様式 C-19

科学研究費補助金研究成果報告書

平成２１年　４月３０日現　在

研究種目：若手研究(B)
研究期間：2006～2008
課題番号：18700141
研究課題名（和文）　日本語－ウズベク語機械翻訳システムの開発

研究課題名（英文）　Development of Machine Translation system from Japanese into Uzbek

研究代表者
小川　泰弘　(Yasuhiro　Ogawa)
名古屋大学・大学院情報科学研究科・助教
研究者番号：70332707

研究成果の概要：
本研究では，日本語からウズベク語への機械翻訳システムを開発した．翻訳システムには，先行して開発した日本語－ウイグル語機械翻訳システムを使用した．このシステムは日本語・ウイグル語・ウズベク語間の文法的な類似性に着目しており，その辞書データなどを日本語－ウズベク語翻訳用のものに置換することにより日本語－ウイグル語機械翻訳を実現した．さらに，ウイグル語とウズベク語との間の語彙の類似度を利用し，日本語－ウイグル語辞書のウイグル語訳をウズベク語に変換することにより，日本語－ウズベク語辞書の拡張も行った．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００６年度	1,100,000	0	1,100,000
２００７年度	900,000	0	900,000
２００８年度	1,000,000	300,000	1,300,000
総　計	3,000,000	300,000	3,300,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：自然言語処理，機械翻訳，対訳辞書，ウズベク語，膠着語，ウイグル語

１．研究開始当初の背景

これまでに機械翻訳に関する研究が様々な言語間で進められているが，日本語－英語間のように数多くの研究がなされている言語がある一方で，あまり研究が進められていない言語もある．

そうした言語の中において，本研究ではウズベク語に着目する．ウズベク語はウズベキスタン共和国の公用語である．ウズベキスタンと日本との間の交流は，これからますます盛んになると予想され，両国の言語を翻訳する需要は高まってきている．実際，本研究代表者はウズベキスタン共和国民商法典起草支援プロジェクトに参加している．このプロジェクトは，社会主義時代の体制から新しい社会体制へと移行しているウズベキスタン共和国に日本の法令を紹介し，同国の法律作成を支援するものである．このプロジェクトにおいては，日本の法令を英語もしくはロシア語に翻訳して紹介している．しかし，日本語からウズベク語への翻訳ができれば，法令の紹介から法律作成の支援までがよりスムーズに進められる．

また，言語学の観点からも日本語－ウズベク語間の翻訳は興味深いものである．ウズベク語と日本語との間には，語順や単語の構成

方法などの点で共通点が多い．よって，日本語－ウズベク語間の翻訳は，日本語－英語間などの文法的な差異が大きな言語と比べて少ない規則で実現できる可能性がある．

本研究に先立ち，本研究代表者はこれまでに，日本語－ウイグル語機械翻訳システムを開発してきた．ウイグル語はウズベク語と同じチュルク諸語に属する言語である．テュルク諸語は，中央アジアを中心に東ヨーロッパからシベリアに至る広大な地域で話されているが，分布の広大さに比べて言語間の差異は小さいとされている．実際，ウイグル語とウズベク語は文法的に類似しているだけでなく，語彙的にも類似点が多い．そこで，日本語－ウイグル語機械翻訳システムを応用することにより，日本語－ウズベク語機械翻訳システムが実現できる．さらに，この方式を応用すれば，他のチュルク諸語への翻訳システムを容易に開発することが期待できる．

本研究代表者が開発してきた日本語－ウイグル語機械翻訳システムは，日本語とウイグル語の文法的類似性を利用し，日本語の形態素解析終了後，構文解析なしで逐語訳を基本として翻訳するものである．その開発の過程において，市販のウイグル語－日本語辞書から機械翻訳用の日本語－ウイグル語辞書(収録語数約 2 万)を構築した．

同様に，日本語－ウズベク語機械翻訳には日本語－ウズベク語辞書が必要となるが，そうした辞書は現在の日本国内においては存在せず，ウズベキスタンにおいて収録語数約 2 千の『和ウズベク語辞典』が存在するだけである．一方，ウズベク語－日本語辞書としては，収録語数約 1 万の『ウズベク語辞典』と，その改訂版である『ウズベク語辞典－新版－』が存在するだけであるが，これらはいずれも電子化された物ではない．そのため，日本語－ウズベク語機械翻訳システムの開発においては，機械可読な対訳辞書を構築する必要がある．

