HPSGにもとづく実用日本語文法について

大谷 朗　宮田 高志　松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{akira-o, takashi, matsu}@is.aist-nara.ac.jp

1 はじめに

自然言語処理において、言語の諸相を見据えた宣言的な文法にもとづく文解析は必須である。そのような言語 (処理) 観から実用的な文法システムを開発すべく、NAIST JPSG という Head-driven Phrase Structure Grammar—HPSG [4, 5] に立脚した日本語句構造文法の構築およびその実装を行った。HPSG は人間の言語活動を情報という観点から形式的に捉えようとした文法理論であり、多くの言語現象を抽象度の高い一般的な原理により説明できる。また特殊な現象も原理間の相互作用の結果として説明するような機構が内在しているため、様々な言語情報が統合的に記述できる。以下では、文法理論の導入によるこのような利点が処理に反映されるように設計を試みた文法システムの概要について述べる。

2 NAIST JPSG の概要

言語情報は (1) のような素性構造で記述される。

(1)
$$\begin{bmatrix} \textit{synsem\_struc} \\ \text{SYN} \begin{bmatrix} \text{HEAD} \begin{bmatrix} \textit{head} \\ \text{CASE} \quad \textit{case} \\ \text{ARG-ST} \quad \textit{list(synsem\_struc)} \\ \text{MOD} \quad \textit{list(synsem\_struc)} \end{bmatrix} \\ \text{VAL} \begin{bmatrix} \text{SUBCAT} \quad \textit{list(synsem\_struc)} \\ \text{ADJACENT} \quad \textit{list(synsem\_struc)} \end{bmatrix} \end{bmatrix} \\ \text{SEM} \quad \textit{sem\_struc} \end{bmatrix}$$

イタリック体による記述は部分情報構造の型に対して付けられた名前である。型同士は上位・下位関係に関して束を成す。*list*(α) は α という型を持つ素性のリストを表す。*synsem_struc* は語または句の持つ情報を記述するための型の総称であり、*phrase, word, lexeme, ...* といった下位型を持つ [5]。また、主辞 (HEAD) 素性の値である *head* は品詞を表し、*noun, verb, ptcl, ...* などの下位型を持つ。格 (CASE) 素性の値である *case* は格に関する情報を表す。語や句・文の意味内容を担う意味 (SEM) 素性およびその値は [4] の記法を踏襲している。

HPSG の抽象的な原理の形式化をそのまま実装することは困難であり、処理に即した調整が必要である。特に I. 語順、II.(音形を持たない空の語彙を仮定した)"省略" や "痕跡" の説明は、原理の体系を実装する際には慎重に対処しなければならない。以下では、I. II に関する原理・制約をはじめ、NAIST JPSG の主だった枠組について述べる。

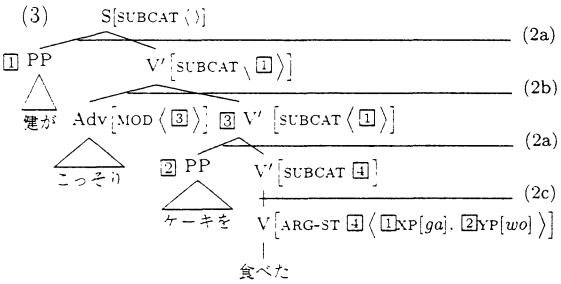
2.1 直接支配原理とスキーマ

HPSG では、I は下位範疇化に関する語彙的情報と線形順序制約、直接支配原理といった普遍的制約によって規定される。これらは構成素間の局所的な順序のみを規定し、制約に違反しない任意の語順を文法的とする。しかし、そのことを解析に反映させるには、上記の制約から全ての句構造規則を事前に派生させておくといった工夫が必要であり、NAIST JPSG では (2a-d) に示す四つのスキーマを導入している。

(2)
a. $[\textit{phrase}] \rightarrow \text{C}[\textit{phrase}] \quad \text{H}$
b. $[\textit{phrase}] \rightarrow \text{A}[\textit{phrase}] \quad \text{H}[\textit{phrase}]$
c. $[\textit{phrase}] \rightarrow \text{H}[\textit{word}]$
d. $[\textit{word}] \rightarrow \text{X}[\textit{word}] \quad \text{H}[\textit{word}]$

