

日本語-ウズベク語機械翻訳

小川 泰弘 福田 ムフタル† 外山 勝彦

名古屋大学大学院情報科学研究科 † 名古屋大学大学院国際開発研究科

yasuhiro@is.nagoya-u.ac.jp

1 はじめに

我々はこれまでに、日本語-ウイグル語翻訳システムの開発を進めてきた [1][2]. 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によって、ある程度の翻訳が可能である.

一方、動詞句については、日本語の場合は活用し、ウイグル語の場合は活用しないという違いがある. しかし、日本語の膠着語としての性質に着目した派生文法 [3] では、活用概念を用いることなく日本語の語形変化を表現している. そこで我々は、日本語だけでなく、ウイグル語も派生文法で記述することによって、複雑な動詞句も逐語的に翻訳可能とした [1].

なお、この翻訳システムでは、派生文法に基づく形態素解析システム MAJO [4] の辞書を、日本語-ウイグル語対訳辞書に置き換えたものを使用している. 派生文法は日本語・ウイグル語以外の膠着語にも適用可能であり、MAJO の辞書を置き換えることにより、他の膠着語への機械翻訳にも応用できる. そうした観点から、現在、我々は日本語-ウズベク語翻訳への応用を進めている. すなわち、市販されているウズベク語-日本語辞書 [5] を電子化し、その逆辞書を MAJO で用いることによる翻訳システムの実現を目指している.

ところで、ウイグル語とウズベク語は使用されている地域が近く、文法だけでなく語彙に関しても類似点が多い. また、ウイグル語とウズベク語は、いずれも計算機で利用可能な言語資源が豊富とは言えない.

そこで、本研究では、ウイグル語とウズベク語の類似点に注目し、日本語との対訳辞書を利用することにより、その類似性を明らかにする. 最終的には、言語間の類似点を利用することにより、対訳辞書の拡充やウイグル語-ウズベク語機械翻訳への応用を図る.

なお、本稿では日本語の表現は「」, ウイグル語およびウズベク語の表現は“”で括り、それらを区別する.

2 日本語-ウイグル語機械翻訳

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、構文解析が不要となる. 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である (図 1).

ここで、図 1 から、「を」と“ni”のような格助詞間や、「た」と“dim”のような動詞接尾辞間にもそれぞれ対応関係があることが分かる.

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 [3] 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」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.

我々に研究により、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ことが明らかになった [1]. それにより、両言語間の接尾辞の対応関係が明確になり、

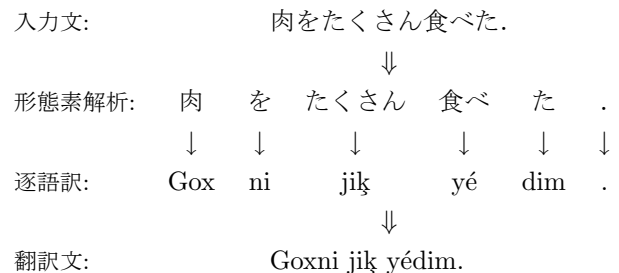


図 1: 日本語-ウイグル語逐語翻訳

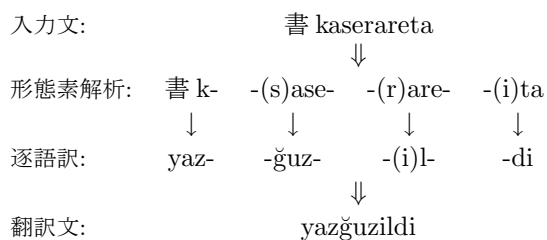


図 2: 派生文法に基づく動詞句の翻訳

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句も逐語訳で翻訳が可能となった(図 2)。

もちろん、日本語とウイグル語の間には相違点もある。例えば、音韻変化規則に関しては、「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」となるといった日本語独自の規則がある。一方、ウイグル語には円唇同化や、後述する母音調和などの日本語にはない種々の音韻変化規則が存在する。また、日本語の格助詞とウイグル語の格助詞の間に 1 対 1 の対応関係が成り立たない場合もある。

