禁止パターンを用いるパージング手法

榊　博史　　橋本　和夫　野垣内　出
(株)国際電信電話　研究所

1．まえがき

Parseという語は研究社新英和辞典1965年版によると「語や語群の品詞・語尾変化を説明する；（文章を）解剖（analyze）する」と訳注されており英語のPartの意味を持つParsというラテン語がその語源であることが示されている。このようにパーズ（Parse）すること、パージングは文を文法的又は統語的に解析することを意味すると解される。本報告では以下パージングという語を文をパーズ木に解析するという狭義の意味で用いることとする。パーズ木を伴なうパージング手法・パーズ木を伴なわないパージング手法双方共数多く報告されている。

さて（狭義の意での）パージングは

$$A \longleftarrow B_1\ B_2\ \text{----}\ B_N \qquad (1)$$

の形式の書き換え規則を用いて行われることは周知のとうりである。この規則は右辺にこの順に並んだ項 B_1 B_2 --- B_N が左辺の1つの項Aに書き換えられることを示す。この場合右辺の項のうちの1つが被修飾項でこれを右辺の他の項が修飾することにより左辺の1つの項が構成されると解される。この被修飾項はガバナー（governer）と呼ばれることがある。式(1)の形式の数多くの規則を制限なく用いてパージングを行なうと数多くのパーズ木（以下木と称する。）が発生し文の解析に非常な困難を来たすことが知られており、これに対する対策として式(1)の形式の書き換え規則に適用制限をもうける方法が種々取られておりこの中で拡張LINGOL"又はそれを土台とする方法が広く用いられている。

本報告は書き換え規則の一適用制限手法に関するものであり、式(1)の形式の書き換え規則の自由な適用により組み上げられた木は各書き換え規則が示す局所的部分で正しいにすぎず、上記木のより広域の部分での正しさは保証されておらず、正しくない部分を含む木は除去しなければならない、という立場に立ちパージングを行なうものである。具体的には、ある式(1)の形式の書き換え規則適用の直後得られた木につき適用された書き換え規則を含みしかもより広域の部分で正しくない部分があるかどうかを調べ、無いならば木の組立てを継続し、あるならばその木全体を消去するという方法を用いる。この正しくない部分が表題の禁止パターンに相当し、本報告中でより具体的に禁止木と呼ぶ部分である。

2．禁止木を用いるパージング手法

本報告で提案する禁止木を用いるパージング手法（以下本パージング手法と略す）の基本的な演算をブロックダイアグラムであらわすと図1のようになる。この図に於て書き換え文法辞書には式(1)の形式の書き換え文法が複数個記憶されている。書き換え規則適用部は現存木群メモリに記憶中の木群の1つを選び、この木群に適用可能な文法を1つだけ書き換え文法辞書より順次提示される文法の中から選びそれを上記選ばれた木群に適用した後、この適用結果新らしく発生した木を含む木群を禁止木包含検

