派生文法による日本語形態素解析

小川泰弘[†] ムフタル マフスット[††]
外山勝彦[†] 稲垣康善[†]

本論文では，派生文法に基づく簡潔な形態素文法を持つ日本語形態素解析システムMAJOを提案する．従来の多くの日本語形態素解析法においては，用言の活用を扱うための複雑な処理が必要であった．一方，派生文法の考え方に従うと，活用という概念を用いないで，形態素文法を作ることができる．派生文法を利用した日本語形態素解析システム開発の報告はすでにいくつかあるが，いずれも既存のシステムを利用して実現されており，そのために派生文法をそのまま用いることができず，変更が加えられている．それに対してMAJOは，派生文法に合わせてシステムを構築したものであり，文法規則の数が少ない簡潔なシステムとなっている．コーパス1,000文に対する形態素解析実験によって解析精度を測定した結果，98.1%の形態素に対して正しく品詞を付与できた．

Derivational Grammar Approach to Morphological Analysis of Japanese Sentences

YASUHIRO OGAWA,[†] MUHTAR MAHSUT,[††] KATSUHIKO TOYAMA[†] and YASUYOSHI INAGAKI[†]

In this paper, we propose a new simple Japanese morphological analysis system named MAJO. The systems which have been proposed so far are quite complicated in order to deal with conjugations of verbs and adjectives. On the other hand, it is known that the derivational grammar, which does not use the concept of conjugations, makes morphological analysis simple. Although there are some literatures on the application of the derivational grammar to Japanese morphological analysis, all of them has modified their grammar to utilize existing systems. MAJO is constructed directly from the derivational grammar, so that it has less grammar rules than the previous systems. We have evaluated its performance on the EDR corpus and succeeded with 98.1% ratio.

1. はじめに

日本語の形態素解析処理の実現において，考慮すべき重要な点は，形態素文法の構成，特に用言の語形変化の取扱いである．

従来の多くの文法では，活用という考え方で用言の語形変化を説明している．たとえば，学校文法では，「書カ」はカ行五段活用動詞「書ク」の未然形とされ，否定の助動詞「ナイ」や使役の助動詞「セル」が接続するとされる．動詞は活用の仕方によって，五段活用動詞，一段活用動詞の2種類に分類され，五段活用動詞は，さらにカ行，ガ行，サ行，タ行，ナ行，バ行，マ行，ラ行，ワ行の8つの活用型に細分される．そのため，学校文法に基づいて形態素解析処理システムを構築した場合，それらの活用の処理が必要となる．とくに，助動詞はその活用が複雑であるため，体系的に取り扱うことが難しくなっている．

そこで近年，形態素解析に使用する文法として，Bloch[1]を源流とする音韻論的分析に基づく文法が注目されている．Blochは日本語の動詞の変形を音韻単位で分析し，語幹と接尾辞の接続で説明している．とくに，動詞語幹の末尾が子音であるか，母音であるかによって，接続する接尾辞が決まることを示しており，その結果，動詞は子音動詞と母音動詞の2つに分類され，活用型による動詞の細分は不要とされている．また，Blochの文法を発展させた寺村[2]は，助動詞も接尾辞の一種としてとらえ，子音動詞に接続する接尾辞と母音動詞に接続する接尾辞を，互いに異形態の関係にあるものとしてとらえている．

一方，久光ら[3]は，音韻論的手法の流れの中で，後

† 名古屋大学大学院工学研究科情報工学専攻
Graduate School of Engineering, Nagoya University
†† 三重大学工学部情報工学科
Faculty of Engineering, Mie University

することによって，形態素解析システムを実現している．その結果，異形態の登録を行うことなく形態素解析が可能になっているが，文法規則とは別に，複雑な音韻規則とそれを扱うシステムが必要となっている．

以上のように，文献 6)〜9) では派生文法に基づいた日本語処理について提案しているが，その実現は既存の解析システムを利用しているため，派生文法の簡潔さを活かしきれていない面がある．

それに対して，本論文では，派生文法を直接的に利用した簡潔な日本語形態素解析システムの構築について述べる．ここで簡潔さは，システムの簡潔さと文法の簡潔さに分けられるが，一般に両者の間はトレードオフがある．たとえば，動詞の語幹，語尾，異形態のすべてを辞書に登録すると，用言に対する活用処理が不要となり，システムが簡潔になるが，その代わりに形態素文法が複雑になり，また，辞書のサイズも大きくなる．一方，音便や不規則動詞などの音韻変化に対して例外処理を導入すると，異形態の登録が不要となり，辞書のサイズが小さくなるが，その分だけシステムが複雑になる．

そこで，本論文では派生文法を用いて，用言の活用処理を行わないですみ，かつ，形態素文法も簡潔になる形態素解析システムを提案する．本論文で提案する手法では，派生文法を音韻論的手法のまま利用するが，音便形に対しては，少数の異形態の登録および音素の補完という例外処理の導入により，動詞の異形態の登録を原則的に不要としている．例外処理の分だけシステムが複雑となるが，従来の手法と比較して文法規則が簡潔なものとなった．この手法を計算機上に実現した形態素解析システム MAJO は，EDR コーパスを用いた実験で，98.1%の形態素に正しい品詞を付与できた．これは，従来の形態素解析システムに匹敵する解析精度である．

本論文は，形態素解析システム MAJO の構築について，その特徴である動詞部分の解析を中心に述べる．以下では，次の 2 章で派生文法について紹介し，3 章では，派生文法を実際に形態素解析に適用する際の問題点の指摘と，MAJO におけるそれらの解決方法の提示を行う．4 章では，MAJO の特徴を既存の形態素解析システムと比較して述べ，5 章では，EDR コーパスを用いた性能評価について述べる．6 章は本論文のまとめである．