(2) は [4, 5] のものと一部異なっている。しかし、その選択が選言的であること、またこれらが伝統的な句構造規則に代わり (3) のような構文木を保証している点においては、統語論に対する考え方に差異はない。

(3)



2.2 結合価原理と語順転換

II は (4) の下位範疇化 (SUBCAT) 素性原理の問題である。

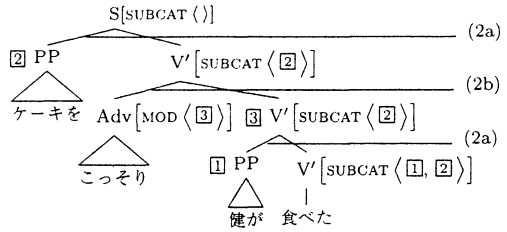
(4)
$$\begin{bmatrix} \text{SUBCAT} \quad \boxed{2} \\ \text{ADJCNT} \quad \langle\rangle \end{bmatrix}$$
$$\boxed{1}[\text{ADJNT} \quad \langle\rangle] \qquad \begin{bmatrix} \text{SUBCAT} \quad \langle \boxed{1} \rangle \oplus \boxed{2} \\ \text{ADJCNT} \quad \langle\rangle \end{bmatrix}$$

この素性が空 (句が飽和している) か否かは、しばしば他の様々な制約と関わる。英語のように省略が少ない言語では句の飽和という概念は重要であるが、日本語ではほとんど全ての要素が省略可能であり、飽和していないと分析できてしまう句が頻出する。[4] ではそのような省略に対して音形を持たない空の語彙を仮定するが、これは解析の効率を著しく低下させる。

また、HPSG では、句構造表示の制約、すなわち言語の階層性を規定するスキーマによって、英語のように語順の制限が厳しい言語と日本語のように語順の制限がゆるい言語を区別していた。それに対し NAIST JPSG は、(2a, b) により日本語の基本的な文の統語構造を語順の制約がゆるい非階層言語のような平坦な構文木とし

て分析するのでもなく、また階層言語のような主語・目的語の非対称性も仮定しない。

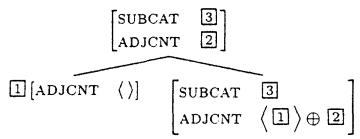
(5)



(5) は (3) の語順転換の一例であるが、どちらも S のノードにおいて全ての SUBCAT 素性が打ち消されている点では同じである。実装では素性の打ち消しに順序を設けないことにより語順転換を説明する。構文木の構造には、いわゆる構成素の階層関係は反映されていないが、そのような情報は SEM 素性の構造で表現している[1]。

NAIST JPSG では下位範疇化に関する素性原理として、(4) の他に隣接 (ADJACENT) 素性も導入している。

(6)



(6) は下位範疇化されている要素の中でも特に隣接したものに関する制約を扱うための素性である。日本語では助詞や助動詞などのように、他の要素を下位範疇化するだけでなく、それらと隣接していることを要求する語が存在する。この素性は 2.3 節で論じるような格に関する様々な現象の説明においても利用される。

2.3 助詞に関する素性と原理

文法の実用面を考慮すると、格助詞については (7) にあげる三つの現象が説明できなければならない。

(7) a. 名詞に後接する場合: 健が 走る。
b. 省略される場合: 健 来た？
c. 動詞に後接する場合: 行くが よい。

NAIST JPSG では (7b, c) に対して、空のガやコトなどの語彙を辞書に仮定した解析はせず、格助詞は名詞と動詞の両方を下位範疇化し得ると考えている。

さらに、説明できなければならない現象として「二つ以上の格助詞を伴えない」ということが挙げられる。

(8) a. *健をが 走る。 b. *行くがが よい。

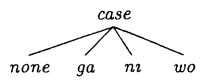
[1] 近年の言語学では統語構造と音韻・形態構造の独立性が考えられており、例えば [2] では、順序連接という制約を用いることにより、それらの間の関係が適切に捉えるような機構を提案している。しかしながら、順序連接は解析システムの構築という点では実装が困難であり、また音韻・形態情報を反映して語順を規定する構文木にもとづく解析の方が不要な可能性の数を制限しやすいため、現時点では (5) のような分析を採用している。

そこで、CASE 素性を下位範疇化で指定し、動詞・名詞・格助詞を (9) のように形式化した。

(9)

