表層的情報と N 近傍ブロック化手法による日本語長文の骨格構造解析

兵藤安昭† 池田尚志†

日本語の構文解析では，格構造など意味情報を用いた処理が広く行われている．しかし，大規模なテキストベースに対して解析を行うためには，広範な領域を覆うことができる精密な意味情報が必要となり，現実的には容易ではない．特に長い文の解析は困難である．本論文は，意味情報を用いずに解析を行う方法を提案するものであり，長文に対する解析実験によって，その有効性を示す．これまでにも，意味情報を用いないで，表層的情報のみを用いた解析方法が提案されている．これらの方法では，係り可能な文節のうち最近接文節を優先するという原則と，それに対する例外規則を用いて文節の係り先を決定して，完全な構文解析木を求めることを追究している．本論文では，N (=3) ブロック内での係り可能性の組合せにより係り先を決定する（ブロック化する）という原則と，それに伴うブロック化のアルゴリズムを提案し，これによって日本語長文に対して高精度な骨格構造解析ができることを示す．骨格構造とは，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，文全体の構造を把握するものである．新聞記事を対象に実験を行ったところ，約94%の文に対して正しく解析された．また，80文字以上の文では，曖昧なブロックとして平均1.3個，並列構造を構成する可能性がある文節は平均2.2個含まれ，それ以外の文節については，係り先が特定されている．

Skeletal Syntactic Analysis of a Long Japanese Sentence Based on Surface Information and N Neighborhood Blocking Technique

YASUAKI HYODO † and TAKASHI IKEDA †

This paper describes syntactic analysis technique for Japanese without using semantic information. Along with a syntactic analysis of Japanese, most system use semantic information such as case structure. However, for dealing with a large scale textbase, it is difficult to give them sufficient detailed semantic information that cover broad area. Another approach is to use only surface information. Methods along this approach detect the unique dependency structure of a sentence by using the nearest possible modifiee principle and exceptional rules. Our system which uses only surface information is implemented with a phrasal blocking technique. By definition a block is composed of several consecutive phrases or blocks. The dependency structure within one block is unique in most cases but is possibly ambiguous in case semantic information are essential to obtain a unique structure. This enables our system to successfully abstract the "skeletal structure" of long Japanese sentences. Skeletal structure is a tree structure that might contain ambiguous blocks in which dependency relations are not able to be decided without semantic information. To evaluate the performance of our system, we have applied it to analyze 300 Japanese sentences from newspaper articles and obtained approximately 94% of accuracy.

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),2)]．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うことがその主な原因である．

† 岐阜大学工学部
Faculty of Engineering, Gifu University

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とする方法がよく行われている[3)]．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われているような意味素性やシソーラスを用いた方法では必ずしも十分ではない．

文が長くなるのは多くの場合，名詞句の並列，連用中止法による並列などの並列構造や連体埋め込み構造

が含まれるためである．文献2）では，単語間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うことにより，長文の解析を効果的に行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，困難な問題も含んでいる．

本論文は，意味情報を用いずに解析を行う方法を提案するものであり，長文に対する解析実験によって，その有効性を示す．

これまでにも，意味情報を用いないで表層的な情報のみを用いて構文解析を行う方法も試みられている[4),5)]．これらの方法では，係り可能な文節のうち最近接文節を優先するという原則と，それに対する例外規則を設けることにより文節の係り先を決定して，完全な構文解析木を求めることを追求している．しかし，意味を考慮しないと曖昧さを排除できない部分についても表層的情報のみを用いて解析することになるため，どうしても誤りを避けることはできない．

本論文では，最近接文節に係るという原則に代えて，N（=3）ブロック内での係り可能性の組合せにより係り先を決定する（ブロック化する）という原則と，それに伴うブロック化のアルゴリズムを提案し，これによって日本語長文に対して高精度な骨格構造解析ができることを示す．

本手法で述べる骨格構造とは，必ずしも完全な係り受けの木構造をなすものではなく，N ブロック内で複数の係り先が考えられる部分や並列構造などのように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．

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．また骨格構造レベルのままのデータでも，全自動的に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（構文的タグの付いたコーパス）の構築に役立てることもできる[8)]．

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る．その際すべての文節について係り可能性を調べることはせず，係り先の範囲を N ブロック以内とする仮説および，表層的情報による他のいくつかの制約条件に基づいて係り先を決定する．このブロック内には，各文節の係り受けパターンにより，係り先が曖昧な文節が含まれている場合もある．この処理を文末に至るまで繰り返す．

今回は，朝日新聞記事より，文字数が30文字以上50文字未満，50文字以上80文字未満，80文字以上の各100文，合計300文に対して実験を行った．骨格構造解析の入力には，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ものを用いている．その結果，$N=3$ とした場合に約94%の文に対して，正しく解析することができた．また，80文字以上（平均20文節）の文では，曖昧なブロックとして平均1.3個，並列構造を構成する可能性がある文節は平均2.2個含まれ，それ以外の文節については，係り先が特定されている．