我々は、そうした問題についても、逐語訳の結果に後処理を加えることによって解決してきた [1][2][7]。今後は、より多くの日本語文をウイグル語に翻訳する実験を通じ、翻訳失敗の例を集めてシステムを改良していくとともに、他の言語への応用を考える。そして、最初の応用として、ウイグル語と比較的近い関係にあるウズベク語への日本語からの翻訳を検討する。

3 日本語-ウイグル語機械翻訳から 日本語-ウズベク語機械翻訳へ

ウズベク語はウズベキスタンの公用語で、ウイグル語と同様、言語学的にはアルタイ語族チュルク諸語南東語群に属する語である。よって、ウイグル語とウズベク語は文法的に類似するだけでなく、語彙的にも類似点が多い。そのため、両言語の話者間では、通訳なしで意志疎通が可能である。

よって、我々が開発してきた日本語-ウイグル語機械翻訳システムを利用すれば、少ないコストで日本語-ウズベク語機械翻訳システムが実現可能と考えられる。我々の日本語-ウイグル語機械翻訳システムは、日本語形態素解析システム MAJO の辞書を日本語-ウイグル語辞書に変更したものを使用している。そこで、この辞書を日本語-ウズベク語辞書に変更すれば、基本的なシステムが実現できると考えられる。

しかしながら、電子化された日本語-ウズベク語辞書はいまだ存在しない。印刷物としても、ウズベキスタン国内での学習者向けに小規模なものが存在するだけである。一方、ウズベク語-日本語辞書に関しては、印刷物として出版された文献 [5] が存在するが、電子化されたものは存在しない。

実は、我々が日本語-ウイグル語機械翻訳の研究をはじめた当初も状況は同じであった。その際、我々は市販のウイグル語-日本語辞書 [6] を電子化し、逆辞書を生成するという手法によって、半自動的に日本語-ウイグル語電子化辞書を作成した [8]。

よって、今回も同様に、ウズベク語-日本語辞書 [5] を電子化し、そこから逆辞書を作成することにより、日本語-ウズベク語電子化辞書を作成する。一方で、動詞接尾辞や格助詞などの接尾辞に関しては、こうした辞書には掲載されていないため、ウズベク語の文法書 [9]などを参考に人手で追加し、基本的な日本語-ウズベク語翻訳システムの実現を目指す。

しかし、ウイグル語とウズベク語に関しては言語資源が足りないという問題がある。そこで次章では、ウイグル語とウズベク語の類似点を利用し、互いを補強することを考える。

4 ウイグル語-ウズベク語変換

4.1 ウイグル語とウズベク語の相似点・相違点

表 1 にウイグル語とウズベク語の対応を示す。“bulut”, “fontan”, “yiring”, “imperator” の 4 単語については、まったく同じ形であり、それ以外に語についても、良く似ていることが分かる。このように、両言語間には語彙的な類似点が多い。しかしながら、もちろん相違点もあり、その原因は以下の三つと考えられる。

なお、以下では〈日本語, ウイグル語, ウズベク語〉の 3 項組で対応する単語を記述する。

4.1.1 表記方法の相違

ウズベク語のアルファベットは、ソ連時代はキリル文字を拡張したものが使用されてきたが、ウズベキスタンの独立に伴ない、現在はローマ字表記を拡張したものが制定されている。その対応を表 2 に示す。

このローマ字表記は、我々日本人が直観的に想起する発音と近いものがあるが、必ずしも一致している訳ではない。また、キリル文字の ш がローマ字表記では

表 1: ウイグル語とウズベク語の対応

日本語	ウイグル語	ウズベク語
雲	bulut	bulut
噴水	fontan	fontan
膿	yiring	yiring
皇帝	imperator	imperator
恐怖	déhxét	dahshat
頭	bax	bosh
ある	bar	bor
花	gül	gul
遺伝	irsiyét	irsiyat
大臣	wézir	vazir
魚	belik	baliq
家	öy	uy
雇う	yallimaq	yollamoq
育てる	östürmek	ochirmoq

sh になるなど、キリル文字では 1 文字であったものが、ローマ字では 2 文字になるなどの違いがある。

一方、ウイグル語では、現在はアラビア文字が利用されているが、対応するローマ字表記もある。ただし、その表記には統一されたものがなく、表 1 中のウイグル語の表記はその一例であり、他の表記方法も存在する。このローマ字表記も我々が想起する発音に近いものがあるが、ウズベク語のそれと一致している訳ではない。例えば、〈恐怖, déhxét, dahshat〉や〈頭, bax, bosh〉の対応を見ると、先程のウズベク語のローマ字表記における sh(キリル文字ш)が、このウイグル語の表記では、x で表されていることが分かる。このように、発音が同じであると推測される場合でも表記が異なる場合がある。

