ウイグル語の複合語と文節の構造について

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

ウイグル語は類型論で膠着語に分類されており、言語表記としてアラビア文字を使用している。アラビア文字は表音文字であり、1 つの音を表す字母が、出現する位置によって独立形、語頭形、語中形、語末形から選択される。テキスト中でこれらの字形の違いにより語の境界が判別できるので、従来の正書法では分かち書きが存在しない。しかし、活字や計算機上でのテキストは分かち書きされている。この分かち書き単位の構造は複雑であり、一般流通辞書[1]の見出し語に一致する語と一致しない語がある。例えば、

(1) u　bëk　kilishkën　yegit　idi.
　u　bëk　kilishkën　yegit　idi.
　代名詞 副詞　形容詞　名詞　助動詞
　彼は　とても ハンサム　男性　だ

(2) tünügün　ata – anilar　yeghini　echildi.
　tünügün　ata – ani + lar　yeghin + i
　副詞　名詞　複数語尾　名詞　三人称
　昨日　両親　たちの　会議が
　ech + il + di.
　動詞　受身形　助動詞
　開く　られ　た

(1)では、文がスペースによって 5 つの語に分解されている。この 5 つの語と一般辞書の見出し語を照らし合わせて見ると完全に一致する。(2)では、文がスペースによって 5 つの語と 1 つの記号に分解されている。スペースにより分解された単位と一般辞書の見出し語を照らし合わせて見ると、この 5 つの語の中で、tünügün「昨日」と ata「父」が一般辞書の見出し語に一致する。しかし、文全体から見ると、ata は ata – ani (ata – ana)「両親」という語の一部である。外見で、ata – ana が ata「父」（ハイフン）ana「母」の形で別々の語のように見えても、分解が出来ない結合語である。

分かち書き単位での語の認定はいろいろな問題を抱えるにも関わらず、現在、計算機によるウイグル語の自動処理などで分かち書きが注目を集めている。すなわち、まず文を一旦分かち書きの単位で分解して、それから分かち書き単位の「語」を解析する研究が行われている[1]。これに対して我々は、ウイグル語の言語学的な特徴と正書法の特徴に着目し、ウイグル語における分かち書きを語の分解単位ではないということを前提にして、ウイグル語の形態素解析方法を提案した[2][3]。

本研究では、形態素解析により得られた結果と分かち書き単位で得られた結果を比較し、辞書の見出し語が実際のテキストでどれぐらいの割合で出現するかを検討する。また、辞書の見出し語に一致しない語の構造も明らかにし、分かち書き単位での処理の欠点を示す。そして、ウイグル語の機械処理に有効な語の単位について考える。

2. ウイグル語の語の構造

現代ウイグル語で使われている語は、語構造に基づいて単一語、合成語、省略語にわけることができる。この 3 種類の語を以下に例文により詳しく説明してゆく。

[1]一般流通辞書とは市販されている人間用の辞書を指す。この原稿では一般辞書と呼ぶ。

2.1 単一語（tüp söz）

自立語の中で意味を持つ最小の単位である。例えば、su 「水」、 kök「青い」、ilgiri「以前)」等。

2.2 合成語（yasalma söz）

2 つ以上の自立語の結合や自立語に派生接辞が付くことにより新しい語が形成される。合成語の構造はかなり複雑である。以下で、例文を取り上げて、合成語の形成パターンを説明する。

(3) közëynëk = köz + ëynëk
眼鏡　　眼　ガラス

(4) ata – bala = ata + bala
親子　　父　子供

(5) miwë – chiwë = miwë + chiwë
果物　　果物　(意味がない語)

(6) arwang – sarwang = arwang + sarwang
ごちゃごちゃ　　(意味がない語)

