条件の付加による到達可能性の改良について

笠 晃一[†] 岡出高徳[†]
弘中大介[†] 横田将生[†]

統語解析においては，不要な部分解析木の発生を抑制するために，到達可能性が使用されることが多い．これは，ある文法範疇から別の文法範疇への成長を予測するものである．しかしながら，到達可能性の効果は万能とはいえず，場合によっては到達可能性がほとんど機能しないことも少なくない．そこで，我々は到達可能性の予測精度を上げるために，先読み情報を用いることにした．すなわち，到達可能性によって文法範疇 A から文法範疇 B が予測されても，範疇 A の後方に特定の語彙範疇が出現しない限り，この予測を無効とするのである．ここで予測を制限するために使用されている語彙範疇は，我々が範疇核と呼んでいるものを通じて求めることができる．範疇核は，言語理論 HPSG の主要部に類似した概念である．ある文法範疇が与えられたとき，これを文脈自由文法の書換え規則によって語彙範疇のみを含む列にする．このとき，規則の適用方法とは無関係に，その列に必ず出現するような語彙範疇があれば，これを範疇核と呼ぶのである．範疇核は，書換え規則から連立集合方程式を作成し，これを解くことによって求められる．この連立集合方程式の解法としては，組合せ的なものも利用できるが時間がかかるので，ここでは逐次近似法による解法を紹介している．

An Improvement of Syntactic Reachability by Adding Conditions

Koichi Ryu,[†] Takanori Okade,[†] Daisuke Hironaka[†]
and Masao Yokota[†]

In syntax analysis, reachability is often used for restraining the generation of useless partial parse trees. It predicts the growth of one grammatical category to another. Nevertheless reachability is not almighty but it sometimes hardly functions. We have been studying a method that uses lookahead information to improve the prediction accuracy of reachability. In our method, even if reachability predicts category B from category A, the prediction is forced to be invalid in case special lexical categories do not appear after category A. The special lexical categories used here are obtained from symbols that we call "categorical kernels". Categorical kernels are found by making simultaneous set equations from rewriting rules of CFG and solving them. In case of solving this equations, we use successive iteration because combinatorial method is much more time-consuming.

1. まえがき

文脈自由文法を用いて自然言語文の統語解析を行おうとする場合，ヒューリスティックな情報なしでは，最終結果に寄与しない不要な部分解析木が大量に発生することが知られている．そこで，このような不要な部分解析木の発生を抑制するための方法がいくつか提案されており，たとえば，代表的なものとして，到達可能性の利用をあげることができる．実際，この到達可能性は，さまざまなシステムに組み込まれ，利用されてきた[1)~3)]．しかしながら，我々の行ったいくつかの実験によると，到達可能性を用いた場合，不要な部分解析木の抑制効果は必ずしも高くないようである．このことは，本論文の最後の方で述べている実験結果にも表れている．

到達可能性を使用しても部分解析木の利用率がそれほど上昇しない原因は 2 つ考えられる．1 つは，予測の精度の問題である．到達可能性とは，端的にいえば，ある文法範疇が完成しているとき，この範疇は将来，別の文法範疇に成長する可能性があるということである．つまり，可能性の問題であって，別の文法範疇への成長を完全に予測しているわけではない．我々の調査によれば，この予測が失敗する場合もかなりあるということが分かっている．もう 1 つの原因は，到達可能性という概念そのものの限界である．つまり，上記のような予測が，たとえ完全なものだとしても，このような概念だけで，不要な部分解析木がすべて除去さ

† 福岡工業大学情報工学部
Faculty of Information Engineering, Fukuoka Institute of Technology

れるということは理論的にありえない．

以上の考察より，部分解析木の利用率を上げるには，到達可能性そのものを改良して予測の精度を上げること，および，到達可能性以外の概念を併用することが考えられる．我々は，このうち前者について検討を行い，到達可能性による予測の精度を上げるために，先読み情報を用いることにした．到達可能性によって，ある文法範疇 X から別の文法範疇 Y が予測されても，文法範疇 X の後方にキーとなる語彙範疇が出現しないかぎり，その予測を無効とするのである．このような先読み情報による制約の付加した到達可能性を，ここでは，条件付到達可能性と呼ぶことにする．

以下，2 章では，条件付到達可能性のもとになった到達可能性の厳密な定義を行う．また，到達可能性の視覚的表現である到達可能性グラフについても説明する．3 章では，準備として，範疇核の定義や核集合方程式の意味について述べ，さらに，核集合方程式の解法についても触れる．4 章では，本論文の主題である条件付到達可能性の意味と記法について述べる．さらに，条件付到達可能性の求め方についても順を追って説明する．5 章では，条件付到達可能性の有効性を検証するために行った実験について述べる．最後に，6 章において，残された問題に対する検討を行う．

2. 到達可能性

2.1 到達可能性の定義

まず，文脈自由文法 (V, T, P, S) に対する若干の制限を行う．ただし，V は非終端記号の有限集合，T は終端記号の有限集合であり，これらは互いに素であると仮定する．本論文では，非終端記号のことを文法範疇と呼ぶ場合もある．また，P は書換え規則の有限集合，S は開始記号である．この文脈自由文法の書換え規則は，次のいずれかの形をとるものとする．

（R1） 語彙規則：$A \to a \quad (A \in V,\ a \in T)$

（R2） 文法規則：$A \to \alpha \quad (A \in V,\ \alpha \in V^+)$

なお，書換え規則に対するこのような制限は，文脈自由文法の能力になんら影響を与えないことが保証されている[4]．また，語彙規則の左辺に現れる非終端記号は特に語彙範疇と呼ばれることがある．

次に，上記の文脈自由文法に対して，到達可能性の定義を行う．

定義 2.1 A，B，C を非終端記号とするとき，到達可能性は以下のように再帰的に定義される．

（D1） A から A へは到達可能である．

（D2） P が，$B \to A\alpha$ $(\alpha \in V^*)$ のような書換え規則を含んでいるとき，A から B に到達可能である．

ap → a	np → s np	d → "実に"
ap → d ap	s → vp	m → "細やかな"
mp → m	s → ppa s	p → "は"
mp → d mp	s → ap s	p → "を"
ppa → np p	vp → v b	v → "持っ"
ppb → np no	vp → vp b	b → "ている"
np → n		a → "しっかり"
np → mp np	n → "日本人"	no → "の"
np → ppb np	n → "神経"	

図 1 日本語書換え規則の例
Fig. 1 An example of Japanese rewriting rules.