さらにウズベク語では、発音と表記が必ずしも一致している訳ではない。例えば、〈ある, bar, bor〉におけるウズベク語 “bor” は、日本語で表記すれば「ボル」に近い発音になるが、〈頭, bax, bosh〉における “bosh” は「バッシュ」に近い発音となる。

4.1.2 発音の相違

ウイグル語とウズベク語の間では、表記だけではなく、発音そのものが異なる場合もある。特に、ウズベク語には母音調和がない点が大きな差異である。

母音調和とはアルタイ諸語やウラル諸語に見られる現象で、単語に現れる母音の組合せが一定になることである。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。それらは、舌の位置の違いから、前母音 é, ü, ö と後母音¹a, 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。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。表 1 における「恐怖」“déhxét” や「雇う」“yallimaq” がその例である。

しかしながら、ウズベク語には前母音・後母音の区別が(少なくとも表記上は)存在せず、母音の数は 6 個に減少している。これにより、ウイグル語とウズベク語の発音に相違が生じる。

それ以外でも、ウイグル語とウズベク語で音韻変化の規則が異なるため、各種の差異が生じている。

4.1.3 単語の由来の相違

最後に、単語の由来の違いによる相違がある。比較的新しい単語などは、他の国経由の外来語としてウイグル語やウズベク語に取り込まれているが、その由来が異なることにより語彙が異なることがある。

例えば、「狙撃兵」に対して、日本語-ウイグル語電子化辞書 [8] には “atquqi”, ウズベク語-日本語辞書 [5] には “snayper” という訳語がそれぞれ掲載されている。ここで、ウズベク語訳の “snayper” は英語の “sniper” と同じ語源をもつと推測される。一方、ウイグル語訳の “atquqi” は「撃つ」という動詞 “atmaq” から派生した語であり、両者は語源が異なることが分かる。なお、“atquqi” に対応するウズベク語としては、動詞「撃つ」“otmoq” から派生した “otuvchi” が存在するが、手元のウズベク語-日本語辞書 [5] には「投げる人；射手」という訳語のみがあり、「狙撃兵」という訳語は掲載されていない。

4.2 ウイグル語とウズベク語の語彙対応

表 1 に示したウイグル語とウズベク語の対応は、人手により作成したものであるが、本研究では、こうした対応の組を自動的に収集する。本研究で作成する日本語-ウズベク語電子化辞書と、我々がすでに作成した日本語-ウイグル語電子化辞書を利用すると、日本語を中間言語と考えることが可能となる。よって、同じ日本語の見出し語の欄にあるウイグル語とウズベク語の訳語を対応づければ、両言語間で対応する語彙が得られる。しかし、実際には、以下のような問題がある。

¹前舌母音・後舌母音とも呼ばれる。また、母音調和する母音の組は言語によって呼称が異なり、モンゴル語では女性母音・男性母音、韓国語では陰母音・陽母音と呼ばれる。

表 2: ウズベク語のキリル文字表記とローマ字表記

a	б	в	г	ғ	д	e	ë	ж	з	и	й	к	қ	л	м	н	о
a	b	v	g	g'	d	ye	yo	j	z	i	y	k	q	l	m	n	o
п	р	с	т	у	ў	ф	х	ҳ	ц	ч	ш	ъ	ь	э	ю	я	
p	r	s	t	u	o'	f	x	h	ts	ch	sh	—	—	e	yu	ya	

