日本語－ウイグル語翻訳掲示板システム

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 ‡名古屋産業大学 環境情報ビジネス学部
yasuhiro@is.nagoya-u.ac.jp

1 はじめに

我々は，これまでに日本語－ウイグル語機械翻訳システムの作成に取り組んできた．ウイグル語は中国の新疆ウイグル自治区で使用されている言語であり，アルタイ語族に分類される．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，ある程度の品質の翻訳が可能である．図 1 にその例を示す．

我々は，日本語の膠着語としての特徴に着目した派生文法 [1] を用いることにより，語順などの構文的な類似だけでなく，動詞句の構成方法などの形態論に関しても，多くの共通点が明らかになることを示した．そして，派生文法に基づく形態素解析システム MAJO[2] を利用した日本語－ウイグル語機械翻訳システムのプロトタイプを開発した [3][4]．さらに，日本語－ウイグル語電子辞書の構築 [5] や，ウイグル語の音韻変化ルールの作成 [6] を通じ，翻訳システムの開発を進めてきた．しかしながら，このプロトタイプは辞書の著作権などの問題から，多くのユーザが利用できるものではなかった．

そこで，本研究では，この翻訳システムを多くのユーザに利用してもらい，システム改善のための様々な情報を得ることを目的として，日本語－ウイグル語翻訳システムを組み込んだウェブ掲示板を作成した．

2 ウイグル語－日本語機械翻訳システムの開発

日本語の使用者とウイグル語の使用者の双方が翻訳掲示板を使うためには，日本語からウイグル語への翻訳だけではなく，ウイグル語から日本語への翻訳も必要となる．そこで，翻訳掲示板の開発に先立って，ウイグル語－日本語機械翻訳システムを開発した．

ただし，我々の日本語－ウイグル語機械翻訳システムは両言語の類似点を利用しているため，ウイグル語－日本語機械翻訳システムの開発においても，翻訳システム自体を新たに開発する必要はなかった．具体的には，日本語－ウイグル語機械翻訳システムにおいて利用している，(1) 対訳辞書 (2) 形態素接続規則 (3) 音韻変化規則，の三つに関して，ウイグル語－日本語翻訳用のものを作成構築することにより，ウイグル語－日本語翻訳システムを開発できた．

本節では，上記の三つに関して簡単に述べる．

入力文:	肉をたくさん食べた.						
	⇓						
形態素解析:	肉	を	たくさん	食べ	た	.	
	↓	↓	↓	↓	↓	↓	
逐語訳:	Gox	ni	jiķ	yé	dim	.	
	⇓						
翻訳文:	Goxni jiķ yédim.						

図 1: 日本語–ウイグル語逐語翻訳

2.1 ウイグル語－日本語対訳辞書

我々の使用する日本語－ウイグル語電子辞書は，そもそも，印刷されたウイグル語－日本語辞書 [7] から逆辞書として構築したものである [5]．つまり，日本語－ウイグル語電子辞書を構築する過程において，我々は既にウイグル語－日本語電子辞書のプロトタイプを構築しており，今回は，この辞書に修正を施して使用した．

ただし，翻訳システムで使用している形態素解析システム MAJO[2] においては，現時点では，形態素の異形態をすべて登録する必要がある．日本語においては語幹が変化する異形態は動詞「来る」など少数であり，異形態の登録にはそれほど問題はなかった．

しかし，ウイグル語においては，語幹が変化する単語が多い．例えば,「切る」を意味する動詞の語幹 “késlé”

に “-mék” という接尾辞が接続すると，“késli̇mék” になる．このように，語幹末尾が é で終わる動詞は，接続する接尾辞によっては末尾の é が i に変化する．すなわち，語幹末尾の é が i になった異形態をもつ．

我々はこれまでの研究において，ウイグル語の音韻変化を調べており [6]，その知見を活かして，語幹から異形態を生成する規則を作成した．この規則を利用して自動的にウイグル語の異形態を生成し，ウイグル語－日本語電子辞書に追加した．

また，接尾辞に関してはより多くの異形態が存在するが，それらはすべて人手で登録した．

その結果，語彙数約 23,000 語 (異形態を含めた形態素数約 25,000 語) のウイグル語－日本語電子辞書を構築した．なお，日本語－ウイグル語電子辞書は，約 20,000 語の語彙 (平仮名表記を含めた形態素数約 36,000 語) を収録している．

機械翻訳システムが，どのくらいの範囲の入力を翻訳できるかは，システムに搭載されている対訳辞書に依存する部分が大きい．今回構築した対訳辞書は，一般の形態素解析システムにおいて使用される辞書と比較して 1 割以下の量である．そのため，一般的な用語はおおよそ含まれているが，実際の機械翻訳において使用するには，翻訳できない語が生じるという問題点があり，対訳辞書の充実が今後の課題である．