以下，2章では前処理となる形態素解析と文節カテゴリについて述べる．3章では，本手法で用いる係り受けブロック化，骨格構造解析アルゴリズムについて述べる．4章では，骨格構造解析の実験結果を示し，$N=2, 3, 4$ とした時の評価を行い，解析成功例，失敗例についても詳しく述べる．

2. 形態素解析と文節カテゴリの付与

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つの内容語と0個以上の機能語）にまとめるところまでを行う．この際，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っている．

次に，各文節に文節カテゴリを付与する（**図1**）．文

原文

両社はこの欠点を克服した新記録膜を開発したことから，これまでに蓄積した技術を持ち寄って，さらに高品質である光ディスクを共同で開発する (66 文字)

((両社 は 体用)(この 副体)(欠点 を 体用)(克服し た 用体)
(新記録#膜 を 体用)(開発し た 用体)(こと から， 体用)
(これ まで に 体用)(蓄積し た 用体)(技術 を 体用)
(持ち寄っ て， 用用)(さらに 副用)(高#品質 体用)
(*である 用体)(光ディスク を 体用)(共同 で 体用)
(開発する 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
Fig. 1 Example text with *bunsetsu* category.

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ライズしたもので，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つに大分類し，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5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を構成する可能性があることを示している．また，文節中に時を表す名詞や，主題を表す機能語（は・では等）が含まれる時は，各々「時用」「は用」「は＊用」として扱う．

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
  1.「彼は東京で仕事をしている」で…体用
  2.「彼は勉強中で，…」で…体用（不正解）

  2.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，用用）…
  というように文節の再構成を行う必要がある．
- 「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．

表1　基本的文節カテゴリ
Table 1　Basic *bunsetsu* categories.

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.
用用	用言に係る用言文節.
用体	体言に係る用言文節.
用終	係り先を持たない用言文節.
副用	用言に係る副詞文節. ただし形容動詞連用形「静かに」や「犬のように」「犬みたいに」等も「副用」とする
副体	体言に係る連体詞文節.
副副	副詞に係る副詞文節. 「程度副詞」で，直後に副詞が続く時.
接用	用言に係る接続詞文節.

表2　その他の文節カテゴリ
Table 2　Other *bunsetsu* categories.

体並	体言並列構造を構成する可能性がある文節. 「名詞＋(，ともやかつ…)」. 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	用言並列構造を構成する可能性がある文節. 「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.「今日」等.
は用	「体言＋は＋，(読点)」 「体言＋[では によると …]＋，(読点)」
は＊用	文節カテゴリ「は用」が出現しない文で「体言＋は」

3. 骨格構造解析

3.1 文節カテゴリに基づく係り受け規則

各文節の係り可能な文節は，文節カテゴリに基づいて決定する（**図2**）．例えば，文節カテゴリ「体用」が付与された文節は「用体」「用用」「用終」「用並」の文節に係り可能である．

図2では，係り先を持たない文節がいくつかある．これらを係り先保留文節と呼ぶ．

- 係り先保留文節 I…「は用」「は＊用」「接用」
  文節カテゴリ「は用」「は＊用」は主題を表す文節であって，一般に後続する広い範囲に勢力をもつ．すなわちその範囲内のいくつもの文節に係る可能性をもっている．また，文節カテゴリ「接用」の係り先も後続する広い範囲に勢力をもつ．これらは，意味を考慮せずに文節カテゴリだけで係り先を特定することは困難である．したがってここでは，これらの文節については係り先を保留しておく．

文節カテゴリ2

文節カテゴリ1 ＼ 2	体用	体体	体終	用体	用用	用終	副体	副用	副副	接用	体並	用並	時用	は用	は＊用
体用				○	○	○						○			
体体	○	○	○								○		○	○	○
体終															
用用				○	○	○						○			
用体	○	○	○								○		○	○	○
用終															
副用				○	○	○						○			
副体	○	○	○								○		○	○	○
副副								○	○						
接用															
体並															
用並															
時用				○	○	○						○			
は用															
は＊用															

○：文節カテゴリ1が2に係り可能である

図2　係り可能性テーブル
Fig. 2　Table of dependency rule.

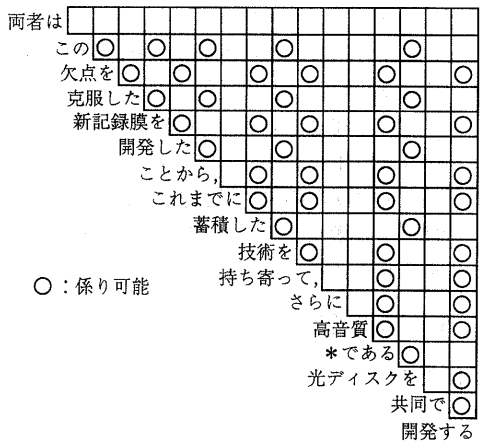


図3 係り受けの三角表
Fig. 3 Triangle table of dependency.

