日本語の派生文法と2レベル規則

三浦 睦美　吉村 賢治　首藤 公昭
福岡大学 工学部

1 はじめに

従来日本語の機械処理において一般に用いられてきた学校文法の活用形は、その整理と分類が五十音図に基づいて行われており、そのことが様々な問題を生じさせる原因となっている [6]。そこで近年、日本語文の形態素解析に用いる文法として、Bloch[2] を源流とする音韻論的分析に基づく文法が取り上げられている。この文法では、学校文法の活用形とは異なり、語幹と接辞の接続により得られた形態を活用形と呼ぶ。ここでは、この種の文法を学校文法と区別して派生文法と呼ぶことにする。これまでに派生文法を用いた形態素解析に関するいくつかの報告があるが、それらの大半は派生文法の特徴の一つである音韻処理を避けている。

音韻処理を行わなければ形態素解析の手続きがその分簡単になるという長所があるが、反面、辞書に動詞などの異形態を登録することが必要になる。一つの単語に対して複数個の異形態が存在しても、解析の場合には特別な処理を必要としないが、生成の場合には異形態から一つを選択する処理が必要になる。

我々は、解析だけでなく生成も行うことができる形態素処理システムの開発を進めている [4]。そこでは音韻処理を前提とする派生文法を、音韻処理を避けずに利用することが望ましい。音韻処理を行う形態素処理のモデルに2レベルモデルがある。2レベルモデルは、生成音韻論で用いられる音韻規則に似た音韻規則を用いるが、生成音韻論とは異なり、表層レベルと語彙レベルの二つのレベルだけを使って規則を記述する点に特徴がある。2レベルモデルで用いられる音韻規則を2レベル規則と呼ぶ。2レベル規則は、表層レベルの形態と語彙レベルの形態の対応関係だけを記述するため規則間の依存関係がなく、各規則を並列に動作する有限状態オートマトンとして表現し、効率的な処理系を実現することができる [3]。

本稿では、まず日本語の派生文法についてその概略を説明し、これまでに日本語文の形態素解析に用いられてきた派生文法の取り扱いを概観した後、筆者らが作成した日本語の2レベル規則について報告する。

2 派生文法

学校文法で活用として扱っている現象を、派生文法は連結母音や連結子音の脱落や内的連声という考え方で説明する。その結果、語彙項目の表記はローマ字となる。語彙項目の表記をローマ字とすることに対しては、一般的なテキストを解析する場合にテキスト中の平仮名をローマ字に変換する前処理が必要になる、平仮名をローマ字に変換した結果、単語辞書の検索開始位置が増加する可能性があるなどの短所が指摘されている [5]。しかし、音便などの規則が簡潔に記述でき、音声認識などでの利用を考えた場合にもローマ字書きの方が適用しやすく、文生成の処理が簡潔になるなどの長所も持つ。筆者らは、生成と解析の両方向性や有限状態トランスデューサを用いた音韻規則の処理に特徴を持つ KIMMO[4] との親和性を考慮して派生文法に基づく機械処理のための文法を作成している。

2.1 品詞体系

派生文法と学校文法の最大の相違点は活用の捉え方である。表2のように、学校文法では動詞や形容詞などの活用語は語幹と活用語尾から成り、それらが例えば「書か」は五段活用動詞の未然形であるという説明をする。一方、派生文法では動詞は「書k」までであり、動詞には様々な接尾辞が接続して、学校文法であれば助動詞に分類される部分までも含んだものが活用形となる。以上のように、派生文法における活用形とは学校文法のものとは異なる概念であり、それを活用形と呼ぶことは誤解を生じるかも知れないが、現在のところは特に名称を改めていない。派生文法における活用形も、接続する接尾辞によって様々な種類に分類されるが、現在のところそれらに名称を与えることはしていない。ただし将来的には、学校文法の活用形のように意味を表現したものと文法上の機能を表現し

