学工学部

本論文では、意味情報を用いない範囲で、すなわち形態素情報と係り受けに関するいくつかの表層上の制約規則のみを用いて、日本語長文の骨格構造を解析する方法について述べる。骨格構造とは、完全な係り受けの木構造をなすものではなく、並列構造など意味に立ち入らなければ解析できない部分は曖昧なブロックとしてそのまま残し、文の全体的な構造を把握しようとするものである。つまり、意味の問題に立ち入らない範囲で可能な最大限度の構文解析を追求した。

Syntactic Skelton Analysis of a Long Japanese Sentence based on Surface Information

Yasuaki Hyodo Takashi Ikeda
Faculty of Engineering, Gifu University

This paper describes "skelton analysis" of a long Japanese sentence. This system does not use semantic information at all but uses only surface information and some restriction rules related to dependency area. Skelton analysis makes a tree that might contain ambiguous blocks in which dependency relations are not decided completely. Ambiguous blocks are such as coordinate structures and noun compounds. Analyzing ambiguous blocks completely will need some semantic informations.

1 はじめに

長い日本語の構文解析は非常に困難であり、80文字以上の文は、ほとんど解析に失敗するという報告もある[1][3]。文が長くなると文節の係り先の可能性が多くなり、係り受けの曖昧さが非常に大きくなってしまふことがその主な原因である。

日本語の構文解析では、格構造を用い、意味を考慮して係り受けの曖昧性の問題を解決しようする方法がよく行なわれている[2]。しかし、意味的な面から係り受けの制約を記述するには、かなり精密な意味情報が必要であり、格構造による処理で行なわれるような意味素性やシソーラスを用いた方法では十分な記述は困難である。

文が長くなるのは多くの場合、名詞句の並列、連体埋め込み文の並列、連用中止法による並列などの並列構造が含まれるためである。文献[3]では、文字列の意味的類似性を調べ、並列構造を抽出してから係り受け解析を行なうことにより、長文の解析を効果的に行なうことができることが報告されている。この手法では、意味的類似性を調べるのに分類語彙表を用いているが、語の意味階層分類は必ずしも一意的に定まるものではなく、観点に依存するものでもあり、多義語の問題も含め困難な問題をはらんでいる。

文献[4]では、意味情報を使わないで表層的情報のみを用いて構文解析を行なう方法を追求している。しかし、このような方法では「AのBのCが」のような場合のAの係り先や、先に述べた並列構造の解析などは、意味情報を用いないで正確に解析することは困難である。

本論文では、意味理解に入り込まない範囲で、すなわち形態素情報と係り受けに関するいくつかの表層的な制約規則のみを用いて、日本語長文の骨格構造を解析する方法について述べる。骨格構造とは、完全な係り受けの木構造をなすものではなく、並列構造など意味に立ち入らなければ解析できない部分は曖昧なブロックとしてそのまま残し、文の全体的な構造を把握しようとするものである。つまり、意味の問題に立ち入らない範囲で可能な最大限の構文解析を追求した。

文の骨格構造が正確に把握できれば、曖昧なまま残された部分に対してのみ、意味情報、文脈情報を用いた詳しい解析を行えばよいので、長文の解析を小さな部分問題に還元することが可能となる。あるいは、このように残された曖昧な部分

に対しては、人間との対話インターフェースを通じて解決する支援型の解析システムも考えられる。また骨格構造のレベルのままのデータでも、全自動的にはほぼ正しく解析できるようになれば、類似文検索など構造を考慮した高度なテキスト検索のための大規模データベースの構築に役立てることができる。

本手法では、まず初めに形態素解析された日本語文に、文節の可能な係り先を示す文節カテゴリを付与する。次に文頭から順に各文節の係り先を調べ、その際すべての文節について依存可能性を調べることはせず、係り先の範囲を3ブロック以内とする仮説により係り先を決定する。このブロック内には、各文節の依存パターンにより、係り先が曖昧な文節が含まれている場合もある。これを文末が見つかるまで繰り返す。今回は、朝日新聞記事より、文字数が50文字以上80文字未満、80文字以上の各50文、合計100文に対して実験を行なった。その結果、50文字以上80文字未満の文で48文、80文字以上の文で47文、正しく解析することができた。