２．研究の目的

本研究における最終的な目標は，日本語からウイグル語・ウズベク語を含むテュルク諸語への機械翻訳システムの開発である．

本研究課題では，日本語－ウイグル語機械翻訳システムの辞書を日本語－ウズベク語辞書に置き換えることによって，日本語－ウズベク語機械翻訳システムを実現する．その際，日本語－ウズベク語電子化辞書が必要となる．本研究では，市販のウズベク語－日本語辞書から日本語とウズベク語を入れ替えて電子化辞書を構築する．それに加えて，ウイグル語とウズベク語との間の語彙的な類似点に着目し，日本語－ウイグル語電子辞書のウイグル語訳をウズベク語訳に変換することにより，日本語－ウズベク語辞書を拡張する．具体的には，ウイグル語単語からウズベク語単語への変換パターンを獲得し，それを利用して半自動的に日本語－ウズベク語辞書を構築する．

以上により，本研究では日本語－ウズベク語電子化辞書と，それを利用した日本語－ウズベク語機械翻訳システムを開発する．

３．研究の方法

本研究では，まず日本語-ウズベク語電子化辞書を構築し，それを用いた日本語-ウズベク語機械翻訳システムを開発する．日本語-ウズベク語電子化辞書を構築するにおいては，まず市販のウズベク語-日本語辞書から日本語とウズベク語を入れ替えて電子化辞書を構築した．さらに，ウズベク語とウイグル語の類似性に着目し，日本語-ウイグル語辞書にあるウイグル語訳語をウズベク語に変換することにより，辞書の拡充を図った．その際に使用したのが翻字と呼ばれる技術である．

以下では，まず，ウイグル語からウズベク語への翻字システムの構築方法について述べ，その後，そのシステムを用いた日本語-ウズベク語辞書の拡張について述べる．最後に，その辞書を使用した日本語-ウズベク語機械翻訳システムの開発について述べる．

(1) ウイグル語-ウズベク語翻字

翻字とは，特定の言語の文字表記を他の言語の文字表記に変換する操作であり，例えば，英語の America を日本語のカタカナ表記である「アメリカ」に変換する操作も翻字である．翻字は広い意味では翻訳の一種であり字訳と呼ばれることもある．ウイグル語からウズベク語への変換は翻訳であるが，両言語の類似性に着目することにより，翻字による変換が可能となる．両言語間の対応を表１に示す．最右列の類似度については後述する．

表 1 において，bulut，fontan の２語は，まったく同じ形であり，それ以外の語も良く似ている．こうした両言語間の類似点に着目し，以下の手順でウイグル語からウズベク語への翻字システムを作成した．

① 対応する単語ペアの作成

今回使用した日本語-ウイグル語電子化辞書は 36,157 語の日本語見出しを含む．一方，今回市販の辞書から構築した日本語-ウズベク語電子化辞書 18,526 語の見出し語を持つ．この二つの辞書から以下の手順で対応するウイグル語とウズベク語のペアを作成した．

(i) 対訳辞書からの抽出

まず，同じ日本語見出し語をもつ辞書項目のペアを３項組として抽出した．なお，日本語見出し語に複数の訳語がある場合には，そ

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	類似度
雲	bulut	*bulut*	1
噴水	fontan	*fontan*	1
頭	bax	*bosh*	2
恐怖	déẖ xét	*dahshat*	2
雇う	yallimaķ	*yollamoq*	2
育てる	östürmék	*o`stirmoq*	2
木	déréh	*daraxt*	3
電気	elektir	*elektr*	3
狙撃兵	atķuqi	*snayper*	-

れらすべてを抽出しているため，３項組の内容は〈日本語見出し語,ウイグル語訳の集合,ウズベク語訳の集合〉である．例えば，日本語見出し語「頭」に対して，日本語-ウイグル語辞書には８語，日本語-ウズベク語辞書には３語の訳語がそれぞれ掲載されていたため,〈頭,{éķil, bax, baxliķ, baxqi, kalla, kattiwax, mingé, uq},{*bosh*, *kalla*, *katta*}〉という３項組となった．そのような３項組は 5,580 組得られた．

(ii) 類似する２項組の抽出

こうして抽出された３項組には，複数のウイグル語とウズベク語が含まれるが，翻字の参考になるのは，ウイグル語とウズベク語が類似するペアだけである．前述の「頭」の例では,〈頭, bax, *bosh*〉や,頭 ，kalla, *kalla*〉の二つの３項組のみが有用である．一方,〈狙撃兵, atķuqi, *snayper*〉のように，日本語訳は同じでも，単語同士が類似しないものは除きたい．そこで，ウイグル語単語とウズベク語単語間の類似度を以下の３段階(類似していない場合を含めれば４段階)に分類した．