●係り先保留文節II…「体並」「用並」
文節カテゴリ「体並」「用並」は，後続する体言，用言文節と並列したり，因果関係を表現する構造を作る（以下では，便宜上これらをすべて並列構造と呼ぶ）．これらについても，意味を考慮せずに文節カテゴリだけで係り先を指定することは困難である．これらについても同様に係り先を保留しておく．

図3は，図2の関係に基づいて文節間の係り受け関係を表示したもの（係り受けの三角表と呼ぶ）である．例えば，図3の文節「この」は「欠点を」「新記録膜を」「ことから,」「技術を」「光ディスクを」に係り可能であることを示す．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「開発した：用体」は「これまでに：体用」に係り可能である．しかし「ことから,：用用」が「これまでに：体用」に係り不可能であるため，非交差条件により「開発した」は「これまでに」に係り不可能となる．

3.2 *N* 近傍による係り受けブロック化規則

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するものとする．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するものとする．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

(両社は
(((((((この 欠点を) 克服した) 新記録膜を) 開発した) ことから,)
(((これまでに 蓄積した) 技術を) 持ち寄って,))
((((さらに 高品質) *である) 光ディスクを) (共同で 開発する))))

図4 完全な構文解析結果
Fig. 4 Complete bracketed structure.

(両社は;
(((((この[1] (欠点を[17] 克服した) 新記録膜を[17]) 開発した) ことから,)
(これまでに[2] (蓄積した[27] 技術を) 持ち寄って,[27]))
((((さらに 高品質) *である) 光ディスクを) (共同で 開発する))))

(A^{1} (B^{17} C) D^{17})：A は B または D のいずれかに係る

図5 骨格構造解析結果
Fig. 5 Skeletal bracketed structure.

するものとする．なお，ここで係り受けという用語は「AとBが遊ぶ」の場合のAとBの関係のように，純粋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 $b1, b2, ..., bn$ 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（$b1\ b2 \ldots bn$）である．係り受け関係が完全に解析されれば，これらは，大きさ2の係り受けブロックの入れ子構造の形になる（**図4**）．

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（**図5**）．

さて，我々のブロック化の方法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（*N* 近傍による係り受けブロック化）．

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4章で述べる）．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応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．

図6に *N* 近傍による係り受けブロック化規則を示す．以下では，$N=3$ として説明する．ブロック内の依存可能性の組合せは $N=3$ の時，$2^6=64$ 通りが考えられる．

この中で，図6(a)-(l)の30通りがブロック化可能なパターンである．このようなパターンが出現した時には，図6の点線で示す範囲でブロック化を行うことにする．

また，ブロック化が可能なパターンの中で図6(h)(i)(j)(l)についてはブロック内に係り先の曖昧なものが含まれる係り受けブロックである（以下では曖昧ブ

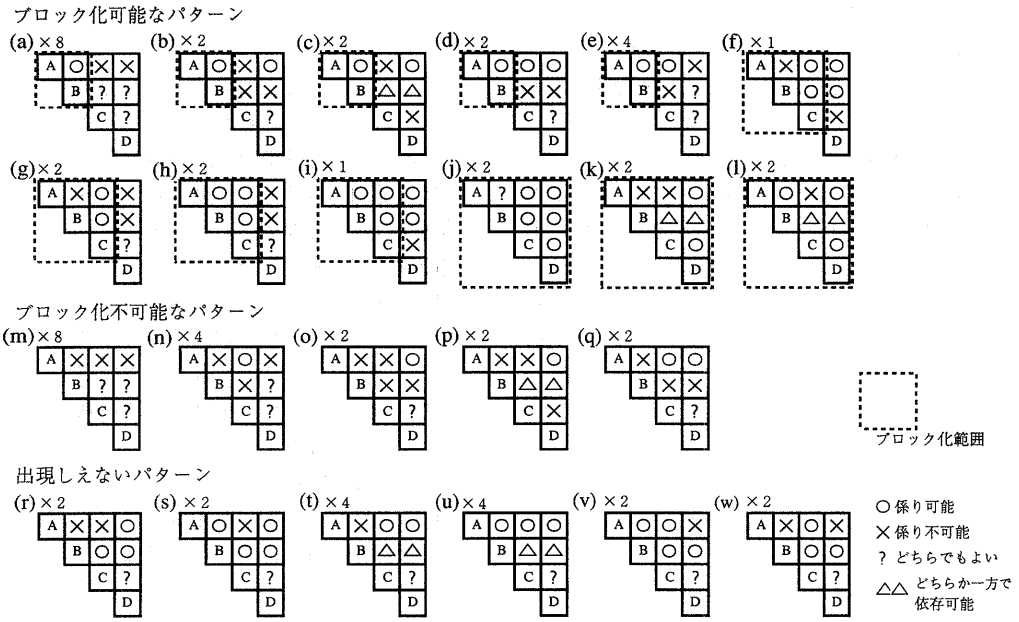


図6 N 近傍による係り受けブロック化規則
Fig. 6 N neighborhood dependency blocking rule.