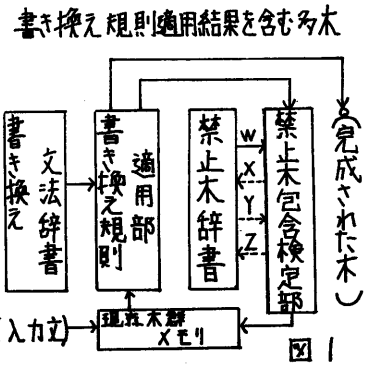


図1

定部へ送る。禁止木包含検定部では送られて来た木群中の新しく今回の書き換え文法適用の結果発生した木(被検定木と称する。)についてこの木を包含する禁止木が禁止木辞書中にあるかどうかを調べる。包含及び禁止木に関しては本節中例を用いて説明する。禁止木包含検定部と禁止木辞書のより具体的な動作について以下述べる。禁止木辞書と禁止木包含検定部間の経路Xに禁止木包含検定部より次情報要求信号が発生する毎に禁止木辞書は経路W上に異った禁止木を順次発生する。禁止木包含検定部はこの禁止木が被検定木を包含するかどうかを検定する。1回の検定が終るごとに次情報要求信号を上記経路X上に発生し次の禁止木の発生をうながす。このような動作がくり返されるのであるが、しばらくして経路Y上に禁止木辞書より、禁止木辞書より発生する禁止木は何もないという終了情報が発生すると、それ迄に被検定木を包含する禁止木が無かった場合は禁止木包含検定部は被検定部を包含する禁止木が禁止木辞書中に無いという判定を行ない被検定木を含む入力木群をそのまま出力として現存木群メモリに送出する。これに反してそれ迄に被検定木を包含する禁止木が有った場合は禁止木包含検定部は出力を発生せずこの木群を抹消する。なお経路Zは禁止木包含検定部より禁止木辞書へ包含又は不包含という判定結果信号すなわち後で述べる能率のよい禁止木辞書走査の助けとなる信号を伝送する経路である。経路W以外は動作制御のための信号であるので点線で示してある。なお書き換え規則は網羅的に適用される。

このようにして図1のループ部分の動作が行われる。なお動作の開始時には現存木群メモリに入力文が導入される。すなわち動作開始時に現存木群メモリに収容されている木群は入力文に相当するものであり1個のみである。動作の中間段階では現存木群メモリには複数個の木群が記憶されている。木群が木として完成したものは順次書き換え規則適用部より取り出される。

以上が一般的な本パージング手法に関する図1のブロックダイアグラムに沿った説明であるが次に例を用いて説明を行なう。

The boy who has cars painted blue is smart. (2)

なる英文をパーズする場合を考える。この場合動作が進行した結果、図2の形状の木群が現存木群メモリより書き換え規則適用部へ入り書き換え文法辞書よりの

NP ←——— NP ed-Part (3)

という書き換え文法の適用をうけて図3の木群が得られこれが禁止木包含検定部へ送られる事態となった場合について考える。禁止木包含検定部は送られて来た木群中の今回の式(3)の書き換え規則適用の結果新たに発生した木である左端の木を被検定木に指定して包含検定を行なう。包含検定とは被検定木中の最上位節点を含む木状の部分が禁止木辞書中に登録されている木状の部分と一致するかどうかの検定である。この場合は図3中線でかこまれた部分が禁止木辞書中に登録されているのでこの場合禁止木包含検定部は被検定木を包含する禁止木が禁止

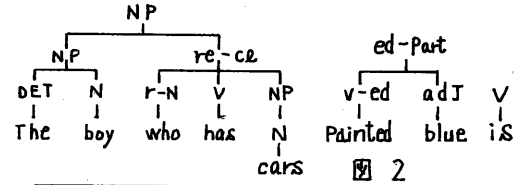


図 2

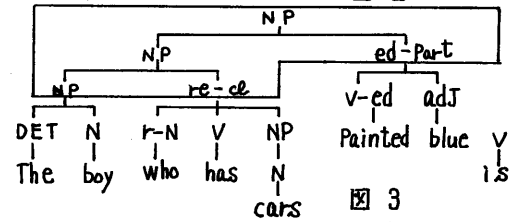


図 3

木辞書中に有るという判定を行ない禁止木包含検定部は図3の木群に関する出力を行わずこの木群を抹消する。以後の図1の本パージング手法の動作の結果、図3中の線でかこまれた禁止木を部分木として含む図4に示す木は出力として発生せず、図3とは異なる木群より発生した図5の木のみが正しいパーズ木として発生する。このような過程で本パージング手法による書き換え文法適用制限が行われる。以上が本パージング手法の基本的な動作である。なお禁止辞書中の禁止木は図6の形式で登録されている。ここでアステリスク"*"は任意化節点と称する部分で1つの節点と同様に表示されており、この部分が任意でよいことを示している。なお図6の禁止木は文法的には関係節(re-d)に後方より修飾された名詞包(NP)をさらに後方より動詞過去分詞(ed-Part)により修飾することは出来ないという禁止規則に対応するものである。次に図1中の禁止木辞書に於ける辞書走査を能率よく行うための準備として次節の説明を行う。

