

TM-0241

CIL による JPSG パーサの
試作

郡司隆男 (大阪大学)
三古秀夫, 奥西稔幸 (ICOT)

November, 1986

©1986, ICOT

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03) 456-3191~5
Telex ICOT J32964

Institute for New Generation Computer Technology

CILによるJPSGパーサの試作

JPSG Parser in CIL

三吉秀夫

Hideo Miyoshi

(新世代コンピュータ技術開発機構)

(ICOT)

奥西稔幸

Toshiyuki Okunishi

郡司隆男

Takao Gunji

(大阪大学)

(Osaka Univ.)

JPSG(Japanese Phrase Structure Grammar)はGPSGやHPSGなどの文法理論の考え方を基礎にした日本語の句構造文法である。本稿ではCILを用いたJPSGパーサの実装方法を述べ、JPSGの特徴である「素性の集合としての統語カテゴリ」や「各種統語原則」がCILの基本機能により比較的容易に実装できることを示す。また簡単な解析例を示す。

1. はじめに

自然言語理解システムに於いて、文法は辞書や推論規則などの知識ベースと同様に、重要なコンポーネントであり、きちんとした言語理論的基礎の上に作り上げられた文法体系は、あらゆる自然言語処理応用システムに有用な基盤となる。JPSG(Japanese Phrase Structure Grammar)は句構造文法に基く新たな日本語の文法体系を開発する、第五世代コンピュータプロジェクトの自然言語処理テーマの1つである。JPSGは既存のGPSG(Generalized Phrase Structure Grammar) [1]やHPSG(Head-driven Phrase Structure Grammar) [7]の基本的な枠組みを利用しつつ、語順の任意性やコントロール現象などに見られる日本語特有の言語現象を扱うべく拡張を行った文法体系である。JPSGはICOTの中期4年間で、日本語における主要な構文をカバーする文法体系を開発することを目指して、JPSワーキンググループにおいて、文法とパーサの開発を行っている。現在我々は2つのアプローチによりJPSGのパーサを開発している。1つは条件付単一化 [3, 8] による方法であり、もう1つは本稿で述べる、既存のロジックプログラミングの枠組を用いる方法である。本稿では、後者の手法によってJPSGの機能の一部を組み込んだパーサを試作したのでその実装方法について述べる。本パーサは、ICOTで自然言語処理ツールとして開発中のPrologの拡張言語であるCIL(Complex Indeterminate Language) [6]を用いて記述される。CILの部分項(Partial Term)、JPSGの特徴である「素性の集合としての統語範疇」は自然な形で実装できる。また、各種の統語原則はCILのゴール呼出しにより実現できる。

2. JPSGの特徴

JPSGはGPSGとHPSGの基本的な枠組みをベースにした日本語の句構造文法である。これらの文法理論は従来変形を用いて説明されていた言語現象を文脈自由文法の枠内で記述しようとするものであり、言語学のみならず計算機科学においてもその有用性が注目されている。これらの文法理論は自然言語に見られる様々な言語現象を記述し易い枠組みを提供しているが、日本語には語順の任意性という特有の性質がある。従って、語順がほぼ固定された英語を対象としたこれらの理論を、そのままの枠組みで日本語に適用するのは賢明でない。JPSGはGPSGで用いられている普遍的な統語原則を用い、HPSG流のSUBCAT素性を更に拡張したSUBCAT素性により、語順の任意性やコントロール現象の記述を扱い易いように改良した、独自の文法体系である。JPSGの理論的枠組みは、現在開発途中でもあるので細部については流動的であり、未決定の部分もある。最近の仕様については文献 [2, 5] を参照されたい。ここでは第3節で必要なJPSGの基本的な特徴を簡単に述べる。

2.1 素性の集合としての統語範疇

統語範疇は(素性, 素性値)対の集合として表現される。素性の例としては、次のようなものが挙げられる。

(例)

POS(Part Of Speech) : いわゆる品詞を表し、V(動詞)、N(名詞)、P(助詞)等をとる。

PFORM(Postposition FORM) : 助詞の形を表し、ga(が)、wo(を)、ni(に)等をとる。

GR(Grammatical Relation) : 文法的な関係を表すもので、SBJ(主格)またはOBJ(目的格)をとる。

SEM(SEMantics) : 統語範疇の意味表現を表す素性で各語彙項目から構成的に作られる。

SUBCAT : 動詞や助詞などが補語としてとる統語範疇の集合を値としてとる素性である。これはHPSGのSUBCAT素性の拡張であり、集合の要素間には原則的に順序関係はない。

