日本語－ウイグル語間機械翻訳におけるウイグル語音韻変化処理の形式化

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
yasuhiro@cse.nagoya-u.ac.jp

1　はじめに

日本語とウイグル語は，言語学においてはともに膠着語に分類され，また語順がほぼ同じであるなどの点で構文的類似性が高い．そのため両言語間の機械翻訳においては，形態素解析した結果を逐語訳することで，ある程度の翻訳が可能である [1]．しかし，両言語における音韻変化には異なる点が多く，そのため，両言語間の機械翻訳において，それぞれの音韻変化を考慮した形態素解析と文生成が必要になる．その際に，言語ごとに別のシステムを構築するのはコストの面から好ましくない．そこで，本稿では，音韻変化を規則化し，両言語を同じシステムで処理する手法を提案するとともに，本手法で作成したウイグル語の音韻変化規則を示す．

なお，本稿では日本語，ウイグル語ともにローマ字で表記する．混同を避けるために，日本語の単語は「」，ウイグル語の単語は“”で括り区別する．

2　日本語とウイグル語間の翻訳

日本語－ウイグル語機械翻訳においては，その構文的類似性を利用することにより，構文解析が不要になる．よって，日本語入力文を形態素解析した段階で各単語を対応するウイグル語に逐語訳することによって，ある程度の翻訳が可能になる (図 1)．

さらに，日本語の膠着語としての性質に着目した派生文法 [2] を用いて日本語とウイグル語を記述すると，動詞句の構成方法など，形態論においても多くの共通点があることが明らかになる．

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」

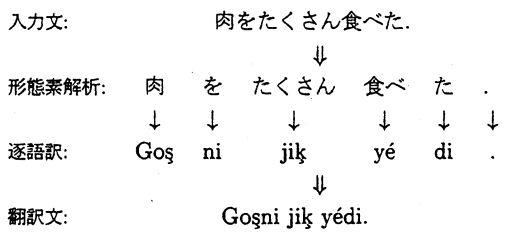

図 1: 日本語－ウイグル語逐語翻訳

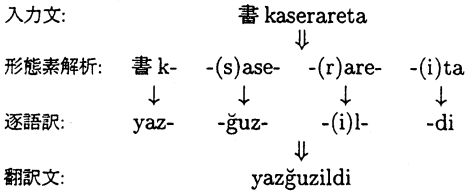

図 2: 派生文法に基づく動詞句の翻訳

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になる．なお，「(r)are」「(i)mas」のように他の接尾辞が後接する接尾辞を派生接尾辞，「(i)ta」のように動詞句の末尾に来る接尾辞を統語接尾辞と呼ぶ．

連結子音・連結母音の欠落や，派生接尾辞と統語接尾辞の区分は，日本語だけでなくウイグル語にも存在するが，派生文法を利用することで，両言語間の接尾辞の対応関係が明確になり，例えば，「書かせられた」のように動詞に複数の接尾辞が接続した文も逐語訳で翻訳が可能になる (図 2)．また，逆方向の翻訳であるウイグル語－日本語翻訳も，同様の手法で実現可能である．

しかしながら，日本語およびウイグル語では，形態素が連続する際に，連結子音・連結母音の欠落以外にも，さまざまな音韻変化を引き起こす．例えば，日本語の場合，「書 k」に過去を表す「(i)ta」が接続する場合，そのまま接続すると「書きた」になるが，実際には，いわゆるイ音便が生じ「書いた」となる．一方，ウイグル語にも母音調和や円唇同化と呼ばれる様々な音韻変化規則が存在する．

我々はこれまでに，派生文法 [2] に基づく日本語形態素解析システム MAJO[3] 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 は日本語の音韻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．つまり，ウイグル語の形態素解析に利用するためには，MAJO のシステム自体を変更する必要があった．

本研究では，そうした音韻変化の処理を規則化することで MAJO 本体から独立させ，他の膠着語も，システムを変更することなく形態素解析可能にした．

本稿では，この手法について簡単に紹介する．さらに，ウイグル語の音韻変化について簡単に述べるとともに，提案手法用に記述したウイグル語の音韻変化規則の一部を示す．

3 音韻変化の規則による解析・生成手法

提案手法では，音韻変化を正規表現を利用して規則化する．ウイグル語音韻変化の具体的な記述方法は，5章において述べる．この音韻変化規則を，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単語について，その音韻変化した形 (異形態) を自動的に生成する．これに，音韻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．