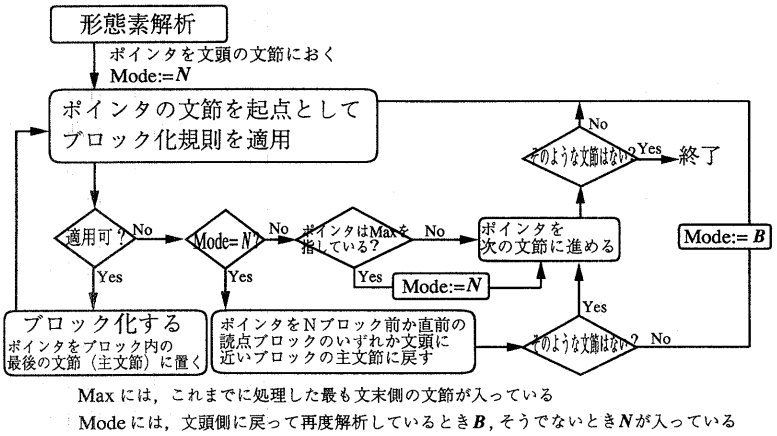


図7 骨格構造解析の手順
Fig. 7 Procedure for skeletal analysis.

ロックと呼ぶ).

その他のパターン（34通り）は，ブロック化が不可能なパターンである．図6(m)の8通りは文節AがB, C, Dのいずれにも係りえないためブロック化が不可能なパターン，図6(n)-(q)の10通りは，非交差条件によりブロック化できないパターン，図6(r)-(w)の16通りは，文節カテゴリの組合せにより出現しえないパターンである．本論文では，文節カテゴリは，係り可能な文節のタイプを一意に定めている．したがって，例えば図6(r)では，文節BはC, Dに係り可能となっているから，C, Dは同じタイプの文節であることになる．一方，文節AはCには係り不可能であるが，Dには係り可能となっているから，C, Dは異なるタイプの文節であることになり矛盾である．よって，このようなパターンは出現しないことになる．

3.3 骨格構造解析

骨格構造解析の手順を**図7**に示す．基本的には，文節カテゴリによる係り受け規則と N 近傍によるブロック化規則に基づく文節列のブロック化処理を文頭から文末に向かって実行していくのであるが，ブロック化が停止してしまった場合や，読点，体言並列ブロックが現れた場合などに関しては，以下の項で述べるよ

うなブロック化処理の制御を行う．

3.3.1 読点によるブロック化の一時停止

読点を含む文節は，あるレベルの意味的なまとまりを示す切れ目に対応するとともに，次に可能な係り受け関係を抑止する働きがあるものと考えられる．そこで，本アルゴリズムではその文節までをブロック化の範囲として解析し，ブロック化を一時中断しておく(読点ブロック)．ここで，N ブロック前か，直前の読点文節のいづれか遠いブロックに戻り，ブロック化が可能な状態に変化しているかどうか順次調べる*．その後，読点ブロックの次の文節を起点とするブロック化処理を行う．ただし，体言並列を表現している場合，つまり「体言＋，(読点)」の場合は，3.3.2 項で述べる体言並列に伴う処理を行う．

また，N ブロック内の係り可能性のパターンが図 6(m)-(q)に該当する時には，ブロック化が行われず，必然的にブロック化が停止する．この場合も，読点で停止した時と同様の処理を行う．

このようにして，N 文節よりも遠い文節と係り受け関係をもつことが可能となるが，この場合でも，ブロックで見るとその関係は N ブロック以内になっている．

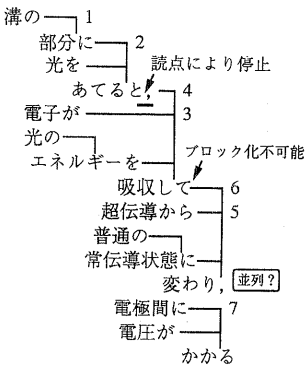


図8 読点によるブロック化の一時停止
Fig. 8 Suspension of blocking by a comma.

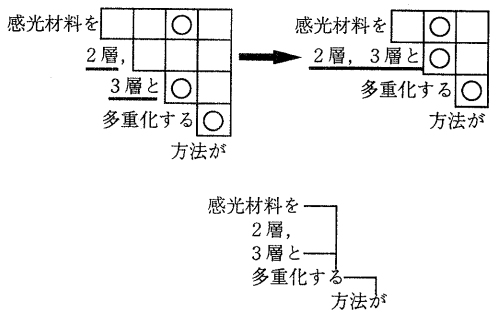


図9 体言並列ブロックの飛び越し
Fig. 9 Pass over noun coordinate block.

