

句構造文法に対する 効率的文解析手法

林達也* 宮俊司** 坂巻利哉** 吉田健一**

* 富士通研究所
** 富士通ソーシャルサイエンスラボラトリ

本論文では、一般の句構造文法に対する効率的な文解析手法について述べる。

この方法は、限定文脈依存文法(RCSG)を受け入れるYAPXR文解析システムを拡張する事によって得られる。従って本方式は、YAPXRと同様に横型トップダウン手法に基づいており、且つ又、YAPXRの特性を継承して効率の良い文解析を行うことが出来る。

日本語のような語順の自由な言語に対して、文脈自由文法ベースの記述形式は必ずしもしくりしない面がある。そして、文節間の係り受け関係も含めて考えると、むしろ句構造文法の方が文脈自由文法ベースよりも素直に日本語の規則を記述出来る事が指摘されている。

一方ギャップ文法も、語順の自由な言語を扱えるが、性能に問題がある。

本方式はこれに対して、文法が与えられた時点で予め全ての位置集合が求められれば、句構造文法に対しても効率的な文解析を行う事が出来るので、此の種の言語に有効な手法であると考えられる。

An Efficient Sentence Analysis Method
for General Phrase Structure Grammars

Tatsuya Hayashi * Syunji miya ** Toshiya Sakamaki ** Kenichi Yoshida**

* Fujitsu Laboratories LTD.

1015 Kamikodanaka Nakaharaku Kawasaki, 211, JAPAN

** Fujitsu Social Science Laboratory LTD.

This paper describes an efficient sentence analysis method for conventional phrase structure grammars.

The expressions based on context free grammars are not particularly suited for languages like Japanese where free word order is allowed. Moreover, it is pointed out elsewhere that when we consider dependency relationships among "Bunsetsus", phrase structure grammars are more suitable for expressing Japanese.

Our method is appropriate for such languages, since it provides the power of phrase structure grammars with improved efficiency, provided all the "position sets" could be calculated in advance.

1. まえがき

筆者等は先に、文脈自由文法を主体にしそれに解析効率に影響を及ぼさない範囲で文脈依存規則を導入したRCSG (Restricted Context Sensitive Grammar) に対する文解析システムYAPXRについて述べた^{1) - 5)}。

RCSGは基本的には、文脈自由文法の枠組みから離れたものではない。そのため、日本語のように語順の自由な言語の場合は、必ずしもRCSGが適切な文法であるとは言えない。

ギャップ文法は語順の自由な言語を簡潔に記述できる特色を持っているが^{6) - 8)}、gapが0を含む任意の長さの部分導出列を代表出来るので、逆に性能上問題が生じる事が予想される。

一方日本語は、語順は自由だが、無条件に自由な訳ではなく、文節を単位として係り受け関係が右向きでかつ非交叉でなければならないと言う条件が存在する。そして、この係り受け関係に対応する非終端記号を導入すると、日本語は一般的な句構造文法(PSG)を用いて素直に記述出来る事が示されている⁹⁾。

そこで、句構造文法に対する効率的な解析手法が課題となるが、これに対しては上述したYAPXRを多少拡張する事で、横型トップダウン方式の句構造文法向けの実用的な解析システムを得る事が出来る。

本論文ではこれについて述べる事にする。以下では、2.で文法記述形式、3.で実現方法、4.で動作例について述べる。

2. 文法記述形式

ここで許される文法は、従来のYAPXRで用いられている機能を全て包含した句構造文法である。基本形式は次に示す通りである。

左辺 → 右辺.

ここで、

左辺 : 左辺項…

左辺項 : 構文要素 [(内部引数, …)]

[(/外部引数, …/)]

右辺 : ε 又は 右辺項

右辺項 : 構文要素 [(内部引数, …)]

[(/外部引数, …/)]

[スラッシュカテゴリ]

[{ 文脈依存制約 }]

[{ 基本制約 }]

但し、外部引数や文脈依存制約、スラッシュカテゴリは本論とは直接関係ないので省略する。(外部引数等については文献2) 4) 5) を参照されたい。

3. 実現方法

記述形式がRCSGに合致する場合は、従来の方法で実現できるため、ここでは、一般の句構造文法規則を対象にして実現方法を述べる。

RCSGを対象とした場合は文法が与えられた時点で、規則右辺の左端要素に対応する核位置を全て容易に求める事が出来る。

しかし句構造文法を対象とする場合は、核位置を求めるのが容易でない場合がある。そこで、ここでは先ず核位置が決定したものとして述べる。尚、文法規則を

$$P_1 P_2 \dots P_m \rightarrow {}^{n_1}Q_1 {}^{n_2}Q_2 \dots {}^{n_n}Q_n$$

($P_i, Q_j \in V, nk \sim$ 位置id)