REFL : 句が日本語の再帰代名詞「自分」を含むかどうかを示す素性である。これは値として、 ϕ (空集合) またはただ1個の要素を持つ集合 {PP[SBJ]} をとる素性である。{PP[SBJ]} をとれば「自分」を

含心ことを表し、要素PP[SBJ]はその束縛に用いられる。

なお、統語範疇をX[Y1;Y2;...]と記すことがあるが、これはPOS 素性値あるいは範疇名がXでその他Y1,Y2,...という素性値を持つ統語範疇であることを示す。

これらの素性を用いることにより、従来、S(文)、VP(動詞句)、TVP(他動詞句)、PP(後置詞句)、などと呼ばれていた統語範疇は次のような素性の集合として表現される。

S = {POS V; SUBCAT ϕ }
 VP = {POS V; SUBCAT {PP[SBJ]}}
 TVP = {POS V; SUBCAT {PP[SBJ], PP[OBJ]}}
 PP = {POS P; SUBCAT ϕ }

また語彙項目も同様に、次のような素性の集合として辞書に記述される。

健: {POS N; SUBCAT ϕ ; SEM ken}
 歩く: {POS V; SUBCAT {PP[SBJ; SEM X]}; SEM walk(X)}
 愛す: {POS V; SUBCAT {PP[SBJ; SEM X], PP[OBJ; SEM Y]}; SEM love(X,Y)}

2.2 句構造規則

JPSGでは、「親範疇(M)は右側のhead(主要部)範疇(H)と左側の他の1個の範疇(D)から構成される」ことを示す次のただ1個の句構造規則を仮定する。

(2.1) $M \rightarrow D H$

この規則は、日本語の構文が図 2.1のような二分木で表現できることを示している。

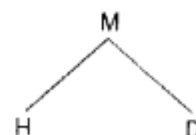


図 2.1 JPSGの句構造

2.3 SUBCAT素性とその原則

SUBCAT素性はhead範疇が補語としてとる統語範疇の集合を素性値として持つ。SUBCAT

素性の原則は、親範疇と娘範疇の間のSUBCAT素性値に関する規約であり、次のように定義されている。

「(2.1)の句構造規則において、MのSUBCATの値はHのSUBCATの値からDと等しいものを取り除いたものと一致する。」

図 2.2にSUBCAT素性の原則の簡単な例を示す。

この素性とそれに関する原則により語順の任意性が扱える。また句構造規則の簡約化が可能になる。

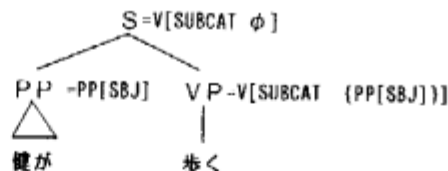


図 2.2 SUBCATとHEAD素性の原則の例

2.4 HEAD素性の原則

これはGPSGでHFC(Head Feature Convention)と呼ばれている基本的な統語原則であり、HEAD素性という一群の素性に関する規約である。JPSGでは、

「親範疇のHEAD素性と、娘範疇の中のhead範疇のHEAD素性は値が等しい。」

と規定されている。JPSGではSEM, SUBCAT, SLASH, REFL以外の素性がHEAD素性に相当し、図 2.1の親範疇Mおよび右の娘範疇HのHEAD素性値が等しいことを規定する。図 2.2はこの原則の例であり、VP(head範疇)のPOS 素性値と親範疇SのPOS 素性値がVに等しいことを示している。

2.5 FOOT素性の原則

これはGPSGでFFP(Foot Feature Principle)と呼ばれている原則で、FOOTと呼ばれる特別な素性の伝わり方を記述する統語原則である。JPSGではとREFLがFOOT素性の1つであり、次のように定義されている。

「(2.1)の句構造規則において、MのFOOT素性は、DのFOOT素性とHのFOOT素性の和集合あるいはそれからPPを一つ取り除いたものに等しい。」

なおJPSGでは、REFLの伝承に関して更に、「『自分』を支配する範疇のSUBCATの要素にPP[SBJ]があれば、それはREFLのPP[SBJ]と束縛可能でその場合その範疇より上にはREFLの値は伝わらない」というコントロールの原則が定義されている。