2. 派生文法

日本語は言語の分類において膠着語であるとされている．膠着語とは，文法的機能を表す接辞が実質的観念を表す語幹に結合することによって単語を形成するという性質を持つ言語の総称である．それに対して，英語などのように，文法的機能が語幹の交替によって表される言語は屈折語と呼ばれる☆．

表 1 動詞幹と動詞接尾辞の接続例
Table 1 Connections between a verbal stem and suffixes

活用形	子音幹の例	母音幹の例	接尾辞
未然形	kak-ana-i	tabe-na-i	-(a)na-i
	kak-are-ru	tabe-rare-ru	-(r)are-(r)u
	kak-ase-ru	tabe-sase-ru	-(s)ase-(r)u
	kak-ou	tabe-you	-(y)ou
連用形	kak-imas-u	tabe-mas-u	-(i)mas-(r)u
	kaφ-ita	tabe-ta	-(i)ta
終止形	kak-u	tabe-ru	-(r)u
連体形	kak-u	tabe-ru	-(r)u
仮定形	kak-eba	tabe-reba	-(r)eba
命令形	kak-e	tabe-ro	-e / -ro,-yo
	kak-una	tabe-runa	-(r)una

活用は屈折の一種であるから，日本語の用言が活用するということは，日本語は膠着語であるにもかかわらず用言に関しては屈折語的特徴を持つ言語である，ということになる．それに対して，活用を前提とせず，日本語の用言の変形も，語幹への接辞の接続という形で記述可能とする派生文法[4),5)]が提案されている．派生文法による日本語の記述を，その特徴が表れる動詞句の形成部分を中心に以下にまとめる．

2.1 連結子音と連結母音

動詞の不変化部分を語幹と呼ぶ．一段活用動詞「見ル」「食ベル」の場合は，不変化の部分「見」「食ベ」が語幹である．すべての一段活用動詞の語幹はその末尾が母音 i か母音 e であるので，母音幹と呼ぶ．一方，五段活用動詞「書ク」“kak-u” の場合は，音韻論的に考えれば，“kak” が不変化部分，すなわち語幹である．また，「買ウ」“ka-u” のようなワ行五段活用動詞の場合には “kaw” を語幹とする．すべての五段活用動詞の語幹は，その末尾が子音であるので子音幹と呼ぶ．また，母音幹と子音幹を合わせて動詞幹と呼び，動詞幹に接続する接尾辞を動詞接尾辞と呼ぶ．

派生文法においては，動詞の変形は動詞幹にいくつかの接尾辞が接続したものとして考える．そのため学校文法におけるほとんどの活用語尾や助詞，助動詞を接尾辞として扱う．それらと学校文法における活用の形との対応を**表 1** に示す．

ここで，終止形 “tabe-ru”，“kak-u” の接尾辞はそれぞれ “-ru”，“-u” である．派生文法ではそれらをまとめて “-(r)u” と表記する．子音 r の有無は語幹の末

☆ その他に孤立語，輯合語などの区分がある．

尾に依存して決まる．たとえば，語幹 “kak-” に接尾辞 “-ru” が接続した場合，子音の連続を避けるために，接尾辞の先頭の r が欠落する．そのような子音を連結子音と呼ぶ．

一方，派生文法では，表 1 に示すように，「書カナイ」「食ベナイ」をそれぞれ “kak-ana-i”，“tabe-na-i” と解析する．ここで，否定を表す接尾辞はまとめて “-(a)na-” と表す．先頭の母音 a は母音が連続する場合に欠落し，そのような母音を連結母音と呼ぶ．

以上から，派生文法においては，動詞幹と動詞接尾辞の接続は以下の 2 つの規則にまとめられる．

接続規則 1： 連結子音を持つ動詞接尾辞が子音幹に接続する場合，連結子音を削除する．

接続規則 2： 連結母音を持つ動詞接尾辞が母音幹に接続する場合，連結母音を削除する．

2.2 統語接尾辞と派生接尾辞

前節の否定の接尾辞 “-(a)na-” は，「書カナカッタ」“kak-ana-katta” のように，後方に他の接尾辞が接続可能である．このことは，動詞語幹に接尾辞が接続することによって，新たな語幹が派生したと考えられる．そのような接尾辞を派生接尾辞と呼ぶ．派生接尾辞には他に “-(r)are-”，“-(s)ase-”，“-(i)mas-”，“-(i)ta-” などがあり，それぞれ受身・可能・尊敬，使役，丁寧，希望の意義を表している．

派生接尾辞に対して，新たな語幹を派生しない接尾辞を統語接尾辞と呼ぶ．前節の “-(r)u” はその例である．統語接尾辞は動詞形の形成の役割を果たす．ここで動詞形とは終止形，連体形，連用形，命令形の 4 形のことである☆．動詞幹に複数の接尾辞が接続する場合には，統語接尾辞が最後に接続する．

2.3 音便形

完了の統語接尾辞 “-(i)ta” が子音幹に接続する場合は，接続規則 1 の例外となる．たとえば，「聞ク」“kik-u” の語幹 “kik-” に接尾辞 “-(i)ta” が接続する場合，接続規則 1 に従えば「聞キタ」“kik-ita” となるが，実際には末尾子音 k が欠落して「聞イタ」“ki-ita” となる．これは音便と呼ばれる特別な語形変化であり，s 以外の末尾子音を持つ子音幹において起こる．その変化は**表 2** に示すように，末尾子音に依存する．なお，表中の ϕ は零記号で，接続によって ϕ に対応する音素が欠落したことを表している．

動詞「行ク」“ik-u” の音便形は，この規則のさらに例外である．表 2 に従えば，末尾子音 k が欠落して「行イタ」“iϕ-ita” となるが，実際には末尾子音 k が t に変化し，「行ッタ」“it-ϕta” となる．

☆ 名称は同じであるが，学校文法の活用形とは異なる．

表 2 動詞の音便形
Table 2 Internal sandhi form of verbs.