とする。単純の引数や制約は除外する。又、 P_i, Q_j に対応する述語をそれぞれ p_i, q_j とする。

引数はYAPXRでは以下の通りであった。

n : 解析バスの先頭要素

Ai : 解析スタック (入力)

(解析バスの残りの要素)

Ao, Ao1 : 解析スタック (出力, 差分リスト)

Oi : 省略スタック (入力)

Oo, Oo1 : 省略スタック (出力, 差分リスト)

Li : 引数スタック (入力)

Lo, Lo1 : 引数スタック (出力, 差分リスト)

Ti : 解析木スタック (入力)

To, To1 : 解析木スタック (出力, 差分リスト)

しかしここでは、本論に関係ない省略スタックは除外する。又、解析木は句構造文法の場合複雑になるので、解析木スタックには代わりに係り受け関係(W1, W2)を格納する事にする。

(W1, W2) は文節W1が文節W2に係る事を示す。

3.1 左端型

構文要素 Q_1 に対応するホーン節は、

$$q_1(11, Ai, [[n2, 11 \uparrow Ai] \uparrow Ao], Ao, \\ Li, [[Li] \uparrow Lo], Lo, \\ Ti, [[Ti] \uparrow To], To).$$

$$q_1(12, Ai, [[n2, 12 \uparrow Ai] \uparrow Ao], Ao, \\ Li, [[Li] \uparrow Lo], Lo, \\ Ti, [[Ti] \uparrow To], To).$$

...

$$q_1(1k, Ai, [[n2, 1k \uparrow Ai] \uparrow Ao], Ao, \\ Li, [[Li] \uparrow Lo], Lo, \\ Ti, [[Ti] \uparrow To], To).$$

となる。11~1kはn1に対応する核位置である。

1jを左位置として持つ要素は一般に、非終端記号に限らないのが句構造文法の場合の特色である。

前述したように一般には、n1に対応する核位置を求める事が容易でない場合がある。但しその場合でも以下の様に2通りの方法のいずれかを用いることによって実現することができる。

まず第1の方法は、解析パスの先頭要素が何であってあたかも核位置であるかのように見なして、解析パスをn2に進めるという方法で、時間・空間的に非効率ではあるが、ともかく解析を正しく行う事が出来るのである。唯、その時の解析は完全なボトムアップになるので、 ϵ 規則は禁止される事になる。

これに対し第2の方法は、 ϵ 規則を許し且つ現実的な方法である。それは次の通りである。

まず最初に、句構造規則の左辺は第1項のみを考える。すると、見かけ上YAPXRの場合と同様に左端要素に対応する核位置は全て求める事が出来る。これを仮核位置と呼ぶ事にしよう。

容易に分かるように、本来の核位置は同時に又仮核位置でもある。

こうしておいて、左端型のアクションを仮核位置に基づいて作成すれば良いのである。

3.2 中間型

Q_2 に対応するホーン節は、

$$q_2(n2, Ai, [[n3 \uparrow Ai] \uparrow Ao], Ao, \\ Li, [[Li] \uparrow Lo], Lo, \\ Ti, [[Ti] \uparrow To], To).$$

で、これは従来と全く同じである。

3.3 右端型

Q_m に対応するホーン節を次のように定める事がポイントである。

$$q_n(nn, [N \uparrow Ai], Ao, Ao1, \\ Li, Lo, Lo1, Ti, To, To1) :- \\ p_1(N, Ai, A1, Ao1, \\ Li, L1, Lo1, T1, To1), \\ p_200(A1, A2, L1, L2, T1, T2), \\ \dots \\ p_m 00(A_{m-1}, Ao, L_{m-1}, Lo, T_{m-1}, To).$$

つまり、左辺の要素全体に対応する（と思われる）部分入力列が検出されたと思なして、記述された順序で連続して述語呼出を行う点が句構造文法の場合の特色である。

これを出発記号 S_0 からの導出と関連づけて見ると、右端型のアクションは、

$$S_0 \Rightarrow \alpha P_1 P_2 \dots P_m \tau \Rightarrow \alpha Q_1 Q_2 \dots Q_n \tau$$