(7) tokyo uniwersiti = Tokyo　uniwersiti
東京大学　　東京　大学

(8) tilshunas = til +　shunash
言語学者　言語　接辞（学者の意味）

(9) biëdëp = bi　+　ëdëp
不徳　否定接頭辞　徳

(3)(4)は、2 つの自立語から構成される複合語であるが、意味的に 1 つの対象を指しているので、分解することが出来ない。(5)(6)は、語の間にスペースとハイフンが入って分かち書きされているが、これらも意味的に 1 つの対象を指しているし、分解した単位で意味を持たない接辞のような語から成り立っている。(7)は、2 つの自立語から形成された複合語であり、分かち書きされている。内容的には 1 つの対象を指しているが、別々の語として考えても問題はない。(8)(9)は自立語の語幹に派生接辞が付くことにより形成された語である。ウイグル語の派生接辞は数が多く 227 個にも達する。これらの接辞が、単一語の語幹に次々と付くことにより第二語幹、第三語幹を派生させる。例えば、(8)の tilshunas という語に、名詞から名詞、名詞から形容詞、あるいは、形容詞から名詞を派生させる派生接辞 liq が付くと tilshunasliq「言語学」 になる。(9)の ëdëp に派生接尾辞 siz を付けることによって、ëdëpsiz「不徳」という新しい語幹が作られる。biëdëp と ëdëpsiz が同じ意味を表しているが、連接する語が異なる。すなわち、ëdëpsiz に派生接辞 lik を付けることによりまた新しい語幹 ëdëpsizlik「不徳なこと」になる。biëdëp には接尾辞が連接しない。

2.3 省略語（qisqartilima söz）

語あるいは文節を含めた各語の第 1 文字をとる形で語が形成される。人名は例外であり、名前の最初の 1 文字と姓を省略せずにそのままで、間にドットを入れて書く。

(10) شنجاڭ ئۇيغۇر ئاپتونۇم رايۇنى= ش ئۇ ئا ر
新疆ウイグル自治区

(11) بىرلەشكەن دۆلەتلەر تەشكىلاتى= ب د ت
国連

(12) ئابدۇرېھىم ئوتكۇر = ئا . ئوتكۇر
アブドレヒミ　オトクリ

以上の例で示したように、ウイグル語の語構造は複雑であり、分かち書きの単位で処理を行った場合は、(4)(5)(6)の複合語と(10)(11)(12)の省略語も分解されてしまう。また、(8)(9)の派生語の処理も問題になる。すなわち、派生接辞により派生された語に付く接辞の数が多くて、すべての派生形を辞書に登録するのは困難である。基本的な派生形を新しい語幹として辞書に登録したとしても、テキスト中でさらに文法的な役割を果たす接辞[2]が次々と後接する。これらの派生形をすべて辞書登録することはできず、派生接辞と文法的な役割を果たす接辞の処理が必要になる。

次節では、実際のウイグル語テキストを使用し、自立語に付く派生接辞と文法的な役割を果たす接辞の構造を明らかにする。

---

2 日本語の動詞の活用語尾や助動詞や助詞に似た形態素を指す。

３．分析

3.1 形態素解析

テキストにおける語の構造を明らかにするために、形態素解析器を使用しウイグル語のテキストを形態素に分割する。汎用形態素解析器として開発された ChaSen と Mecab にウイグル語の辞書を追加したものを用いた[2][3]。ウイグル語の語の形成規則は複雑なので、辞書の構築について簡単に説明する。

辞書を作成する際に以下の方法で見出し語を認定した。前節で説明した 3 種類の語の中で、単一語と省略語はそのまま使用した。合成語は、(3)(4)(5)(6)の語形を見出し語として登録した。

派生接辞から形成される語の登録は困難な問題である。現在市販されているウイグル語の辞書を見てみると、派生語の登録状況は、出版社や辞書のサイズにより異なる。すわなち、辞書によって使用頻度が高い語の派生形を登録したり、あるいは、意味を基準にして、派生接辞により品詞と意味が変わる語だけを登録したりしている。我々は、ウイグル語国語辞書[4]に登録されている派生語を分析し、生産性・品詞・意味を考慮し、227 個の派生接辞から 47 個の接辞を別項の形態素として登録した。その他の文法的な役割を果たす語も別項の形態素として登録した。

3.2 言語資料

言語資料としてウイグル語の国語教科書・短編小説（Teghritagh Zhurnali、2005年）・詩・インターネットニュースから1878文を選んで、形態素解析を行った。これらは、記号等も含めて述べ数で62944個の形態素に分解された。異なり形態素の数は4484個である。この中で自立語は4201個で、接辞類の形態素は283個である。