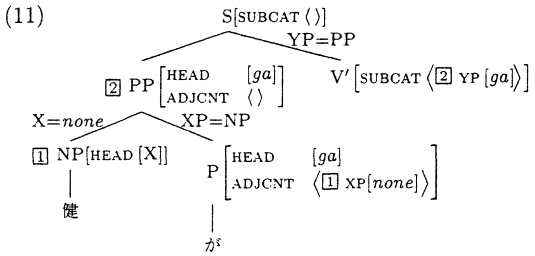
ここで XP[α] は (10a) の素性構造の略記である。

(10) a. [phrase HEAD [CASE α]] b.



none および *ga* は 2 節で説明した型 *case* の下位型であり、(10b) のような型階層を形成しているとする。すでに格の情報が指定されているものにさらに格の指定をすることは、格助詞の ADJCNT 素性に CASE 素性が *none* である要素を指定することで禁止している。

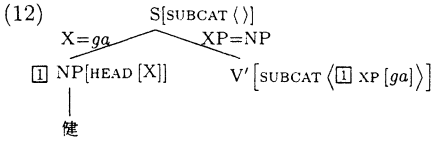
(11)



(11) では、格助詞ガの HEAD 素性の一部である CASE 素性の値 *ga* が、同じく PP の HEAD 素性の一部である CASE 素性に主辞素性原理によって受け継がれている。この情報が、下位範疇化素性原理において、PP が動詞の SUBCAT 素性の中のガ格を持つ要素と単一化する際には制約として機能している。

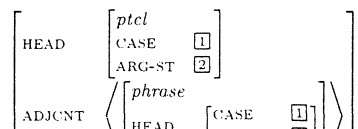
また格助詞が明示されるか否かの任意性は、名詞の CASE 素性の指定方法の差異によって捉えている。(12) にガ格が省略された場合の素性構造を示す。

(12)



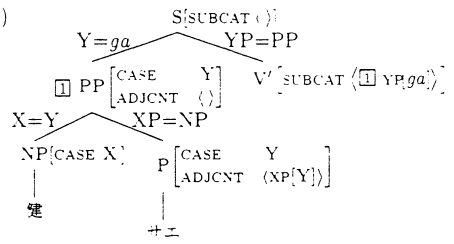
「健が奈緒美を誉めた」では、ガ格を伴うことで「健」は主語であることを明示するが、「健サエ奈緒美を誉めた」でも、それは主語である。このことからガ格を伴うことで明示していた文法機能は、(13) に示す取り立て助詞サエにより、形態的に表出していないと考えられる。

(13)



また、サエ自身はヲ格などの代わりに用いることができるので、動詞はサエに関係なく主語の助詞句を下位範疇化していると考えられる。

(14)



(14) ではサエの CASE 素性が変数 Y になっており、PP 全体としても Y の値は指定されない。この状態で、下位範疇化原理に従って動詞の SUBCAT 素性のガ格を持つ要素と PP が単一化すれば、Y の値が *ga* となり助詞句はガ格句と同じ素性の指定を持つ句として解析される。

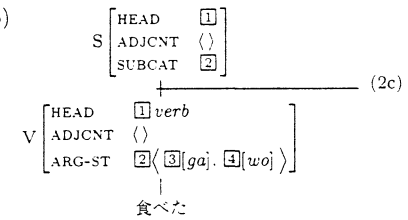
2.4 頻出構文に関するスキーマ

日本語では、いわゆるゼロ代名詞とよばれる、項の省略が頻繁に起こる。動詞が要求する項が一部表出していなくても句となり得るし、極端な場合、項が全く表出していなくても、その動詞は句または文と成り得る。

(15) a. 健がケーキを食べた。　b. 健が ∅ 食べた。
c. ∅ ケーキを食べた。　d. ∅ ∅ 食べた。

NAIST JPSG では、I. 部分的に飽和した動詞句がごく自然に現れる、II. 仮に全ての項が表出しても、それらの間に任意の付加語が入り得る、という二点を考慮し (2c) を導入した。(2c) に続けて (2a) を再帰的に適用していけば任意の個数の項を打ち消すことができる。(16) はそのような解析を行なった (15d) の構文木である。