まず，類似度１となるのは，〈雲, bulut, *bulut*〉のように単語の綴が完全に一致する場合である．

類似度 2 となるのは，以下の変換アルゴリズムによってウイグル語単語とウズベク語単語が一致した場合である．例えば bax と *bosh* はこのアルゴリズムにより，ともに fix に変換されるため，類似度２となる．

a) ウズベク語中の sh を x に，ch を q に，x を h に変換.
b) 子音が後続する ng を n に変換.
c) 以下の文字を対応する文字にそれぞれ変換.
   i, e, a, o, u, é, ö, ü, y, o` → i
   k, q, g, ķ, ğ, g` → k
   f, p, b, w, v → f
   t, d → t
   s, z → s
   n, m → n
d) ウズベク語単語中に残った`を除去
e) 同じ文字の連続を一つにまとめる.

この変換アルゴリズムの適用後，さらに
a) 一方の末尾を削除したものが，もう一方と一致.
b) 一方の末尾が子音・母音・子音であり，その母音を削除したものが，もう一方と一致.

という条件を満たした場合を類似度３とした．木 ，déréh, *daraxt*〉や〈タンク，tanka, *tank*〉などが類似度３となる．

表１の最右列は，このように計算した類似度である．なお,〈狙撃兵, atķuqi, *snayper*〉の場合は，いずれの方法でも一致しないため,類似していないと判断される．

この類似度による判定を用いた結果，ウイグル語とウズベク語が類似している３項組を 3,575 組獲得した．さらに，日本語見出し語だけが異なるものをまとめ，2,660 組のウイグル語とウズベク語の２項組を獲得した．

② ウイグル語-ウズベク語翻字

こうして獲得した 2,660 組のペアを利用してウイグル語からウズベク語への翻字を行った．翻字手法としては，人手で規則を作成する方法と，統計的な手法の二つを試みた．

(i) 人手で作成した規則による翻字

翻字規則を人手によって作成する場合のベースラインは，ウイグル語の各文字を対応するウズベク語の文字に１文字ずつ置き換える手法である．

そして，このベースラインとなる規則に，人手による規則を追加して変換精度の向上を図った．例えば，単語末尾の mék は *moq* になるという規則を追加した．このように，変換規則は正規表現にマッチした部分を置換する形式で作成した．

(ii) 統計的手法による翻字

統計的な翻字は，各文字を単語と見做せば,統計的翻訳として捉えることができる．そこで，統計的翻訳用に公開されている各種のツールを使用してウイグル語からウズベク語への統計的翻字を試みた．具体的には，翻訳モデルの学習に GIZA++，言語モデルの学習に SRILM，デコーダに Moses を使用した．

翻訳モデルの学習には，前章で得られた２項組 2,660 組を使用した．言語モデルの学習には，日本語-ウズベク語辞書に含まれるウイグル語訳語 23,494 語を使用した．

(2) 日本語-ウズベク語辞書拡張実験

日本語-ウズベク語電子化辞書の作成には,ウズベク語-日本語辞書である『ウズベク語辞典－新版－』を使用した．幸いにも辞書の

著者から原稿となるデータを入手することができたため，それに修正・変換を施して日本語–ウズベク語電子化辞書を作成した．修正の際には，ウズベク語ネイティブに依頼し，細かな綴間違いなどを修正した．その結果，18,526 語の見出し語を持つ辞書を作成した．

しかし，この見出し語数は機械翻訳用の辞書としては不十分である．そこで，(1)で作成した統計的翻字モデルを用いて，対訳辞書の拡張する．対訳辞書の拡張方法を以下に示す．例えば，日本語–ウイグル語辞書には〈ミルク,süt〉が存在するが，日本語–ウズベク語辞書には「ミルク」の項目はない．しかし，süt を翻字モデルで変換したところ，*sut* が出力され，これは日本語–ウズベク語辞書にある〈牛乳,*sut*〉と一致した．つまり，翻字によって，〈ミルク,*sut*〉という項目を日本語–ウズベク語辞書に追加可能となったのである．この手法を実現し，日本語–ウズベク語辞書の拡張を試みた．