に於いて、 α の最後の要素の右位置から τ の最初の要素の左位置に解析パスを進めることに他ならない。

$p_200 \sim p_m 00$ は入力文に対応する述語と同様に、解析スタック（入力）は一般に複数の解析パスから成る（但し、リストの最後が変数である点が異なる）。従って、以下のホーン節を設けて、インターフェイスを合わせるものとする。

$$p_j 00([[N \uparrow Ai] \uparrow Ao], Ao1, \\ [Li \uparrow Lo], Lo1, \\ [Ti \uparrow To], To1) :- \\ p_j(N, Ai, Ao, Ao1, \\ Li, Lo, Lo1, \\ Ti, To, To1),$$

!

p_j 00(Ao, Aol, Lo, Lol, To, Tol).
p_j 00(Aol, Aol, Lol, Lol, Tol, Tol) :- !.

仮核位置方式の場合も右端型のアクションは同様に、左辺の要素に対応する述語を逐次的に呼び出せばよい。

但しこの場合の相違点は、核位置方式では最終的な解析パスが常に存在するが、仮核位置方式では逐次的な述語呼出の過程で、解析パスが消滅してしまう場合があると言う事である。

つまり、仮核位置方式は核位置方式よりは効率が良くないが、単純なボトムアップ方式に比べると効率的であり、両者の中間に位置する方法であるといえる。

4. 動作例

本節では簡単な日本語の文法を例にとり上げて、パーサの動きを説明しよう。

使用する文法例を図1に示す。これは文献9)を参考にして少し手を加えたものである。

[e] はピリオドを表す。[p] は文節を示す終端記号で以下の引数を持つものとする。

W : 自立語+付属語,

C1 : 自立語の品詞区分,

C2 : 付属語の品詞区分又は活用形

relは係り受け関係を示す非終端記号である。但し、relのみを左辺に持つ規則はなく、従ってそれから導出される終端記号は存在しない点を指摘しておく。

簡単な為、s や prseqには引数を持たせない事にする。規則1, 4, 5, 6 の checkは構文的側面から係り受けの可否を調べる簡単な述語である。又、putは係り受け関係(W1, W2)を解析木スタックに格納する命令である。規則1の基本制約は最後の文節が終止形でなければならない事を指定している。

この例では、左端要素⁹ [p]、¹² [p]、¹⁶ [p]に対応する核位置には、b, 8, 10, 14, 11, 15, 17の7個が存在する。

この文法に基づくprologのプログラム(Quintus版)を付録に示す。

入力文を「tarouga jirouni renrakusuru kotowo yakusokusuru.」としよう。

又、本システムへの入力述語は、

yopen(A1, L1, T1),

p0 (A1, A2, L1, L2, T1, T2, [tarouga, n, ga]),

p0 (A2, A3, L2, L3, T2, T3, [jirouni, n, ni]),

p0 (A3, A4, L3, L4, T3, T4, [renrakusuru, vt, cn]),

p0 (A4, A5, L4, L5, T4, T5, [kotowo, n, wo]),

p0 (A5, A6, L5, L6, T5, T6,

[yakusokusuru, vt, end]),

e0 (A6, A7, L6, L7, T6, T7, ['.', end, end]),

yclose(A7, L7, T7).

としよう。ここで、

n : 名詞,

vt : 他動詞,

cn : 連体形,

end : 終止形,

yopen : スタックの初期設定を行う述語,

yclose : 結果(係り受け構造)を出力する述語,

である。入力述語のホーン節も付録に示しておく。

解析チャートの数はこの場合22個であるが、係り受け構造で見ると異なり数は図2に示すように5通りとなる。

(1)に対応する解析チャートの一つを示すと図3のようになる。余談であるが、この例は文脈処理を行わなければ曖昧性が解消しない。

本方式はLR手法をベースにしているのので、従来のYAPXRと同様に、解析チャートはボトムアップに作成される。

図3の場合、先ず規則6により(kotowo, yakusokusuru)に対応する部分チャートが生成され、述語p, rel, pが呼び出される。先頭のpはやはり規則6により右端型アクションを行って(renrakusuru, kotowo)に対応する部分チャートを生成し、再びp, rel, pを呼び出す。以下同様の手順を経て図のような解析チャートが生成される訳である。