2 骨格構造解析

2.1 形態素解析と文節カテゴリ

まず初めに、入力文に形態素解析処理を施す。形態素解析処理では、入力文を単語単位に分割し、単語列を1つの文節(1つの自立語と0個以上の機能語)にまとめるところまでを行なう。この際、複合語は1つの自立語として考える。また名詞文「Nだ」、「Nである」などについては「だ、である」を内容語「*である」として、いわゆる準体助詞の「の」は内容語「*の」として文節の再構成を行なっている。

次に、各文節に文節カテゴリを付与する(図1)。文節カテゴリとは、文節自身のタイプと係りうる文節のタイプによりカテゴリ化したもので、文節自身のタイプを、体(名詞)、用(動詞・形容詞・形容動詞)、副(副詞・連体詞)、接(接続詞)の4つに分類し、これらの組合せにより10種の基本的文節カテゴリを設けた(表1)。

その他に、表2に示すような4つの文節カテゴリを設けた。文節カテゴリ「体並、用並」は、その文節が並列構造の可能性を示している[3]。また、文節中に時を表す名詞や、主題を表

す機能語(はでは)が含まれる時は、各々「時用」「は用」として扱う。

体用	用言に係る体言文節。
体体	体言に係る体言文節。
体終	係り先を持たない体言文節。
用体	体言に係る用言文節。
用用	用言に係る用言文節。
用終	係り先を持たない用言文節。
副体	体言に係る副言文節(連体詞)。
副用	用言に係る副言文節(副詞等)。 ただし形容動詞連用形「静かに」「犬のように」「犬みたいに」なども「副用」とする
副副	副言に係る副詞文節。 「程度副詞」で、その後に副言が続く時。
接用	用言に係る接言(接続詞)文節。

表 1: 基本的文節カテゴリ

体並	体言並列の可能性のあるもの。「名詞+ (, とも や か かつ …)」。ただし「時詞+ ,」の場合は「時用」とする。「時詞+ ,」が連続する場合は「体並」とする。
用並	用言並列の可能性のあるもの。「文節の連用形+ ,」「用言+」ならびにあるいはおよびまたはもしくはとともに…」。
時用	体言が時を表す名詞の時。例「今日」
は用	主題を表す機能語が含まれる時。「体言+ (はでは)」等。

表 2: その他の文節カテゴリ

以下に述べる骨格構造解析では、正しく文節カテゴリが付与されたものを用いるが、文節カテゴリを正しく特定できない場合としては、以下のような例が挙げられる。

- 「体言+で」の文節カテゴリ
 - 「彼は東京で仕事をしている」で…体用
 - 「彼は勉強中で、…」で…体用(不正解)
- の「で」は、助動詞「だ」の連用形である。この場合は、
(彼は 体用)(勉強# 中 体用)(*だ 用終)
というように文節の再構成を行なう必要がある。

- 「体言+, (読点)」の文節カテゴリ
 - 「年間約60億円出荷、輸出しているが…」
出荷…体並(不正解)

「出荷」はサ変動詞が名詞化したもので、この場合は用言と考えないと「年間、約60億円」の係り先を正しく得ることはできない。

原文

あらかじめ掃除する経路を教えておかなくても、ロボットが壁などに沿って移動し、掃除する範囲を自分で記憶、そのあとは往復動作をくり返しながら掃除をし、終わると自動的に止まる (84文字)

((あらかじめ 副用)(掃除する 体用)(経路を 体用)
(教えるておくないても、 用用)(ロボットが 体用)
(壁などに 体用)(沿うて 用用)(移動する、 用並)
(掃除する 体用)(範囲を 体用)(自分で 体用)(記憶、 用並)
(その 副体)(あとは 体用)(往復#動作を 体用)
(くり返しながら 用並)(掃除を 体用)(する、 用並)
(終わると 用用)(自動#的に 体用)(止まる 用終))

: 複合語を示す

図 1: 文節カテゴリを付与したテキスト例

2.2 三角表による係り受け関係の表示

2.1 で述べた文節カテゴリに基づいて、すべての文節の可能な係り先を求め、三角表上に表示する(図2)[2]。例えば、図2の「あらかじめ」の係り先は、「掃除する、教えておかなくても、沿って」等であることを示す。ただし、明らかに非交差条件に反する係り先は除く。例えば、文節カテゴリから考えると「掃除する:(用体)」は「ロボットが:(体用)」には依存可能である。しかし「教えておかなくても:(用用)」が「ロボットが:(体用)」に依存不可能であるため、非交差条件により「掃除する」は「ロボットが」へ依存不可能になる。