* ただし，この場合は読点ブロックであっても停止せずに処理を続行する．

図8の木につけた番号は，以上の処理によってこの順にブロック化が行われることを示す．

3.3.2 体言並列ブロックの飛び越し

体言並列のカテゴリに属するブロックは，後続する体言ブロックと並列するブロックである．つまり，後続の体言ブロックと重ね合わされて1つのブロックを構成するとみなすことができる．そこで，ここでは N 近傍として数え上げる際，数え上げの対象からはずしている．**図9**に示す例では，「2層,」「3層と」を1つのブロックとみなすことにより「感光材料を」が「多重化する」に係ると解析される．

3.3.3 解析手順例

図1の例文の解析手順を以下に示す(**図10**)．文頭より解析を始めるが，文節「両者は」の係り先は保留されているため，次の文節「この」よりブロック化を行う[図10(1)]．可能な限りブロック化を行うと，文節「ことから,」には読点が含まれるため，この文節でブロック化を停止する[図10(2)]．この結果，文節「この」から「ことから,」までが1つにブロック化される．通常は読点が出現すると，文頭側に戻ってブロック化を再び行うが，この場合は文節「両者は」しか存在しないので，つぎの文節より処理を続ける．この時点で，「この欠点を…ことから,」の係り先は「蓄積した」と「持ち寄って,」の2通りが考えられる[図10(3)]．次に「これまでに」から「持ち寄って,」までが1つにブロック化され，再び読点が出現するので，文頭側に戻って「この欠点を…ことから,」と「これまでに…持ち寄って,」が1つにブロック化される[図10(4)]．これにより，「この欠点を…ことから,」の係り先は必然的に「持ち寄って,」に決定される．以下，同様に解析を進める．解析結果を**図11**に示す．

4. 実験結果

実験は2回に分けて行った．まず始めに，朝日新聞記事より文字数が30文字以上50文字未満，50文字以上80文字未満，80文字以上の各50文，合計150文に対して実験を行い，ブロック化処理のアルゴリズムの検討を行った．これによって得られたシステムの結果が実験1(**表3**)である．次に，これを上記と同種の別の150文に適用した．その結果が実験2(**表4**)である．実験2の結果が，実験1の結果とほぼ同様であり，また，かなりの精度の結果が得られたことから，このブロック化処理のアルゴリズムは，ほぼ妥当なものであると考えている．

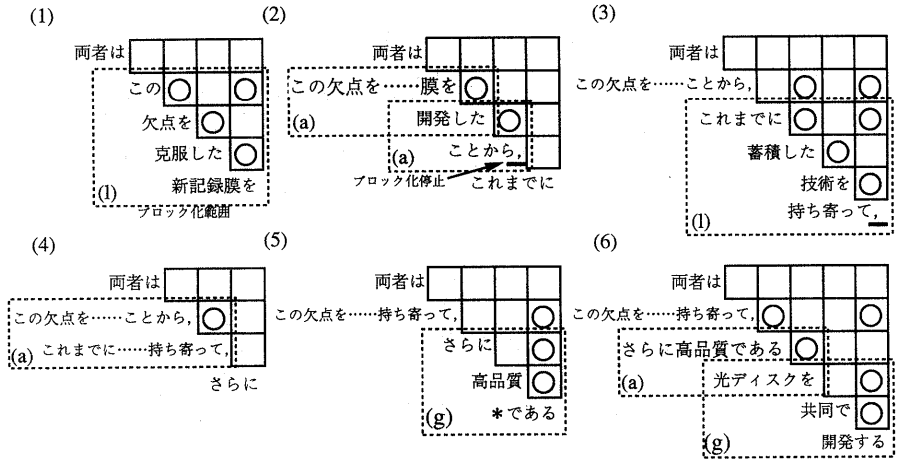


図 10 例文解析過程
Fig. 10 An example of skeletal analysis process.

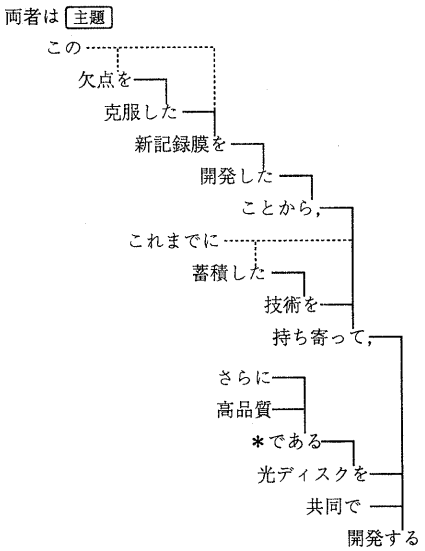


図 11 例文解析結果
Fig. 11 Result of skeletal analysis.

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[6] 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）[7] を用いた．この形態素解析システムでは，文節の認定に失敗した文は 300 文中 34 文，また文節カテゴリの付与に失敗した文は，文節の認定に成功した文（266 文）中 13 文であった．形態素解析に失敗した文については，解析インタフェースを用いてそれらの誤りの訂正を行い，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ものを骨格構造解析の入力として用いている．この解析インタフェースは，ウインドウベースで構築されており，簡単な操作で，文節の切り直しや文節カテゴリの訂正，未知語登録を行うことができる．

4.1 $N=2, 3, 4$ とした時の評価

本実験では，各々の例文について，$N=2, 3, 4$ とした時の骨格構造解析の評価を行った．表 3，表 4 には，$N=2, 3, 4$ とした時の各々50 文の正解文数，1 文あたりの曖昧ブロックの出現数，1 文あたりの体言・用言並列構造を構成する可能性がある文節の出現数（括弧内は最大数）を示す．ここで述べる正解とは，解析された曖昧ブロックを含む係り受けの木構造の中に，正しい木構造が含まれているものをいう．また，以下では，実験 1, 2 の解析結果を 1 つのものとして，まとめて扱うことにする．