（D3） A から B に到達可能であり，B から C に到達可能であれば，A から C に到達可能である．

なお，以後，A から B に到達可能であることを，$A \Rightarrow B$ と書くことにする．

2.2 到達可能性グラフ

定義 2.1 の**（D2）**から，任意の書換え規則に対して，それぞれ，1 つの到達可能性が得られることが分かる．したがって，以下のような手続きを，すべての書換え規則に対して実行することにより，有向グラフを作成することができる．

（G1） 書換え規則 $B \to A\alpha$ から，到達可能性 $A \Rightarrow B$ を求める．

（G2） 非終端記号 A と B に対応する点を作成する．ただし，対応する点が，すでにグラフ上に存在しているときには，新たな点を作成するようなことをせず，すでにある点を利用する．

（G3） A に対応する点から B に対応する点に向けて有向線を引く．ただし，このような線がすでに存在する場合には何もしない．

このとき得られるグラフについて考えてみると，定義 2.1 により，グラフ理論の意味での到達可能性[5]が，現在議論している統語的な到達可能性に一致することが分かる．つまり，非終端記号 A に対応する点から B に対応する点へ，グラフ理論的に到達可能ならば，$A \Rightarrow B$ がいえるし，この逆も成り立つ．そこで，このようにして作成した有向グラフを，到達可能性グラフと呼ぶことにしよう．書換え規則の例を**図 1**☆に，この規則に対する到達可能性グラフを**図 2** に示す．

3. 範疇核と核集合方程式

3.1 範疇核の定義

言語理論において，HPSG[6]という文法理論が知ら

☆ 無意味な文を生成する規則を一部含んでいるが，4 章の説明の都合上このようにしている．

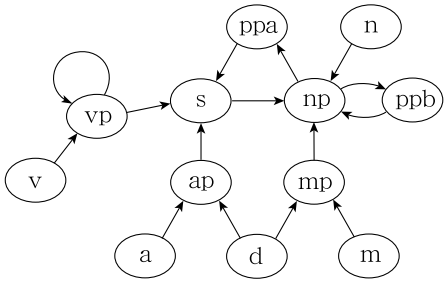


図 2 図 1 の書換え規則に対する到達可能性グラフ
Fig. 2 A reachability graph for rewriting rules in Fig. 1.

Ker(ap) = { a } ∩ ({ d } ∪ Ker(ap))
Ker(mp) = { m } ∩ ({ d } ∪ Ker(mp))
Ker(ppa) = Ker(np) ∪ { p }
Ker(ppb) = Ker(np) ∪ { no }
Ker(np) = { n } ∩ (Ker(mp) ∪ Ker(np))
∩ (Ker(ppb) ∪ Ker(np))
∩ (Ker(s) ∪ Ker(np))
Ker(s) = Ker(vp) ∩ (Ker(ppa) ∪ Ker(s))
∩ (Ker(ap) ∪ Ker(s))
Ker(vp) = ({ v } ∪ { b }) ∩ (Ker(vp) ∪ { b })

図 3 図 1 の書換え規則に対する核集合方程式
Fig. 3 Kernel set equations for rewriting rules in Fig. 1.

れているが，一般の文脈自由文法においても，HPSG における主要部あるいは主辞に類似したものを求めることができる．すなわち，多くの文法範疇は，これを書き換えて語彙範疇だけにしたとき，特定の語彙範疇を必ず含んでおり，これは計算によって求めることが可能である．ここで，任意の文法範疇に対する，上記のような語彙範疇を範疇核と呼ぶことにする．範疇核の正確な定義は次のとおりである．

定義 3.1 文脈自由文法 (V,T,P,S) において，書換え規則の集合 P は，前節で述べたように，語彙規則と文法規則に分類されるものとする．文法範疇 $X \in V$ に対して，文法規則を 0 回以上適用して，語彙範疇だけを含む列にしたとき，規則の適用の仕方とは無関係に，この列が語彙範疇 A を必ず含むならば，そしてこのときに限り，A は X の範疇核であるという．

3.2 核集合方程式の作成

この節では，範疇核を求める方法について述べる．今，文法範疇 X の範疇核をすべて集めたものを X の範疇核集合と呼び，$Ker(X)$ で表すことにしよう．さらに，文脈自由文法 (V,T,P,S) における文法範疇の集合 V を，語彙範疇の集合 L とそれ以外の文法範疇の集合 C に分割する．ここに，集合 L と集合 C は，以下のような要素から成るものとする．

$$L = \{A_1, A_2, \ldots, A_D\},$$
$$C = \{X_1, X_2, \ldots, X_M\},$$
$$L \cup C = V,\ L \cap C = \phi \quad (1)$$

すると，$A_d \in L$ に対しては，定義 3.1 により，明らかに次の関係式が成立する．

$$Ker(A_d) = \{A_d\} \quad (d = 1, \ldots, D) \quad (2)$$

一方，$X_i \in C$ に対しては，X_i を左辺に持つ文法規則をすべて集めてくる．ここでは，以下のような N_i 個の規則が集まったと仮定しよう．

$$X_i \rightarrow Y_{i,j,1}\ Y_{i,j,2}\ \ldots\ Y_{i,j,n_{ij}} \quad (j = 1, \ldots, N_i) \quad (3)$$

