ウイグル語接辞の頻度について

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

ウイグル語は言語学的に、日本語と同じく膠着語に分類されており、語幹に接頭辞・接尾辞がくっつくことによって語が形成される。正書法では、アラビア式のウイグル文字を使用しており、文が空白によって分かち書きされる。しかし、分かち書きされた単位は形態素と一致しないことが多い。そのために、現在存在するウイグル語処理システムでは、語の認定が多く異なっている。例えば、空白で区切られた文字列全体を一つの語として扱っていることもある (UyghurEdit, n.d.)。この方法では、自立語に後接する多くの接辞を語の一部として扱うことになり、一つの語の多数の形を辞書に記述することになる。これに対して、我々はウイグル語の膠着言語であるという特徴を前提にし、伝統的なウイグル語で用いられている派生文法規則に基づいて語を認定している (アブドハリリ, 2004)。そして、ウイグル語と日本語の言語学的な特徴と正書法の類似性に着目し、ウイグル語における分かち書きの単位を無視し、日本語の形態素解析システム「茶筌」を利用して語を同定している。

ウイグル語では音韻調和規則等により、接頭辞・接辞辞は種類が多く、テキストにおける出現形も複雑である (アブドハリリ・伝・土屋, 2005, 2006)。すなわち、語幹に接辞が連接することにより語の一部が変わったら、音韻規則の制約を受けたりする。さらに、ある名詞の語幹に接尾辞が付くことにより形容詞になったり、さらにこの語に新しい接尾辞が付くことにより名詞になったりする。そのため、形態素解析用の辞書を作成する際にどの単位まで一つの語にするか、品詞を基準にするか、あるいは意味を基準にするかという複雑な問題に直面する。さらに、語を認定する際に細か過ぎても問題がある。ウイグル文字は日本語のように漢字、ひらがな、カタカナなど有力な特徴を持っていないので、未知語の処理も困難であると考えられる。

本論文では、ウイグル語テキストにおける各接尾辞の出現頻度を計算し、語を形成する際の接辞の重要さと接辞の取り扱いについて検討を行なう。

2　形態素解析における語の認定の問題

2.1　分かち書き単位での語の認定

ウイグル語の文は、空白によって分かち書きされているが、２つの空白の間の単位が一つの形態素から成り立つ場合だけでなく、複数の形態素から成り立つ場合もある。基本的な構造は以下の通りである。動詞の場合は、「動詞語幹＋派生語尾＋助動詞＋人称助詞・格助詞＋疑問を表す接辞」の形をとる。名詞の場合は、「接頭辞＋語幹＋複数語尾＋人称助詞（限定語尾）＋格助詞」の形をとる。

UyghurEdit (n.d.) の研究で、以上のパターンに基づいて一つの動詞を派生させると 2014 個、名詞の場合は 200 個以上になることが報告されている。この数はあくまでも論理的なものであり、この中でどれぐらいの物が実際のテキストで出現するかを試したことはまたない。このことから、空白単位の語の認定が、ウイグル語の特徴に不適切なものであると考えられる。

2.2　語幹を中心した語の認定

ウイグル語は日本語と同じく膠着語に分類されているが、日本語と違って活用の概念を用いていない。語の形成規則として派生文法を用いている。派生文法というのは、音韻規則に基づいて語幹に接辞をつけることにより新しい語を形成する方法を示す。例えば、“yaz” (書く) という動詞の第二語幹に派生する形は以下の通りである。

この第二語幹をさらに派生させることができる。語幹を語として認定する方法は、第一の語幹を自立語として記述し、他の派生語尾を接辞として記述する方法である。こうしたら自立語の部分が単純になるが、接辞の数が多くなり機能語が複雑になってしまう。ウイグル語