注: ғ, қ, ў, ҳ の文字は正しくはそれぞれ Unicode の U+0493, U+049B, U+045E, U+04B3 である。

まず、同じ日本語に対する訳語が複数存在する場合である。先程〈頭, bax, bosh〉を例に挙げたが、実際には日本語-ウイグル語辞書には「頭」の訳語として、“bax”の他に“ékil”, “baxliq”, “baxqi”, “kalla”, “kattiwax”, “mingé”, “uq”が挙げられている。この合計8語の中からウズベク語の bosh に対応する語を選択する必要がある。

また、〈狙撃兵, atquqi, snayper〉は、その由来が異なるため、対応する組として抽出するのは、後の利用を考えると望ましくない。そのため、こうした対応語の特定には表記に関する単語間の類似度が必要となる。

今回の研究では、修正編集距離や、soundex[10] や metaphone[11] などの発音が似た単語に同じキーを割り当てるアルゴリズムを、ウイグル語とウズベク語用に修正することを考える。

例えば、「頭」のウイグル語訳の“bax”, “ékil”, “baxliq”, “baxqi”, “kalla”, “kattiwax”, “mingé”, “uq”のそれぞれについて、ウズベク語訳の bosh との編集距離を計算すれば、“bax”がもっとも適切な単語であることが分かる。また、ウズベク語の単語との修正編集距離が一番短い単語が、必ずしも対応するウイグル語とは限らない。例えば、〈狙撃兵, atquqi, snayper〉のような組が最短となったとしても、それを対応するものとは見做したくない。よって、修正編集距離を正規化したものや soundex を考慮するなどして対応する組を限定する。

4.3 ウイグル語-ウズベク語変換モデル

前節で抽出したウイグル語とウズベク語の組から、ウイグル語-ウズベク語変換モデルを作成する。人手でルールを作成する方法も考えるが、雑音のある通信路モデルを使用する統計的手法も考慮する。そのようにして作成したウイグル語-ウズベク語変換モデルを使用すれば、日本語-ウイグル語辞書にあって、日本語-ウズベク語辞書にない単語の訳語を推定し、対訳辞書の拡充を図ることが可能になる。さらには、ウイグ

ル語-ウズベク語翻訳システムへの応用も考えられる。

5 おわりに

日本語-ウイグル語翻訳システムを利用した日本語-ウズベク語翻訳システムの構想について述べた。今後はシステムの実現に向けて研究を進めていく。

また、本研究は名古屋大学大学院法学研究科で進められているアジア法整備支援とも関連しており、ウズベキスタンに日本語の法制度を紹介する上での貢献も目指している。

本研究の一部は、文部科学省科学研究費補助金(課題番号 18700141)による。

参考文献

- [1] 小川泰弘, ムフタル・マフスット, 杉野花津江, 外山勝彦, 稲垣康善: 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] ムフタル・マフスット, 小川泰弘, 稲垣康善: 日本語-ウイグル語機械翻訳のための格助詞の変換処理, 自然言語処理, Vol. 8, No. 3, pp.123-142 (2001).
- [3] 清瀬義三郎則府: 日本語文法新論-派生文法序説-, 桜楓社 (1989).
- [4] 小川泰弘, ムフタル・マフスット, 外山勝彦, 稲垣康善: 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).
- [5] 小松格: ウズベク語辞典-新版-, 大阪ウズベキスタンの会 (2004).
- [6] 飯沼英二: ウイグル語辞典, 穂高書店 (1992).
- [7] 小川泰弘, ムフタル・マフスット, 杉野花津江, 稲垣康善: 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, 言語処理学会 第8回年次大会 発表論文集, pp.29-32 (2002).
- [8] ムフタル・マフスット, 小川泰弘, 杉野花津江, 稲垣康善: 日本語-ウイグル語辞書の半自動作成と評価, 自然言語処理, Vol.10, No.4, pp.83-108 (2003).
- [9] 小松格: やさしいウズベク語-新版-, 大阪ウズベキスタンの会 (2003).
- [10] Donald Knuth: Sorting And Searching, The Art Of Computer Programming, vol.3, Addison-Wesley (1973).
- [11] Lawrence Philips: Hanging on the Metaphone, Computer Language, Vol.7, No.12 (1990).