表 1: 派生文法の品詞体系

派生文法	例	学校文法
動詞	*kak*(書く)	動詞
形容詞	*taka*(高い)	形容詞
名詞	本	名詞
形式名詞	こと	形式名詞
形容動詞	*kirei*(きれいだ)	形容動詞
連体詞	この	連体詞
副詞	時々	副詞
接続詞	しかし	接続詞
繋辞	*da*(だ),*desu*(です)	
文法接尾辞	*ita*(た)	助動詞
派生接尾辞	*rare*(られる)	
格接尾辞	が、を	格助詞
副助辞	さえ	副助詞
提題助辞	は	
接続助辞	けれども	接続助詞
終助辞	ね	終助詞
接頭辞	未	接頭語

たものとを混在させた名称ではないものを確立したいと考えている。表3には、表2にほぼ対応する動詞と接尾辞の組み合わせを示した。

表1に派生文法における品詞体系を示す。動詞は例の *kak*（書く）のような子音動詞、*mi*（見る）のような母音動詞、「する」、「来る」のような特殊動詞に下位分類される。動詞には、文法接尾辞、派生接尾辞などが接続する。文法接尾辞は、動詞に接続して文節を形成することができる。派生接尾辞は、動詞に接続して二次的な意味を添えるものである。文法接尾辞は派生接尾辞にも接続する。例えば、「書 kaserareta」は動詞「書 k」、派生接尾辞 *sase* と *rare*、文法接尾辞 *ita* の接続の結果、*sase* の *s*、*ita* の *i* が脱落して得られた形である。文法接尾辞、派生接尾辞、及び名詞に接続する繋辞を合わせたものが概ね学校文法における助動詞に相当する。

表 2: 学校文法の活用

動詞		活用形
語幹	活用語尾	
書	か	未然
	こ	
	き	連用
	く	終止
	く	連体
	け	仮定
	け	命令

表 3: 派生文法の活用

動詞	接尾辞	活用形
書 k	*anai*	書 *kanai*
	you	書 *k(y)ou*
	imas	書 *kimas*
	ru	書 *k(r)u*
	reba	書 *k(r)eba*
	e	書 *ke*

2.2 音韻規則

前節でも見たように、形態素の接続の際、音韻変化が生じることがある。音韻変化は音韻規則で記述する。ここで、音韻規則は次のような形式で記述する。

$$CPS \;\; op \;\; LC \;\; \_\_ \;\; RC$$

CPS： 変化する文字対のリスト

op： op には右辺の環境で左辺の変化が必ず起きることを示す⇔(義務的規則) と右辺の環境で左辺の変化が起きてもよいことを示す⇒(選択的規則) がある。

LC： LCにはCPSを条件づける文頭側の文字対のリストを記述する。

RC： RCにはCPSを条件づける文末側の文字対のリストを記述する。

ここで、文字対とは語彙レベルの文字 L と表層レベルの文字 Sの対であり、$L:S$の形で表す。特に L と Sが等しい場合には、省略して L と表す。また、形態素の境界を+、空記号を 0 で表す。

基本的な音韻規則として、接続の結果子音の連続又は母音の連続が起こった場合後ろの子音又は母音が脱落するという規則がある（表4)。 ここで、C_1は子

表 4: 基本的な音韻規則

規則	例
$[+:0, C_2:0] \Leftrightarrow [C_1] \_\_ []$	*kak+sase+ru* → *kak_aseru*
$[+:0, V_2:0] \Leftrightarrow [V_1] \_\_ []$	*mi+ana+i* → *mi_nai*

音 $\{k, g, s, t, n, b, m, r, w\}$、$C_2$は子音 $\{r, s, y\}$、V_1は母音 $\{i, e\}$、V_2は母音 $\{a, i, u\}$ を表している。例えば、動詞「書 k」に非完了態の文法接尾辞 ru が接続した

場合、後ろの子音rが脱落して表層側では「書ku」という文字列が得られる。ただし、wに関してはさらに次の規則も立てなければならない。

$$[w:0,+:0] \Leftrightarrow [] \_\_ [V_3] \quad (1)$$

ここで、V_3は$\{i,u,e,o\}$を表す。規則(1)は、表4の規則を適用した後にも働く。例えば「会w」にruが接続して「会u」の形を得るのは、まず表4の一行目の規則が適用されて$aw+u$になった後、規則(1)が適用されてwが脱落した結果である。このように、実際には規則が多段階に適用されて表層文字列を得るのだが、KIMMOでは音韻規則を2レベルモデルで表現するのでwについては表4と規則(1)を一つの規則として記述しておく必要がある。従って、wは表5のような音韻規則を持つ。 表6は、音便に関する音