そこで，まず，j 番目の規則のみが存在する場合を考えると，$Y_{i,j,k}$ $(k = 1, \ldots, n_{ij})$ から導かれる語彙範疇列に必ず含まれる語彙範疇は，X_i から導かれる語彙範疇列にも必ず含まれるから，次の式が成立することが分かる．

$$Ker(X_i) = \bigcup_{k=1}^{n_{ij}} Ker(Y_{i,j,k}) \quad (4)$$

しかし，実際には X_i で始まる規則は N_i 個あり，どの規則が使用されるかは不明である．したがって，X_i から導かれる語彙範疇列に必ず含まれる語彙範疇は，式 (4) で求められる集合の共通部分ということになり，次の式が得られる．

$$Ker(X_i) = \bigcap_{j=1}^{N_i} \bigcup_{k=1}^{n_{ij}} Ker(Y_{i,j,k}) \quad (i = 1, \ldots, M) \quad (5)$$

最後に，この式 (5) の右辺に出現する文法範疇 $Y_{i,j,k}$ が語彙範疇の場合には，式 (2) を用いて $Ker(Y_{i,j,k})$ を $\{Y_{i,j,k}\}$ で置き換える．このようにして，範疇核集合に関する M 元の連立集合方程式が得られるが，これを，範疇核集合方程式，あるいは，単に核集合方程式と呼ぶことにする．この方程式を解けば，式 (2) とあわせて，すべての文法範疇に対して範疇核集合が求まることになる．**図 3** に，図 1 の書換え規則に対する核集合方程式を示す．

3.3 核集合方程式の解法

上で求めた核集合方程式は，組合せ的な方法で解くこともできるが，逐次近似法を用いて，比較的短時間に解くことも可能である．以下，この方法を説明しよう．まず，核集合方程式は式 (5) に式 (2) を代入したものなので，$Ker(A_d)(A_d \in L)$ を含まず，形式的に次のように記述できる．

$$Ker(X_i) = f_i(Ker(X_1), \ldots, Ker(X_M)) \quad (i = 1, \ldots, M) \quad (6)$$

Ker(n) = { n }	Ker(ap) = { a }
Ker(d) = { d }	Ker(mp) = { m }
Ker(m) = { m }	Ker(ppa) = { n, p }
Ker(a) = { a }	Ker(ppb) = { n, no }
Ker(p) = { p }	Ker(np) = { n }
Ker(no) = { no }	Ker(s) = { v, b }
Ker(v) = { v }	Ker(vp) = { v, b }
Ker(b) = { b }	

図 4 図 1 の書換え規則に対する範疇核集合
Fig. 4 Categorical kernel sets for rewriting rules in Fig. 1.

さらに，$Ker(X_i)$ $(i = 1, \ldots, M)$ を下のようにまとめて書くことにする．

$$\mathbf{X} = (Ker(X_1), \ldots, Ker(X_M)) \tag{7}$$

すると，式 (6) はより簡単な次の形に記述される．

$$\mathbf{X} = \mathbf{f}(\mathbf{X}) \tag{8}$$

そこで，次の式に基づいて，$\mathbf{X}^h (h = 1, 2, \ldots)$ を次々に求めることを考えよう．ただし，L はすべての語彙範疇の集合である．

$$\mathbf{X}^0 = (L, L, \ldots, L) \tag{9}$$

$$\mathbf{X}^{h+1} = \mathbf{f}(\mathbf{X}^h) \quad (h = 0, 1, 2, \ldots) \tag{10}$$

実は，このようにして求められる $\mathbf{X}^h$ は，有限回の反復で収束し，しかも，その収束値が方程式 (8) の最大解になっていることを証明できる．その証明については付録で述べる．なお，図 3 の方程式を逐次近似法で解くことにより，図 1 の書換え規則に対する範疇核集合をすべて求めることができるが，その結果を**図 4** に示す．また，反復回数，すなわち $\mathbf{f}$ の計算回数については，集合 L の大きさを D としたとき，MD 以下になることが分かっている．これに対し，組合せ的な解法では，$\mathbf{f}$ を 2^{MD} 回だけ計算しなければならない．

4. 条件付到達可能性

4.1 直接的な条件付到達可能性

まずは，直接的な到達可能性というものを定義しておこう．

定義 4.1 A と B を文脈自由文法 (V, T, P, S) の非終端記号であるとする．このとき，A を右辺の先頭に持ち，B を左辺に持つような書換え規則が P の中にあるならば，そしてこのときに限り，A から B に直接的に到達可能であるという．

そこで，文法範疇 X から文法範疇 Y に直接的に到達可能であると仮定する．定義 4.1 により，X を右辺の先頭に，Y を左辺に持つ文法規則が存在するので，このような規則の 1 つをとってくる．

$$Y \rightarrow X\ X_1\ X_2\ \ldots\ X_m \tag{11}$$

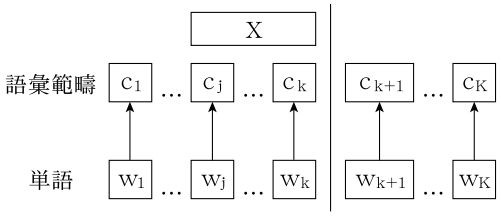


図 5 文法範疇 X を構成する語彙範疇
Fig. 5 Lexical categories which compose grammatical category X.