2.2 ウイグル語形態素接続規則

形態素接続規則とは，形態素間の接続しやすさを表したものであり，形態素解析の精度を大きく左右する．現在は，コーパスを利用して形態素の接続規則を学習する手法が広く用いられているが，ウイグル語に関しては学習のためのコーパスが用意できなかったため，今回は手動で作成した．

ウイグル語は図 1 で示した例から分かるように，分かち書きされる言語であるが，その単位は日本語の文節にほぼ相当する．すなわち，語幹と接尾辞は連続して記述されるため，語幹と接尾辞を分割する形態素解析が必要になる．そのため今回は，主に語幹と接尾辞間の接続を考慮して形態素接続規則を作成した．

なお，分かち書きにおいて使われる空白も形態素の一つとして考えている．しかし，現在の形態素接続規則は形態素の前後 1 語しか考慮しない．そのため，「形容詞の後に名詞が出現しやすい」といった空白を越える規則を表現できない．こうした点の改良は今度の課題である．

2.3 日本語音韻変化規則

ウイグル語－日本語機械翻訳システムにおいて，形態素解析システムの出力は，翻訳された日本語形態素の列である．これを接続して翻訳結果となる日本語文が生成される．その際に，いわゆる動詞の活用を考慮する必要がある．我々のシステムは派生文法 [1] に基づいており，日本語音韻変化規則についても，派生文法に基づいて作成した．

3 翻訳掲示板の作成

日本語－ウイグル語及びウイグル語－日本語機械翻訳システムの作成に続いて，これらを組み込んだウェブ掲示板を作成した．

作成した翻訳システムの概要を図 2 及び図 3 に示す．翻訳掲示板としては，日本語と韓国語の間で実現した enjoy Korea[1] が有名である．enjoy Korea は，基本的には単言語表示である．すなわち，日本語の使用者が閲覧する場合，韓国語の使用者が投稿した内容も日本語訳され，日本語の使用者には日本語の文章しか表示されない．

しかしながら，我々が作成した掲示板では，(1) 翻訳の精度がそれ程高くない (2) 当初の利用者は日本語－ウイグル語翻訳の研究に携わる者が多い，という二つの理由から，日本語とウイグル語の双方を同時に表示した．具体的には，図 2 において示すように，左側に日本語，右側にウイグル語の文章を表示している．ここで左側には，日本語の使用者が投稿した文章と，ウイグル語の使用者が投稿した文章の日本語訳が表示される．

掲示板への投稿においては，辞書の語彙数が少ないため，翻訳できない語がある．図 2 の画面で [送信] ボタンを押すと，図 3 の投稿画面が表示される．その際，投稿内容が自動的に翻訳され右側に表示される．しかし，図 3 では，「レポート」という単語の翻訳が対訳辞書になかったため，ウイグル語へ翻訳された結果には，「レポート」がそのまま残っている．ユーザは，左側の欄の「レポート」を他の言葉，例えば「論文」などに修正し，下の [再翻訳] ボタンを押すことにより翻訳できるかどうかを確認できる．投稿者がウイグル語を理解可能であるならば，ウイグル語訳を直接修正することも可能である．これにより，対訳辞書になくて翻訳ができない場合でも，ある程度の翻訳が可能である．