表 1　動詞の形成規則

yaz	基本形	yaz + ghuche	連用比較形
yaz + di	完了形	yaz + mas	中止未完了形
yaz + ma	否定形	yaz + ala	可能形
yaz + a	三人称願望形	yaz + sun	三人称命令形
yaz + sa	条件形	yez + iwal	連用状態形
yaz + ay	意志形	yez + ip	連体中止形
yaz + ghan	連体完了形	yez + il	受身形
yaz + dur	使役形	yez + ish	共同形
yaz + ghin	二人称願望形	yez + iwat	連体未完了
yaz + ghach	方向-理由形	yez + iwer	連用状態形
yaz + ghili	連用目的形	yez + ing	二人称命令形

正書法辞書によると現代ウイグル語で一般に使われている語５万（専門用語を含むと７万）語と言われている (Yaqup, 1999)。この５万語の中に多くの派生語が入っている。

3　語の形成規則

ウイグル語も他の言語と同じく自立語と付属語に分けられている。ウイグル語の付属語は役割により２つのグループに分けられる。１つ目は、新しい語を派生させる接辞 (Söz yasighuchi qoshumchæ)、２つ目は、文法的な役割を果たす接辞 (Söz türlighüchi qoshumchæ) である。

3.1　派生接辞

派生接辞というのは、語幹に付くことによって新しい語を派生させることを示す。現在の国語教科書で使われている派生規則により、ウイグル語における派生語尾と語幹の連接するパターンを以下のようなパターンにまとめることができる。

(1) bash (頭・首長) + liq = bashliq (首長)
　　名詞　　派生語尾　名詞

(2) æqil (知恵) + liq = æqilliq (聡明な)
　　名詞　　派生語尾　形容詞

(3) bash (頭) + siz = bashsiz（首長の不在)
　　名詞　　派生語尾　形容詞

(4) æqil (知恵) + siz = æqilsiz (愚鈍な)
　　名詞　　形容化語尾　形容詞
　　+ liq = æqilsizliq (愚鈍さ)
　　派生語尾　名詞

例文 (1) のように語幹に派生語尾が付くことにより品詞が変わらず、語を形成することもある。例文 (2) のように例文 (1) と同じ派生語尾が付くことにより品詞が別の語に派生させることもある。例文 (3) のように特定の品詞だけに付くことにより語を派生させる接辞がある。例文 (4) の場合は一段と複雑であり、一つの語幹から品詞が別の語に派生させたら、その次また、派生語尾が付くことによりさらに新しい語を派生させることができる。以上のような派生語尾とされているものの数は 104 であり、音韻調和規則に適用した異形態を合わせたら 227 個もある。

3.2　文法的な役割を果たす接辞

文法役割を果たす接辞は、日本語の文法で使われて「機能語」に想定するものである。ウイグル語の文法的な役割を果たす接辞も多数存在し、日本語で存在していない人称助詞や名詞の単数形・複数形を表す助詞がある。例えば、

(5) kitab (本) + lar + ning (の)
　　語幹　　複数　格助詞
　　= kitablarning (たくさんの本の)

(6) kitab (本) + lar + im (私の)
　　語幹　　複数　人称助詞
　　= kitablarim (私のたくさんの本)

以上のような文法的な役割を果たす接辞の数は 33 個であり、異形態を含むと 112 個にもなる。

3.3　接辞の曖昧さ

ウイグル語の付属語が以上で説明したように分けれているが、派生接辞と文法的な役割を果たす接辞の間に曖昧な点も少なくない。例えば、派生語尾と文法的な役割を果たす接辞の境界ははっきりしてない。すなわち、一般に、接辞の連接によって語幹の基本的な意味が変わらないまま語が形成されるとその連接した接辞は文法的な役割を果たす接辞であると言われている。しかし、例文 (1) を見ると語幹に派生語尾とされている ”liq” が連接することにより新しい語が形成さているが、意味があまり変わらないにも関わらず、派生語尾とされている。しかし、これは例文 (5) の複数形助詞や、(6) 人称助詞とあまりかわらない役割を果たしている。そのため、本研究では、学校文法で使われている規則を基にして語の認定を行なう。すなわち、語の意味と第一の語幹を考慮した上で、形態素解析を前提し、第一の語幹を中心する方向で辞書を記述する。