たとえば，上記のような規則をとってきたとしよう．右辺は X だけのこともあるが，ここでは，X の後にも 1 個以上の非終端記号が続くものと仮定する．ここで，上昇型のアルゴリズムを用いて下記のような単語列の統語解析を行うことを考える．

$$w_1\ w_2\ \ldots\ w_K \quad (w_i \in T) \tag{12}$$

この単語列に対して語彙規則を適用し，次のような語彙範疇列が得られるものとする．

$$c_1\ c_2\ \ldots\ c_K \quad (c_i \in L) \tag{13}$$

さらに，この語彙範疇列の一部 c_j〜c_k を用いて，式 (11) の規則に出現する文法範疇 X が完成されていたとしよう．この様子を**図 5** に示す．一方，式 (11) の規則は，文法範疇 X から文法範疇 Y に到達可能であること，すなわち，文法範疇 X が完成していれば，将来的に文法範疇 Y が完成する可能性のあることを表していた．しかしながら，文法範疇 X を構成する語彙範疇の後に続く語彙範疇 c_{k+1}〜c_K の中に，たとえば，文法範疇 X_1 の範疇核 A が出現しなかったとしたらどうだろう．A が文法範疇 X_1 の範疇核であるということは，X_1 を完成させるのに語彙範疇 A が必要だということである．したがって，X を構成する語彙範疇の後に A がなければ，X_1 を完成させるのは不可能であり，式 (11) の規則を用いて文法範疇 Y を完成させることも不可能になる．以下，この考え方を精密化しよう．

まず，X を構成する語彙範疇の後に続く語彙範疇を用いて，次のようなリストを作成する．

$$R = [c_{k+1}, c_{k+2}, \ldots, c_K] \tag{14}$$

さらに，簡単のために，式 (11) の規則において $m = 1$ であると仮定する．このときは，文法範疇 X_1 の範疇核集合 $Ker(X_1)$ のみを考えればよい．そして，上記の考察により，リスト R に $Ker(X_1)$ の要素が 1 つでも出現しなければ，式 (11) の規則を用いて文法範疇 Y を完成させることは不可能である．言い換えれば，式 (11) の規則によって文法範疇 Y が完成するための必要条件は，リスト R に $Ker(X_1)$ のすべての要素が出現することである，ということ

になる．ここでは，この必要条件をリストを用いて表現しよう．それには，集合 $Ker(X_1)$ をリスト表現に変換する必要がある．集合の要素は順序を持たないので，全要素のあらゆる順列を作成しなければならない．つまり，$Ker(X_1)$ の大きさを N_1 とするとき，全部で $N_1!$ 個のリストができるので，これらを，$Q_{1,k}$ $(k=1,\ldots,N_1!)$ によって表そう．すると，上記の必要条件は，$Q_{1,k}$ の少なくとも 1 つがリスト R の部分リストであると言い換えることができる．ただし，$A=[a_1,a_2,\ldots,a_s]$，$B=[b_1,b_2,\ldots,b_t]$ のとき，B が A の部分リストであるとは，以下の条件を満たす $j_1,j_2,\ldots,j_t$ が存在することである．

$$1 \le j_1 < j_2 < \ldots < j_t \le s \ \text{ かつ } \ b_k = a_{j_k} \quad (k=1,2,\ldots,t)$$

次に，$m \ge 2$ の場合を考える．このとき，式 (11) の規則によって文法範疇 Y が完成するための必要条件は，リスト R の最初の方に $Ker(X_1)$ のすべての要素が出現し，その後に $Ker(X_2)$ のすべての要素が出現し，ということを繰り返して，最後に $Ker(X_m)$ のすべての要素が出現することである．そこで，$Ker(X_i)$ の大きさを N_i で表し，また，この集合に対する前述のリストを $Q_{i,k}$ $(k=1,\ldots,N_i!)$ で表すことにする．そして，これらのリストを結合して，次のようなリストを作成する．

$$S_{k_1,k_2,\ldots,k_m} = Q_{1,k_1} \cdot Q_{2,k_2} \cdot \ldots \cdot Q_{m,k_m}$$
$$1 \le k_1 \le N_1!, \ldots, 1 \le k_m \le N_m! \qquad (15)$$

ここに，ドット記号「$\cdot$」はリストの連結を表している．これを用いると，上記の $m \ge 2$ の場合の必要条件は，$S_{k_1,k_2,\ldots,k_m}$ の少なくとも 1 つがリスト R の部分リストであるということになる．ちなみに，式 (11) の規則の右辺に X のみが出現する場合には，$S_{k_1,k_2,\ldots,k_m}$ として 1 個の空リストを考えればよい．ところで，X を右辺の先頭に，Y を左辺に持つ文法規則は，一般に複数個ある．これらを，次のようにすべて集めてこよう．

$$Y \to X\ X_{i,1}\ X_{i,2}\ \ldots\ X_{i,m_i} \quad (i=1,\ldots,N) \qquad (16)$$

このとき，i 番目の規則に対しても式 (15) のようなリストが作成できるので，これを $S^i_{k_1,k_2,\ldots,k_{m_i}}$ で表す．すると，式 (16) の i 番目の規則によって文法範疇 Y が完成するための必要条件は，i を固定したときの $S^i_{k_1,k_2,\ldots,k_{m_i}}$ の少なくとも 1 つがリスト R の部分リストであるということになる．ここで，$S^i_{k_1,k_2,\ldots,k_{m_i}}$ に通し番号をつけたものを，K_1〜K_G で表すことにしよう．ただし，i と k_1〜k_{m_i} はすべて動かすものとする．文法範疇 Y は式 (16) のどの規則を使用して完成させてもよいから，結局，文法範疇 X が完成していたときに文法範疇 Y が完成するための必要条件は，K_1〜K_G の少なくとも 1 つがリスト R の部分リストであると結論できる．このことを，ここでは次のように表記する．

$$X \Rightarrow Y \quad K_1|K_2|\ldots|K_G \qquad (17)$$

そして，このとき，X から Y に条件付で到達可能であるということにする．なお，式 (17) で用いた縦棒の記号は選言を表しているが，この記号を用いて記述された部分，すなわち，式 (17) の後半を特に条件部と呼ぶことにしよう．そして，到達可能性グラフの弧にこの条件部を付加したものを，条件付到達可能性グラフと呼ぶことにする．図 1 の規則に対する条件付到達可能性グラフを**図 6** に示す．

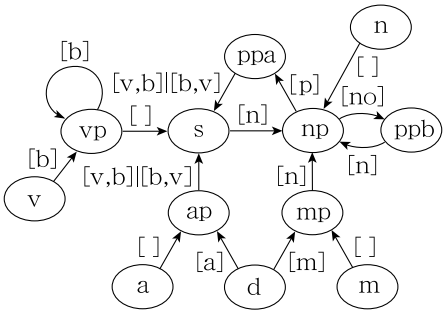


図 6 図 1 の書換え規則に対する条件付到達可能性グラフ
Fig. 6 A conditional reachability graph for rewriting rules in Fig. 1.