語幹末子音		音便形		語例	備考
-k	-(i)ta	-ϕ	-ita	聞いた	イ音便
-g	-(i)ta	-ϕ	-ida	泳いだ	イ音便＋連濁
-r	-(i)ta			切った	
-t	-(i)ta	-t	-ϕta	立った	促音便
-w	-(i)ta			買った	
-b	-(i)ta			飛んだ	
-n	-(i)ta	-n’	-ϕda	死んだ	撥音便＋連濁
-m	-(i)ta			読んだ	
ik	-(i)ta	it	-ϕta	行った	例外
-s	-(i)ta	-s	-ita	貸した	音便変化なし

音便形を形成する接尾辞は-(i)t-で始まる統語接尾辞だけであり，“-(i)ta” のほかには “-(i)te”，“-(i)tara”，“-(i)temo” などがある．なお，希望の派生接尾辞 “-(i)ta-” は-(i)t-で始まるが音便形を形成しない．

2.4 形状動詞

派生文法においては，学校文法での形容詞は動詞の一種として扱われ，形状動詞と呼ばれる．形状動詞においても活用を考える必要はない．たとえば，「寒イ」“samui” は形状動詞幹 “samu-” に形状動詞接尾辞 “-i” が接続したものである．

動詞接尾辞と同様に，形状動詞接尾辞にも統語接尾辞と派生接尾辞が存在する．前述の “-i” は終止形を形成する統語接尾辞である．また，「寒ガル」“samu-gar-u” における “-gar-” は形状動詞幹に接続する派生接尾辞であり，子音幹を派生する．一方，否定の派生接尾辞 “-(a)na-” と希望の派生接尾辞 “-(i)ta-” は動詞幹に接続して形状動詞幹を派生する．たとえば，「書カナイ」“kak-ana-i” では，動詞 “kak-” の語幹に派生接尾辞 “-(a)na-” が接続することによって形状動詞幹が派生され，その後に形状動詞接尾辞 “-i” が接続している．

さらに，形状動詞においても音便変化を起こす場合がある．“gozar-” は形状動詞接尾辞 “-ku” の後に接続して丁寧の意義を表す補助動詞であるが，この場合はつねに**表 3** に示すような音便変化を起こす．たとえば，“taka-ku” に “gozar-” が接続する場合には “taka-ku” が “takou” と変化し，「タコウゴザル」“takou-gozar-u” となる．

2.5 不規則動詞

動詞語幹への接尾辞の接続規則には，若干の例外が存在する．学校文法におけるカ行変格活用動詞「来ル」とサ行変格活用動詞「スル」はそれぞれ例外の 1 つであり，不規則動詞と呼ばれる．派生文法では，不規則動詞は複数の語幹を持つ動詞として扱う．「来ル」は

表3 形状動詞の音便形
Table 3 Internal sandhi form of qualitative verbs.

末尾母音		音便形	語例		
−a + ku	→	ou	小さい	→	小そう
−i + ku	→	yuu	大きい	→	大きゅう
−u + ku	→	uu	明るい	→	明るう
−o + ku	→	ou	広い	→	広う

注：語幹末尾に母音eを持つ形状動詞は存在しない．

"ko-"，"k-"，"ku-" の3つ，「スル」は "se-"，"s-"，"si-"，"su-" の4つの語幹を持ち，以下の規則3.1〜3.4により接尾辞と接続する．

接続規則3.1： "-(i)ta" などの連結母音iを持つ接尾辞や，"-(u)mai" などの連結母音uを持つ接尾辞は，語幹 "k-"，"s-" に接続する．

接続規則3.2： "-(r)u"，"-(r)eba" など連結子音rを持つ接尾辞は，語幹 "ku-"，"su-" に接続する．

接続規則3.3： 否定の派生接尾辞 "-(a)na-" と命令形の派生接尾辞 "-ro" は，語幹 "si-" に接続する．

接続規則3.4： 上記以外の接尾辞は，語幹 "ko-"，"se-" に接続する．

この他にも不規則な変化をするいくつかの動詞がある．動作主に対する話者の敬意を示す尊敬語の動詞「仰ル」"ossyar-u" に "-(i)mas-" などのiで始まる接尾辞が接続するとき，語幹末尾のrが欠落し，「仰イマス」"ossyaϕ-imas-u" という形をとる場合がある☆．そのような動詞は，派生文法では変則動詞と呼ばれる．また，「アル」"ar-u" は子音幹動詞であるが，否定の接尾辞 "-(a)na-" は接続しない．さらに，前述の「行ク」も音便変化において不規則な変化をする動詞である．

3. 形態素解析システムMAJO

前章で述べた派生文法に基づき，日本語形態素解析システムMAJO（Morphological Analyzer of Japanese based On derivational grammar）を構築した．MAJOは派生文法が持つ簡潔さを活かすために，入力文を音韻単位で解析する．そのため，入力文として漢字仮名混じり文が与えられた場合，MAJOはその平仮名の部分を自動的に日本式ローマ字表記に変換して解析する．

派生文法は文生成を念頭においた文法であり，形態素解析に利用する際にはいくつかの問題点がある．以下では，形態素解析システムへ適用する際の問題点と，それに対してMAJOで採用する解決方法を示す．

☆ 音便変化を起こす "-(i)ta" の場合には音便変化が優先される．

表4 MAJOの接続行列（一部）
Table 4 Connectability matrix in MAJO.