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・解析が可能になる．我々は，このシステムを用いて日本語の音韻変化も取り扱っているがそれについては文献 [4] を参照されたい．

4 ウイグル語の音韻変化

本章では，ウイグル語の音韻変化の主なものについて，文献 [5] を参考にしつつ，派生文法 [2] の考えも適用して簡単に説明する．

4.1 ウイグル語の音素と文字

ウイグル語には，アラビア文字に似た 32 の文字があり，文は右から左へと書かれる．それとは別に，ローマ字表記を用いる場合もあり，本論文では，便宜上，こちらのローマ字表記を使用し，不足する文字については ç, é, ğ, ḫ, ķ, ö, ş, ü, ź と表記する．

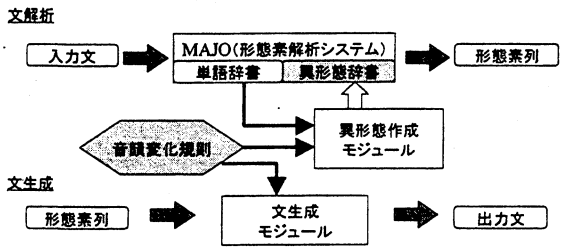

図 3: 形態素解析システムの概要

ウイグル語表記に使用される文字は表音文字であるが，ほとんどの言語がそうであるように，ウイグル語でも，文字と実際の発音には一致しない点がある．しかし，本研究では，実際に記述されたウイグル語文を解析・生成するため，原則として文字単位でウイグル語の音韻変化規則を論じる．

4.1.1 ウイグル語の母音

ウイグル語には a, é, u, ü, o, ö, i, e の 8 個の母音が存在する．ここで，舌の位置の違いから，前母音 é, ü, ö と後母音 a, u, o に区別される．なお，i と e は前母音・後母音の区別に関しては中立である．また，口の丸め方から，ü, u, ö, o は円唇母音と呼ばれ，それ以外の é, a, i, e は非円唇母音と呼ばれる．さらに，口の開き方から，é, a は広母音と呼ばれ，ö, o, e はやや狭い母音，i, ü, u は狭母音と呼ばれる

é については，表記上は e に良く似た音に見えるが，音素としては後母音 a に対応する前母音であり，ä と表記される場合もある．

4.1.2 ウイグル語の子音

ウイグル語の子音は全部で 24 個あり，有声音 b, d, g, ğ, j, l, m, n, r, w, y, z, ź, ng と無声音 f, h, ḫ, k, ķ, p, ç, s, t, ş に分類される．なお，母音はすべて有声音である．

4.2 接尾辞の音韻変化

日本語の「へ」に相当するウイグル語の格助詞 “-ğa” や，ウイグル語で過去を表す統語接尾辞 “-(i)p” は，それぞれ図 4や図 5のように変化する．これは，“-ğa” は，“gé”, “ké”, “ğa”, “ķa”, “ga” の 5 つ，“-(i)p” は，“ip”, “p”, “up”, “üp” の 4 つの異形態を持つと考えられる．これらの異形態のうち，どの形態が接続するかは，接尾辞が接続する単語との関係から以下のような性質が分かっている．これ以降，接尾辞が接続する単語を語幹と呼ぶ．

köz	+	-ğa	→	közgé	「目へ」
ķol	+	-ğa	→	ķolğa	「手へ」
és	+	-ğa	→	éské	「記憶へ」
baş	+	-ğa	→	başķa	「頭へ」
tağ	+	-ğa	→	taķķa	「山へ」
til	+	-ğa	→	tilğa	「舌へ」
kim	+	-ğa	→	kimgé	「誰へ」
ata	+	-ğa	→	atiğa	「父へ」
mén	+	-ğa	→	manga	「私へ」

図 4: 格助詞 "-ğa" の語形変化

4.2.1 有声音，無声音の区別

"gé" と "ké" のように，異形態に有声音で始まるものと無声音で始まるものがある接尾辞においては，語幹の末尾の音素が有声音の場合には有声音で始まる異形態が接続し，末尾が無声音のときには無声音で始まる異形態がが接続する．

ただし，この規則には例外があり，末尾が g か ğ の語幹に "-ğa" が接続する場合，全体が無声音化する．例えば "tağ" 「山」 に "-ğa" が接続する場合は "tağğa" ではなく，"takka" になる．

4.2.2 母音調和

ウイグル語では，原則として，一つの単語の中に前母音と後母音が混在することはない．さらに，接尾辞に前母音を含む異形態が存在する場合には，前母音を含む語幹には前母音を含む異形態が接続する．こうした現象は母音調和と呼ばれ，ウイグル語を含むトルコ系の言語の他，韓国語などでも見られる現象である