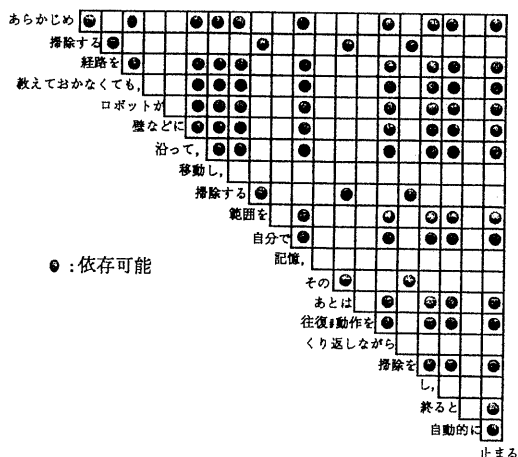


図 2: 係り受けの三角表

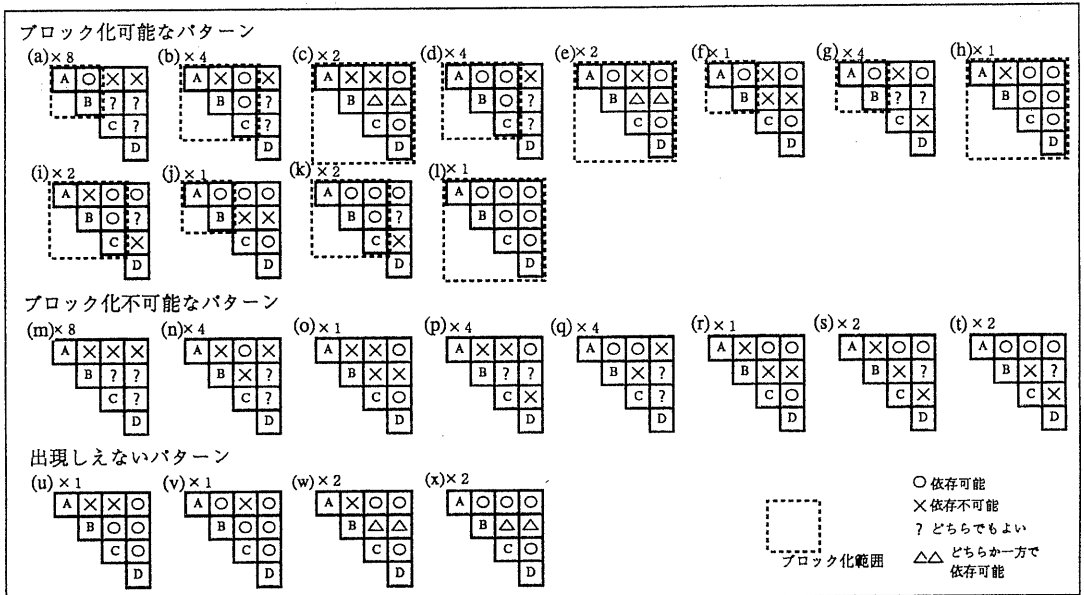


図 3: 係り受けブロック化ルール

2.3 係り先範囲の制約に基づく 係り受けブロック化

係り受けブロックという用語を、その範囲内で係り受けが行なわれる文節ないし係り受けブロックの列として定義する。

係り受けブロック $(B1, B2)$ は、係り受けブロック $B1$ が係り受けブロック $B2$ に係るということを表現し、この時の係り受けブロックの大きさは 2 である。また、係り受けブロック $(B1, B2, B3)$ は、係り受けブロック $B2$ は $B3$ に係るが、 $B1$ の係り先は $B2$ または $B3$ のいずれかであるという曖昧な状態を表現し、この時の係り受けブロックの大きさは 3 である。さらに、 $B1$ は、 $B1$ の係り先が不明であり、これ以上の解析はしないことを表現する。なお、ここで係り受けという用語は、「A と B が遊ぶ: $A \rightarrow B$ 」の場合の様に、純粹の係り受けでない場合に対しても用いる。

形態素解析された文節の列が $b_1, b_2, \dots, b_n$ であったとすると、構文解析開始時の初期係り受けブロックは $(b_1, b_2, \dots, b_n)$ である。係り受け関係が完全に解析されれば、 $(b_1, b_2, \dots, b_n)$ はブロックの入れ子構造の形になり、その時のすべて