(3) 日本語–ウズベク語機械翻訳システム

(2)で作成した辞書を使用した日本語–ウズベク語機械翻訳システムを作成した．今回の日本語–ウズベク語機械翻訳システムは日本語–ウイグル語機械翻訳システムをベースにしたものである．しかし，このシステムだけでは一般のユーザに使用してもらうには不十分である．そこで，広く一般のユーザに使用してもらうための翻訳掲示板システムも開発した．この掲示板システムは，日本語–ウイグル語機械翻訳システムと日本語–ウズベク語機械翻訳システムの両方で使用可能な形で実装した．

４．研究成果

前節で挙げたそれぞれの研究方法に対応して，その成果を述べる．

(1) ウイグル語–ウズベク語翻字

人手で作成した規則による翻字と統計的手法による翻字の成果を得た．

まず，人手で作成した規則では，ベースラインの規則だけを用いた場合，２項組 2,660 組に含まれるウイグル語単語に対して，正しいウズベク語単語に変換できたのは 992 個，すなわち変換精度は 37.3%であった．このベースラインとなる規則に，人手による規則を 21 個追加したところ，1,190 個(44.7%)が正しく変換できた．さらに規則を追加して変換精度を上げることは可能であるが，現在の規則に新たな規則を追加しても，正しく変換できるようになるのは数個がせいぜいであり，人手によるこれ以上の精度向上は容易ではないことが判明した．

一方，統計的手法による翻字の場合，クローズド・テストでは変換精度は 77.9%であった．また，Moses では，複数の結果を順位付きで出力可能なことから，Moses に上位 10 語までの候補を出力させた．その中に正解が含まれる割合を調べたところ，84.7%であった．翻訳モデルの学習用データを分割し，10 分割交差検定を試みた場合の変換精度は 73.0%であり，この場合も人手による規則作成よりも良い結果が得られた．

ウイグル語–ウズベク語翻字は，世界でもこれまで試みられたことがなく，本研究が最初である．また，今回の統計的手法を用いた実験では，細かなパラメータの調整が不十分であるが，その調整によっては精度が更に向上する可能性もある．こうした研究を進めることにより，翻字ベースでのウイグル語–ウズベク語機械翻訳の実現も考えられる．

(2) 対訳辞書拡張実験

上記の統計的手法を用いた翻字システムを用いて，日本語–ウイグル語辞書にはあるが，日本語–ウズベク語辞書にない項目 862 語に対し，そのウイグル語訳語をウズベク語に変換した．上位 10 個までの出力を調べたところ，532 語(61.7%)に対して適切なウズベク語単語が含まれていた．変換精度としては充分ではなく，今後翻字モデルの改良により更なる精度向上を目指す必要がある．また，変換に失敗した単語の中には，ウズベク語とウイグル語の単語間に対応関係がなく，翻字に基づく本手法ではそもそも変換できないものもあった．しかしながら，対訳辞書に無い語の 3 語に 2 語は正しい訳語候補が得られることから，対訳辞書の拡張には有効であることが示せた．特にウイグル語やウズベク語のように計算機で利用可能な言語資源が少ない言語においては有用である．今後は，日本語–ウズベク語辞書を使用して日本語–ウイグル語辞書を拡張することにも本手法を適用し，その有効性を検証する．

(3) 日本語–ウズベク語機械翻訳システム

電子化した日本語–ウズベク語辞書に，(2)で正しく変換できた日本語–ウズベク語のペアを追加した辞書を構築し，それを利用した日本語–ウズベク語機械翻訳システムを開発した．なお，翻訳システム自体は日本語–ウイグル語機械翻訳システムと基本的には同じである．ただし，接尾辞や細かな音韻変化規則に関して足りない部分があり，それらを追加している最中である．

また，機械翻訳システムを広く使用してもらうために，機械翻訳システムを組み込んだ翻訳掲示板を作成した．図１にそれを示す．なお，図１では日本語–ウイグル語機械翻訳システムを使用した翻訳掲示板を示してい