$N=3$ とした時，解析に失敗したのは，16 文であった．このうちの 1 文は $N=4$ とした場合には正しく解析できたが（**図 12**），残りの 15 文は $N=4$ とした場合でも解析に失敗している．これらの失敗例については 4.3 節で詳しく分析する．

また，$N=2$ とした時には，この 16 文に加えて 7 文，合計 23 文が誤って解析された．**図 13** に示したものは，その 7 文の中の 1 つである．$N=2$ とした時のブロック化規則は**図 14** のようになる．そのため図 13 の例では，「打ち出した」の係り先は「外国の」「不公正貿易慣行に対する」に係り可能であると解析され（正解…「対抗策の」），「アメリカ政府が」の係り先が「強い」であると誤って解析されている（正解…「臨んでいるだけに」）．

$N=4$ としたときの方が $N=3$ とした時より正解数

表3 実験結果1
Table 3 Result of experiments 1.

	30–50 文字 (50 文)			50–80 文字 (50 文)			80 文字以上 (50 文)		
	正解 (文)	曖昧	並列	正解 (文)	曖昧	並列	正解 (文)	曖昧	並列
$N=2$	47	0.24(2)		46	0.40(3)		43	0.72(2)	
$N=3$	47	0.44(2)	0.66(2)	48	0.80(3)	1.36(4)	46	1.32(3)	2.24(5)
$N=4$	47	0.48(2)		48	0.94(3)		46	1.48(4)	

表4 実験結果2
Table 4 Result of experiments 2.

	30–50 文字 (50 文)			50–80 文字 (50 文)			80 文字以上 (50 文)		
	正解 (文)	曖昧	並列	正解 (文)	曖昧	並列	正解 (文)	曖昧	並列
$N=2$	50	0.3(2)		47	0.52(3)		44	0.78(6)	
$N=3$	50	0.6(3)	0.92(3)	48	0.78(4)	1.54(5)	45	1.26(6)	2.74(6)
$N=4$	50	0.84(3)		48	1.1(4)		46	1.64(6)	

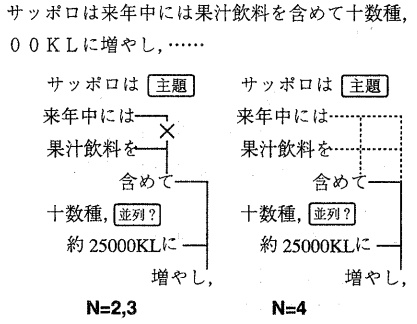


図12 $N=4$ とした時の解析成功例
Fig. 12 A successful result in case of $N=4$.

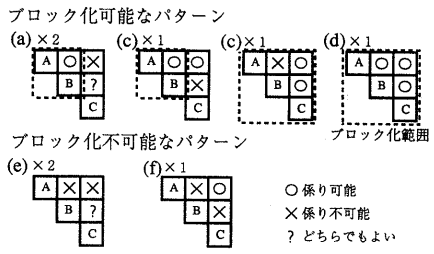


図14 $N=2$ とした時の係り受けブロック化規則
Fig. 14 Dependency blocking rule in case of $N=2$.

九月に大統領が打ち出した外国の不公正貿易に対する対抗策の一環であり，アメリカ政府が強い姿勢で臨んでいるだけに，単に業界の問題だけでなく……

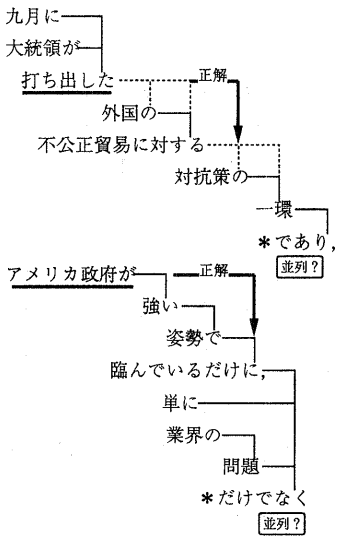


図13 $N=2$ とした時の解析失敗例
Fig. 13 A failed result in case of $N=2$.

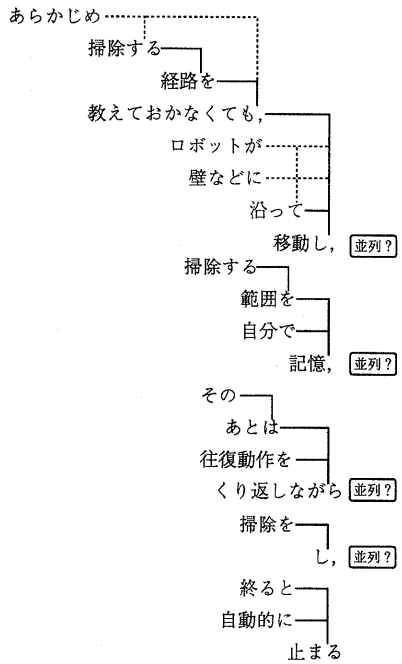


図15 解析例(1)
Fig. 15 An example of successful result (1).