4.2 一般の条件付到達可能性

条件付到達可能性グラフが求まれば，一般の場合の条件付到達可能性を求めるのは容易である．以下，このやり方を説明しよう．まず，文法範疇 X から文法範疇 Y に到達可能であると仮定する．すると，条件付到達可能性グラフにおいて，X から Y に至る通路が存在するはずであるから，これらのうちの 1 つを W で表そう．ここに，通路（path）というのはサイクルを含まない遊歩道（walk）のことであるが，このような制限を課してよい理由については次節で説明する．通路 W は，次のような文法範疇によって構成されるものとする．

$$X \to Z_1 \to Z_2 \to \ldots \to Z_n \to Y \qquad (18)$$

この通路は，次のような条件付到達可能性の集まりと見なすことができる．

$$\begin{aligned} X &\Rightarrow Z_1 \quad K_{01}|K_{02}|\ldots|K_{0G_0} \\ Z_1 &\Rightarrow Z_2 \quad K_{11}|K_{12}|\ldots|K_{1G_1} \\ &\ldots \\ Z_n &\Rightarrow Y \quad K_{n1}|K_{n2}|\ldots|K_{nG_n} \end{aligned} \qquad (19)$$

ここで，上昇型の統語解析アルゴリズムを用いた解析

によって，文法範疇 X がすでに完成しているものとする．そして，X を構成する語彙範疇の後に続く語彙範疇のリストを再び式 (14) のように R で表そう．式 (19) によると，X が完成しているときに Z_1 が完成するための必要条件は，K_{01}〜K_{0G_0} の少なくとも 1 つが，リスト R の部分リストとなることである．他の条件付到達可能性も同様であるから，X が完成しているときに Y が完成するための必要条件は，リスト R の中に，最初 K_{01}〜K_{0G_0} のどれかの全要素が出現し，次に，K_{11}〜K_{1G_1} のどれかの全要素が出現し，ということを繰り返して，最後に K_{n1}〜K_{nG_n} のどれかの全要素が出現することである．そこで，式 (19) の条件付到達可能性の条件部のリストを結合して次のようなリストを作成する．

$$H_{k_0,k_1,\ldots,k_n} = K_{0,k_0} \cdot K_{1,k_1} \cdot \ldots \cdot K_{n,k_n}$$
$$1 \le k_0 \le G_0, \ldots, 1 \le k_n \le G_n \qquad (20)$$

これを用いれば，文法範疇 X が完成しているときに文法範疇 Y が完成するための必要条件は，$H_{k_0,k_1,\ldots,k_n}$ の少なくとも 1 つが，リスト R の部分リストとなることである，ということがいえる．

ところで，X から Y に至る通路は一般に複数個あるから，これらを W_i $(i = 1, \ldots, N)$ によって表すことにする．そして，それぞれの W_i に対する式 (20) のようなリストを $H^i_{k_0,k_1,\ldots,k_{n_i}}$ と書く．X から Y を作成する場合，どの通路を通ってもよいから，文法範疇 X が完成しているときに文法範疇 Y が完成するための必要条件は，結局，$H^i_{k_0,k_1,\ldots,k_{n_i}}$ の少なくとも 1 つが，リスト R の部分リストとなることである，ということになる．ここでも，i と k_0〜k_{n_i} はすべて動かすものとする．このことは，次のように表すことができる．

$$X \Rightarrow Y \quad | \underset{i,k_0,k_1,\ldots,k_{n_i}}{} H^i_{k_0,k_1,\ldots,k_{n_i}} \qquad (21)$$

たとえば，図 6 の条件付到達可能性グラフにおいて，文法範疇 d から文法範疇 s に至る通路には，次の 2 つのものがある．

$$\mathrm{d} \to \mathrm{ap} \to \mathrm{s}$$
$$\mathrm{d} \to \mathrm{mp} \to \mathrm{np} \to \mathrm{ppa} \to \mathrm{s} \qquad (22)$$

したがって，弧に付加された条件を用いて，d から s への条件付到達可能性が次のように求められる．

$$\mathrm{d} \Rightarrow \mathrm{s} \quad [\mathrm{a}, \mathrm{v}, \mathrm{b}]|[\mathrm{a}, \mathrm{b}, \mathrm{v}]|$$
$$[\mathrm{m}, \mathrm{n}, \mathrm{p}, \mathrm{v}, \mathrm{b}]|[\mathrm{m}, \mathrm{n}, \mathrm{p}, \mathrm{b}, \mathrm{v}] \qquad (23)$$

4.3 条件部の簡略化

条件付到達可能性の条件部は，簡略化可能な場合がある．たとえば，次のような条件付到達可能性を考えてみよう．

$$X \Rightarrow Y \quad K_1|K_2|\ldots|K_G \qquad (24)$$

この条件付到達可能性の条件部において，K_i が K_j の部分リストであると仮定する．ただし，$i \neq j$ である．そうすると，式 (14) のようなリスト R に対して，K_j が R の部分リストならば，明らかに，K_i は R の部分リストである．ところが，論理学によれば，「A ならば B である」という命題が真のとき，「A または B である」という命題と「B である」という命題は同値になる．したがって，条件付到達可能性の定義に従えば，K_i が K_j の部分リストであるとき，条件部から K_j を削除してよいことが分かる．特に，条件付到達可能性の条件部が空リストを含むとき，条件部は空リストだけになる．