4 分析

4.1 形態素解析について

最近、ウイグル語を表記しているアラビア式のウイグル文字（UEY）が Unicode のアラビア語領域にエンコードされたことと Windows Vista にウイグル語フォントとウイグル語 IME が追加されたことを受け、1999 年から議論になってきたコンピュータにおけるウイグル文字の処理問題が決着した。そして電子化されたテキストも増えつつある。

数年前から我々は、日本語の形態素解析システム「茶筌」を使ってウイグル語の形態素解析の設計について努力しているが、十分な電子化された言語資料がないという理由により実用的なウイグル語の形態素解析システムを完成するまで至っていない。

今回形態素解析システムとして汎用形態解析素システム Mecab を利用した。Mecab を利用した理由は、以下である。ウイグル語は日本語と異なって活用の概念を利用してないので、活用の処理を行なう必要がない。しかし、ウイグル語では日本語に存在していない音韻調和・音韻変換現象がある。我々は、(アブドハリリほか, 2005, 2006) でこの現象を扱う方法を提案したが、以前我々が使用した形態素解析システム「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、音韻論的属性をすべて品詞の下位分類とみなした。このため、品詞が細かくなり 1932 個にもなった。品詞が細かくなるとデータスパースネスの問題が生じる。細い品詞階層と粗い品詞階層の確率値を混ぜるといった処理について「茶筌」よりも Mecab の方が効率的で, また少ない学習データでも高い性能で学習ができるといる報告がある (Mecab, n.d.). また、「茶筌」では未知語処理もハードコーディングされており自由に定義することはできないのに対して、Mecab ではユーザが未知語処理を自由に定義することができる。そのために Mecab を利用した。

4.2 方法

ウイグル語の国語教科書と雑誌（Teghritagh Zhurnali2005 年）から 1066 文を人手で形態素に分解し品詞情報を与えた。形態素に分解する際に、第一の語幹を中心にして語の認定を行なった. その結果延べ語数は 26595 個、異なり語数は 2723 個になった。この中で接辞の数は 264 個であり、テキストに出現するほぼすべての接辞が入っていると予測した。辞書やコーパスをアラビア文字式のウイグル文字で入力し、Mecab で学習を行ない、ウイグル語の形態素解析システムを設計した。その次に、設計したウイグル語解析システムを用いて、解析システムの性能と接辞の頻度をシミュレートをするために、Web ページと雑誌（Teghritagh Zhurnali 2005 年）から、データを集めた。集めたデータは短編小説である。

4.3 結果

設計したウイグル語形態素解析システムを用いて解析を行なって以下のような結果が得られた。分析用コーパス 4880 文、175055 個の形態素に分解された。この中には未知語や記号等も含む。解析で使用されたコーパスはタグを与えていないコーパスであるために、この中でどれぐらいの語が正しく解析されているかをこのままで同定できない。そのため、uniq コマンドを使って解析データの中から異なり語を抽出した。異なり語は未知語を含んで 1790 個しかなかった。この数は学習コーパス中の異なり語数よりも少ない。各語の出現頻度をみて見ると自立語の中で動詞、指示代名詞、数詞の頻度が高いことが分かった。例えば、動詞の “bol” (なる・できる) が 2420 回、指示代名詞の “u” (彼・彼女) が 1715 回出現している。

接辞の出現総数は 86510 回であり、全形態素の 33% を占めた。以下の表で出現頻度が高い接辞を示す。表で示したのは異形態を一つにまとめたものである。たとえば、動詞の中止形 “p”, “ip”, “up”, “ü p” を “p” としてまとめた。