3.3 分析方法

分解した形態素から*n*-gramを抽出した。ウイグル語の正書法では、自立語の前にスペースが入るため2つの自立語の間を1つの単位にまとめることができる。ただし、少数の助動詞は語幹に付かず、スペースによって区切って書かれる。また、後置詞の多数は、語幹に付かず、スペースによって区切って書かれる。このような語に関しては、自立語と同様に扱った。

3.4 分析結果

N-gram を用いた語の構造は表 1 の通りである。表 1 の*N*の値は、単位を構成する形態素の数を示す。すなわち、最後の接辞から語幹あるいは接頭辞までの形態素の数を表す（つねに単位内の最後の接辞から数え部分列の *n*-gram は考えない)。ここでの接辞は、派生接辞と文法的な役割を果たす接辞を合わせたものである。この分析データでは、第一の語幹に最も長いとき 6 個の接辞が付いた。表 2 は、各グラムにおける品詞別の出現頻度を表している。$N = 1$ のとき、すなわち、語の辞書形では、名詞の出現頻度が一番高いのに対して、接辞が付くのは動詞が一番多い。$N = 2, 3, 4$ の語形は動詞、名詞、形容詞で同じく高くなっており、その次のグラムでは急に減少している。

4. 考察

この結果から、ウイグル語のテキストにおける分かち書きを語の境界にした場合、半分近くの語が一般辞書の見出し語に当てはまることが分かる。すなわち、$N = 1$ は自立語に何も接辞が付いていない状態であり、全形態素の 43%が一般辞書の見出し語に一致する。$N = 2$ から $N = 7$ の割合が 57%占めており、何らかの処理によって辞書形を認定する必要がある。

表 2 で示したように、品詞別で見ると、動詞が単独で出現する頻度は少なく、名詞と形容詞の頻度が高いことが分かる。このことから、ウイグル語は、同じ膠着語である日本語よりも強い膠着性の特徴を持っているとも言える。

5. おわりに

今回使用した言語資料の量は少ない。ジャンル別で見ると国語教科書・雑誌・詩・ニュースであり、国語教科書と詩では、基本的に簡単な言葉が使われている。分析データの量とジャンルを変えると結果が変わるかも知れない。すなわち、論文や長編小説などに出現する難しい語では、もっと長い接辞の連接が生じる可能性がある。今後、データの量とジャンルを増やして検討したい。

参考文献

[1] Bliqiz Muhemmedniyaz. (2005). Uyghur tili til materiyal ambirdiki sozlarni aptumatik statstika qilishta uchraydighan bir qanche muhim mesile tughrisida. *Language and Translation*, *2*, 18-28.

[2] 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).

[3] アブドレイム・アブドハリリ・伝康晴・土屋俊. (2006). 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. 言語処理学会第12回年次大会発表論文集(pp.188–191).

[4] Yaqup, A. (1999). *Uyghur tilning izahliq lughiti*. Urumchi: Shinjang Heliq Neshiryati.

表１　*N*-gram による語構造

各グラムにおける例文（頻度が高い順）		頻度	述べ数 29715	割合	異なり数 10072	割合
N= 1	bir	408	12463	43%	2170	22%
N= 2	dë+ p	148	9503	32%	3016	30%
N= 3	bol+i+du	46	5792	19%	3317	32%
N= 4	til+shunas+liq+i	29	1551	5%	1207	12%
N= 5	qara+sh+lir+im+ni	5	320	1%	281	3%
N= 6	bol+ma+ydighan+liq+i+ni	3	76	0%	72	1%
N= 7	izdi+n+iwat+qan+liq+i+ni	1	9	0%	9	0%

表２　品詞別頻度

n-gram / 品詞	*N*= 1	*N*= 2	*N*= 3	*N*= 4	*N*= 5	*N*= 6	*N*= 7
動詞	26	3314	2733	976	265	72	9
名詞	7721	5508	2865	502	53	4	
形容詞	2486	443	175	70	2		
副詞	836	170	12	3			
感動詞	44	2					
擬態詞	8	1					
接続詞	1008	18					
後置詞	334	47	7				