タグ付きコーパスを用いたウイグル語テキストの文法間違い発見手法

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

1　はじめに

ウイグル語を初め、アルタイ諸語に属する言語には、形態素と形態素が連接する際に、母音と母音・子音と子音の組み合わせが一定の制約を受ける (竹内, 1991)。この制約により、一つの語幹に対して同じ意味を持つ幾つかの語尾の異形態が存在し、語幹に含まれる母音と子音のタイプにより連接可能な形態が選ばれる。さらに、選ばれた形態素の影響で語幹の音韻が変化することもある。ウイグル語の文字は表音文字であり、これらの音韻的違いは綴りにも影響する。

実際のテキストにおいて、方言や言語使用の差により、異形態選択や音韻変化の記述が間違っていることが少なくない。そのため、現在一般に使用されているウイグル語処理システムでは、以下の 2 つの方法を用いている。1 つ目は、語幹と語尾を別個に、音韻論的な情報を与えず辞書登録する方法である。この方法では、連接の制約がまったくないので、語幹と語尾・語尾と語尾の正しくない連接を許容してしまう。これはウイグル語への機械翻訳などで問題になる (UyghurEdit, n.d.)。2 つ目は、語尾を含めた派生形、あるいは、より長い文節の単位で辞書を構築する方法である。ウイグル語の派生形のバリエーションはかなり多く、この方法では辞書のサイズが大きくなりすぎて、解析速度が遅いという指摘がある (Uighursoft, n.d.)。

我々は、この現象に対して、異形態選択・音韻変化を扱える形態素解析システムを提案した (アブドハリリほか, 2005)。具体的には、音韻論的な属性を導入し、上記の現象を正しく反映したタグ付きコーパスを作成し、異形態選択・音韻変化規則を学習させた。本研究では、この形態素解析システムを利用して、異形態選択・音韻変化の誤用を発見する手法を提案する。

2　ウイグル語の形態論

2.1　音韻調和

ウイグル語を表記している文字は、母音と子音に分かれる。その中で、母音は調音位置によって前母音・後母音・中立母音に分かれ、子音は有声子音・無声子音に分かれる。形態素と形態素が連接する際に、前後の形態素の母音もしくは子音の組み合わせがこのグループによって制約を受ける。この現象を音韻調和と呼ぶ。

音韻調和により、先行する形態素のタイプに応じて、同じ意味を持つ後続形態素の異形態が 2 つあるいは 4 つ存在する。これらを 3 つのタイプに分けて説明する。

2.1.1　母音だけで選択可能なタイプ

母音の調和規則はかなり強く、外来語以外、前母音と後母音の混合ができない。例えば、名詞の複数形を表す語尾には “lar” と “lær” があるが、名詞末尾の母音に応じて以下のように使い分けられる。

(1) 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。従って、先行する形態素の最後の母音のタイプと連接する形態素の最初の母音のタイプを記述すれば選択できる。

2.1.2　子音だけで選択可能なタイプ

子音の場合も、先行する形態素の末尾子音と連接する形態素の先頭子音が同じグループに属する必要がある。

(2) a. asman + din = asmandin
　　空　格助詞　空から
　 b. dølæt + tin = dølættin
　　国　格助詞　国から

日本語の格助詞「から」に当たる “din” と “tin” がある。これらの母音は中立母音で、前母音とも後母音とも共起できる。従って、母音による選択は不可能である。子音を見てみると、それぞれ有声子音・無声子音で始まる。“asman” の末尾子音は有声なので、有声子音で始まる “din” が選ばれる。“dølæt” の場合は、末尾子音が無声なので、無声子音で始まる “tin” が選ばれる[*1]。従って、先行する形態素の最後の子音のタイプと連接する形態素の最初の子音のタイプを記述すれば選択できる。

2.1.3 母音と子音の組み合わせで選択可能なタイプ

母音あるいは子音だけでは、選択が不可能なタイプもある。例えば、方向を表す助詞（日本語の「に」）には、“gha”, “qa”, “gæ”, “kæ” の 4 つの異形態がある。このうち、“gha” と “qa” はともに後母音、“gæ” と “kæ” は前母音で綴られているので、母音だけでは選択できない。同様に、“gha” と “gæ” はともに有声子音、“qa” と “kæ” は無声子音で綴られているので、子音たけでも区別ができない。そのため、母音と子音のタイプを同時に考慮しなければならない。表 1 のように、母音に加えて子音のタイプにより後接形態素の形が決まる。

