日本語－ウイグル語機械翻訳特有の諸問題について

小川泰弘　ムフタル・マフスット†　杉野花津江　稲垣康善
名古屋大学大学院工学研究科　†名古屋大学大学院国際開発研究科
yasuhiro@nuie.nagoya-u.ac.jp

1　はじめに

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 [1][2]．この手法に基づいて，我々は日本語–ウイグル語機械翻訳の研究を進めており，日本語の論説文136文を対象に翻訳実験を行ったところ，文節単位で判定して，82.4%の正訳率を得た．ここで，翻訳に失敗した事例について分析したところ，訳語選択における失敗や，未登録語の処理など，あらゆる言語間の機械翻訳に共通の問題がある一方で，日本語における「A の B」の表現や，ウイグル語における人称接尾辞など，言語に特有の問題も見られた．本稿では，日英翻訳との比較も混じえながら，こうした日本語–ウイグル語機械翻訳に特有の問題について分析する．

なお，本稿では日本語の表現は「」，ウイグル語の表現は“”で括り区別する．

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，構文解析が不要になる．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になる (図 1)．

ここで，「を」と “ni” といった格助詞間や，「た」と “dim” といった動詞接尾辞間にもそれぞれ対応関係があ

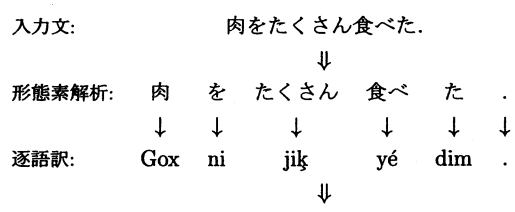

図 1: 日本語–ウイグル語逐語翻訳

あることが分かる．我々は，こうした点に注目し，派生文法 [3] に基づく日本語形態素解析システム MAJO[4] を用いて日本語入力文の解析と逐語訳を行っている．この結果に動詞接尾辞の選択処理 [1] と格助詞の選択処理 [2] を組み合わせることで，日本語–ウイグル語機械翻訳における動詞接尾辞と格助詞の問題を解決している．しかし，これ以外にも様々な問題が残っており，そのうち日本語とウイグル語の性質に起因する問題について検討する．

3　日ウ機械翻訳における問題点

3.1　格助詞「の」の翻訳

我々は文献 [2] において，格助詞の翻訳に取り組んできた．そこでは，格格助詞についてデフォルトの訳を与えるとともに，例外的な訳が必要となる場合には，格助詞を含む名詞句の係り先となる動詞に格助詞の訳語情報を付加した．しかし，格助詞「の」については，名詞–動詞間ではなく，名詞–名詞間の関係を示しているため，文献 [2] の手法では扱えなかった．

日本語における「A の B」という表現に関しては数多くの研究がなされているが，翻訳という観点からは，文献 [7] および文献 [8] の研究がある．[7] では，「A の B」という名詞句の英訳を 10 種類の型に分類するとともに，朝日新聞のコラム「天声人語」から抽出した 4925 例をこの型で分類している．また，[8] では和英辞書から抽出した 1104 例に対応する英語表現を 57 種類に分類している．

それに対し，本稿では日本語の表現「A の B」をウイグル語に翻訳した場合，どのような訳が与えられるかという観点から分析してみた．

日本経済新聞の社説などから取り出した環境問題に関する論説文 3 編 136 文中に出現した「A の B」の事例 197 例が，人手により，どのようなウイグル語に翻訳されたかを表 1 に示す．以下では逐語訳という観点から，「の」の訳語が，格助詞になる場合，(格助詞以外の) 名詞接尾辞になる場合，訳出されない場合，その他の 4 つに分けて考察する．

3.1.1 「の」が格助詞に翻訳される場合

表1から「A の B」という日本語表現の66.3%が，“A' ning B'”というウイグル語表現に翻訳されていることが分かる．この“ning”はウイグル語の格助詞であり，「家の庭」“öyning ḫoylasi”の例にあるような所有の意味や，「資源の枯渇」“bayliḳning ḳuruğdilixi”の例のように，名詞 A' が名詞 B' の主体になることを示しており，その機能は日本語格助詞「の」と同じである．