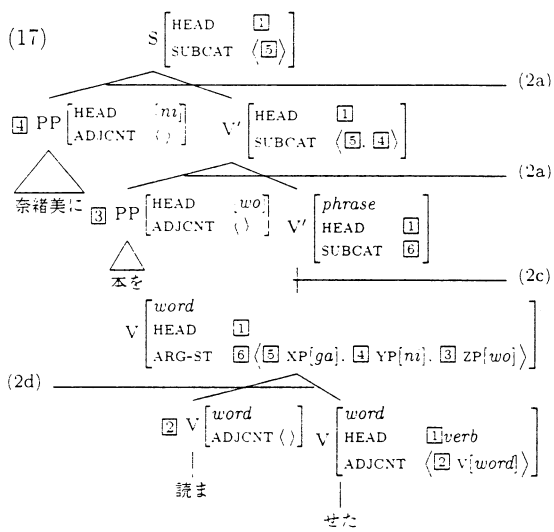
(16)



「食べた」の HEAD 素性が主辞素性原理によって S の HEAD 素性に受け継がれ、また ARG-ST 素性は項顕在化原理によって S の SUBCAT 素性に変換されている。

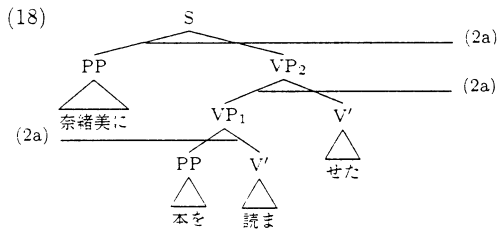
格助詞欠落と並ぶ特徴的な現象としては、動詞に対する助動詞の生産的な付加があげられる。「奈緒美と違って健は絵本を 読ませられたがらなかった。」のような文を扱うために (2d) を導入した。例えば使役構文「健が奈緒美に本を読ませた」における複合動詞「読ませた」は (17) のように分析される。

(17)



(2d) を導入することの利点は、助動詞の ADJCNT 素性の指定で左にくる姉妹を *word* に制限することで、解析における複合動詞句の曖昧性を抑制できる点である。例えば、(17) の助動詞「せた」の ADJCNT 素性が指定する要素を *word* に制限しないと (2a) のスキーマが適用され、(18) のようにも分析できるようになる。

(18)



しかし、この分析では「奈緒美に」と「本を」が語順転換できることが説明できず、また「奈緒美に」と「本を」の間に付加語が挿入されると VP_1 と VP_2 のどちらに付加されるかで曖昧性が組み合わせ的に増大する。

それに対し、(17) の分析を採用すると、いくつかの語順転換やゼロ主語などの省略を構文木の情報を反映した形で直接的に扱えるという利点がある。ただし、構文木と意味素性の構造は必ずしも一致しないが、先に述べたように NAIST JPSG では音韻・形態情報は構文木に、階層関係などの統語情報は SEM 素性にそれぞれ反映しているので、特にこのことを問題としない。

3 統計的係り受け情報の利用

スキーマだけでは任意の二分木を解析木として許してしまうため、全ての解析木を生成した後で原理に違反していないか否かを調べるのは効率が悪い。そこで、文節同士の係り易さに関する統計情報を利用して「もっともらしい」解析木から調べる方法が考えられる。これには多

くの場合、上位 n 個の候補だけを保持するビーム探索が用いられるが、原理に違反しない解析木がその n 個の中にない場合は解析に失敗してしまう。本研究では CKY アルゴリズムを改良し、(潜在的には全ての解析木 (係り受け) を候補として出力するが) 確率が高い解析木から順に出力する係り受け解析アルゴリズムを提案する。

3.1 統計的係り受け情報の問題点

(19) のような、構造的な曖昧性をもった文を考える。

(19) a. 自転車で 通り過ぎた奈緒美に声をかけた。
b. ...健の 妹の就職...

