膠着語形態素解析における音韻変化処理

3M−02

小川泰弘† ムフタル・マフスット‡ 杉野花津江† 稲垣康善†

†名古屋大学大学院工学研究科 ‡名古屋大学大学院国際開発研究科

1 はじめに

我々はこれまでに，派生文法 [1] に基づく日本語形態素解析システム MAJO[2] の開発を進めてきた．派生文法は日本語の膠着語としての特徴を捉えており，その点に着目した MAJO は，他の膠着語の形態素解析にも応用可能である．しかしながら，これまでの MAJO は日本語の音便変化に対する処理を含んでおり，そのままでは，他の膠着語の解析に適用できなかった．

本研究では，そうした日本語特有の処理を，音韻変化を規則化することで MAJO 本体から独立させ，他の膠着語についても，システムを変更することなく形態素解析可能とした．

本稿では，このシステムの概要と，実際に日本語について記述した音韻変化規則について述べる．

2 派生文法に基づく膠着語形態素解析

派生文法は，日本語の形態素を音韻単位で設定し，例えば「売られました」は「売 r+(r)are+(i)mas+(i)ta」の 4 つの形態素から構成されると考える．ここで，括弧内の音素は連結子音・連結母音と呼ばれ，連結子音 (母音) は，子音 (母音) に後接する場合に欠落する．これにより，「売 r+(r)u」「食 be+(r)u」のように，子音で終わる動詞 (子音幹動詞) と母音で終わる動詞 (母音幹動詞)，に同じ接尾辞が接続していると考えることが可能である．なお，「(r)are」「(i)mas」のように後に他の接尾辞が接続する接尾辞を派生接尾辞，「(i)ta」のようにそれで動詞句が終わる接尾辞を統語接尾辞と呼ぶ．

連結子音・連結母音の欠落は日本語だけでなく，他の膠着語にも存在する．また，膠着語では，単語に接辞が接続することによって意味が付加されるが，他の膠着語にも派生接尾辞・統語接尾辞の区分があり，さらに，異なる膠着語間で，それぞれ対応する接辞が存在することがある．そうした場合，日本語の接辞も派生文法に基づいて音韻単位で設定した方が，対応関係が明確になり，膠着語間での機械翻訳が簡単になる [3]．

そこで，派生文法を日本語以外の膠着語にも適用し，それに基づいて形態素を設定することによって，膠着語に対する汎用的な形態素解析が可能になる．

Phonological Change Approach to Morphological Analysis of Agglutinative languages
Yasuhiro, OGAWA†, Muhtar MAHSUT‡, Kazue SUGINO† and Yasuyoshi INAGAKI†
†Graduate School of Engineering, Nagoya University, Japan
‡Graduate School of International Development, Nagoya Univ.
yasuhiro@inagaki.nuie.nagoya-u.ac.jp

なお，音韻単位で形態素を設定するため，日本語の場合は，漢字かな混じり文のひらがな部分をローマ字表記にした，漢字ローマ字混じり文を入力とする．

3 派生文法に基づく従来の日本語形態素解析

日本語形態素解析の主たる問題は活用処理にある．例えば動詞句「売 tta」が入力された場合，「売 r」の形で現れないため，単純に接続関係を調べるだけは解析ができない．派生文法では，動詞「売 r」に過去を表す統語接尾辞「(i)ta」が接続する際に，「売 tta」になったと考えるが，形態素解析では，「売 tta」を「売 r+(i)ta」として解析する必要がある．

従来，MAJO では後方からの探索および音素の補完という内部処理で，この問題を解決してきた．しかし，この処理は日本語特有のもので，他の言語の音韻変化を扱うには別のプログラムが必要であった．

それに対して，本研究では，言語ごとに個別のプログラムを書くのではなく，音韻変化を規則として記述し，それに基づいて処理する方法を提案する．

4 音韻変化規則の作成

提案手法では，音韻変化規則を正規表現で記述し，先程の「売 r+(i)ta」が「売 tta」となる音便変化を次のように表現する．

`{r|t|w}+(i)t → {t|t|t}0000t`

ここで，'+' は形態素の区切，'0' は対応する文字が消滅することを示している．矢印の左が音韻変化する前の文字列であり，右が変化後の文字列となる．また '()' を連結子音・連結母音を示すのに使用しているため，正規表現における演算子の適用順序の変更は '{}' で示す[1]．上記の規則は，末尾が 'r', 't', 'w' のいずれかである形態素に，'(i)t' ではじまる形態素が接続すると，それらが 'tt' となることを示している．

日本語には，形態素の形は同じでも音韻変化が異なる場合がある．希望を表す派生接尾辞「(i)ta」は過去を表す統語接尾辞「(i)ta」と同じ形をしているが，撥音便の変化をせず，例えば「売 r+(i)ta+i」は「売 ritai」となる．そこで，音韻変化規則に品詞情報を付加し，それらを区別できるようにした．

また，統語接尾辞「(i)tutu」は '(i)t' ではじまるが，音便変化を起さないので，これも区別した．

以上の結果，撥音便の音韻変化規則は表 1 のようになる．表 1 の撥音便の欄に記述された 3 つの規則を上

[1]連結子音・連結母音を示す () が無くても，音韻変化規則を記述することは可能であるが，今回はあえて明記した．

表 1: 日本語音韻変化規則 (一部)