少し違うタイプもある。例えば、動詞の命令形を表す “ghin” には、“qin”, “gin”, “kin” の異形態がある。この 4 形態を見てみると、母音はすべて同じ中立母音で、子音も有声のものが 2 つ、無声のものが 2 つあるので、母音と子音を組み合わせても選択できない。このようなタイプを記述するため、事例をリストアップし特徴を調べた。その結果、“ghin” が後母音と有声子音、“gin” が前母音と有声子音、“qin” が後母音と無声子音、“kin” が前母音と無声子音で綴られた形態素に連接することがわかった。従って、“ghin” と “qin” に後母音タイプ、“kin” と “gin” に前母音タイプを与えれば “gha” の異形態と同じように扱える。

表 1　“gha” の異形態とその調和タイプ

先行形態素	異形態	調和タイプ
gul	gha	後母音・有声子音
yantaq	qa	後母音・無声子音
üzym	gæ	前母音・有声子音
mæktæp	kæ	前母音・無声子音

[*1] ただし、先行形態素が母音で終わる時は、つねに有声子音から始まる形態素が連接する。

表 2　音韻変化とその生起要因

変化する形態素	後続形態素	語末変化	生起要因
kæt	sæ	なし	
ket	ing	弱化	アクセント移動
bala	ng	なし	
bali	si	弱化	アクセント移動
jisim	gha	なし	
jism	i	脱落	後続狭母音
imla	da	なし	
imlay	im	挿入	後続狭母音

2.2 音韻変化

ウイグル語の音韻調和は連接形態素の異形態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。形態素と形態素が連接する際に、後続する形態素の影響で先行する形態素が強勢を受けたりすると、表 2 のように音韻変化（母音の弱化・脱落・子音の挿入）が生じる。詳細は (アブドハリリほか, 2005) 参照。

3 形態素解析システムによる実装

異形態選択・音韻変化を扱えるウイグル語形態素解析システムを (アブドハリリほか, 2005) で提案した。ここでは、その後の変更点を反映させた実装について、とくに異形態選択に関わる点を中心に説明する。

3.1 辞書記述

辞書記述の例を表 3 に挙げる。異形態選択を扱うために語頭変化型・変化形・変化結合型という属性を、音韻変化を扱うために語末変化型・変化形・変化結合型という属性を導入する。変化型・変化形はその形態素の形を記述するための属性であり、結合型は隣接する形態素に与える制約を記述するための属性である。

まず、“kør” は異形態を持たず、音韻変化も生じないので、変化型・変化形はいずれも空である。“kæt” は異形態は持たないが弱化を生じ、弱化形に “ket” がある。そこで、“kæt” と “ket” の語末変化型に「弱化・子音型」と記述し、前者の語末変化形には「基本形」、後者には「弱化形」と記述する。“bala” も同様に弱化形の “bali” を持つ。“kæt/ket” が子音で終わるのに対して、“bala/bali” は母音で終わる。前者は母音で始まる語尾から、後者は子音で始まる語尾から影響を受ける。弱化型の「子音型」「母音型」の区別はそのためである。

次に、異形態がある形態素を考える。異形態がある

表 3　辞書記述の例

出現形	辞書形	品詞	語頭 変化型	語頭 変化形	語頭変化 結合型	語末 変化型	語末 変化形	語末変化 結合型
kør	kør	動詞			前・有声			
kæt	kæt	動詞			前・無声	弱化・子音型	基本形	
ket	kæt	動詞			前・無声	弱化・子音型	弱化形	
bala	bala	名詞			後・有声	弱化・母音型	基本形	
bali	bala	名詞			後・有声	弱化・母音型	弱化形	
ghan	ghan	動詞語尾	調和型	後・有声	後・有声	弱化・子音型	基本形	母音型アクセント移動
ghin	ghan	動詞語尾	調和型	後・有声	後・有声	弱化・子音型	弱化形	母音型アクセント移動
qan	ghan	動詞語尾	調和型	後・無声	後・有声	弱化・子音型	基本形	母音型アクセント移動
qin	ghan	動詞語尾	調和型	後・無声	後・有声	弱化・子音型	弱化型	母音型アクセント移動
gæn	ghan	動詞語尾	調和型	前・有声	後・有声			
kæn	ghan	動詞語尾	調和型	前・無声	後・有声			