(19a) には、「自転車で通り過ぎた」と「*自転車で声をかけた」の二つの構造的な曖昧性があるが、後者は意味的な制約で排除される。このような制約は現在の統計的係り受け解析においても語の共起頻度によってモデルに取り込めるが、(19b) は同様に扱えない。「健の妹」も「健の就職」も単独では意味的な制約には違反していないが、後者の場合は「妹の」が「就職」に係るので、「健」と「妹」の「就職」に対する意味役割が衝突する。このことから*(健の (妹の 就職)) という構造は排除できるが、意味役割の衝突といった制約は二文節間の単純な共起では捉えられない。もちろん、このような制約をより複雑な確率モデルで捉えることは原理的には可能かもしれないが、データの過疎性の問題から現実的ではない。

[1] はこのような問題を避けるため、部分解析および冗長解析という考え方を提案する。これは一文に対する完全な係り受け構造を計算するのではなく、ある文節の係り先の決定を保留したり、複数の候補を出力したりすることで適合率をあげる手法である。ただし、このような部分的な情報を統語構造を構成する時にどのように利用するかについては検討を要する [3]。

3.2 漸進的解析アルゴリズムの概略

四つの文節 (u_1, u_2, u_3, u_4) を含む入力文に対し、提案するアルゴリズムの実行過程を (20), (21) に示す。

(20)

$(((u_1u_2)u_3)u_4) \leftarrow$ $(u_1(u_2(u_3u_4)))$	$(u_2(u_3u_4)) \leftarrow$ $((u_2u_3)u_4)$	(u_3u_4)	u_4
$((u_1u_2)u_3) \leftarrow$ $(u_1(u_2u_3))$	(u_2u_3)	u_3	
(u_1u_2)	u_2		
u_1			

(20) では、各係り受けが対応する句構造として表現している。表の左上のエントリを (1,4), 左下のエントリを (1,1) などとよぶことにすると、「(x, y) は u_x から u_y を覆う係り受けを格納している」ということができる。アルゴリズムは最初に、確率最大の係り受けを bottom-up に計算する。この段階では各エントリにおいて第一位の係り受けしか計算しない。(20) の (1,4) が二つしか係り受けを含まないのはそのためである。

次の段階では、第二位の係り受けを計算するために、第一位の係り受けの各部分をより確率の低い別の候補に置き換える。ここで (1,4) にある $(u_1(u_2(u_3u_4)))$ という係り受けは必ずしも第二位の係り受けであるとは限らないことに注意されたい。

(21)

$(((u_1u_2)u_3)u_4) \leftarrow$ $((u_1(u_2u_3))u_4) \Leftarrow$ $(u_1(u_2(u_3u_4)))$	$(u_2(u_3u_4)) \leftarrow$ $((u_2u_3)u_4)$	(u_3u_4)	u_4
$((u_1u_2)u_3) \leftarrow$ $(u_1(u_2u_3)) \Leftarrow$	(u_2u_3)	u_3	
(u_1u_2)	u_2		
u_1			

(21) では第一位の係り受けの前半部分 $((u_1u_2)u_3)$ を次に確率の高い (それは (1,3) を見ればわかる) $(u_1(u_2u_3))$ に置き換えている。このようにしてできた $((u_1(u_2u_3))u_4)$ と最初の段階で作った $(u_1(u_2(u_3u_4)))$ の確率を比べて真の第二位の係り受けを得る。

上の説明では第一位の係り受けの前半部分しか置き換えなかったが、一般には後半部分の置き換えも必要である。さらに、置き換えるべき部分の第二位は上のアルゴリズムを再帰的に適用して求める。

4 おわりに

NAIST JPSG では、日本語の特徴的な言語現象を局所的な制約として記述するように文法の原理、スキーマおよび素性を設計した。また、コーパス調査の知見も考慮されているが、特に係り受け情報を参照した漸進的解析アルゴリズムの実装は、連体修飾の係り関係の特定に起因する処理の曖昧性を低減するための一助となった。

参考文献

[1] Masakazu Fujio and Yuji Matsumoto. Japanese dependency structure analysis based on lexicalized statistics. In *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing*, pp. 88–96, Granada, Spain, Jun 1998.

[2] Takao Gunji. On lexicalist treatments of Japanese causatives. In Robert D. Levine and Georgia M. Green, editors, *Studies in Contemporary Phrase Structure Grammar*, chapter 3, pp. 119–160. Cambridge University Press, Dordrecht, 1999.

[3] 今一修, 松本裕治, 藤尾正和. 統計情報と文法制約を統合した統語解析手法. 自然言語処理, Vol. 5, No. 3, pp. 67–83, Jul 1998.

[4] Carl J. Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago, 1994.

[5] Ivan A. Sag and Thomas Wasow. *Syntactic Theory: A Formal Introduction*, Vol. 92 of *CSLI Lecture Notes Series*. CSLI Publications, Stanford, California, 1999.