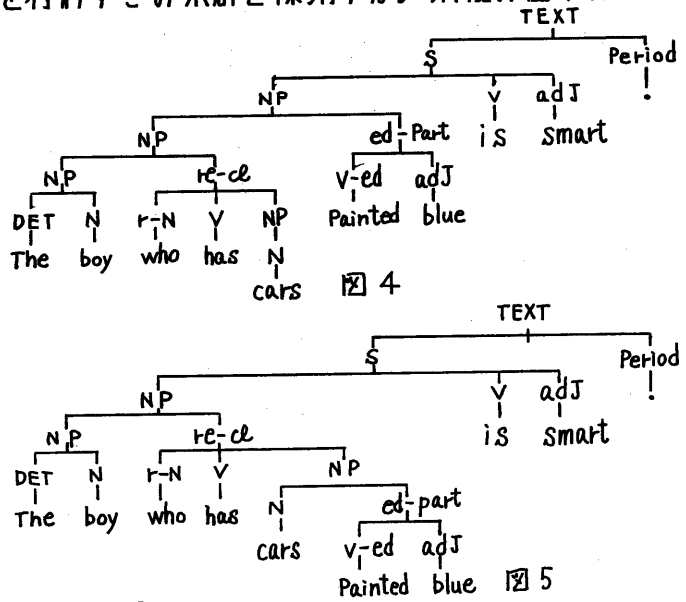


図4

図5

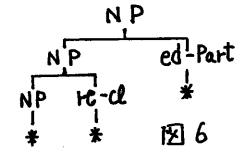


図6

3. 木に関する二、三の性質

この節及び次節に説明することのある部分はデータベースの理論と平行した議論になると思われるが自然言語用パーズ木に関する議論としてコンパクトに展開することを試みる。まず包含の定義を正式に導入する。

〔定義1〕木Aが木Bを包含するとは木Bに於て存在する0個以上の部分木のそれぞれが木Aに於てそれぞれ1つの任意化節点に置き換えられていることを称する。

但し、部分木に関しては周知のことであるが以下の定義2が与えられる。

〔定義2〕部分木とはある木の部分をなす非分離の部分で木状の形状をなすものを称する。

又定義1の系として

〔定義3〕ある木はそれ自身を包含する。

ことがわかる。なお定義1に示した任意化節点に対しては、

〔性質1〕相隣る複数の同位な任意化節点は1つの任意化節点にまとめられる。逆に1つの任意化節点は相隣る同位な任意化節点に分割できる。

という性質を与えることにする。前節で述べた例に関しては図3の木群に於ける左端の木に於けるDETとthe, Nとboy, r-Nとwho, Vとhas, NPとNとCars, V-edとpainted 及びadJとblue から成るそれぞれ部分木をそれぞれ1つの任意化節点に変更し次いで性質1を用いてまとめることにより図6の木が得られるところから図6に示す禁止木は図3の左端の木を包含することが

示されこの場合の包含検定結果は包含であることが判明する。このように任意化節点を用いた禁止木の表現による定義1の意味での包含検定表現は第2節に示した包含検定の厳密な定義である。次に包含関係に関する次の定理を導入する。

〔定理1〕木Aが木Bを包含するとき木Aに包含されない木は木Bにも包含されない。

（証明）定義1に示す操作は可逆な操作である。このことより木Aに於ける任意化節点の一部分又は全部を任意の木に変更することによって木Cが得られない。所で木Bが木Cを包含するということは木Aの任意化節点の指定されたいくつかを指定された木に変更して木Bを得た後他の任意化節点の一部又は全部を任意の木に変更して木Cが得られるということであるが既に木Aの任意化節点の一部又は全部を全く任意的に任意の木に変更しても木Cが得られないことがこの証明の前半より明らかであるのでこのことが成立しないことは明らかである。（証明終）