形態素には「調和型」という変化型を与える。例えば、“ghan” には、“qan”, “gæn”, “kæn” の異形態があり、この 4 形態はそれぞれ、後母音・有声子音、後母音・無声子音、前母音・有声子音、前母音・無声子音で始まる。そこで、これら母音・子音タイプを語頭変化形として記述する。さらに、このうちの “ghan” と “qan” には、後続する形態素によって弱化が生じる。そこで、この 2 形態素とそれぞれの弱化形である “ghin”, “qin” には、上述と同様にして語末変化型・変化形を記述する。

さらに、隣接形態素に制約を与える可能性のある形態素には変化結合型を記述する。例えば、名詞・動詞・語尾などは、後続形態素に異形態がある場合にその選択に際して末尾母音・子音のタイプが制約を与えるため、末尾母音・子音タイプを語頭変化結合型として記述する[*2]。同様に、“ghan” など一部の語尾は先行形態素の語末に変化を引き起こすので、語末変化結合型を記述する。

3.2　連接規則の学習

人手で形態素に分割し品詞を与えたコーパスに対して、上記の音韻論的属性を追記し、その連接関係を学習した。本研究で形態素解析に用いる『茶筌』では、連接関係の検査の際に、出現形と品詞しか利用できないため、音韻論的属性はすべて品詞の下位分類とみなした。ただし、これでは品詞が非常に細かくなり、データスパースネスの問題が生じる。そのため、学習時に品詞をグループ化する手法 (浅原・松本, 2002) を用いた。たとえば、語頭変化型・変化形は後続形態素に対しては検査する必要があるが、先行形態素に対しては検査する必要がない。そこで、連接規則の先行形態素としては、語頭変化型・変化形の違いを捨象して連接コストを学習する。これにより、十分に汎化された連接規則が学習される。

4　実験

前節の形態素解析システムがウイグル語テキスト中の異形態選択・音韻変化の誤用を発見するのに有効であることを実験によって検証する。

4.1　方法

■言語資料　形態素解析辞書の学習に用いた言語資料は、(ウイグル語国語教科書, 2003) から人手で入力したウイグル語テキスト 265 文（延べ語数 6791 語、異なり語数 1667 語）である。学習データに対する解析精度は再現率・精度ともに 99.8% であり、データの内的整合性は十分に高い。このうち、異形態選択・音韻変化処理の必要な形態素を少なくとも 1 つ含む文は 248 文（全体の 93.6%）あり、これらを実験対象とした。なお、語彙は全学習データから獲得し、未知語はないものとした。

■手続き　実験対象となる元データから異形態選択・音韻変化の誤用を含むデータを以下の手順で作成した。

- 語頭変化型を持つ形態素のうち、出現形が辞書形（もっとも誤用されやすい形）と異なるものについて、出現形を辞書形と同じ形にした。
- 語末変化型を持つ形態素のうち、出現形が辞書形（基本形）と異なるもの（音韻変化が生じているもの）について、出現形を辞書形と同じ形にした。