の係り受けブロックの大きさは 2 である。従来の構文解析は、すべてのブロックの大きさが 2 となるように、与えられた文節列のブロック化を遂行するものであった。これに対して我々の構文解析の目標は、意味解析を用いずに、できるだけ小さなブロックからなるように与えられた文節列のブロック化を遂行するものである。

さて、我々のブロック化の方法の基本は次の原則に基づく。

- (1) 文頭側から順次ブロック化していく。
- (2) N ブロック先までをブロック化の範囲として調べる。

我々の実験によれば、 N は、具体的には 3 が適切であった (詳しくは 2.5 節で述べる)。ブロック化ということは構造の理解の過程であり、上記の (1), (2) は人間の理解の過程に相應しているものと考えている。(1) は理解度の低い自包的 (self-embedding) な文構造を避けることに通じ、(2) は、人間の短期記憶の深さがあまり大きくないということに通じるものと推定できる。具体的なブロック化のアルゴリズムは次節で述べるが、図 3 に基

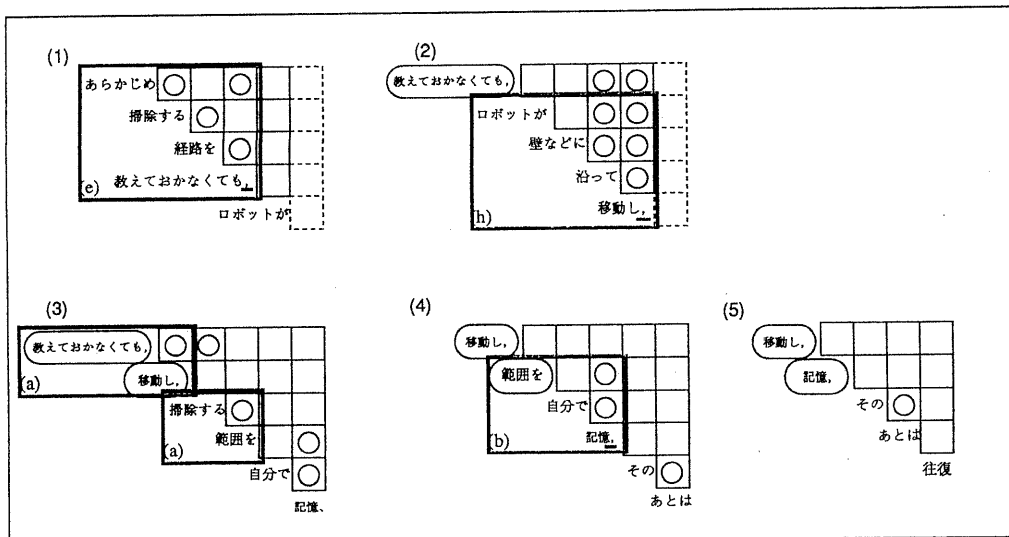


図 4: 解析例

本的なブロック化ルールを示す。ここでは $N = 3$ とした。

$N = 3$ の時、ブロック内の依存可能性の組合せは $2^6 = 64$ 通りが考えられる。これらすべての組合せを図示すると図 3 のようになる。この中で、図の (a)–(l) の 32 通りがブロック化可能なパターンである。また、ブロック化が可能なパターンの中で (d,e,h,k,l) については曖昧な状態を表現する係り受けブロックである (曖昧ブロック)。

その他のパターン (32 通り) は、非交差条件によりブロック化不可能なパターン [(m)–(t)] (26 通り) と出現しえないパターン [(u)–(x)] (6 通り) である。

2.4 骨格構造解析

骨格構造解析の手順は以下の通りである。今回の実験では $N = 3$ ブロック先までをブロック化の範囲とした。

- (1) 入力文を形態素処理し、文頭のブロックを IB とする。
- (2) IB に対して、前節で述べた方法により係り受け解析を行ないブロック化処理を可能な限り遂行する。ただし、係り先が表 3 に示すブ

ロック化停止文節 I, あるいは II となった場合には、そこでブロック化を停止する。

- (3) ブロック化が停止した次のブロックを IB とする (次のブロックが無ければ終了)。
- (4) IB より $N = 3$ ブロック前の係り受けブロックを PB とする。 PB に対して、(2) と同様の処理を行なう。ただし、ブロック化を停止するのは、ブロック化停止文節 I の場合のみとする。ブロック化が停止した時、次のブロックが IB より文頭側のブロックであったら、それを PB として (4) を繰り返す。そうでなければ、それを IB として (2) に戻る (次のブロックが無ければ終了)。