ここで図7(1)~(21)に示す21個の木及び図8に示す木を導入する。これらは

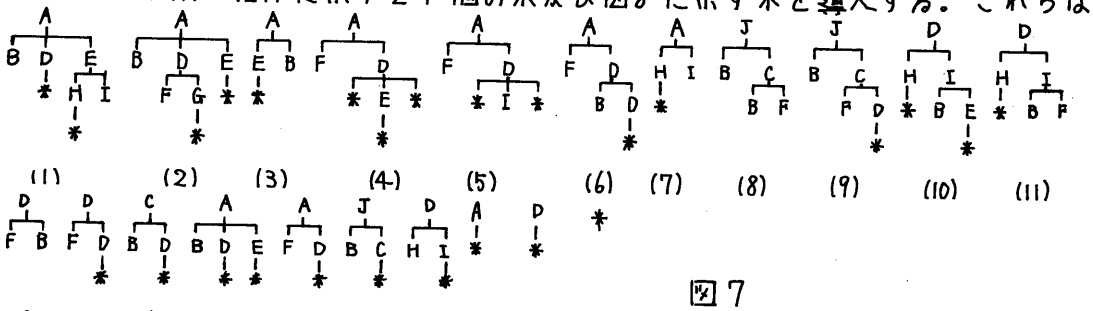


図7

主に次節に於ける説明にも用いるが一部ここでの説明に使用する。図7(16)の木は図7(4)~(6)のそれぞれの木を包含する。そして図7(16)の木は図8の木を包含しないので図7(16)が包含する上記3つの木も図8の木を包含しない。これが上記定理1が成立する一例である。次に木の交差及び交差木に関する以下の定義4を導入する。

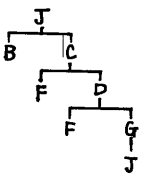


図8

〔定義4〕木Aと木Bが交差であるとは、木Aに於てそれに含まれる任意節点を必要に応じ性質1を用いて複数個の任意化節点に増加させた後、それらの任意化節点を必要に応じ空木を含む任意の木に置きかえて得られた木が木Bに於て同様な処置を行って得られた木と同一に出来る場合を言う。又このようにして得られた木を木Aと木Bの交差木と呼ぶ。又木Aと木Bが交差でないとき木Aと木Bは非交差であると言う。

定義4に於ける空木はもちろん木が無いことを意味する。定義4を例を用いて説明する。図7(1)の木に於て節点Dの下位の任意化節点を性質1により2つの任意化節点に増加させ、このようにして得た左側のものを節点下で又右側のものを節点Gの下位に任意化節点が接続されたものに置き換えると図9の木を得る。一方図7(2)の木に於てEの下位に接続されている任意化節点に同様な操作を加えるとやはり図9の木を得る。このため図7(1)の木と図7(2)の木は互いに交差であり図9に示す木がこれらの交差木となる。定義4の記述に関し以下の定理2が成立する。

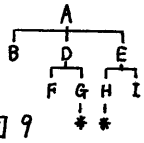


図9

〔定理2〕木Aと木Bの交差木は木Aにも木Bにも包含される。又逆に木AにもB

にも包含される木はこれらの交差木又は交差木中の任意化節点が任意の木に置き換った形式の木のみである。

（証明）この定理の前半をまず証明する。上記木A、Bの各最上位節点を含み又終端節点の一部を含む木状の部分で木A、Bが一致しており、又木A及木Bの間で上記一致部分から下位への接続位置が全く一致しておりただ上記接続位置の各々に於て接続される下位の部分が木Aと木Bで異っているのみであり、さらにこの際木A及び木Bの一方に於ける上記下位の部分が木状形状をなす場合は他方に於ける上記下位の部分が任意化節点であるという状態に於てのみ交差木が得られ他の場合は得られないことが定義4より判明する。もちろん交差木は上記各接続位置に於て木A又はBの一方に於ける任意化節点の部分を他方の対応する接続位置に接続されている木に置き換えて得られる。以上述べたことと包含関係に関する定義1から本定理の前半が証明される。以下後半の証明を行う。

上記接続位置の1つに於て木A又Bの一方に於ける任意化節点の部分を他方の木の対応する接続位置に接続されている木以外のものに置き換えると交差木以外の木を得るがこれは木A又はBのどちらかに包含されない。なおある1つの接続位置に於て任意化節点をそのまま放置しておき他の全ての接続位置に於て交差木を発生する変更を行って得た木は木A又は木Bのうち任意化節点を持たない方の木を包含するが逆に包含されない。このようなことを行う接続位置を複数に増加させた場合も同様である。以上のことから本定理は証明される。なおこのようにして得られた交差木中に含まれる任意化節点を木状の構造に変更したものも木A及びBに包含されることは自明である。（証明終）