ウイグル語においては，図 4に見られるように，前母音を含む "köz"「目」，"és"「記憶」には前母音を含む "gé"，"ké" が，前母音を含まない "kol" 「手」，"baş" 「頭」には前母音を含まない "ğa"，"ķa" が接続する．

ただし，i を含む場合は単語ごとに異なり，例えば，"til" 「舌」 に接続するのは "ğa" であるが，"kim" 「だれ」 に接続するのは "gé" である．

4.2.3 連結子音・連結母音の欠落

日本語の統語接尾辞「(r)u」「(i)ta」に見られる連結子音・連結母音はウイグル語にも存在する．ウイグル語で過去を表す統語接尾辞 "-(i)p" の (i) もその一つであり，語幹末尾が子音である動詞 "qiķ-" 「出る」 には異形態 "ip" が，語幹末尾が母音である動詞 "oķu-" 「読む」 には異形態 "p" が接続する．

4.2.4 円唇同化

語幹の最後の音節に含まれる母音が円唇母音である場合，接尾辞の中に含まれる非円唇母音 i が円唇母音 u に変わることがある．この現象を円唇同化と呼ぶ．

çiķ-	+	-(i)p	→	çiķip	「出て」
oķu-	+	-(i)p	→	oķup	「読んで」
yul-	+	-(i)p	→	yulup	「抜いて」
kül-	+	-(i)p	→	külüp	「笑って」
bar-	+	-(i)p	→	berip	「行って」
çiķar-	+	-(i)p	→	çiķirip	「出して」

図 5: 統語接尾辞 "-(i)p" の語形変化

図 5の例では，"-(i)p" が "yul-" 「抜く」 に接続する場合に "up" となっている．さらに，円唇同化した母音は母音調和の規則にも従うため，"kül-" 「笑う」 に接続する際には "üp" になる．

4.3 語幹の音韻変化

前節の音韻変化は，語幹が後接する接尾辞に影響を与え，接尾辞を変化させていた．それとは逆に，接尾辞が語幹を変化させる場合がある．

形態素解析の観点から見れば，接尾辞はその種類が限られているため，異形態が少数となり人手で登録可能であるが，語幹は単語数が多いため，異形態の数も多くなり，提案手法のように自動生成が必要になる．

4.3.1 母音の同化・弱化

図 5の例では，"bar-" に "-(i)p" が接続すると "berip" となり，"çiķar-" に "-(i)p" が接続すると "çiķirip" となるように，語幹の最後の母音 a が変化している．この現象は，後の狭母音 i が前の母音を同化していると考えられ，母音の同化と呼ばれる．

また，図 4の例では，"ata" に "-ğa" が接続した場合に "atiğa" となる．これは，語幹末尾の a が弱められていると考えられ，母音の弱化と呼ばれる．

この母音の同化・弱化は，接尾辞が接続したものを音節で区切った際に，"be·rip"，"çi·ķi·rip"，"a·ti·ğa" のように，接尾辞を含む音節の直前の音節が母音で終わるときに発生し，"bax·ğa" のように音節の最後が子音の場合には起きない．そして，母音の同化・弱化が起きる場合，最後の音節に含まれる a もしくは é が，語幹が単音節の場合は e に，複音節の場合には i に変化する．なお，母音の同化・弱化は，語幹が名詞の場合にも動詞の場合にも発生する．

4.3.2 特定の単語に接続する際の音韻変化

ウイグル語の "mén"，"sén" は，それぞれ日本語の「私」「あなた」に相当するが，いくつかの接尾辞が接続する場合に特別な変化を起こす．例えば，"-ğa" が後接する場合，母音調和の規則に従えば "-ğa" が "gé" になるが，実際には名詞の母音の方が変化し，さらには "-ğa" も特別に "ga" となり，結果的にそれぞれ "manga"，"sanga" になる．

5 提案手法による音韻変化規則の記述

この章では，前章のウイグル語の音韻変化を提案手法で処理する方法について述べる．なお，プログラム上では é, ğ, ḫ, ķ, ö, ü, ẓ の文字を直接は扱えないため，それぞれ，!e, !g, !h, !k, !o, !u, !z のように，アルファベットの前に ！を付けて表現した．ただし，ç と ş については，q と x をそれぞれ使用した．