	動詞	子音幹接尾辞	母音幹接尾辞	名詞	名詞接尾辞	格接尾辞
子音幹	-	5	-	-	-	-
母音幹	20	-	5	20	-	20
連体接尾辞	20	-	-	10	-	25
連用接尾辞	10	-	-	20	-	-
名詞幹	20	-	-	15	5	5
格接尾辞	10	-	-	10	-	30

3.1 MAJOの品詞と形態素文法

MAJOでは，品詞を左連接属性と右連接属性の組合せで表現する．左（右）連接属性は左（右）側に連接する形態素を規定する属性である．ここで，動詞接尾辞の左連接属性が動詞接尾辞であるとき，接続規則1および2に従って連結子音・連結母音を扱う必要がある．文生成の際にはそれらの規則を用いて簡単に動詞句を構成できるが，解析においてそのまま適用すると，欠落した音素の復元が必要となる．そこで，MAJOにおいては，接続規則をそのまま適用するのではなく，音素の欠落した異形態を辞書に登録する手法を採った．たとえば，連結子音を持つ統語接尾辞 "-(r)u" に対しては，"-ru" と "-u" を，連結母音をもつ派生接尾辞 "-(a)na-" に対しては，"-ana-" と "-na-" をそれぞれ辞書に登録した．また，それにともない，接尾辞の左連接属性を母音幹接尾辞と子音幹接尾辞とに細分した．たとえば，"-ru" と "-na-" の左連接属性は母音幹接尾辞であり，"-u" と "-ana-" の場合は子音幹接尾辞である．

以上の結果，MAJOは24種類の左連接属性と26種類の右連接属性を持ち，その組合せで表される品詞は59種類となった．MAJOで定義している品詞の一覧を，付録の**表9**に示す．

MAJOにおける形態素文法は，右連接属性と左連接属性との間の連接可能性で記述する．ただし，単に接続可能性を記述するだけでなく，接続コスト最小法[15),16)]を用いて接続可能性に順序を付ける（表4）．たとえば，表4の第1行第2列の数5は，子音幹という右連接属性を持つ品詞の直後に，子音幹接尾辞という左連接属性を持つ品詞が接続する場合のコストが5であることを示している．コストが小さいほど，2つの品詞は接続しやすい．また表4の中の'-'は接続が不可能であることを示している．

3.2 音便処理

MAJOでは動詞幹と接尾辞の接続は，活用処理を

行うことなく，接続行列に記された形態素間の接続の可否を調べるだけで解析できる．しかし，音便は語幹と接尾辞が2.3節で述べたような変化をするため，そのままでは解析できない．

従来の手法[7]では，たとえば，“書k-”，“買w-”に対してそれぞれ“書-”，“買t-”を異形態として登録していた．この手法は単純ではあるが，登録すべき異形態の数は膨大なものになる．

それに対して，MAJOでは，後方からの探索および音素の補完という手法を採用し，動詞の異形態を登録しないこととした．

異形態の登録が必要となるのは，音便変化をした動詞が辞書に登録された形と異なり，単純な辞書引きでは検索ができなくなるからである．しかし，入力文を後方から探索した場合，子音幹動詞の異形態より先に統語接尾辞の異形態が見つかる．たとえば，“-(i)ta”が接続した動詞句の場合，表2の音便形に示してある“-ita”，“-ida”，“-tta”，“-n'da”の4種類の異形態のいずれかが見つかる．そこで，MAJOではこれらの異形態を見つけたときに，失われたと思われる子音を**表5**に従って補う．この一連の処理の流れを**図1**に示す．“-tta”，“-n'da”に対しては補う子音の候補が複数あるが，その場合にはすべての候補を補い，辞書引きによって適当なものを選択する．

さらに，接尾辞の異形態を他の品詞と区別するため，

表5 動詞幹への子音の補完
Table 5 Supplementing a consonant into a verbal stem.

接尾辞	補う子音	解析例		
-ita	k	聞 ita	→	聞 k-ita
-ida	g	泳 ida	→	泳 g-ida
-tta	r	切 tta	→	切 r-tta
	t	立 tta	→	立 t-tta
	w	買 tta	→	買 w-tta
-n'da	b	飛 n'da	→	飛 b-n'da
	n	死 n'da	→	死 n-n'da
	m	読 n'da	→	読 m-n'da
-ita	ϕ	貸 sita	→	貸 s-ita

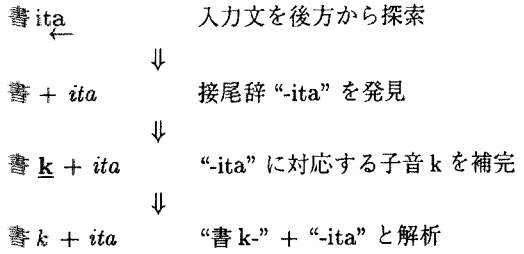


図1 MAJOにおける音便処理
Fig. 1 Treatment of internal sandhi by MAJO.

「イ音便」「促音便」「撥音便」「濁音便」という左連接属性を設け，それと「連体接尾辞」「連用接尾辞」という右連接属性の組合せで異形態用の品詞を作成した．これは表9における最後の8種類の品詞に相当し，たとえば，“-ita”の品詞は「イ音便/連体」となる．なお，音素補完処理後は，これらの左連接属性を「子音幹接尾辞」として扱うため，形態素文法では異形態のため品詞を考慮する必要はない．

この手法により，動詞の異形態の登録は不要となった．なお，接尾辞に関しては異形態の登録は依然として必要となるが，そのような接尾辞は-(i)tで始まる統語接尾辞に限られるため，登録する異形態は少数ですむ．

また，「行ク」“行 k-u”は音便形が特殊であるが，これは数が限られているため，異形態“行 t-”を子音幹動詞として辞書に登録することで対処した．接尾辞の異形態については，子音の補完処理のために左連接属性を用いてマークする必要があったが，異形態“行 t-”については，特別な処理は必要ないためマークの必要はなく，通常の形態素と同様に扱うことが可能である．MAJOでは，後述のように，他にも若干の異形態を辞書に登録するが，これらについても，特別な処理は必要ないので，他の形態素と区別する必要はない．