表5: wの2レベル規則

規則	例
$[+:0,C_2:0] \Leftrightarrow [w] \_\_ [a]$	aw_aseru
$[w:0,+:0,C_2:0] \Leftrightarrow [] \_\_ [V_3]$	$aw+ru \rightarrow a\_u$
$[w:0,+:0] \Leftrightarrow [] \_\_ [V_3]$	$aw+e \rightarrow a\_e$

韻変化の例である。音便は、動詞に文法接尾辞itaやiteが接続する際に起こる現象である。例えば、動詞「書k」にitaが接続する時にはkが脱落して「書ita」という形を得る。 表7は、連声に関する音韻変化

表6: 音便に関する音韻変化

音便	例
イ音便	$kak+ita \rightarrow ka\_ita$(書いた)
	$oyog+ita \rightarrow oyo\_ida$(泳いだ)
はつ音便	$sin+ita \rightarrow sin\_da$(死んだ)
	$nom+ita \rightarrow non\_da$(飲んだ)
	$tob+ita \rightarrow ton\_da$(飛んだ)
促音便	$tat+ita \rightarrow tat\_ta$(立った)
	$kir+ita \rightarrow kit\_ta$(切った)
	$kaw+ita \rightarrow kat\_ta$(買った)
	$ik+ita \rightarrow it\_ta$(行った)

の例である。連声は、学校文法における形容詞の連用形に補助動詞「ござる」などが接続した時に起こる現象である。一番上の例では、形状動詞「寒」に文法接尾辞ku、派生接尾辞$gozaimas$、文法接尾辞ruが接続した結果、「寒$ugozaimasu$」という形になる。これらの音韻変化も、wの規則と同様に2レベル規則で記述することができる。

表7: 連声に関する音韻変化

samu+ku+gozaimas+ru → samu_ugozaimas_u
samisi+ku+gozaimas+ru → samisyu_ugozaimas_u
tiisa+ku+gozaimas+ru → tiiso_ugozaimas_u
hiro+ku+gozaimas+ru → hiro_ugozaimas_u

3 従来の研究

日本語文の形態素解析に派生文法を用いた研究には、久光[5]、渕[9]、西野[8]、小川[7]、Alam[1]などがある。

久光は、子音動詞の効率的な処理についての考察を行っている。その方法は、音韻論的分析を基本としながら漢字仮名混じり文に直接適用するというものである。そのため、子音動詞語幹の末尾子音を屈折接辞の先頭に付加している。

渕は、派生文法を基本として、日本語の文字の単位（漢字仮名混じり文）で解析することを目的とした工夫を行っている。それは、「動詞語幹の末尾の子音」と「動詞語幹に接尾する接尾辞の始まり部分」を組み合わせて動詞の活用語尾とするというものである。

また、西野も派生文法を基本として、縮退した形の接尾辞を辞書に登録することで音韻処理を不必要にするという方法を提案している。例えば、我々の使用する派生文法における非完了態の文法接尾辞ruに相当するものとして、ここではruとuの二つの形態を辞書に登録している。また、音便処理については、語幹末子音を変形した語を辞書に登録しておくという手法を取っている。例えば、kikに対してkiを、itaに対してta、ida、daを異形態として登録している。同様に、特殊動詞についても複数個の異形態を登録している。

小川は、西野と同様に縮退した形の接尾辞を辞書に登録しておくことにより音韻処理を不要にしている。ただし、動詞については西野と異なり、異形態を辞書に登録せずに文末側から接辞を探索して省略された動詞末尾の子音を表から補完する方法を取っている。

Alamは、Blochの形態音韻論的な規則を2レベル規則で記述して、KIMMOで生成と解析を行っている。音便処理は2レベル規則で記述し、音韻処理を行っている。特殊動詞については、複数個の異形態を立てている。