これにより、実際のウイグル語テキストで生じている可能性のある誤用をシミュレートした。

[*2] ただし、制約を与えるのは出現形ではなく辞書形である。

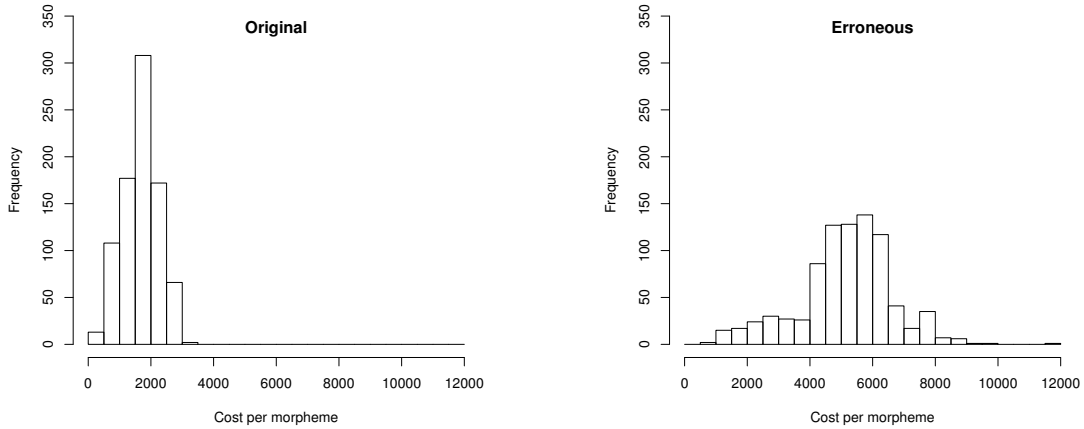

図 1　元データ（Original）と誤用データ（Erroneous）におけるブロックごとの（形態素あたり）コスト値の分布（$N = 846$）

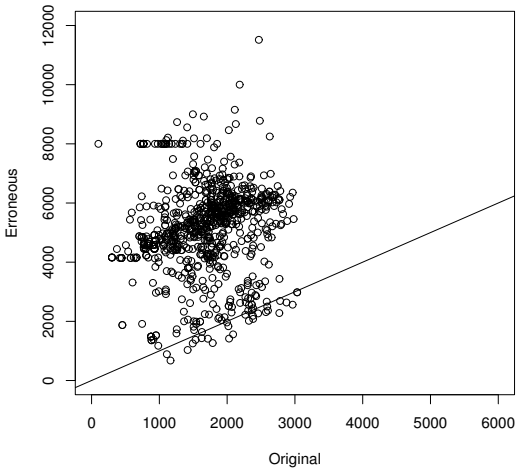

図 2　誤用に置き換える前後のブロックごとの（形態素あたり）コスト値の散布図（$N = 846$）

4.2　結果

元データと誤用データをそれぞれ『茶筌』で解析した。未定義連接コストは 10000 とした。元データと誤用データの形態素解析結果の差分（`diff` コマンドの出力）における各ブロックに対して、誤用に置き換える前と後の（形態素あたり）コスト値をそれぞれ算出した。コスト値に対する度数分布を図 1 に示す。誤用に置き換える前後のコスト値の平均はそれぞれ 1675.2（範囲: 100.5〜3033.0）と 5150.0（範囲: 675.0〜11516.0）であり、誤用に置き換えた後のほうがかなり大きかった。

各ブロックのコスト値の変化を散布図として示したのが図 2 である。直線は誤用への置き換え前後でコスト値が等しくなる点を示している。多くの点がこの直線の左上領域にあることから、個々の事例についても誤用への置き換えによりコスト値が増大するものが大半であることが確認できる。ただし、コスト値がほとんど変化しないものや、若干ながら減少するものも多少含まれる。

5　考察

実験結果から、正しい異形態選択・音韻変化を学習した形態素解析システムを利用することが、テキスト中での誤用を発見するのに有効であることがわかった。誤用を含む部分は一般にコスト値が高かった。かりにコスト値が 3000 以上であるか否かを判断基準として用いるとすると、846 件中 758 件（89.6%）の誤用を発見できる。

その一方で、10% の誤用は正用例として見逃されてしまう。これを避けるために閾値を引き下げることが考えられるが、図 1 から明らかなように、コスト値 3000 付近にはすでに多くの正用例が分布しており、これらの事例を誤用とする虚反応が多くなる。図 2 からわかるように、コスト値 3000 以下の誤用例の多くは正用例とコスト値があまり変わらないものである。これらの事例の特徴を分析し、対処策を考えることが今後の課題である。

参考文献

アブドレイム・アブドハリリ・伝康晴・土屋俊. (2005). ウイグル語形態素解析における母音調和の扱い. 言語処理学会第 11 回年次大会発表論文集 (pp. 787–790).

浅原正幸・松本裕治. (2002). 形態素解析のための拡張統計モデル. 情報処理学会論文誌, *43*, 685–695.

竹内和夫. (1991). 現代ウイグル語四週間. 大学書林.

ウイグル語国語教科書（第 9 版）. (2003). 新疆教育出版社.

Uighursoft. (n.d.). `http://www.uighursoft.com/`.

UyghurEdit. (n.d.). `http://kenjisoft.homelinux.com/uyghuredit/ukyindex.html`.