形状動詞の音便形についても，動詞の場合と同様に処理する．音便変化するのは補助動詞“gozar-”が接続する場合だけであるから，形状動詞幹末尾の母音，形状動詞統語接尾辞“-ku”，“gozar-”の3つをまとめた形を異形態とする．この異形態と補う母音の対応を**表6**に示す．このとき，異形態は形状動詞幹に接続し，子音幹を派生するので，表9からその品詞は形子派生接尾辞となる．

3.3 子音wの補完処理

派生文法に限らず，一般に動詞の語形変化を音韻論的に解析する場合，語幹末尾がwである子音幹動詞に対しては固有の問題を生じる．派生文法では動詞「カウ」の語幹は“kaw-”であるが，「カウ」を解析しようとした場合，そのローマ字表記は“kau”であり，動詞幹の末尾子音wが表記されない．そのため，そのまま

表6 形状動詞幹への母音の補完
Table 6 Supplementing a vowel into a qualitative verbal stem.

接尾辞	補う母音	解析例
-ougozar-	a	小 sa-ougozar-
-yuugozar-	i	大 ki-yuugozar-
-ugozar-	ϕ	明 ru-ugozar- 広-ugozar-

では解析ができないという問題が生じる．MAJOでは，音便処理を応用してこの問題を解決した．

すなわち，音便処理と同様に入力文を後方から探索し，左連接属性が子音幹接尾辞である形態素を発見し，かつ直前に母音がある場合に子音wを補完する．“ka-u”の例では，接尾辞“-u”の左連接属性が子音幹接尾辞であり，かつ，その直前が母音aであるため，子音wを補完することにより“kaw-u”と解析できる．また，MAJOでは漢字ローマ字混じり文を扱っているため，直前の語が漢字である場合（たとえば，直前の語が“買”である場合）もwを補完する．

3.4 不規則動詞の処理

2.5節で述べたように，不規則動詞も語幹の形が変化するため，そのままでは解析できない．それに対しては，不規則動詞の種類が限られているので，変化した語幹を異形態として辞書に登録して解決する方法が考えられる．しかし，“k-”，“s-”という子音1音だけの語幹を登録すると，解析パターンが増えて形態素解析は効率的に行えなくなる．そこで，MAJOでは，「来ル」に対しては“ko-”，“ku-”，“ki-”を，「スル」に対しては“se-”，“su-”，“si-”をそれぞれ語幹とした．その結果，たとえば“kita”は“k-ita”ではなく，“ki-ta”と解析される．これにより，不規則動詞の語幹末尾は母音となるから母音幹動詞として辞書に登録できる．このとき，不規則動詞と接尾辞との間の接続規則3.1～3.4が反映されなくなる．しかし，そのような規則が必要となるのは文生成の場合であり，解析の場合には規則に反する形は入力文に現れないと考えられるので，MAJOでは取り扱わなかった．

ところが，そのようにすると，たとえば“kinai”の解析の際に，“ki-”を母音幹動詞「着」ではなく，不規則動詞「来」と解析する可能性がある．この問題に対しては，不規則動詞に対して通常の動詞より大きなコストを付与し，解候補としての優先度を低くすることで対処した．

この他にも「行ク」や変則動詞が不規則な変化をする．「行ク」は音便変化が不規則であり，変則動詞はiで始まる接尾辞が接続する場合に，末尾のrが欠落するという点で不規則である．しかし，EDR日本語単語辞書において「行ク」と同種の動詞は14個，変則動詞は18個しか登録されていない．そこで，それらの動詞に対しては，異形態を辞書に登録することで対処した．なお「アル」も否定の接尾辞“-ana-”が接続することがないという点で不規則であるが，解析時にはそのような入力は現れないと考え，単なる子音幹動詞として辞書に登録した．

4. MAJOの特徴

本章では，従来の形態素解析システムと比較しながらMAJOの特徴について述べる．MAJOには以下に示す(1)～(4)の特徴がある．

(1) 形態素文法が簡潔である．
(2) 辞書登録する単語数を削減することができる．
(3) 口語的表現への対処が容易である．
(4) ローマ字表記を用いることで解析効率が低下する．

4.1 文法規則の簡潔さ

形態素文法の簡潔さについて，よく参照される形態素解析システムJUMAN[10]とMAJOとを比較する．JUMANは文献14)に基づいて形態素文法を作成しているが，文献14)では動詞を活用形によって分類しているため，JUMANでは個々の活用形ごとに接続規則が必要となっている．それに対して，MAJOの形態素文法は派生文法に基づいているため，そのような細かな規則は不要である．

文法規則の複雑さの指標となる接続行列は，JUMANで使用されている文法では195×165の大きさであった．また，派生文法をJUMANに適用した方法[6]では，行列の大きさは63×71と改善されている．しかし，それには活用形17種と活用型16種が導入されたことによる増加分が含まれており，活用型による動詞の細分を必要としないMAJOでは，接続規則は24×26とさらに小さくなる．

接続コスト最小法[15),16)]を利用する場合，文法規則に付加したコストを調整しなければならない．接続コストは，一部を変更すると他の部分にも影響を与えるため，その調整が困難であるが，MAJOは文法規則が少ないため，接続コスト調整の点でも有利である．また，品詞の設定によっては，複数の文法規則のコストを同時に変更することが必要になる．たとえば，活用を考慮した品詞分類を行っている場合，動詞語幹と接尾辞の接続コストを変更する際に，カ行五段動詞語幹，サ行五段動詞語幹などの区別があると，それらの語幹とそれに接続可能な接尾辞間の接続コストをすべて変更する必要がある．しかし，MAJOにおいては，子音幹–子音幹接尾辞間および母音幹–母音幹接尾辞間の接続コストを変更するだけですむ．さらにMAJOでは，解析精度向上のために，「アル」「イル」は補助動詞である，などの意味的な品詞分類を追加しても，もともとの文法規則が少ないため，その接続コストの設定が簡単である．