ここで、終端要素¹⁷ [p]が左端要素¹⁶ [p]の核位置になっている所が従来のRC SGとは異なる点である。

5. あとがき

本論文では、横型トップダウン文解析システムY

A P X R を拡張して一般の句構造文法に対する効率的な文解析システムが得られる事を示した。

日本語のように語順が自由な言語の場合、文脈自由文法の枠組みで記述する事は必ずしもしっくりしない所がある。ギャップ文法では、記述力は一応あるとしても、能率の良い解析が期待出来ない。

日本語は係り受け関係も含めて考えると、句構造文法の方が素直に表現出来るので、本方式の有効性が期待出来る。

つまり、ここに述べた方法は、従来の Y A P X R と同様に、

- (1) 左端要素に対応する核位置あるいは仮核位置を全て、文法が与えられた時点で求める事が出来るので、実行時のトップダウン予測が不要である。
- (2) prolog 処理系の特性を生かして、第 1 引数をアトムにしてダブルハッシュを可能にしている。
- (3) 実行時にはバックトラックを生じない。

等により効率の良い解析システムになっているからである。本文では基本手法とその有効性の片鱗を簡単な例で示したに過ぎないので、機械翻訳や自然言語インターフェイス等の実用規模の日本語への適用が今後の課題と考えている。

参考文献

- 1) 林達也：論理型言語による構文解析法 Y A P について、情報処理学会論文誌， Vol. 29, No. 9, pp. 835-842(1988)。
- 2) 林達也：拡張 C F G とその構文解析法 Y A P X について、情報処理学会論文誌， Vol. 29, No. 5, pp. 480-487 (1988)。
- 3) 林達也：Y A P X の効率的実現法，情報処理学会論文誌， Vol. 30, No. 10, pp. 1354-1356(1989)。
- 4) 林達也：横型トップダウン文解析システムの実現と評価，情報処理学会，自然言語処理研究会 Vol. 74-9, pp. 65-72(1989)。
- 5) 林達也：文解析システム YAPXR の実現と評価，情報処理学会論文誌（投稿中）(1989)。
- 6) Dahl, V. & Abramson, H. : On Gapping Grammars, Proc. 2nd International Conference on Logic Programming, pp. 77-88(1984)。
- 7) Dahl, V. : More on Gapping Grammars, Proc. FGCS, PP. 669-677(1984)。
- 8) Popowich, F. : Unrestricted Gapping Grammars, Proc. IJCAI 85, pp. 765-768(1985)。
- 9) Sugimura, R. : Logical Dependency Grammar and Its Constraint Analysis , ICOT technical memorandum TM-0679, pp. 10 (1989)。

```

0.  so          -->  b s • |
1.  s           -->  1 prseq      2 [p ] (w, c1, c2) {c2 = end}
                        3 [e ] .
2.  prseq      -->  4 [p ] (w, c1, c2)      5 rel .
3.  prseq      -->  6 [p ] (w, c1, c2)      7 rel      8 prseq .
4.  [p ] (w1, c11, c12) rel [p ] (w2, c21, c22) -->
                        9 [p ] (w1, c11, c12) 10 prseq
                        11 [p ] (w2, c21, c22) {chech(c11, c12, c21), put(w1, w2) } .
5.  [p ] (w1, c11, c12) rel [p ] (w2, c21, c22) -->
                        12 [p ] (w1, c11, c12) 13 rel 14 prseq
                        15 [p ] (w2, c21, c22) {chech(c11, c12, c21), put(w1, w2) } .
6.  [p ] (w1, c11, c12) rel [p ] (w2, c21, c22) -->
                        16 [p ] (w1, c11, c12)
                        17 [p ] (w2, c21, c22) {chech(c11, c12, c21), put(w1, w2) } .