音韻変化	音韻変化規則	
撥音便	{r|t|w}+(i)ta	(動詞+派生接尾辞)
	→ {r|t|w}+0i0ta	(動詞+派生接尾辞)
	{r|t|w}+(i)tutu	(動詞+連用接尾辞)
	→ {r|t|w}+0i0tutu	(動詞+連用接尾辞)
	{r|t|w}+(i)t	(動詞+ALL)
	→ {t|t|t}0000t	(動詞促音便+ALL)
連結子音	C+({r|s|y})	(動詞+ALL)
	→ C00{0|0|0}0	(動詞+ALL)
	+({r|s|y})	(ALL+ALL)
	→ 00{r|s|y}0	(ALL+ALL)
連結母音	C+({a|i|u})	(動詞+ALL)
	→ C+0{a|i|u}0	(動詞+ALL)
	+({a|i|u})	(ALL+ALL)
	→ 00{0|0|0}0	(ALL+ALL)
w の欠落	w+{i|u|e|o}	(動詞+ALL)
	→ 00{i|u|e|o}	(子音幹動詞+ALL)

注：C は{k|g|r|t|w|b|n|m|s}を表す.

から順番に適用することによって，正しい音便変化形が得られる．イ音便，促音便に関しても同様に記述できる．なお，品詞欄に'ALL'とある場合は，あらゆる品詞に対して規則が適用されることを意味する．

派生文法の特徴である，連結子音・連結母音の欠落も表 1 に示したように記述できる．なお，母音幹動詞の末尾の音素は i か e だけであるが，実際の入力では「出+(i)ta」のように漢字で表記される場合もある．そこで，子音が接続する場合の規則を先に適用し，それが適用されないときに次の規則を適用することで，漢字表記にも対応した．ここで，'+' の左に文字がない場合，その規則は前の形態素に関わりなく適用される．

その他，「買 w+(r)u」が「買 u」となる w の欠落規則，不規則動詞「来る」「する」，音便形が特殊な「行く」，「下 sar+(i)」が「下 sai」となる r の欠落規則などについても音韻変化規則を記述した．現在のところ，日本語の音韻変化を 50 個の規則で記述している．

5 音韻変化規則を使用した解析・生成

提案手法では，前章の音韻変化規則を形態素解析システム MAJO で直接用いるのではなく，別に作成した異形態生成モジュールで利用する．このモジュールでは，音韻変化規則から有限状態オートマトンを作成し，音韻変化する可能性のある動詞について，その音便変化した形 (異形態) を生成する．これに，音便変化規則に記された異形態の品詞情報を付加して，MAJO の辞書に追加し，形態素解析に使用する．

この手法には，辞書への登録単語数が膨大になるという欠点があるが，現在では，TRIE 法などの登録単語数に影響されない高速な辞書検索アルゴリズムが考案されており，また，記憶装置も安価で大容量のものが存在するため，大きな問題では無くなっている．

本研究では，異形態を登録することによって，形態素解析システム内での音韻変化処理を不要とし，システムを簡単化した．

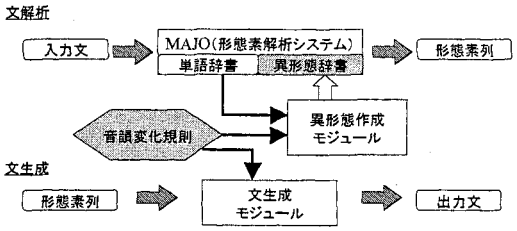

図 1: 形態素解析システムの概要

また，この音韻変化規則は，生成規則であるから，解析だけでなく生成にも利用できる．そこで，音韻変化規則を使用する文生成モジュールも作成した．このモジュールは，入力される形態素列から音韻変化を考慮した文を生成して出力する．

以上で説明した音韻変化規則と解析・生成システムの関係は図 1 のようになり，同じ規則で膠着語文の生成・解析が可能となる．

6 関連研究

汎用的な形態素解析に関する研究としては，PC-KIMMO[4] がある．また，PC-KIMMO を改良し，派生文法に基づく日本語規則を記述した研究として，文献 [5][6] がある．PC-KIMMO は，音韻変化規則を正規表現で記述し，オートマトンで解析するものであり，解析・生成の両方が可能である．

こうした点は提案手法と同じであるが，提案手法は，音韻変化規則の記述方法と，あらかじめ異形態をすべて登録する点が異なる．

7 まとめ

派生文法に基づく汎用的な膠着語形態素解析システムの枠組について述べた．これにより，各膠着語ごとに，辞書，品詞データ，形態素間の接続コストおよび音便変化規則を用意すれば，同じシステムを使用して形態素解析が可能となる．

現在は，本システムを使用したウイグル語の形態素解析を進めるとともに，日本語–ウイグル語間の機械翻訳システムについても研究している．

参考文献

[1] 清瀬義三郎則府：日本語文法新論-派生文法序説-, 桜楓社 (1989).

[2] 小川泰弘, ムフタル・マフスット, 外山勝彦, 稲垣康善：派生文法による日本語形態素解析, 情報処理学会論文誌, Vol. 40, No. 3, pp.1080–1090 (1999).

[3] 小川泰弘, ムフタル・マフスット, 杉野花津江, 外山勝彦, 稲垣康善：派生文法に基づく日本語動詞句のウイグル語への翻訳, 自然言語処理, Vol. 7, No. 3, pp.57–78 (2000).

[4] Koskenniemi, K. : Two-level model for morphological analysis, IJCAI-83, pp.683–685, (1983).

[5] 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).

[6] 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).