ウイグル語形態素解析における母音調和の扱い

アブドレイム・アブドハリリ　　伝 康晴　　土屋 俊
千葉大学大学院自然科学研究科　　千葉大学文学部　　千葉大学文学部

1　はじめに

ウイグル語はアルタイ語諸語に属する言語であり、正書法ではアラビア文字によく似た文字を使用している。本来は、正書法の分かち書きが存在しないが、現在は計算機上でのテキストにおいて文節ごとにスペースを挿入する形で書かれている。キーワード検索、索引づけといったテキスト処理を行なう場合は、前段階として文や文節を単語に分割する処理が不可欠である。

ウイグル語は日本語と同じ膠着語であり、語幹に接頭辞や接尾辞が接合する。文の構造も日本語に類似し、文構造は基本的に名詞－目的語－動詞の形をとる。たとえば、

(1) mæn　mæktæpkæ　barmajmæn
　　私は学校に行きたくない

という例文は以下のように形態素に分けられる。

(2) mæn　mæktæp kæ　bar　ma　j　mæn
　　私　学校　に　行く ない　たい　私
　　代名詞 名詞　格助詞 動詞 助動詞 助動詞 代名詞

この文では、正書法的な特徴から見ると文が 2 つのスペースにより分解しており、3 つの形態素のように見えるが、実際には 7 つの形態素から成り立っている。

我々は、ウイグル語と日本語の同じ膠着語であるという特徴と正書法の分かち書きが存在しないという特徴に着目し、日本語の形態素解析器である「茶筌」を用いて、ウイグル語の形態素解析を行った (アブドレイム, 2004)。

ウイグル語には、以上のように日本語と共通する特徴とともに、日本語に存在しない特有の母音調和という現象がある。この現象については、日ウ機械翻訳システム (小川ほか, 1999) 等で、出力の最後の整形段階で扱っている場合もあるが、形態素解析処理では機械翻訳のそれとは異なり、入力中にすでに存在しており、まず最初に対処しなければならない。

そこで、本稿では、母音調和により発生する母音の弱化・脱落・子音の挿入の形態素解析システムでの扱いについて述べる。

表 1　母音の分類

	円唇母音	非円唇母音
前母音	ø　y	æ
中立母音		e　l
後母音	o　u	a

2　ウイグル語の形態論

2.1　母音調和

ひとつの語の中に一緒に現われることのできる母音の組合せに一定の制約があることは母音調和と呼ばれている。ウイグル語を表記している文字は 32 個であり、そのうち母音が 8 個、子音が 24 個ある。8 個の母音は調音位置と口唇の形によって表 1 のように分けられる。

単語自体の、あるいは、形態素と形態素の母音が表 1 のグループによって調和される。中立母音グループは前・後母音グループと混在することができるが、前母音グループと後母音グループは互いに混在できない。そのため、同じ文法的な役割を持つ接辞の中に、前母音グループに付く形と後母音グループに付く形、また、円唇母音グループに付く形と非円唇母音グループに付く形が存在し、接合する際に語幹末尾の母音によってどちらかが選ばれる.

たとえば、名詞の複数形を現す語尾には “lar” と “lær” があるが、語幹である名詞末尾の母音に応じて以下のように使い分けられる

(3)　a. kitab + lar　　= kitablar
　　本　複数形語尾　（複数の）本
　　b. dølæt + lær　　= dølætlær
　　国　複数形語尾　（複数の）国

“kitab” の最後の母音 a は後母音なので、同じ後母音

を持つ “lar” が選ばれる。“dølæt” の場合は最後の母音æが前母音なので、前母音を持つ “lær” が選ばれる。母音調和は接辞の選択に影響するだけでなく、語幹の形を変化させることもある。以下、母音の弱化・脱落・子音の挿入について説明する

2.2 母音の弱化

ウイグル語では、形態素と形態素が接合する際に、母音の弱化が発生する。伝統的なウイグル語の文法では母音の弱化現象は以下の規則によって記述される。

規則 1 後母音 a や前母音 æ をもつ語幹に接辞がつくとき、語幹の音節が開音節になりアクセントが後に移るならば、a または æ が e に弱まる (竹内, 1991)。

たとえば、

(4) a. jaz + ip = jezip
書く 中止形 書いて
b. at + i = eti
名前 三人称語尾 彼の名前

規則 2 語幹に 2 つ以上の母音があるとき、語末の a または æ が開音節になりアクセントが後に移るならば、a または æ が i に弱まる (竹内, 1991)。

たとえば、

(5) a. jasa + di = jasidi
作る 完了形 作った
b. bala + si = balisi
子供 三人称語尾 彼の子供

2.3 母音の脱落

子音字で終了する 2 つの音節から成り立っている語幹に接辞が付くと、音節数を保つために、語幹末尾の音節の i, u, y が脱落する。

(6) a. orun + um = ornum
場所 一人称語尾 私の場所
b. ømyr + i = ømri
人生 三人称語尾 彼の人生

2.4 子音の挿入

母音字で終了する語に母音字で始まる接辞が付くと、母音連続の発音が難しいために、j という子音が挿入される。

(7) a. polu + um = polujum
ご飯 一人称語尾 私のご飯
b. bina + ing = binajing
建物 二人称語尾 あなたの建物