```

図1 句構造文法例

- | | | | |
|-----|-------------------------|-------------------------|------------------------|
| (1) | (tarouga, yakusokusuru) | (jirouni, yakusokusuru) | (jirouni, renrakusuru) |
| | (renrakusuru, kotowo) | (kotowo, yakusokusuru) | |
| (2) | (tarouga, yakusokusuru) | (jirouni, yakusokusuru) | |
| | (renrakusuru, kotowo) | (kotowo, yakusokusuru) | |
| (3) | (tarouga, yakusokusuru) | (jirouni, renrakusuru) | |
| | (renrakusuru, kotowo) | (kotowo, yakusokusuru) | |
| (4) | (tarouga, renrakusuru) | (jirouni, renrakusuru) | |
| | (renrakusuru, kotowo) | (kotowo, yakusokusuru) | |
| (5) | (tarouga, yakusokusuru) | (tarouga, renrakusuru) | (jirouni, renrakusuru) |
| | (renrakusuru, kotowo) | (kotowo, yakusokusuru) | |

図2 例文の解析結果

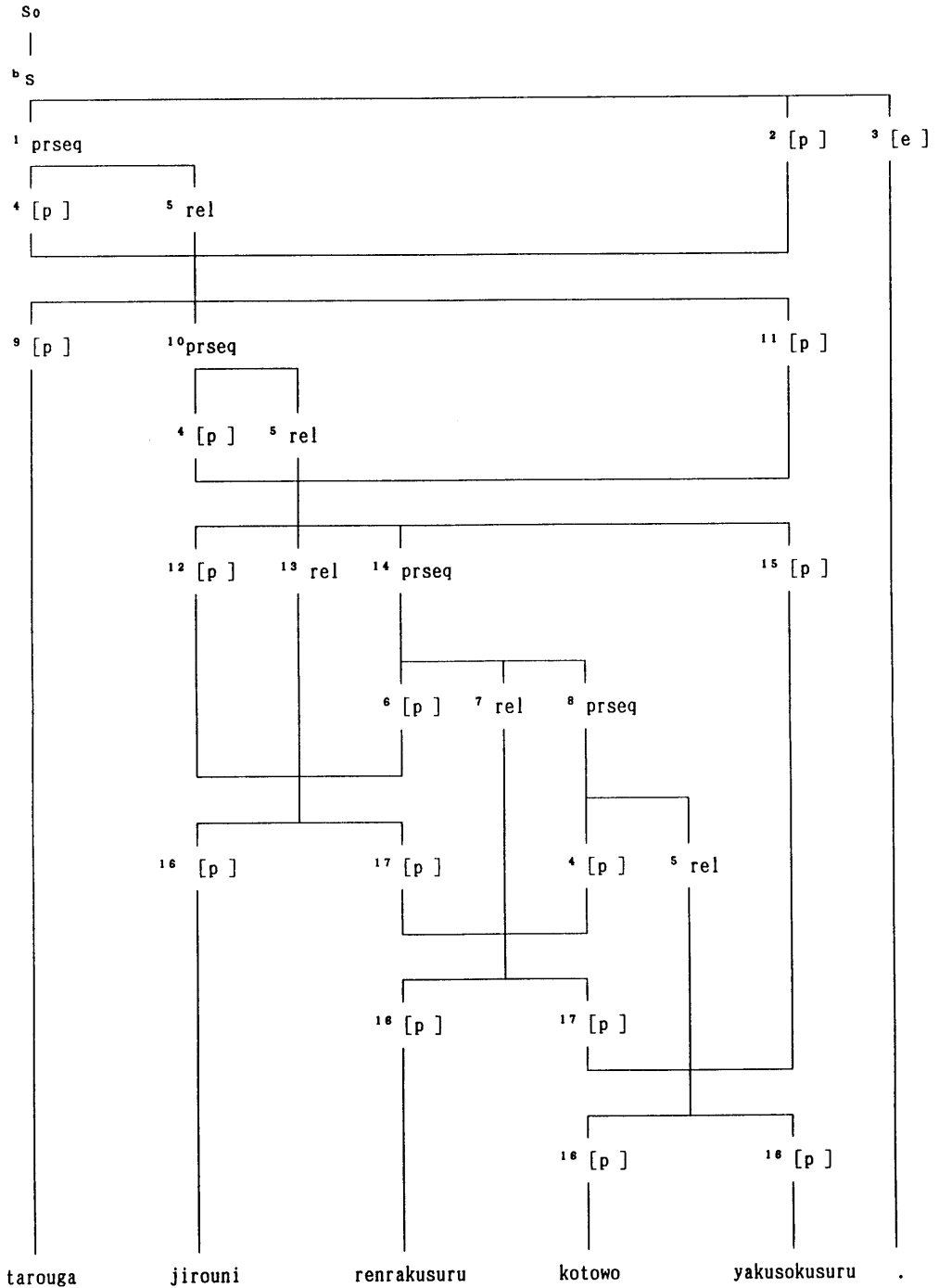


図3 例文の解析チャート