3. CILによる実装

本節では第2節で述べたJPSGの基本原理解に基くJPSGパーサをCILを用いて実現する方法を述べる。CILは各種の自然言語処理システムを開発することを目的として開発された、Prologの拡張言語である。基本的な動作はPrologと同様バクタラックベースの逐次実行型であるが、基本データとして部分項(Partially Specified Term)と呼ばれるものを持ち、また変数に値が束縛されるまで実行を凍結するフリーズ機能が組込まれている。部分項は通常の項(term)の拡張であり、属性/属性値の集合として表現される。これは一種のフレームとみなすことができる。部分項は次の形式をしている。

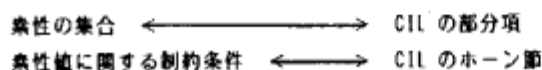
$$\{\alpha_1/\beta_1, \dots, \alpha_n/\beta_n\} \quad n \geq 0$$

各 α_i が属性を表し、 β_i が α_i の属性値を表す。部分項同士には拡張されたユニフィケーションが定義されている。たとえば、 $\{a/1, b/2, c/3\} = \{a/1, c/X, d/4\}$ を実行すると、ユニフィケーションは成功し、その結果、

$$\{a/1, b/2, c/3, d/4\} \text{ 及び } X = 3$$

が得られる。またCILには各種の略記法が定義されており、たとえば「A!B」は部分項Aの属性Bの値を表す。なお、部分項の値が再び部分項であってもよい。

JPSGのような、ユニフィケーションベースの文法は、「素性の集合」と「それらに関する制約(constraints)」で形式化することができる。第2節で述べたJPSGの枠組みはすべて素性の集合としての統語範疇とM、D、Hの箇のそれらの値に関する制約条件としてとらえられる。本パーサではこれらの特徴をCILのプリミティブと次のように対応付けている。



この考え方により、たとえば固有名詞「健」、動詞「愛す」はそれぞれ次のような部分項として表現される。

<pre>{ pos/n, subcat/[], refl/[], sem/ken }</pre>	<pre>{ pos/v, subcat/[[pos/p, gr/sbj, pform/ga, subcat/[], sem/Sem1], pos/p, gr/obj, pform/wo, subcat/[], sem/Sem2]], sem/love(Sem1, Sem2) }</pre>
---	---

一方、CILの基本動作はPrologと同じであるので、既に論理型言語で開発されている構文解析手法がそのまま適用可能である。ここではGALOP(BUP) [4]を拡張することを考える。GALOPでは1箇の句構造規則が1箇のホーン節に対応し、非終端項に相当する述語の引数に各種属性を与えることができる。ここで示す手法は、素性の集合としての部分項を述語の引数として与え、SUBCAT素性の原則やHEAD素性の原則のような各種原則を句構造規則中の付加手続きとして埋込む方式である。これにより(2.1)の句構造規則は次のようなCILの節形式で表される。

```
nt(d(FD), G, X0, X) :-
    t(h(FH), X0, X1),
    sfp(FD, FH, FM), hfc(FH, FM), ffp(FD, FH, FM),
    nt(m(FM), G, X1, X)
```

この節の中の変数FD、FH、FMがそれぞれ(2.1)のD、H、Mの素性の集合を表す部分項である。3行目の述語sfp、hfc、ffpがそれぞれSUBCAT素性の原則、HEAD素性の原則、FOOT素性の原則に対応する付加手続きである。SUBCAT素性の原則は拡張ユニフィケーションを用いて次のような節で書ける。

```
sfp(FD, FH#(subcat/SH), FM#(subcat/SM)) :-
    del(FD, SH, SM)
```

述語delは、第2引数のリストから第1引数の要素を除いたものが第3引数であることを表す。hfcやffpも似たような形式で定義されている。

本プログラムによる構文解析結果を次に示す。本JPSGパーサは逐次型推論マシンPSI [9]上にインプリメントされている。解析結果はPSIのビットマップディスプレイ上に構文解析木及び木の中の任意のノードの素性の集合を表示することができる。入力分ち書きされた単語リストで与えられ、形態素解析は行わない。図3.1は「奈緒美が健に富夫が自分を愛すと言う。」に対する構文解析木と最上位の文に対応するノードの素性の集合を示す。