4 不規則変化の取り扱い

特殊動詞「来る」、「する」については従来の研究のほとんどが複数個の異形態を立てることにより不規則な変化を記述している。しかし、先に述べたように、生成の処理を考慮すると特殊動詞についても単一の形態を定めたい。そこで、従来使われている特殊動詞の各形態について、規則の数などを検討した。ここでは、$V+V$で右のVが脱落しないこともあるので2章で示した一般的な音韻規則は用いることができない。そのため、実際には品詞に対する制約などを音韻規則に加える必要がある[4]。

まず、「来る」についてはko、ki、ku、kなどが語彙項目の候補として考えられる。koの場合、音韻規則として以下のものが必要となる。

$$[+:0, a:0] \Leftrightarrow [k, o] \_\_ [\{z, n\}]$$

$$[o:0, +:0] \Leftrightarrow [k] \_\_ [i]$$

$$[o:0, +:0] \Leftrightarrow [k] \_\_ [u, m, a, i]$$

$$[o:u, +:0] \Leftrightarrow [k] \_\_ [r, \{u, e\}]$$

同様に、kiを選択した場合には7個、kuの場合には6個、kでは5個の規則が必要となる。そこで、我々は規則の数が最も少くなるkoを語彙項目とした。

「する」については、se、si、su、sなどが語彙項目として考えられる。「来る」と同様に音韻規則とその数を調べた結果、seを選択した場合には10個の規則が必要になり、siでは9個、suでは11個、sでは10個になった。規則の数が最少となるのはsiの場合であるが、これにはsiから命令の意味を表す活用形$seyo$を生成するために次のような規則(2)が含まれている。

$$[i:e, +:0] \Leftrightarrow [s] \_\_ [y, o] \qquad (2)$$

この規則は、siにyouが接続して意志を表す活用形$siyou$の生成を妨げる。sについても同様の問題が生じる。一方seを語彙項目とした場合、命令を意味する$seyo$についてはseとyoの接続は音韻変化を伴わないので規則を記述する必要がなく、意志の$siyou$のために次の規則がある。

$$[e:i, +:0] \Leftrightarrow [s] \_\_ [y, o, u]$$

この場合、命令の接尾辞yoの後ろにuで始まる形態素が接続するとs、siのときに類似した問題が生じるが、そのようなことは稀であると考えられる。そこで、「する」の語彙項目としては、seを選択した。

5 おわりに

日本語の派生文法に基づく2レベルの音韻規則について報告した。今回作成した規則は、「来る」や「する」などの特殊動詞に対しても異形態を必要としない点に特徴がある。今後は、この音韻規則と組み合わせて用いる構文規則を記述していく予定である。

参考文献

[1] Y. Sasaki Alam. A two-level morphological analysis of japanese. *Texas Linguistic Forum*, 22:229–252, 1983.

[2] B. Bloch. Studies in colloquial japanese, part i, inflection. *Journal of the American Oriental Society*, 66, 1946.

[3] L. Karttunen. Kimmo: A general morphological processor. *Texas Linguistic Forum*, 22:165–186, 1983.

[4] 吉村賢治, 三浦睦美, 首藤公昭. 2レベルモデルに基づく日本語の形態素処理. 言語処理学会第*3*回年次大会講演論文集, 1997.

[5] 久光徹, 新田義彦. 日本語形態素解析における効率的な動詞活用処理. 情報処理学会研究会報告 *94-NL-103-1*, 1994.

[6] 寺村秀夫. 日本語のシンタクスと意味 *II*. くろしお出版, 1984.

[7] 小川泰弘, 稲垣康善, ムフタル・マフスット. 派生文法に基づく日本語動詞接尾辞の形態素解析. 情報処理学会研究会報告 *96-NL-116-2*, 1996.

[8] 西野博二, 鷲北賢, 石井直子. 派生文法による日本語構文解析. 情報処理学会研究会報告 *92-NL-87-6*, 1992.

[9] 渕武志, 米澤明憲. 日本語形態素解析システムのための形態素文法. 自然言語処理, 2(4), 1995.