提案手法では，音韻変化規則を正規表現を用いて記述する．例えば，4.3.2節で述べた “mén” と “sén” の変化を次のように記述する．

+{m|s}!en+-!ga → +{m|s}0an000ga

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になる．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．上記の規則は，“mén” もしくは “sén” に，接尾辞 “-ğa” が接続する場合に，語幹の中の母音が a に変化するとともに，接尾辞が “ga” に変化することを示している．なお，形態素の形は同じでも，品詞により音韻変化が異なる場合を考慮し，音韻変化規則に品詞情報を付加するが，今回は省略する．

次に，接尾辞 “-ğa” が “ké” に変化する場合であるが，4.2.1節と 4.2.2節の規則を考慮して，以下のように記述した．

!{e|o|u|i}$A*$U+-!ga → !{e|o|u|i}$A*$U00k!e

ここで，$A，$U は，それぞれ任意の記号 1 文字，無声音 1 文字を示し，* は閉包を示している．すなわち，語幹のいずれかに前母音 é, ö, ü が出現し，末尾が無声音である場合に，接尾辞 “-ğa” が “ké” になる．i は単語によって異なるが，前母音を含む語が後接する語幹に含まれる i を!i と記述して区別した．つまり，この場合には，実際の表記と，音韻変化規則中で用いられる記号が異なり，例えば “kim” を k!im と記述する．

さらに，接尾辞 “-ğa” の “ķa” への変化は，そのまま考えると，語幹に前母音が出現せず末尾が無声音である場合に起きるとなるが，出現しないという条件を記述するのは複雑なため，規則に適用順序を設けた．すなわち，接尾辞 “-ğa” が “ké” になる規則の適用後に，

$U+-!ga → $U00!ka

を適用することで，接尾辞 “-ğa” が “ké” になる場合と “ké” になる場合を区別した．

同様にして，図 4, 図 5に示された格助詞 “-ğa” および統語接尾辞 “-(i)p” の音韻変化を，表 1のように記述した．表 1において上にある規則ほど，先に適用される．

現在のところ，ウイグル語の音韻変化のために 186 個の規則を記述している．

表 1: ウイグル語音韻変化規則 (一部)

格助詞 “-ğa”	統語接尾辞 “-(i)p”
+{m|s}!en+-!ga	+{$C\|}!e$C-+-(i)
→ +{m|s}0an000ga	→ +{$C\|}0e$C000!i0
!{e|o|u|i}$A*!g+-!ga	+{$C\|}a$C-+-(i)
→ !{e|o|u|i}$A*!k00k!e	→ +{$C\|}e$C0000i0
!{e|o|u|i}$A*$U+-!ga	!e$C-+-(i)
→ !{e|o|u|i}$A*$U00k!e	→ 0i$C000!i0
!g+-!ga	a$C-+-(i)
→ !k00!ka	→ i$C0000i0
$U+-!ga	!{o|u}$C-+-(i)
→ $U00!ka	→ !{o|u}$C000!u0
!e+-!ga	{o|u}$C-+-(i)
→ 0i00g!e	→ {o|u}$C0000u0
!{e|o|u|i}$A*+-!ga	{a|i|u|e|o}-+-(i)
→ !{e|o|u|i}$A*00g!e	→ {a|i|u|e|o}000000
a+-!ga	-+-(i)
→ i00!ga	→ 0000i0

注：$A は任意の 1 文字，$U は任意の無声音 1 文字，$C は任意の子音 1 文字を表す．

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[6] がある．また，PC-KIMMO に基づいてウイグル語と良く似たトルコ語の音韻変化規則を記述した研究として文献 [7] 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

日本語 – ウイグル語間での機械翻訳のための音韻変化規則の取り扱い手法の提案と，それに基づくウイグル語の音韻変化規則の記述について述べた．これにより，日本語とウイグル語を同じシステムで解析・生成が可能になる．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，ウイグル語 – 日本語機械翻訳の実現を目指している．

参考文献

[1] 小川, ムフタル, 杉野, 外山, 稲垣 : 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).

[2] 清瀬: 日本語文法新論 – 派生文法序説 –, 桜楓社 (1989).

[3] 小川, ムフタル, 外山, 稲垣 : 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).

[4] 小川, ムフタル, 杉野, 稲垣 : 膠着語形態素解析における音韻変化処理, 情報処理学会第 64 回全国大会講演論文集, (2002).

[5] 竹内: 現代ウイグル語四週間, 大学書林 (1991).

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

[7] Oflazer, K.: Two-level Description of Turkish Morphology, Literary and Linguistic Computing, Vol. 9, No. 2, (1994).