定理2より直ちに次の定理3を得る。

〔定理3〕木Aと木Bが互いに非交差の関係にあるとき一方に包含される木は他方に包含されない。

（証明）木Aと木Bが互いに非交差である場合、交差木を持たない。しかしながら交差木及びそれに含まれる任意化節点を木状の構造に置き換えたもののみが木A及び木B双方に包含される。このことから本定理が証明される。（証明終）

4. 階層木の導入

前節で得た結果を用いる階層木の導入を行う。これは次の定義5により定義されるものである。

〔定義5〕階層木は木状の構造を持ち、その各節点に対して1つの木が対応する。そしてある階層木節点に対応する木はそれより下位の階層木節点に対応する木を包含する。なお同位な階層木節点の左右の相互位置は任意に変更できる。

木（すなわちパーズ木）と階層木を区別するため、後者に於て定義5に於て示すように木、節点等木を表わす語の前に“階層木”の語を付加する。階層木は何種類かのラベルが付加された、あらかじめ与えられた項目木に対応する項目木階層木節点と階層木を構成するために任意に導入できる分類階層木節点の2種類の節点を持つ。階層木の役目は、別に任意に与えた被検定木を包含する項目木が上記与えられた項目木中にあるかどうかを能率よくしらべることにある。説明の簡単化のためにさし当り項目木のラベルを1種類とし、それが禁止木の役割を示すラベルである場合を考える。ここでは一般的な項目木でなく禁止木を取り扱かう。

この場合の階層木の構成法は与えられた禁止木の各々に対応する禁止木階層木

節点を導入し、これらの禁止木のある集団を包含する木があればそれに対する分類階層木節点を導入し、上記集団に対応する禁止木節点の上位に置き線上記集団の構成要素と線で結び、さらにこのようなことを他の禁止木の集団についても行い、さらにこのようにして得られた分類階層木節点に対しても上記多層化を行い1つの木を形成するというものである。

以下例を用いて説明する。図10がこれである。これは先に導入した図7の木のうち図7(1)～7(14)が禁止木として与えられた場合の階層木の例である。図10中の各階層木節点はこれらに付番された番号と同一の図7中の副番号の木と対応する。図10に於て下方が黒い節点は与えられた禁止木階層木節点であり、その他の節点が階層木を構成するために導入した分類階層木節点である。このように項目木のラベルが一種類の場合は項目木、すなわちここでは禁止木、は全て階層木に於ける終端節点に対応する。なお図10に於ける節点21に対応する図7(21)の木は任意化節点のみから成る木であり最上位節点以下全てが任意の木に相当する。定義5及び定理1より以下の定理4が成立する。

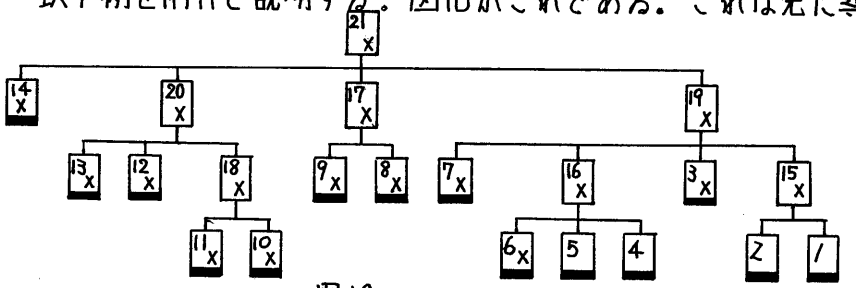


図10

〔定理4〕ある階層木節点Aに対応する木が、ある別の木Bを包含しないならば、木Bは階層木節点Aの下位にある階層木節点に対応する木には包含されない。

この証明は自明である。ここで階層木に於ける同位非交差階層木節点の定義を導入する。

〔定義6〕ある階層木節点に対応する木が、この階層木節点と同位の位置にあるどの階層木節点に対応する木とも非交差の関係にあるとき、その階層木節点を同位非交差階層木節点と称する。