4.2 登録単語数の削減

派生文法を構文解析に利用した方法[7]では，異形態の登録によって音便変化を取り扱っている．その結果，システムの動作は単純となるが，語幹末尾がs以外の子音幹動詞と同数の異形態を辞書に登録する必要がある．実際，EDR日本語単語辞書においては，登録されている動詞26,091個（うち子音幹動詞14,353個）に対して9,247個の異形態の登録が必要となる．また，音便形に対する文法規則も別に用意する必要があり，形態素文法が複雑になる．

それに対して，MAJOでは後方からの探索および音素の補完という手法で音便形を解析し，異形態の登録を接尾辞と不規則な変化をする動詞だけにおさえている．実際，MAJOは音便処理のために40個の異形態を登録しているにすぎない．また，音素を補完した後は，通常の動詞と同様に扱うことができるため，文法規則の追加も必要ない．

なお，派生文法の音韻規則を扱っている手法[8]では，形態素文法はMAJOより簡潔になると予想されるが☆，音韻規則の追加とそれを扱うシステムが必要となるので，全体としてはMAJOより複雑になると考えられる．

4.3 口語的表現への対処

MAJOにおいては，音韻論的なアプローチである派生文法を用いているため，口語的表現への対処が容易である．特に，多くの口語的表現が，それぞれ1つの形態素を辞書に追加するだけで解析可能となる．いわゆる「ら抜き言葉」である「食べレル」は，MAJOでは可能の派生接尾辞 "-(r)e-" を辞書に登録することにより "食 be-re-ru" と解析できる．また，"-(r)e-" の登録により，たとえば「書ケル」も "書 k-e-ru" と解析できる．これにより，学校文法では「書ク」とは別の動詞とされてきた可能動詞「書ケル」の登録が不要となる．さらに，「書カセル」「食ベサセル」に対する「書カス」「食ベサス」といった表現も，派生接尾辞 "-(s)as-" を登録することにより，それぞれ "書 k-as-u"，"食 be-sas-u" と解析できる．

ただし，本手法では "-(r)e-" や "-(s)as-" が接続可能な語幹を接続不可能な語幹と区別しないため，たとえば "届 keru" を "届 k-e-ru" と誤って解析する可能性もある．この問題に対しては接続コストの調整により対処した．正しい解析である "届 ke-ru" と誤った解析である "届 k-e-ru" では，"届 k-" （右連接属性：子音幹）と "-e-" （左連接属性：子音幹接尾辞）との接続コストの分 "届 k-e-ru" の方が接続コストが大きくなるため，接続コスト最小法により，正しい解析である "届 ke-ru" が選択される．

なお，同様の手法は派生文法をJUMANに適用した方法[6]でも行われているが，可能の派生接尾辞 "-(r)e-" を登録する際に，活用形の追加が必要であり，単語を辞書に登録するだけで実現可能なMAJOに比べると，口語的表現への対処は複雑である．

4.4 ローマ字表記による解析効率の低下

MAJOでは入力文における平仮名の部分をローマ字表記にするため，解析対象の文字列の長さが増大する欠点を持つ．文字列の長さが増大すると辞書との照合開始位置となる可能性のある箇所が増え，解析効率が低下する．MAJOでは，ダブル配列法によるトライ検索[13]を導入することにより，辞書検索の高速化を図っているが，この問題の根本的な解決とはなっていない．

5. 性能評価

EDR日本語コーパスのうち，1,000文を用いてMAJOの解析精度を評価した．MAJOでは接続コスト最小法により解析結果に順序を付けるが，今回は優先度の最も高い解析結果を1つだけ出力させ，それを人手による解析結果と比較した．なお，接続コストは人手によって与えた．

解析精度の目安として，文献17)と同様に，正解の全形態素数に対する解析誤りの出現数の比（エラー率）を算出した．品詞の設定基準が異なるため，エラー率を単純に比較することはできないが，同じ基準でエラー率を算出している文献6)および17)との比較を**表7**に示す．MAJOは簡潔な形態素文法を持ちながらも，従来のシステムに匹敵する精度を達成しているといえる．

MAJOの解析誤りの原因の内訳は**表8**示すとおりである．MAJOは後方からの解析，および子音の補完という独自の手法を採ってているが，これが原因となる解析誤りは今回の実験では発見されなかった．なぜなら，MAJOでは接続コスト最小法を用いて，それぞれの形態素解析結果についてコストを計算するため，後方から解析した場合も，前方から解析した場合

表7 実験結果の比較
Table 7 Comparison of morphological analyzers.

システム	入力文	形態素数	誤り個数	エラー率
MAJO	1,000	25,012	469	1.88%
文献17)	1,016	29,024	687	2.36%
文献6)	10,000	207,547	2,040	0.98%

☆ 文献8)では形態素文法については述べられていない．

表8 解析誤りの原因の内訳
Table 8 Classification of alalysis errors.