表 3: ブロック化停止文節

I.	(a) 「体言+は+、(読点)」
	(b) 「体言+(では、にとっては、としては、によると)+、(読点)」
	(c) (a),(b) が出現しない文で、「体言+は」
	(d) 「接続詞+,」
II.	(a) 文節カテゴリ「体並、用並」
	(b) 読点が含まれる文節。

図1の例文の場合は以下になる(図4). 文頭より解析を初め, 図3-(e)を適用して文節「あらかじめ」から「教えておこななくても,」までを係り受けブロックとする. 文節「教えておこななくても,」には読点があるのでここでブロック化を停止し次の文節に進む(図4-(1)). 次に「ロボットが」から「移動し,」までに図3-(h)を適用する. ここでも読点が出現するのでブロック化を停止し次の文節に進む(図4-(2)). この時, 以前にブロック化した2つの文節に図3-(a)を適用する(図4-(3)). この後は, 図3-(a),(b)を適用することにより, 図4-(5)になる. この例文の解析結果を図5に示す. この例では, 文節「あらかじめ」「教えておこななくても」「ロボットが」「壁などに」の係り先に曖昧さがあり, 文節「移動し,」「記憶,」「繰り返しながら」「し,」が並列の可能性があると解析される.

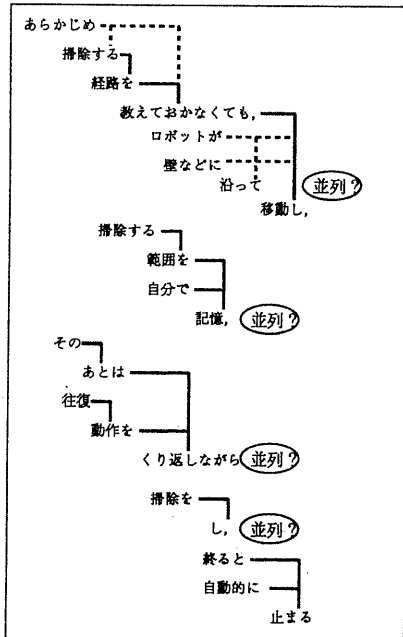


図5: 骨格解析結果

2.5 実験結果

実験は, 朝日新聞記事より文字数が50文字以上80文字未満, 80文字以上の各50文, 合計100文に対して行なった(表4).

今回の実験では, 形態素解析用の辞書として, 自立語辞書についてはEDR日本語単語辞書より単語の見出し・品詞情報のみを取り出したものを,

機能語辞書については我々が実際のテキストベースから収集し拡張, 整理した複合機能語を含む辞書(約200のグループ, 見出し語数約1500語)[5]を用いた. また骨格構造解析には正しく形態素解析され, 正しく文節カテゴリが付与された例文を用いている.

表4: 実験結果

	50-80文字(50文)		80文字以上(50文)	
	正解(文)	曖昧	正解(文)	曖昧
$N=2$	46	0.4(3)	44	0.72(2)
$N=3$	48	0.8(3)	47	1.32(3)
$N=4$	48	0.94(3)	47	1.48(4)

曖昧: 1文あたりの曖昧ブロック数

[()]内は1文中の曖昧ブロックの最大数

$N=3$ (2.3節参照)とした時, 骨格構造解析が正しく行なわれたのは, 50文字以上80文字未満の文で48文, 80文字以上の文で47文である. また, $N=2$ とした時には, $N=3$ で正しく解析された中の4文が誤って解析された. これは $N=2$ としてブロック化を行なうと, 図3-(e)の文節Aは文節Bに, 図3-(l)の文節Aは文節B, Cに係り先が決定されてしまうからである. そのため図6の例では, 文節「打ち出した」「アメリカ政府が」の係り先が誤って解析される. 同様の理由で, 先に示した図1の例文においても「あらかじめ」が「掃除する」に係ると解析されてしまう.