表で示した分析比率というのは、分析用コーパスの述べ語数の中での出現比率を表したものであり、学習比率というのは学習コーパスの述べ語数の中で出現比率を表したものである。分析比率と学習比率を比較して見ると、学習データと分析用コーパスの内容が同じにも関わらず、一部の接辞の出現頻度が大きく異なっている。特に、接辞の単位が短いもので、分析比率の方が高くなっているのが多い。これらは誤解析を含んでいる可能性が高い。

5 考察

分析結果から見ると出力された異なり語の数が少ないことが分かる。特に未知語の数が少なかった。分析用コーパスは学習コーパスより 4 倍ぐらい大きいにも関わらず、出力語の数が少なかったことは、未知語として出

表 2　接辞の出現頻度

品詞	接辞	頻度	分析比率 (%)	学習比率 (%)
人称助詞	i	10363	5.92	4.66
複数形	lar	5119	2.92	2.66
動詞語尾	p	4989	2.85	3.28
格助詞	gha	4366	2.49	2.62
動詞語尾	a	4097	2.34	1.05
動詞語尾	sh	3907	2.23	2.02
格助詞	ning	3878	2.22	2.27
動詞語尾	ghan	3853	2.20	2.41
格助詞	ni	3603	2.06	2.36
格助詞	da	3453	1.97	1.76
助動詞	di	2341	1.34	1.42
語尾一般	lik	2300	1.31	1.53
動詞語尾	ma	2220	1.27	0.91
動詞語尾	y	2166	1.24	0.86
助動詞	du	2095	1.20	1.22
格助詞	din	1983	1.13	1.26
動詞語尾	n	1955	1.12	0.45
人称助詞	m	1758	1.00	1.85
動詞語尾	l	1514	0.86	0.41
動詞語尾	t	1376	0.79	0.21
動詞語尾	r	1349	0.77	0.05
人称助詞	q	1246	0.71	0.11
後置詞	mu	1234	0.70	0.73
動詞語尾	sa	1198	0.68	0.62
後置詞	la	770	0.44	0.38
格助詞	diki	767	0.44	0.51
動詞語尾	dighan	686	0.39	0.42
動詞語尾	ala	612	0.35	0.23
動詞語尾	dur	545	0.31	0.28
人称助詞	ng	534	0.31	0.11
助動詞	dur	496	0.28	0.08
名詞語尾	chi	454	0.26	0.22
動詞語尾	ghu	429	0.25	0.14
人称助詞	miz	427	0.24	0.35
格助詞	dek	424	0.24	0.31
形容詞語尾	ri	418	0.24	0.01
後置詞	ghu	416	0.24	0.03
動詞語尾	ra	400	0.23	0.00
動詞語尾	wat	387	0.22	0.32
助動詞	tti	377	0.22	0.30
助動詞	idi	360	0.21	0.32
形容詞化	iy	358	0.20	0.18
形容詞化	qi	331	0.19	0.10
動詞語尾	ar	327	0.19	0.07
後置詞	üchün	322	0.18	0.23

力されるはずのところが、出現頻度が高い既知語として出力されたことが考えられる。実際一つの未知語として出力されると期待したところが、細かい単位で分解されてしまったことが多かった。表 2 の分析から短い接辞での誤解析が多いと思われ、これらが未知語の部分文字列と一致してしまっているのではないかと思われる。

この結果から、ウイグル文字のような表音文字の場合は、辞書登録する語の単位が細か過ぎると未知語の処理等で問題が生じることが分かった。対策として、出現頻度が高く、短い単位の接辞を語幹に結合して扱うことが考えられる。しかし、2.1 で述べたように、接辞を語幹にくっつけた形で辞書登録すると、辞書のサイズが急激に大きくなる。解析精度と辞書サイズのバランスをとりながら作業を進める必要がある。今後、データを増やして、いろいろな側面から分析することで有効な対策を考えていきたい。

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.

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).

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).

Mecab. (n.d.). http://mecab.sourceforge.net/.

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

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