誤りの原因	誤り数	エラー率
未登録語	114	0.46%
名詞分割誤り	79	0.32%
引用の接続助辞 "to"	66	0.26%
名詞+"de"	47	0.19%
平仮名表記の単語	31	0.12%
副詞	19	0.08%
音便変化	4	0.02%
その他	109	0.43%
合計	469	1.88%

も，接続コストの和が最小となる解析結果は同じものになるからである．また，子音の補完については，補完する条件を動詞接尾辞の異形態が発見された場合か，子音幹に接続可能な接尾辞の直前に母音がある場合に限定しているため，誤解析になる例が現れなかったと考えられる．

誤りの原因のうち，最も多いのが未登録語に起因するものであり，114個（0.46%）あった．未登録語の中には，漢字表記の単語は辞書に登録されているが，平仮名表記☆のものが登録されていなものが7個存在した．たとえば，今回の実験では，「山本五十六ひきいる連合艦隊」という文を解析する場合，辞書に "率 i-ru" は登録されていたが，"hikii-ru" が登録されていなかったため，正しく解析できなかった．その問題に対処するため，文献6) では漢字表記の単語を平仮名で表記したものも辞書に登録しており，その結果，文献6) では未登録語に起因する誤りは0.42%となっている．辞書登録されている単語の総数を比較すると，文献6) では約50万語であり，これはMAJOにおける登録単語の総数約25万語の2倍となっている．

しかし，いたずらに平仮名表記の単語を辞書に登録すると，解析候補が増え誤解析の原因となる場合もある．今回の実験でも「閉山シタタメ」という入力文に対して，本来ならば "閉山-si-ta-tame" と解析すべきところ，辞書に「シタタメ-ル」"sitatame-ru" が登録されていたため，"閉山-sitatame" と誤解析した例があった．そうした平仮名表記の単語に起因する誤りは，今回の実験では31個（0.12%）あった．その点を考慮し，MAJOでは漢字表記の単語の平仮名表記をすべて辞書に登録する手法は採用しなかった．

動詞に関しては，11個の誤りがあった．その内訳は，上述の "hikii-ru" のように，動詞の平仮名表記が辞書に登録されていなかったものが5個，誤って可能動詞と解析されたもの2個，音便形に起因する誤りが4個であった．4.3節で述べたとおり，MAJOは派生接尾辞 "-e-" を辞書に登録することで，可能動詞の辞書登録を不要としている．しかし，「包マレタ」"包 m-are-ta" を "包 mar-e-ta" と誤解析してしまう可能性があり，そのような誤りが今回は2個あった．なお，いわゆる可能動詞は今回のコーパスには14ヵ所出現したが，すべて正しく解析できた．音便形に起因する誤りは，「行ッタ」が "行 t-tta"（イッタ）であるか "行 w-tta"（オコナッタ）であるか，または「イッタ」が "iw-tta"（言ッタ）であるか "it-tta"（行ッタ）であるかの区別を誤ったものである．これらの曖昧さを区別するためには，単語の意味的な情報が必要であるため，現在のMAJOでは正しく解析することができない．

以上のような形態素レベルの情報だけでは判別できない誤りや，接続コストの調整不足が原因の誤りに対しては，個々の例外処理の導入や接続コストの最適化などによって解析精度を上げることが必要となる．

6. おわりに

本論文では派生文法に基づく日本語形態素解析システムMAJOを提案した．MAJOの開発では，派生文法に準拠した形態素文法を作成し，それに基づいてシステムを構築したので，派生文法の持つ簡潔さを活かして，文法規則数の少ない形態素文法で日本語の解析を可能としている．また，辞書に登録する異形態を従来の音韻論的手法よりも少なくでき，さらに可能動詞などの登録を省略できている．その結果，システムの辞書も簡潔なものとなっている．

MAJOはその解析精度においてまだ改良の余地がある．接続コストについては，現在は人手により調整し決定しているが，学習により自動的に決定する手法も試みている．

また，MAJOは音韻論手法を用いているため，口語的表現への対処が容易であると思われる．そこで，今後ネットニュースやWWW上の文章に見られる口語表現を調査し，MAJOで使用する口語文法を作成することも検討中である．

その他にも，派生文法は日本語の膠着語としての性質に着目しているため，同じ膠着語である韓国語，ウイグル語などとの間の機械翻訳への応用が期待できる．筆者らは現在，MAJOを利用した日本語-ウイグル語機械翻訳[18]の研究に取り組んでいる．

末筆ながら，有意義ないくつかのご指摘をいただいた査読委員に感謝いたします．

☆ MAJOの辞書上ではローマ字表記．

参考文献

1) Bloch, B., 林 栄一（訳）：ブロック日本語論考，研究社 (1975).
2) 寺村秀夫：日本語のシンタクスと意味 II，くろしお出版 (1984).
3) 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，NL 103-1, pp.1–7 (1994).
4) 清瀬義三郎則府：日本語文法新論—派生文法序説，桜楓社 (1989).
5) 清瀬義三郎則府：日本語学とアルタイ語学，明治書院 (1991).
6) 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，Vol.2, No.4, pp.37–65 (1995).
7) 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，NL87-6, pp.43–50 (1992).
8) 三浦睦美，吉村賢治，首藤公昭：日本語の派生文法と 2 レベル規則，言語処理学会第 3 回年次大会発表論文集，pp.59–62 (1997).
9) 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 回年次大会発表論文集，pp.425–428 (1997).
10) 松本裕治，黒橋禎夫，宇津呂武仁，妙木 裕，長尾 真：日本語形態素システム JUMAN 使用説明書 version 3.0，テクニカルレポート，奈良先端科学技術大学院大学 (1996).
11) 田中穂積，佐藤泰介，元吉文男：自然言語処理のためのプログラム・システム—拡張 LINGOL について，電子通信学会論文誌，Vol.J60-D, No.12, pp.1061–1068 (1977).
12) Koskenniemi, K.: Two-level model for morphological analysis, *Proc. IJCAI-83*, pp.683–685 (1983).
13) Aoe, J., Morimoto, K. and Sato, T.: An Efficient Implementation of Trie Structures, *Softw. Pract. Exper.*, Vol.22, No.9, pp.695–721 (1992).
14) 益岡隆志，田窪行則：基礎日本語文法—改訂版，くろしお出版 (1992).
15) 久光 徹，新田義彦：接続コスト最小法による日本語形態素解析の提案と計算量の評価について，電子情報通信学会技術研究報告，NLC 90-8, pp.17–24 (1990).
16) 中村順一，吉田 将，今永一弘：接続コスト最小法による日本語形態素解析の評価実験，電子情報通信学会技術研究報告，NLC 91-1, pp.1–8 (1991).
17) 丸山 宏，萩野紫穂：正規文法に基づく日本語形態素解析，情報処理学会論文誌，Vol.35, No.7, pp.1293–1299 (1994).
18) 小川泰弘，ムフタル・マフスット，外山勝彦，稲垣康善：派生文法に基づく日本語–ウイグル語機械翻訳—動詞接尾辞の変換，情報処理学会研究会報告，NL 120-1, pp.1–6 (1997).