が1文多いが，その代わり曖昧な状態を含む係り受けブロックの数が多くなるので，今回の実験では $N=3$ が適切であると判断した．

4.2 解析成功例

$N=3$ とした時の解析成功例を示す．

例文1：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（図15）

この例文では「教えておかなくても，」の係り先は「移動し，」であると解析されているが，実際の係り先は「移動し，」「記憶，」「(掃除を) し，」「止まる」が正解である．しかし，これは，この後の処理で（本手法では扱っていないが）並列解析を行うことにより正しい係り先を求めることが可能となる．また，この例文で $N=2$ として解析を行うと，「あらかじめ」の係り先は「掃除する」と誤って解析される．

例文2：一度汚れた環境を元に戻すことの難しさを身をもって体験している日本は，技術援助の先進国となり，破壊されつつある地球環境にストップをかけることに一役買う義務があるはずだ（図16）

本システムでは，主題を表す「は，」は係り先保留文節とし，係り先を特定していない．実際には，この例での「は，」は「なり，」「かける」「買う」「あるはずだ」に係ることになる．

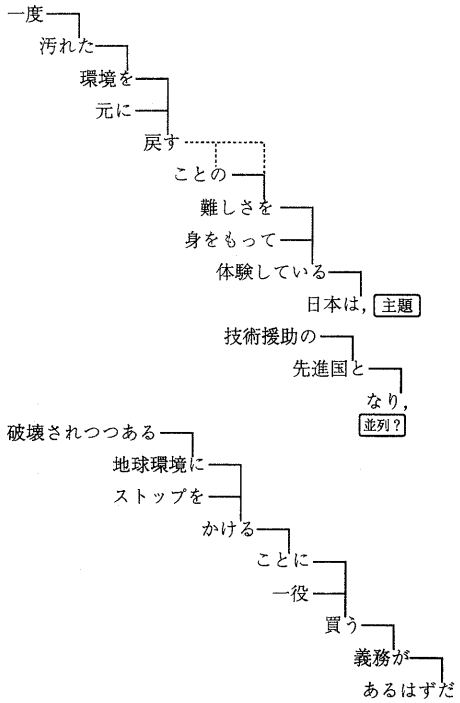


図16 解析例(2)
Fig. 16 An example of successful result (2).

4.3 解析失敗例

$N=3$ とした時の解析失敗文は16文であった．以下にこの失敗の内容について分類する（**表5**）．16文中以下の(1)，(2)に分類される10文については本手法の枠内で解決できる可能性がある．

(1) **係り先保留文節の改善により解決の可能性があるもの**

図17の例文では係り先保留文節「は用」に「用言＋(に) は，」という規則を加えることにより，正解を求めることが可能である．現在，係り先保留文節は，単独の文節のみにより決定しているが，前後の文節列のパターンを見て決定することなどを含め，係り先保留文節を改善することによって，解析の精度を上げる余地がある．

(2) **文節カテゴリ「用用」の細分類により解決の可能性があるもの**

本実験では，動詞，形容詞，形容動詞をすべて用言として扱っている．そのため**図18**の例文のように「用用」というカテゴリが付与された文節「技術提携して」は形容詞「新しい」にも係り可能であるとして処理が行われるため解析に失敗している．形容詞については別のカテゴリを与えることにより，この例は解決可能である．

表5 $N=3$ とした時の解析失敗文（16文）の内訳
Table 5 Classification of failures in case of $N=3$.

解決の可能性があるもの……10文	
係り先保留文節の改善	7文
係り先ルールの改善	3文
解決が困難なもの……6文	
Nを大きくしないと正しく解析できない文	3文
「は」の係る範囲が正しく特定できない文	1文
人間が読んでも係り先が判然としない文	2文

これまでのコンピュータでこうした機能を実現するには，膨大なデータを蓄えて，それを逐一計算するしかなかったので，高速の大型計算機で長時間処理していた

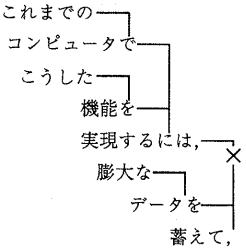


図17 解析失敗例(1)
Fig. 17 An example of failed result (1).

また 両者は技術提携して新しい産業用ロボットの開発や，
大日機工でのＥＣＤ社の製品の委託生産を行う

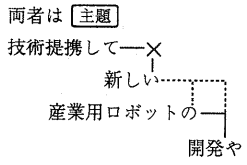


図18 解析失敗例(2)
Fig. 18 An example of failed result (2).