投稿内容及び翻訳結果を確認した後，[投稿] ボタン

[1] http://www.enjoykorea.jp/

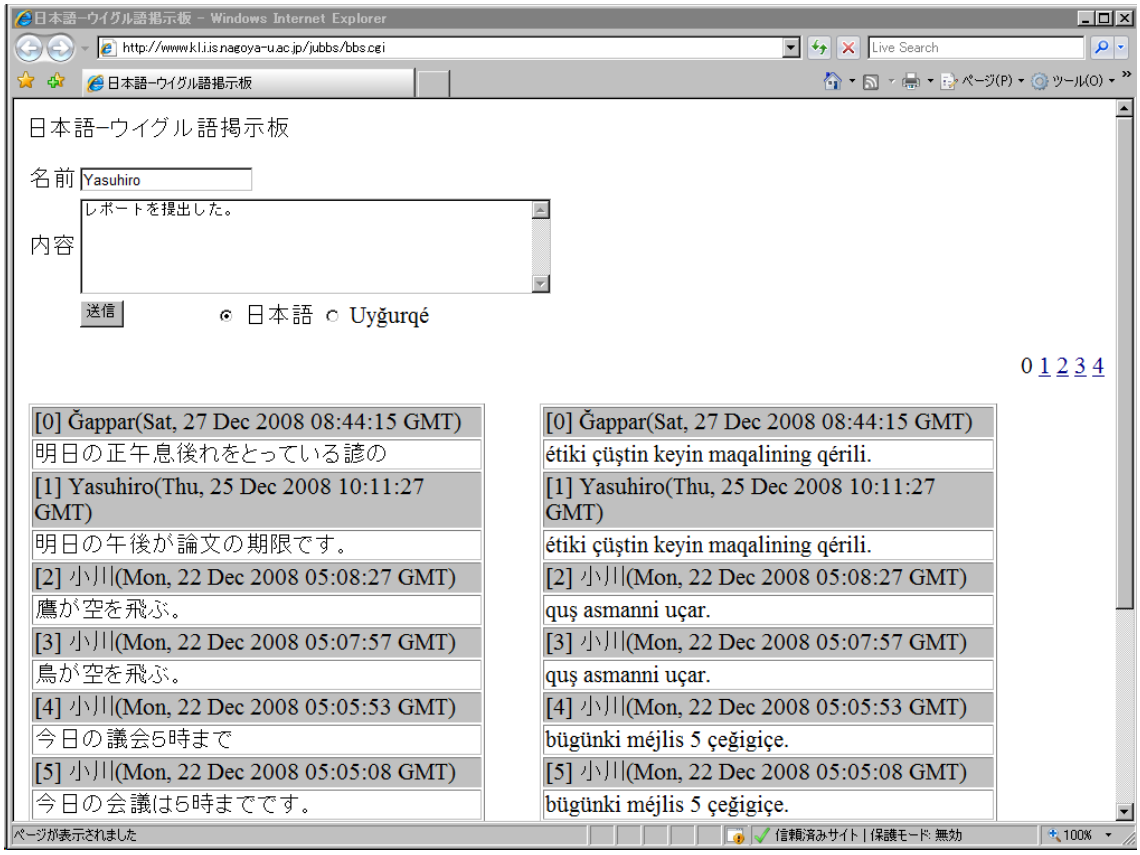

図 2: 日本語–ウイグル語翻訳掲示板の外観

を押すことにより，掲示板にデータが送信され，図 2 の画面に戻る．

4 おわりに

本稿では，日本語とウイグル語間の双方向の翻訳システムを組み込んだ翻訳掲示板について述べた．現在

`http://www.kl.i.is.nagoya-u.ac.jp/jubbs/`

において，本翻訳掲示板を試験運用中である．

今後の課題としては，まず翻訳システムの精度向上が挙げられる．ウイグル語－日本語翻訳に関する課題については，既に第 2 節で述べたが，日本語－ウイグル語，ウイグル語－日本語のいずれの翻訳においても辞書を充実させることが第一である．

それに関しては，掲示板に辞書登録機能を組み込むことを検討している．また，ユーザが対訳語を知らない場合でも，実際にユーザが使用した際に翻訳できなかった語を収集することにより，辞書の拡充を図る．

なお，ウイグル語を表示する際の文字体系に，どのようなものを採用するかを現在検討中である．現在の新疆ウイグル自治区では，アラビア文字を使用した文字体系が採用されている．しかし，インターネット上での利用を考えると，アラビア文字を読める人はラテン文字を読める人に比べて少ないという問題がある．

そこで，我々はラテン文字を使用した文字体系の一種を採用して辞書及び翻訳システムを開発している．しかし，このラテン文字を使用したウイグル語の文字体系に関し，我々が研究を始めた当初は統一されていなかった．すなわち，ウイグル語を扱う文献ごとに，同じ音に対して異なる文字が使われることがあった．また，現在の我々が採用しているラテン文字の中には，オンライン上で入力が容易でない文字も含まれている．

そうした状況を踏まえ，特にインターネット上の利用を意識した新しいウイグル語の文字体系である LSU(Latin-Script Uyghur)[2] [3]が提案され，広く使用されてきている．アラビア文字と LSU の対応表を表 1 に示す．

[2]http://www.uyghurdictionary.org/excerpts/An%20Introduction%20to%20LSU.pdf

[3]LSU とは別に ULY(Uyghur Latin Yéziqi) と呼ばれることもある．なお，yéziqi はウイグル語で「文字」を意味する．

図 3: 掲示板への投稿画面

表 1: アラビア文字と LSU の対応表

ئا	ب	د	ە	ف	گ	ھ	ى	ج	ك	ل	م	ن	ۇ	پ	ق	ر	س	ت	ۆ	ۋ	خ	ي	ز	ې	ۈ	ۇ	چ	غ	ش	ژ	ڭ
a	b	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	w	x	y	z	é	ö	ü	ch	gh	sh	zh	ng

そこで，本翻訳掲示板においても，LSU の採用を検討している．修正を最小で済ませるには，我々が採用している文字体系と LSU との間で変換フィルターを通せば良い．しかし，将来的な展望を考えると，対訳辞書及び音韻変化規則も LSU に対応するように変更するのが望ましく，その点についても検討中である．

謝辞 本研究の一部は，財団法人日東学術振興財団の助成による．

参考文献

[1] 清瀬義三郎則府: 日本語文法新論–派生文法序説–, 桜楓社 (1989).

[2] 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).

[3] 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).

[4] 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).

[5] 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).

[6] 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).

[7] 飯沼英二: ウイグル語辞典, 穂高書店 (1992).