付録 MAJO で定義されている品詞

表 9 MAJO で定義されている品詞
Table 9 Parts of speech used in MAJO.

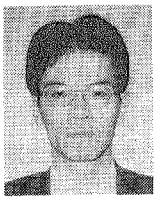
品詞名	左連接属性	右連接属性
名詞	名詞	名詞幹
形状名詞	名詞	形状名詞幹
形式名詞	形式名詞	形式名詞幹
形式形状名詞	形式名詞	形状名詞幹
数詞	数詞	数詞
格接尾辞	格接尾辞	格接尾辞
副助辞	副助辞	副助辞
助数詞	助数詞	助数詞
繋辞/終止	繋辞	繋辞/終止
繋辞/連用	繋辞	繋辞/連用
繋辞/連体	形状繋辞	繋辞/連体
繋辞ニ	形状繋辞	繋辞/連用
子音幹動詞	動詞	子音幹
母音幹動詞	動詞	母音幹
形状動詞	形状動詞	形状幹
補助母音幹動詞	補助動詞	母音幹
補助子音幹動詞	補助動詞	子音幹
補助形状動詞	補助動詞	形状幹
子統語接尾辞/連体	子音幹接尾辞	連体接尾辞
子統語接尾辞/連用	子音幹接尾辞	連用接尾辞
母統語接尾辞/連体	母音幹接尾辞	連体接尾辞
母統語接尾辞/連用	母音幹接尾辞	連用接尾辞
形状統語接尾辞/連体	形状幹接尾辞	連体接尾辞
形状統語接尾辞/連用	形状幹接尾辞	連用接尾辞
名名派生接尾辞	名詞接尾辞	名詞幹
名母派生接尾辞	名詞接尾辞	母音幹
名子派生接尾辞	名詞接尾辞	子音幹
名形派生接尾辞	名詞接尾辞	形状幹
母名派生接尾辞	母音幹接尾辞	名詞幹
母母派生接尾辞	母音幹接尾辞	母音幹
母子派生接尾辞	母音幹接尾辞	子音幹
母形派生接尾辞	母音幹接尾辞	形状幹
子名派生接尾辞	子音幹接尾辞	名詞幹
子母派生接尾辞	子音幹接尾辞	母音幹
子子派生接尾辞	子音幹接尾辞	子音幹
子形派生接尾辞	子音幹接尾辞	形状幹
形名派生接尾辞	形状幹接尾辞	名詞幹
形母派生接尾辞	形状幹接尾辞	母音幹
形子派生接尾辞	形状幹接尾辞	子音幹
形形派生接尾辞	形状幹接尾辞	形状幹
接続助辞	接続助辞	接続助辞
終助辞	終助辞	終助辞
名詞接頭辞	接頭辞	体言接頭辞
数詞接頭辞	接頭辞	体言接頭辞
動詞接頭辞	接頭辞	動詞接頭辞
形状接頭辞	接頭辞	形状接頭辞
連体詞	連体詞	連体詞
副詞	副詞	副詞
接続詞	接続詞	接続詞
句読点	句読点	句読点
記号	記号	記号

continued on next page

continued from previous page

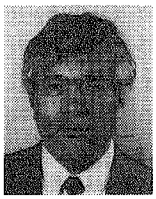
品詞名	左連接属性	右連接属性
イ音便/連体	イ音便	連体接尾辞
濁音便/連体	濁音便	連体接尾辞
促音便/連体	促音便	連体接尾辞
撥音便/連体	撥音便	連体接尾辞
イ音便/連用	イ音便	連用接尾辞
濁音便/連用	濁音便	連用接尾辞
促音便/連用	促音便	連用接尾辞
撥音便/連用	撥音便	連用接尾辞

(平成10年1月9日受付)
(平成10年11月9日採録)



小川 泰弘（学生会員）

1995年名古屋大学工学部情報工学科卒業．1997年同大学院工学研究科情報工学専攻修士課程修了．現在，同博士課程在学中．自然言語処理に関する研究に従事．言語処理学会会員．



ムフタル マフスット（正会員）

1983年新疆大学数系卒業．1996年名古屋大学大学院工学研究科情報工学専攻博士課程満了．同年，三重大学助手．自然言語理解に関する研究に従事．人工知能学会会員．

外山 勝彦（正会員）

1984年名古屋大学工学部電気学科卒業．1989年同大学大学院工学研究科情報工学専攻博士課程満了．同大学助手，中京大学講師，助教授を経て，1997年名古屋大学大学院工学研究科助教授．工学博士．論理に基づく知識表現と推論，自然言語理解に関する研究に従事．電子情報通信学会，人工知能学会，言語処理学会，日本認知科学会各会員．

稲垣 康善（正会員）

1962年名古屋大学工学部電子工学科卒業．1967年同大学大学院博士課程修了．同大学助教授，三重大学教授を経て，1981年より名古屋大学工学部教授．1997年より同大学大学院工学研究科長・工学部長．工学博士．オートマトン・言語理論，計算論，ソフトウェア基礎論，代数的仕様記述法，人工知能，自然言語処理に関する研究に従事．1992～1994年本学会理事，1995～1997年本学会調査研究運営委員会委員長．電子情報通信学会，人工知能学会，日本ソフトウェア科学会，日本OR学会，言語処理学会，IEEE，ACM，EATCS各会員．