又この定義の系としてささいなことではあるが以下の定義7を導入する。

〔定義7〕同位な階層木節点を持たない階層木節点は同位非交差階層木節点である。

定義6と定理3より以下の定理5が成立する。

〔定理5〕ある同位非交差階層木節点に対応する木Aが他のある木Bを包含するならば同位の他の階層木節点は木Bを包含しない。

図10の階層木に於てX印の付いた節点は同位非交差階層木節点である。

さて、図10にその例を示すような階層木が与えられた場合、それが持つ項目木階層節点に対応する項目木のなかで与えられた被検定木を包含するものが有るかどうか、有ればそれはどれかを調べるための定理4及び5を応用したアルゴリズムを項目木に関する一般的な表現で箇条に分けて以下述べる。

〔被検定木を包含する項目木を発見するための階層木走査アルゴリズム〕

(i)ある終端節点以外の階層木節点に対応する木に関する被検定木の包含検定を行った際、その木が被検定木を包含するなら、次に上記階層木節点の直下位の階層木節点のうち最右端のものに対応する木に関する包含検定を行う。

(ii)ある階層木節点に対応する木に関する被検定木の包含検定を行った際、その階

層木節点が終端節点であるか、又はその階層木節点が終端節点以外の節点でありそれに対応する木が被検定木を包含しないならば、次にその階層木節点より出発して階層木上を上位方向にさかのぼりそのさかのぼる経路上より最初に左方に分岐する経路上の最初の階層木節点に対応する木に関する被検定木の包含検定を行う。

但し上記の出発階層木節点に於ける包含検定結果が非包含の場合は上記出発階層木節点を除き、上記出発階層木節点が終端節点であり包含検定結果が包含の場合は出発階層木節点を含めて上記さかのぼる経路上にある同位非交差階層木節点の直上位の節点より左方に分岐することは出来ない。

(iii)包含検定は最上位階層木節点より始める。但し最上位階層木節点に比してより適切な階層木節点があればそれより始める。

(iv)本アルゴリズムは、(ii)項に示す左方への分岐が最上位階層木節点に至るも不可能である場合終了する。

(v)包含検定に関して被検定木を包含する項目木が発見されたらそれを記録する。

(vi)本アルゴリズム(iv)項終了後本アルゴリズム(v)項に於て累積記録された結果を出力する。記録結果が無い場合は項目木中に被検定木を包含するものが無いという出力を発生する。

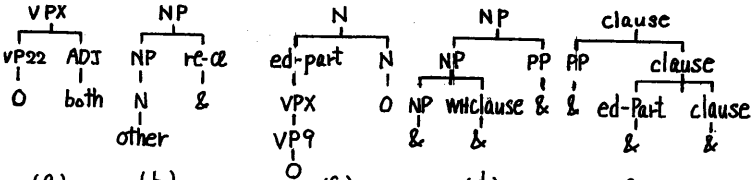
以上がアルゴリズムである。同位な階層木節点の相互位置は自由に変えられるので上記のアルゴリズムのように走査順序を定めても一般性は保たれる。(ii)項の但し以下の部分が定理5により他の部分は全て定理4より導かれる。図10の例に於て図8の被検定木につき包含検定を行う場合、上記アルゴリズムに従うと、階層木節点21、19、17、8及び9に関する5回の包含検定を行うことにより階層木節点9に対応する木である図7(9)の項目木(この場合は禁止木)のみが図8の被検定木を包含することが判明する。これに反し階層木を用いない場合は21個の項目木全部に対する21回の包含検定が必要である。このように階層木を用いると包含検定の回数を項目木の数の対数に比例した程度迄少く出来る。

ここで述べた階層木による木の整理法を用いて図1中の禁止木辞書を構成するとここで述べた包含検定の減少効果を利用することが出来る。この場合禁止木辞書と禁止木包含検定部をむすぶ経路Xに次情報要求信号が禁止木辞書へ入力すると共に経路Z上に前回の包含検定結果を同じく入力せしめると禁止木辞書は上記アルゴリズムに従い次に包含検定すべき木を経路W上に発生することが出来る。この場合包含検定は計算機上に於てS式を用いて木の重ね合わせと同様な操作を行うことにより行われる。なおこの場合より一般的な記述である上記アルゴリズム(v)項の方式が行われるのではなく被検定木を包含する禁止木が発生したということのみを禁止木包含検定部が記憶することが行われる。