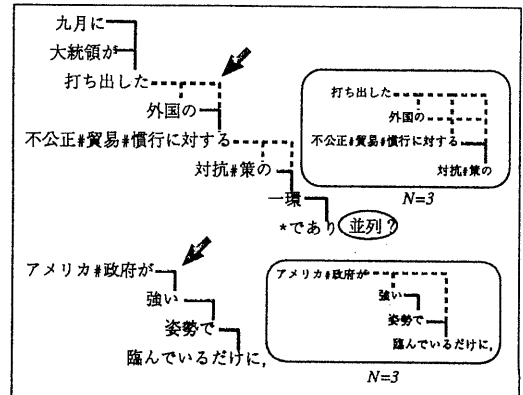


図6: $N=2$ とした時の解析失敗例

$N=4$ とした時は, $N=3$ と同数の例文を正しく解析することができた. しかし, $N=4$ の場合は $N=3$ よりも, 曖昧な状態を含む係り受けブロックが多くなる. これらの理由から, $N=3$ が適切であると判断した.

以下に $N = 3$ とした時の解析例を示す。

- (1) 規格が統一されず混乱した現行のVTRの反省にたつて、関連機器の統一を図る(図7)

解析例(1)では、「規格が」～「混乱した」と「混乱した」～「反省に」が曖昧ブロックとして解析される。

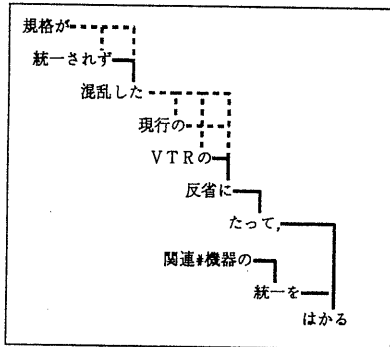


図7: 解析例(1)[正解]

- (2) …、情報産業の発展で、将来、米国の市場規模がさらに拡大すると予想されることから現地生産に踏み切った(図8)

(2)では、文節「発展で、」将来、」の係り先が誤って解析されている。これらの文節には読点が含まれているからである。読点が出現すると2.4節で述べたように、ブロック化を停止して次のブロックに進み、可能な限りブロック化を行なった後に、停止した文節のブロック化を行なうため「発展で、」「将来、」の係り先が「踏み切った」になると解析される。

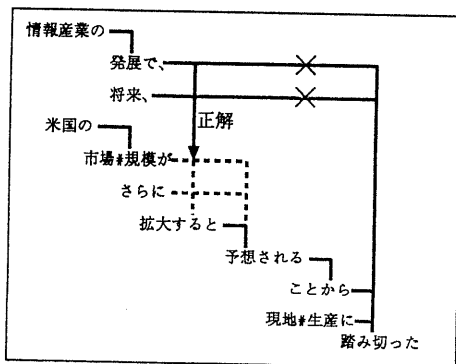


図8: 解析例(2)[不正解]

- (3) 日本側も、北洋での操業規制が強まったため新しい水産事業に魅力を感じているが、…(図9)

(3)では、文節「強まったため」の係り先が誤って解析されている。これは「強まったため」に読点がないので、続けてブロック化を行ない「新しい」に係ると解析される。読点があれば、正しく解析することができる。

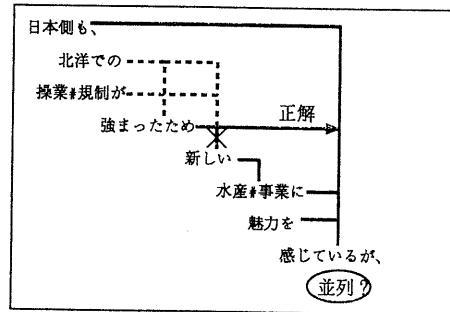


図9: 解析例(3)[不正解]

3 解析インターフェース

2章で述べた骨格構造解析は、すべてインターフェース上で行なうことができる(図10)[6]. ユーザが解析したいテキストを選択すると、計算機が自動的に形態素解析処理を施し、文節カテゴリを付与する。そして、これらの結果を元に各文節の係り受け可能な文節を示す三角表がインターフェース上に表示される。三角表では、各文節の依存可能な文節と不可能な文節が区別できるように色を変えて表示した。

形態素解析の後編集はウィンドウ上で、誤って切っている箇所、切っていない箇所にマウスを合わせてボタンを押すという操作で文節の切り直しを行なうことができる。未知語が出現した場合には、ウィンドウ上で品詞名を選択することにより辞書登録される。また文節カテゴリが誤っている場合にも、簡単な操作で訂正することが可能である。

その後、正しく形態素解析された例文を元に骨格構造解析が行なわれ、解析結果がウィンドウ上に木構造表示される。ここでは係り先が曖昧なものについては点線で、並列の可能性があるものについては色を変えて表示される。