光通信は高速で大量の情報を送れるが，これまでは導線を
光ファイバーに変えただけで信号は電子に変えて処理して
いたために利点が十分に生かされなかった

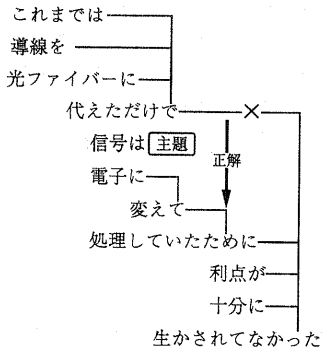


図19 解析失敗例(3)
Fig. 19 An example of failed result (3).

ハーツの争奪戦には，日本長期信用銀行系の企業も加わっ
たが，フォード側が予想を上回る１２億ドルの買値を付け
たため，敗退した

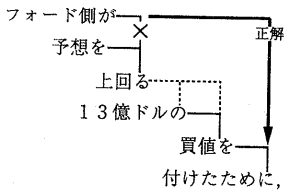


図20 解析失敗例(4)
Fig. 20 An example of failed result (4).

伝送速度は最高毎秒９６００ビットで家庭内や外部の電話
線とつながったパソコン通信にも使える

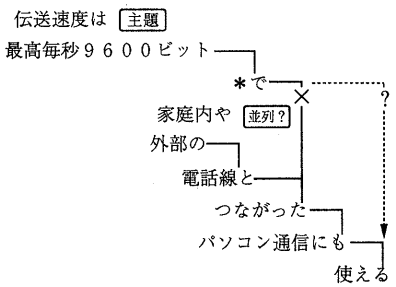


図21 解析失敗例(5)
Fig. 21 An example of failed result (5).

(3)主題を表す「は」の係る範囲が誤っているため正しく解析できない文

図19の例文では「代えただけで」の係り先が誤って解析されている．これは「信号は」が主題を表す「は」であり，現在のルールでは「は」の係る範囲を以後の文全体に及ぶとして解析しているため，「代えただけで」は「は」の係る範囲内にある「処理していたために」に係り不可能となる．「は」の係る範囲を特定する必要があるが，これは意味的な情報を考慮しないと困難な問題である．

(4)Nを大きくするか意味的な係り受けルールを導入しないと正しく解析できない文

本手法では，係り受けルールは単純に文節カテゴリを見て決めている．したがって図20の例文は「フォード側が」の係り先を誤っている．この場合，$N=5$とすれば曖昧ブロックの形で正しく解析できるが，Nの値をあまり大きくすることは係り先が特定されない曖昧な文節が増え，解析結果も無意味なものになり好ましくない．この例文の場合「フォード側が→(予想を)上回る」の係り受け関係があり得ないことを規則化できない限り解決は困難である．それには意味的な関係を調べなければ，解決することはできない．

(5)人間が読んでも係り先が判然としない文

人間が読んでも係り先が判然としない文がある．図21の例文では，「＊で」の係り先が「つながった」なのか「使える」なのかが曖昧である．

5. おわりに

本論文では，意味情報を用いないで，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た．Nブロック内の係り可能性の組合せにより係り先を決定するという原則と，それに伴うブロック化アルゴリズムを用いることにより日本語長文に対して高精度な骨格構造解析が可能となった．$N=3$として解析を行ったところ，約94%の文に対して正しく解析することができた．

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[8]のための大規模データベース（構文的タグの付いたコーパス）の構築に役立てたいと考えている．

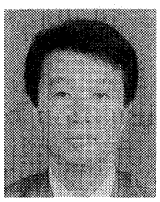
参 考 文 献

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, pp. 1018-1028 (1994).
2) 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, pp. 1022-1031 (1992).
3) 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, 情報処理学会自然言語処理研究会報告, 92-9, pp. 65-72 (1992).
4) 亀田：簡易日本語解析系 Q-JP, 情報処理学会自然言語処理研究会報告, 94-4, pp. 25-32 (1993).
5) 山下, 安原：形態素情報による日本語の係り受け解析, 情報処理学会自然言語処理研究会報告, 98-2, pp. 9-16 (1993).
6) 日本電子化辞書研究所：日本語単語辞書評価版第 2.1 版 (1994).
7) 兵藤, 池田：スロット表現による複合機能語の処理, 第 45 回情報処理学会全国大会論文集, 5 F-10, pp. 197-198 (1992).
8) 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 TWIX, 電子情報通信学会論文誌, Vol. J 77, D-II, No 5, pp. 1028-1030 (1994).

(平成 6 年 11 月 30 日受付)
(平成 7 年 6 月 12 日採録)

兵藤 安昭（学生会員）

1967 年生. 1991 年岐阜大学工学部電子情報工学科卒業. 1993 年同大学院修士課程電子工学専攻修了. 現在, 同大学院博士後期課程電子情報システム工学在学中. 自然言語処理の研究に従事. 言語処理学会会員.



池田 尚志（正会員）

1944 年生. 1968 年東京大学教養学部基礎科学科卒業. 同年工業技術院電子技術総合研究所入所. 制御部情報制御研究室, 知能情報部自然言語研究室に所属, 主任研究官. 1991 年岐阜大学工学部電子情報工学科教授. 工学博士. 主として人工知能, 自然言語処理の研究に従事. 人工知能学会, 電子情報通信学会, 言語処理学会, ACL 各会員.