なお，条件付到達可能性グラフにおいて，文法範疇 X から文法範疇 Y に至る遊歩道を考える場合，これを通路に限定してもよいことを前節で述べたが，これも，上述の事実がその理由になる．すなわち，X から Y に至る遊歩道でサイクルを含むものがあったとして，これを W_S で表そう．このような遊歩道は，以下のようにしてサイクルの部分を取り除くことで，通路にすることが可能である．

$$\ldots \to A \to B_1 \ldots B_n \to A \to C \to \ldots$$
$$\Downarrow$$
$$\ldots \to A \to C \to \ldots \qquad (25)$$

このようにして作成された通路を W_A としよう．そうすると，前節におけるリスト $H^i_{k_0,k_1,\ldots,k_{n_i}}$ の作成法から明らかなように，W_A から作成されるリストは，必ず W_S から作成されるリストの部分リストになる．すなわち，上述の議論によれば，W_S から作成されるリストは不要であるということになり，結局，サイクルを含む遊歩道は除外して考えてよいことが分かる．

5. 比較実験

条件付到達可能性の有効性を検証するために，有用な部分解析木の比率，および，解析時間の両方について，従来の到達可能性との比較実験を行った．書換え規則としては，日本語によるデータベース問合せシステム honed-Base[7] で使用した規則のサブセット約 100 個を用いた．実験に使用した例文の数も 100 個である．なお，実験に用いたコンピュータは Apple 社の Power Macintosh G3 DT233（クロック周波数 233 MHz）であり，64 MB の主記憶を搭載している．

5.1 部分解析木の利用率に関する実験

実験に先立ち，文脈自由文法によって記述された書換え規則を，条件付到達可能性の組み込まれたパー

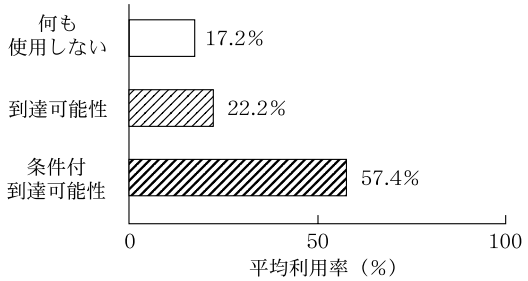
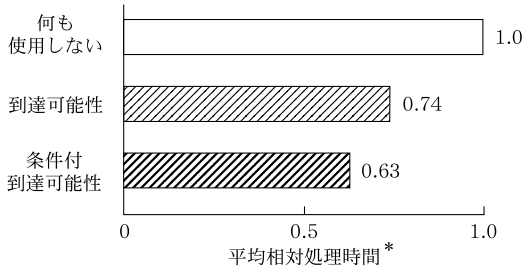


図 7 各方法に対する平均利用率の比較
Fig. 7 A comparison of average utilization ratio for each method.



*各例文において，何も使用しないときの処理時間を1とした．
図 8 各方法に対する処理時間の比較
Fig. 8 A comparison of processing time for each method.

ザへと変換するためのトランスレータ[8]を作成した．このパーザには，オブジェクト指向的な考え方を取り入れているが，基本的には上昇型のチャート法に基づいており，到達可能性を利用することもできるようになっている．我々は，実験を，(1) 到達可能性を使用した場合，(2) 条件付到達可能性を使用した場合，(3) 何も使用しなかった場合の 3 つに分けて実施したが，部分解析木の平均的な利用率は**図 7** のようになった．ただし，利用率というのは，次の式によって求められるものであり，有用な部分解析木というのは，最終的な解析結果に寄与する部分解析木のことである．

$$\text{利用率} = \frac{\text{有用な部分解析木の個数}}{\text{作成されたすべての部分解析木の個数}}$$

図 7 によれば，到達可能性を使用しても，何も使用しなかった場合に比べ，利用率は 5 パーセント程度上昇しているのみである．これに対し，条件付到達可能性を使用すると，何も使用しなかった場合に比べ，利用率が約 40 パーセントも上昇している．もちろん，これは一例にすぎないので即断はできないが，条件付到達可能性が到達可能性よりも大きな効果を有するという傾向だけは読み取れよう．

なお，当然のことではあるが，実験に用いたすべての文において，条件付到達可能性を使用した場合の方が，到達可能性を使用した場合よりも部分解析木の利用率が高かった．到達可能性を使用した場合の利用率の最低値と最高値は，それぞれ，3.9 パーセントと 46.4 パーセントであり，条件付到達可能性を使用した場合の利用率の最低値と最高値は，それぞれ，8.8 パーセントと 100 パーセントであった．ただし，条件付到達可能性を使用した場合の利用率が 10 パーセントを切ったのは，8.8 パーセントに対する 1 文のみであり，他の 99 個の文において，利用率はすべて 30 パーセントを上回っていた．実際の数を示すと，作成された部分解析木の平均個数は，有効な部分解析木の平均個数 104.0 に対して，何も使用しなかった場合が 952.4，到達可能性を使用した場合が 704.9，そして，条件付到達可能性を使用した場合が 259.4 であった．

5.2 解析時間に関する実験

実験は解析に要する時間についても行った．何も使用しなかった場合の最小の解析時間は 0.22 msec であったが，このときの到達可能性を使用した場合の解析時間は 0.28 msec，条件付到達可能性を使用した場合の解析時間は 0.11 msec であった．一方，何も使用しなかった場合の最大の解析時間は 182.5 msec であり，このときの到達可能性を使用した場合の解析時間は 98.6 msec，条件付到達可能性を使用した場合の解析時間は 150.8 msec であった．続いて，各例文につき何も使用しないときの解析時間を 1 として，到達可能性を使用したときの時間と条件付到達可能性を使用したときの時間を，これに対する比率として計算した．さらに，全例文に対して各比率の平均をとった．その結果を**図 8** に示すが，解析時間においても，条件付到達可能性の方が効率的であるということが分かった．ただし，個々に見れば，上に示した最大解析時間のデータにも現れているように，到達可能性の場合よりも条件付到達可能性の場合の方が時間がかかるケースもあり，特に文の単語数が増えるほど，この傾向が顕著であった．