3 形態素解析における母音調和の扱い

3.1 方針

2 節で述べた母音調和の規則は言語学分野で提案されたものであり、この現象を自然言語処理で試したものはない。ウイグル語の語彙は、地理的な位置により外来語の影響を強く受けているので、単語の構造が多様になりつつある。とくに、外来語の場合は伝統的なウイグル語文法で使われている母音調和の規則に適合しないことが多く、処理がかなり面倒である。そのため、現在のウイグル語処理システムでは、語の派生形、あるいは、文節の単位で辞書を構築している。

日本語と同様に、ウイグル語も自立語と付属語に分けられ、自立語の中で動詞と名詞（名詞的な語を含む）に語尾が付き、語形が変化する。この中で動詞の変化が一番多い。正書法では、動詞の語幹に活用語尾、その次に人称語尾が付き、さらに疑問を表す接辞が付く。そのため、日本語の「書く」にあたる “jaz” という動詞の変化形をリストアップして見るとその数は 150 を超える。これらすべてを見出し語として登録するのは、見通しも悪く、辞書の量も大きくなりすぎる。このような屈折語的な扱いはウイグル語には不適切なようと思われる。

そこで、ウイグル語の膠着語的な特徴にふさわしい処理方法を提案する。単語ごとに 100 を越える変化形を辞書登録する代わりに、語尾を独立した形態素として登録する。この場合問題になるのは、語幹と語尾の間に見られる母音調和をいかにして扱うかである。

2 節で述べたこの現象は、日本語の助数詞などに見られる音の変化に類似している。たとえば、「六回」では「ロク → ロッ」のように前部要素の末尾音が変化し、「三本」では「ホン → ボン」のように後部要素の先頭音が変化する。伝ほか (2002) は、このような現象を扱うために、名詞・助数詞の変化形（促音化・濁音化など）とその要因となっている隣接語のタイプ（後続語が子音で始まる・先行語が撥音で終わるなど）との対応関係を記述し、形態素解析後の発音変化処理で利用している。この方式はウイグル語の母音調和にも適用できる。以下、詳しく見ていく。

3.2 母音の弱化

動詞や名詞に語尾が結合した場合に生じる母音の弱化を表 2・3 のように記述する。まず、母音の弱化を起こす語には「語末弱化型」という変化型を与える。次

表 2　動詞の記述（母音の弱化）

見出し語	kæt		
品詞	動詞		
変化型	語末弱化型		
出現形	隣接形態素	変化形	変化要因
kæt	ϕ	基本形	
kæt	sæ	基本形	
kæt	mæ	基本形	
ket	ing	弱化形	アクセント移動
ket	ip	弱化形	アクセント移動
ket	imæn	弱化形	アクセント移動

表 3　名詞の記述（母音の弱化）

見出し語	bala		
品詞	名詞		
変化型	語末弱化型		
出現形	隣接形態素	変化形	変化要因
bala	ϕ	基本形	
bala	m	基本形	
bala	ng	基本形	
bali	lar	弱化形	アクセント移動
bali	ni	弱化形	アクセント移動
bali	miz	弱化形	アクセント移動

に、各語の基本形（“kæt” や “bala”）と弱化形（“ket” と “bali”）がどのような形になるかを記す。さらに、それぞれの形がどのような隣接語が来たときに起こるかを記す。2.2 節で見たように、母音の弱化は後続形態素がアクセント移動を引き起こすときに生じる。たとえば、“ing”, “ip”, “imæn” などの活用語尾はそのような形態素であり、“sæ”, “mæ” などはそうでない。表 2・3 には、いくつかの隣接形態素の例と、それぞれが動詞や名詞に変化を引き起こすかどうかに関する情報（変化要因）が書かれている。

3.3　母音の脱落と子音の挿入

母音の脱落と子音の挿入も、母音の弱化と同様に扱う。母音の脱落で、語幹に連接する連接形態素が子音で始まるとこの現象は生じない。狭母音の i で始まる形態素が接続するとこの現象が起こる。そこで、後者に対しては先行語は脱落形を取る。子音の挿入でも、脱落の変化要因と同じく、連接形態素が狭母音で始まると先行語は挿入形を取る。

表 4　母音の脱落

見出し語	jisim		
品詞	名詞		
変化型	語末脱落型		
出現形	語尾助詞	連接形態素	変化要因
jisim	ϕ	基本形	
jisim	gha	基本形	
jisim	dim	基本形	
jism	i	脱落形	狭母音
jism	ing	脱落形	狭母音