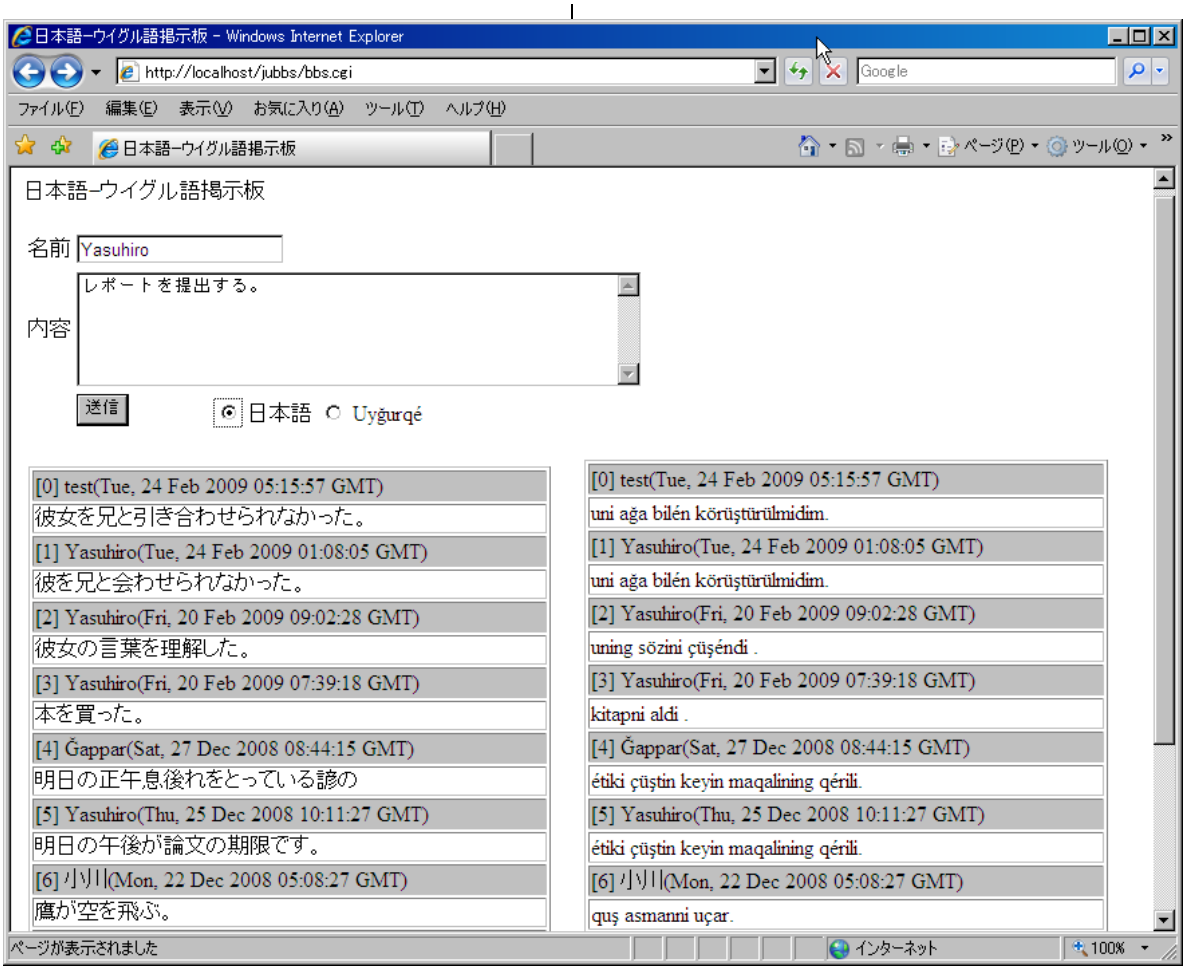

図 1 日本語－ウイグル語翻訳掲示板

るが，翻訳システムを日本語−ウズベク語翻訳システムと交換すれば，日本語−ウズベク語翻訳への対応も可能であり，近日中の公開を予定している．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）
(1)OGAWA Yasuhiro, FUKUDA Muhtar, TOYAM Katsuhiko: Transliteration from Uighur to Uzbek for Expansion of Japanese Translation Dictionary, International Conference on Asian Language Processing 2008, (査読有), pp.182-188 (2008)

〔学会発表〕（計 2 件）
(1) 小川泰弘：日本語－ウイグル語翻訳掲示板システム，言語処理学会第 15 回年次大会，2009 年 3 月 3 日，鳥取大学

(2)小川泰弘：日本語対訳辞書拡張のためのウイグル語からウズベク語への翻字手法，言語処理学会第 14 回年次大会，2008 年 3 月 19 日，東京大学

〔その他〕
ホームページ等
日本語－ウイグル語翻訳掲示板システム
http://www.kl.i.is.nagoya-u.ac.jp/jubbs

６．研究組織
(1)研究代表者
小川 泰弘 (Yasuhiro Ogawa)
名古屋大学・大学院情報科学研究科・助教
研究者番号：70332707