6. 検　　討

この研究で明らかになった事柄は次の 2 点である．

(A1) 条件付到達可能性は，従来の到達可能性に比べて不要な部分解析木の発生を抑制する力が強い．

(A2) 範疇核は，書換え規則から得られる連立集合方程式を，逐次近似法を用いて解くことにより求められる．さらに，範疇核から条件付到達可能性が求まる．

また，以下のような 2 つの問題点が未解決で残されている．

(P1) 不要な部分解析木の発生率が低ければ，部分

解析木の記録に必要な記憶量も少なくて済むが，条件付到達可能性の場合，新たに条件部の記憶量が問題になる．

（P2） 同様に，条件付到達可能性の場合，部分解析木の処理に必要な時間も少なくて済むようにみえるが，新たに条件部の検査が必要になる．

これらの問題点のうち，P2 に関しては，条件付到達可能性を使用した場合の方が到達可能性を使用した場合よりも平均的な解析時間が短くなる条件について考察を行った．その結果，さらに詳しい検討が必要ではあるが，近似的に次のような不等式を得ている．ただし，この式の導出については煩雑になるので別の機会に発表したいと考えている．

$$\frac{\beta}{\alpha} > \frac{qNC}{4(p-q)} \tag{26}$$

ここに，この式の各変数の意味は次のとおりである．

α： 式 (14) で定義されるリスト R の 1 つの要素を検査するのに必要な時間

β： 1 つの部分解析木を処理するのに必要な時間からリスト R の検査に必要な時間を除いたもの

p： 条件付到達可能性を使用した場合の部分解析木の平均的な利用率

q： 到達可能性を使用した場合の部分解析木の平均的な利用率

N： 解析すべき文の単語数

C： 条件付到達可能性の条件部における選言数の平均値

なお，統語解析と同時に意味解析も行うシステムの場合，意味解析に必要な時間も β に含まれる．また，α と β は一定であると仮定している．式 (26) より，文の単語数 N が増加すると条件が満たされなくなることが分かるが，これは実験の結果とも一致する．また，書換え規則数が増加すれば，選言数の平均値 C も増加すると思われるが，この場合も式 (26) の条件は満たされなくなる．もちろん，統語解析と意味解析を同時並行的に行うシステムでは，逆に条件は満たされやすくなるが，いずれにしても，リスト R の検査にかかる時間を短縮することは必要であろう．

問題解決の見通しとしては，条件部を TRIE 構造で記述することが P1 と P2 の両方に対して有効であると考えられ，現在研究中である．また，範疇核の出現順序に関する制限を利用する方法についても検討を行っている．

7. むすび

到達可能性の自然な拡張である条件付到達可能性について述べた．条件付到達可能性は不要な部分解析木を完全に除去できるわけではない．最初にも少し触れたが，不要な部分解析木をさらに除去するためには，到達可能性以外の概念，たとえば，連接可能性などを併用するのが効果的と考えられる．我々は，すでに連接可能性の拡張である条件付連接可能性の概念を完成しており，条件付到達可能性と条件付連接可能性の併用実験も小規模ながら実施している．これらについては，さらに大規模な実験を行った後，また別の機会に発表したいと考えている．

参考文献

1) Pratt, V.R.: LINGOL – A Progress Report, *Proc. 4th IJCAI*, pp.422–428 (1975).
2) Matsumoto, Y., Tanaka, H., Hirakawa, H., Miyoshi, H. and Yasukawa, H.: BUP: A Bottom-Up Parser Embedded in Prolog, *New Generation Computing*, Vol.1, No.2, pp.145–158 (1983).
3) 松本裕治，杉村領一：論理型言語に基づく構文解析システム SAX，コンピュータソフトウェア，Vol.3, No.4, pp.4–11 (1986).
4) Aho, A.V. and Ullman, J.D.: *The Theory of Parsing, Translation and Compiling*, Vol.1, Parsing, p.541, Prentice-Hall, Englewood Cliffs (1972).
5) Harary, F.: *GRAPH THEORY*, Addison-Wesley (1968).
6) Pollard, C.J. and Sag, I.A.: *Head-Driven Phrase Structure Grammar*, Unversity of Chicago Press (1994).
7) 笠 晃一，小林修二，白石正人，横田将生：自然言語問合せ文の意味表現方法とその応用，情報処理学会論文誌，Vol.34, No.5, pp.925–933 (1993).
8) 笠 晃一，横田将生：条件付到達可能性の組込まれたパーザを生成する CFG トランスレータについて，平 10 電気関係学会九州支部連合大会 (1998).

付 録

A.1 核集合方程式を逐次近似法で解く場合の最大解への収束について

まず，次のような有限集合 L に関する連立集合方程式を考える．

$$X_i = f_i(X_1, \ldots, X_N) \quad (i = 1, \ldots, N) \tag{27}$$

ただし，f_i は L の部分集合と変数 X_j に対して和集合と積集合の演算のみで構成されるものとする．また，解 X_j の候補として，L の部分集合のみを考える．ここでの目的は，式 (27) を満足するそのような集合のうち最大のものを求めることである．なお，

$\mathbf{X} = (X_1, \ldots, X_N)$ と書けば，式 (27) は次の形に記述できる．

$$\mathbf{X} = \mathbf{f}(\mathbf{X}) \tag{28}$$

次に，$\mathbf{Y} = (Y_1, \ldots, Y_N)$ に対し，$X_i \supseteq Y_i$ がすべての i について成り立つとき，かつそのときに限り，これを $\mathbf{X} \supseteq \mathbf{Y}$ と書くことにすれば，次の補題が成立する．

補題 A.1.1 $\mathbf{X} \supseteq \mathbf{Y}$ ならば $\mathbf{f}(\mathbf{X}) \supseteq \mathbf{f}(\mathbf{Y})$ である．

（証明）$A_1 \supseteq B_1$，$A_2 \supseteq B_2$ のとき，集合の性質より次の関係が成り立つ．

$$(A_1 \cap A_2) \supseteq (B_1 \cap B_2),$$
$$(A_1 \cup A_2) \supseteq (B_1 \cup B_2) \tag{29}$$