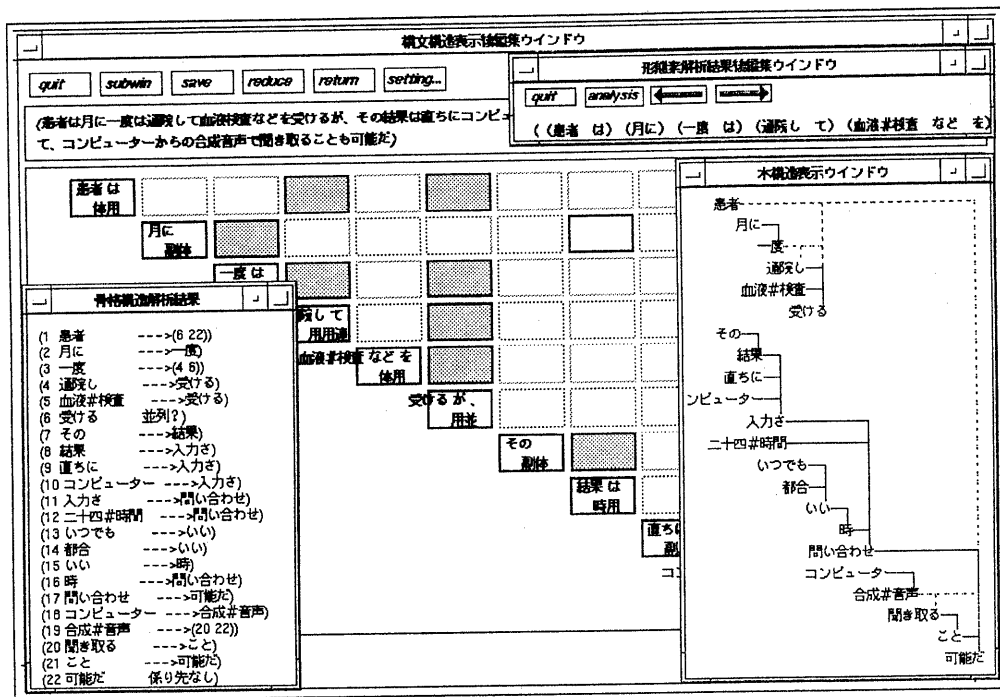


図 10: 解析インターフェース

4 おわりに

形態素情報と係り受けに関するいくつかの表層上の制約規則のみを用いて、日本語長文の骨格構造を解析する方法について述べた。本手法では、意味に立ち入らなければ解析することができない部分に対しては係り先を保留したままの曖昧なブロックとして解析する。

骨格構造解析により、長文の解析を小さな部分問題に還元することが可能となる。今後は、曖昧ブロックとして解析された部分に対してのみ、意味情報、文脈情報を用いて詳しく解析する方法について検討していきたい。また、骨格構造のままでも大量データに適用することによる応用の可能性を検討したい。具体的には、類似文検索など構造を考慮した高度なテキスト検索[7]のための大規模データベースの構築に役立てたいと考えている。

{ 形態素解析では、「日本語単語辞書評価版第 2.1 版」
株式会社日本電子化辞書研究所}を使用した。 }

参考文献

- [1] 金, 江原: 日英機械翻訳のための日本語長文自動短文分割と主語の補完, 情報処理学会論文誌, Vol.35, No.6, 1994
- [2] 黒橋, 長尾: 格構造解析への評価関数の導入による統語的曖昧性の解消, 情報処理学会 N L 研, 92-9, 1992
- [3] 黒橋, 長尾: 長い日本語文における並列構造の推定, 情報処理学会論文誌, Vol.33, No.8, 1992
- [4] 山下, 安原: 形態素情報による日本語の係り受け解析, 情報処理学会 N L 研, 98-2, 1993
- [5] 兵藤, 池田: スロット表現による複合機能語の処理, 情報処理学会第 45 回全国大会, 1992
- [6] 兵藤, 河田, 青山, 浅井, 池田: 構文テキストベースと意味分類コードを用いた類似例文検索への応用, 情報処理学会 N L 研, 100-13, 1994
- [7] 兵藤, 池田, 係り受け構造の照合に基づく用例検索システム T W I X, 電子情報通信学会論文誌, Vol.J77,D-II, No5, 1994