ここで項目木に2種類のラベルを付ける一つの方式について述べる。それはある階層木節点のラベルを"禁止木"に相当するラベルとし、その下位にある階層木節点のラベルを"推薦木"に相当するラベルとする方法である。これにより禁止の範囲をよりきめ細かく表現できる。

5. 禁止木の例

典形的な禁止木の例を図11の各図にあげる。この図に於てNPは名詞句、Nは名詞、VPXは文法項を一部欠いた動詞句、VP9及びVP22はホーンビーの動詞分類で分類された動詞[2]、PPは前置詞句、ed-Partは過去分詞、re-d.wHclause



(a) (b) (C) (d) (e)

図11

はそれぞれ関係詞節で、前者は関係詞を欠いてもよいもの、後者は欠いてはいけないものである。clauseは一般的な節である。&及び○は今迄述べて来た任意化節点＊に相当するもので前者は非終端節点、後者は終端節点である。図11(a)は形容詞に分類されるbothが式(2)中のPaintのようなVP22形動詞の目的補語にならないことを示し、図11(b)は名詞に分類されるOtherが関係詞節の先行詞にならないことを示し、図11(C)はthat clauseを取るsuppose decide等のVP9の動詞は過去分詞の形で前方から名詞を限定しないことを示し、図11(d)は後方から関係詞節によって限定される名詞句はさらに前置詞句によって限定されないことを示し、図11(e)は分詞構文句と主節から成る節は前方から前置詞句に限定されないことを示す。

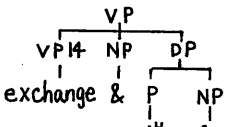


図12

禁止木による書き換え規則適用制限の一変形として推薦木の方法がある。これは特定の項目が隣接して存在するときに特定のパーズ結果のみを生成させるものであり単語間親和性を記述するための方法である。例えば図12はexchange NP with NPの項目がこの順に発生する場合、図12に示す部分木を含む木のみを発生させる方法であり、図12の木が何れかの位置で分割されたものが自動的に禁止木として作用する方法である。推薦木はそれが含む単語が読み込まれた時に起動するので一般の禁止木とは別に各単語毎に整理、収容される。

6. 実験結果、まとめ

ＣＣＩＴＴ文書より100文を選びパーズ実験を行った。これらの平均単語数は26であり、これらのうち複文は57、重文3、単文中修飾句を含むもの13であった。全文のパーズに成功し、この際の1文当り平均発生木数は2.0であり、禁止木(含推薦木)総数は958にのぼった。使用した書き換え規則は著者の1人である橋本が拡張ＬＩＮＧＯＬのために作成したものをそのまま使用した。(3) この規則は498個の規則より成る。

以上の実験の結果、ここで述べたパージング手法により実用的なパージングを行うことが出来ることが判明した。禁止規則導入の手数はほぼ拡張ＬＩＮＧＯＬに於けるそれと同程度のように見える。同一文を解析する計算時間は拡張ＬＩＮＧＯＬの60%増程度である。これは本パージング手法が出来上った木を抹消することを行っているためであると思われる。本手法の最大の特徴は禁止規則を階層木により整理できること及び禁止規則が正しくない部分木と1対1対応をなしているため規則の見通しが良くパーズ不能の防止、原因発見が容易なことである。なお階層木による木の整理は先に筆者らが発表したＫＰＰ法にも適用できる。(4)

〈謝辞〉筆者らは、機械翻訳研究開始時より御指導頂く鍛治ＫＤＤ研究所長、寺村同副所長、中井同次長、川井第1特別研究室長に感謝する。又実際の作業を担当された㈱ソフトウエアコンサルタント、山城、田中、築地、井上各氏、㈱日本ＩＲ、山本、若菜氏に感謝する。

○文献1. 田中、計算機による自然言語意味処理に関する研究、電子総研報告797号、昭和54年7月、文献2. ASホーンビー(著)、伊藤(訳)、英語の型と語法、オックスフォード大、文献3 橋本他、英語構文解析のための文法 作成、昭和58年度信学会全大1331、文献4 神、橋本、パターンマッチングによる英日機械翻訳の一手法、自然言語処理研 32-2(1982.8.2)