表 5　子音の挿入

見出し語	imla		
品詞	名詞		
変化型	語末挿入型		
出現形	連接形態素	変化形	変化要因
imla	ϕ	基本形	
imla	din	基本形	
imla	da	基本形	
imlaj	im	挿入形	狭母音
imlaj	ing	挿入形	狭母音

表 6　語尾・助詞が母音調和に影響するパターン

	ing	um	lar
母音の弱化	○	×	○
母音の脱落	○	○	×
子音の挿入	○	○	×

3.4　接辞の記述

今まで主に自立語の記述について述べた。次は接辞の記述について考えてみよう。接辞 “ing” は表 2（母音の弱化）・表 4（母音の脱落）・表 5（子音の挿入）のすべてで基本形以外の隣接形態素として現れる。一方、語尾 “lar” は表 3 以外では基本形以外の隣接形態素としては現れない。ある接辞が母音の弱化・脱落・子音の挿入のいずれに関わるかを表にすると、表 6 のようになる。語尾・助詞の辞書には、表中で○がついているすべての現象に関わる変化要因を記述しておく必要がある。

表 7　接辞の変化

見出し語	ma		
品詞	助動詞		
変化型	語末弱化型		
出現形	隣接形態素	変化形	変化要因
ma	ϕ	基本形	
ma	ng	基本形	
ma	j	基本形	
mi	di	弱化形	アクセント移動
mi	sa	弱化形	アクセント移動
mi	ghaj	弱化形	アクセント移動

3.5　接辞に対する母音調和

伝統的なウイグル語の文法では、母音の弱化は語幹に接辞が後接する際に発生すると言われているが、実際には、接辞に接辞が後接する際にも母音の弱化が発生する。たとえば、

(8) jaz + ix + ma + sa + ng
書く　共同形　否定形　仮定形　人称語尾
= jezixmisang
あなたがたが書かなければ

では、まず “jaz” に共同形 “ix” が後接する際に母音の弱化が起こり（jaz → jez）、さらに、否定の助動詞 “ma” が付いて、その次に仮定形接尾 “sa” が付く際にも母音の弱化が現れる（jezixma → jezixmi）。

この現象は母音調和の規則を左から順に再帰的に適用すれば扱える。この際、表 2・3 の動詞や名詞と同様に、接辞の変化形も表 7 のように記述する。

3.6　接辞に対する子音調和

2.1 節で、母音調和の規則により、同じ文法的な役割を持つ接辞に対応する幾つかの異形態があることを述べた。これについてもう少し考えてみよう。ウイグル語には、日本語の格助詞に当たる曲形というものがある。たとえば、方向を表す助詞（日本語の「に」）は “gha” である。これに対応して、“qa”, “gæ”, “kæ” の 3 つの異形態がある。これらは、先行名詞の末尾音節の子音と母音のタイプによって使い分けられる。表 8 の名詞・出現形欄はこの使い分けの例を示している。

これらの例から見ると、母音調和に加えて子音調和の規則により後接接辞の形が決まっていることが分かる。子音調和は先行語末の子音と接辞頭の子音の間で起こる。子音・母音それぞれの調和性は変化形・変化要因欄に示した通りである。少数の外来語や、先行語が母音で終わる場合の後接はこれとは異なるが、大部分の語がこの規則により接合する。この規則性を辞書記述に含めることにより、解析の効率を向上することができる。

表 8　子音調和

見出し語	gha		
品詞	曲用助詞		
活用型	語頭子音母音調和型		
隣接形態素	出現形	変化形	変化要因
gul	gha	基本形	後母音・有声子音
jantaq	qa	後母音形	後母音・無声子音
yzym	gæ	有声子音形	前母音・有声子音
mæktæp	kæ	無声子音形	前母音・無声子音

4　おわりに

ウイグル語の自然言語処理に関する研究は極めて不足している。しかし、時代の発展は伝統的なウイグル語を自然言語処理技術レベルで実現することを求めている。そのために、伝統的なウイグル語文法で使用されている規則を自然言語処理技術に応用できる形で改良しなければならない。本稿で提案した母音調和の辞書記述は、茶筌などの形態素解析システムで容易に実装できる。ただし、茶筌では、形態素の接続を検査する際に品詞情報しか参照しないので、変化形・変化要因などの情報をいかにして接続規則に取り込むか工夫する必要がある。

参考文献

アブドレイムアブドハリリ. (2004). 現代ウイグル語の形態素解析. 修士論文, 千葉大学大学院文学研究科.

伝康晴・宇津呂武仁・山田篤・浅原正幸・松本裕治. (2002). 話し言葉研究に適した電子化辞書の設計. 第 2 回話し言葉の科学と工学ワークショップ講演予稿集 (pp. 39–46). 東京.

小川泰弘・マフスットムフタル・外山勝彦・稲垣康善. (1999). 日本語-ウイグル語機械翻訳における単語接続関係を用いたウイグル語文の生成方法. 言語処理学会第 5 回年次大会発表論文集 (pp. 454–457).

竹内和夫. (1991). 現代ウイグル語四週間. 東京: 大学書林.