関数 $f_i(\mathbf{X})$ は L の部分集合 $C_1, \ldots, C_{m_i}$ と変数 X_j に対して和集合と積集合の演算を再帰的に適用したものであり，さらに，任意の k に対して $C_k \supseteq C_k$ が成り立つので，次の関係式が成立する．

$$f_i(X_1, \ldots, X_N) \supseteq f_i(Y_1, \ldots, Y_N)$$
$$(i = 1, 2, \ldots, N) \tag{30}$$

これは，$\mathbf{f}(\mathbf{X}) \supseteq \mathbf{f}(\mathbf{Y})$ を意味している．（証明終了）

そこで，次の式に基づいて，$\mathbf{X}^h$ $(h = 1, 2, \ldots)$ を次々に求めることを考えよう．

$$\mathbf{X}^0 = (L, L, \ldots, L) \tag{31}$$
$$\mathbf{X}^{h+1} = \mathbf{f}(\mathbf{X}^h) \quad (h = 0, 1, 2, \ldots) \tag{32}$$

このとき，次の補題が成立する．

補題 A.1.2 $\{\mathbf{X}^h\}$ は有限回で方程式 (28) の解に収束する．

（証明）L は全体集合であるから，$\mathbf{X}^1$ の各要素は L の部分集合になる．すなわち，$(L, L, \ldots, L) \supseteq \mathbf{X}^1$ であり，これは式 (31) より $\mathbf{X}^0 \supseteq \mathbf{X}^1$ を意味する．補題 A.1.1 を用いれば，$\mathbf{f}(\mathbf{X}^0) \supseteq \mathbf{f}(\mathbf{X}^1)$ となるが，これは式 (32) より $\mathbf{X}^1 \supseteq \mathbf{X}^2$ を意味している．そして，補題 A.1.1 を繰り返し適用することにより，$\{\mathbf{X}^h\}$ は非増加列を形成することが分かる．ところが，$\{\mathbf{X}^h\}$ は下界 $(\phi, \ldots, \phi)$ を持っており，しかも，$\mathbf{X}^h$ の各要素は有限集合であるので，$\{\mathbf{X}^h\}$ は有限回で収束することがいえる．ここで，K 回で $\{\mathbf{X}^h\}$ が $\mathbf{X}^K$ に収束したと仮定しよう．これは，次の式を意味している．

$$\mathbf{X}^K = \mathbf{X}^{K+1} = \mathbf{f}(\mathbf{X}^K) \tag{33}$$

つまり，$\mathbf{X} = \mathbf{X}^K$ は方程式 (28) の解である．（証明終了）

最後に次の定理を証明する．

定理 A.1.1 式 (31) と式 (32) によって与えられる逐次近似法により，$\{\mathbf{X}^h\}$ は有限回で方程式 (28) の最大解に収束する．

（証明）補題 A.1.2 により，$\{\mathbf{X}^h\}$ が有限回で方程式 (28) の解に収束することは分かっているので，ここでは求まった解が最大解であることを示せばよい．いま，逐次近似法によって求められた解を $\mathbf{X} = \mathbf{A}$ とし，方程式 (28) の任意の解を $\mathbf{X} = \mathbf{B}$ としよう．L は全体集合であるから，$\mathbf{B}$ の各要素は L の部分集合になる．すなわち，$(L, L, \ldots, L) \supseteq \mathbf{B}$ であり，式 (31) を用いれば $\mathbf{X}^0 \supseteq \mathbf{B}$ となる．ここで，補題 A.1.1 を用いると $\mathbf{f}(\mathbf{X}^0) \supseteq \mathbf{f}(\mathbf{B})$ となるが，$\mathbf{B}$ は方程式 (28) の解であり，また，式 (32) が成り立つので，$\mathbf{X}^1 \supseteq \mathbf{B}$ が得られる．この操作を繰り返せば，任意の k に対して，$\mathbf{X}^k \supseteq \mathbf{B}$ がいえるが，特に k を $\{\mathbf{X}^h\}$ の収束回数 K に固定すると，$\mathbf{X}^K \supseteq \mathbf{B}$ となる．これは，$\mathbf{A} \supseteq \mathbf{B}$ を意味するので，逐次近似法によって求められた解は，方程式 (28) のどの解よりも小さくないことが分かる．（証明終了）

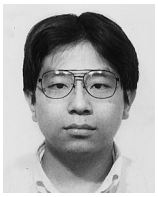
(平成 10 年 12 月 9 日受付)
(平成 12 年 1 月 6 日採録)

笠　晃一（正会員）

昭和 58 年九州大学大学院工学研究科電子工学専攻修士課程修了．昭和 61 年（株）日本データベースネットワーク研究所入社．平成 9 年より福岡工業大学情報工学部講師．計算言語学，自然言語の意味論，ニューラルネットワーク等の研究に従事．博士（工学）．

岡出　高徳（学生会員）

平成 9 年福岡工業大学工学部情報工学科卒業．平成 11 年同大大学院工学研究科修士課程修了．現在，同大大学院工学研究科博士後期課程に在学中．自然言語理解や文生成等の研究に従事．



弘中　大介（学生会員）

平成 9 年福岡工業大学工学部情報工学科卒業．平成 11 年同大大学院工学研究科修士課程修了．現在，同大大学院工学研究科博士後期課程に在学中．自然言語理解や地図理解等の研究に従事．

横田 将生（正会員）

昭和 47 年九州工業大学工学部電子工学科卒業．昭和 52 年九州大学大学院博士課程修了．同年同大学工学部助手．昭和 54 年同大学医学部附属病院講師．現在，福岡工業大学情報工学部教授．自然言語処理，図形処理，医療情報処理等に従事し，自然言語理解システム IMAGES 等を試作した．現在，メディアに依存しない意味表現を媒介とするマルチメディア統合理解の研究を行っている．工学博士．電子情報通信学会，人工知能学会，言語処理学会各会員．