「A の B」を英語に翻訳する場合と比較すると，文献[8]では，〈A' の所有格 B'〉となるのが37.7%，〈B' of A'〉となるのが17.2%とされている．また，文献[7]の実験では，〈B' of A'〉となるのが24.5%，〈A' の所有格 B'〉となるのが5.8%以下[1]とされている．翻訳事例の収集方法が異なるため一概には比較できないが，日ウ翻訳においては“A' ning B'”のパターンが6割以上を占めることから，日英翻訳の場合に比べて高い類似性が確認される．よって，日ウ機械翻訳においては，格助詞「の」をデフォルトで“ning”に翻訳するだけでも，ある程度の精度が期待できる．

次に「の」が，日本語の格助詞「を」に相当するウイグル語の格助詞“ni”に翻訳される場合が，7.0%あった．これは，動作を表す名詞に対して，その目的語が格助詞“ni”で示されているのであり，実際，今回の実験で“ni”に翻訳された14例では，Bはすべて動作名詞(いわゆるサ変名詞)であった．

しかし，Bが動作名詞であるからといって，「A の B」が必ずしも“A' ni B'”に翻訳される訳ではない．先述の「資源の枯渇」“bayliḳning ḳuruğdilixi”の例においては，「枯渇」“ḳuruğdilixi”が動作名詞であるにも関わらず，「の」は“ning”に翻訳されていた．これは，AとBの関係に依存しており，「AがBする」の場合と「AをBする」の場合で，「の」の訳語が異なるのである．よって，名詞AとBの関係を判定することになるが，この判定はウイグル語で行う必要がある．

例えば，「技術の開発」のウイグル語訳には，“tehnikani aqix”，“tehnikaning aqilix”の2種類が考えられる．“aqix”は「開発する」を意味するウイグル語動詞“aq”に名詞化接尾辞“ix”が接続したものであるが，“aqilix”は，同じ「開発する」“aq”に受身を表す接尾辞“il”と名詞化接尾辞“ix”が接続したものである．すなわち，同じ「技術の開発」に対するウイグル語訳であるが，前者は「技術を開発すること」を，後者は「技術が開発されること」を意味する．

[1]文献[7]では，他のパターンも一緒に分類されているため，〈A' の所有格 B'〉だけの具体的な割合は不明．

表1: 日–ウ翻訳における「A の B」の翻訳型

型	例文	頻度
-ning	家の庭 öyning ḫoylasi	
	資源の枯渇 bayliḳning ḳuruğdilixi	132 (66.3%)
-ni	水の吸収 suni singdürüx	
	焼却炉の開発 oqaḳni aqix	14 (7.0%)
-din	紙の利用 ḳéğézdin paydilinix	1 (0.5%)
-diki	江戸時代の日本 Edo déwrdiki Yaponiyé	
	大量の銅 zor miḳdardiki mis	12 (6.0%)
-ki	現在の行政 ḫazirḳi mémuriyét	
	外の世界 sirtḳi dunya	6 (3.0%)
-lik	滅亡の危機 ḫalakétlik kirizis	1 (0.5%)
何もなし	100万年の間 milyon yil arliḳta	5 (2.5%)
形容詞	多くの国 köpligén dölét	
	人工の一大世界 sün'i dunya	24 (12.1%)
別の表現	過度の焼き畑農業 ḫéddidin artuḳ ormanliḳ köydürüp yér aqix	4 (2.1%)
合計		197

そのため，Bが動作名詞である場合には，ウイグル語訳の A' と B' の関係を調べる必要がある．ただ，動詞「開発する」“aq”が他動詞であるのに，派生された「開発される」“aqil”が自動詞であることから，B' の元となる動詞が他動詞か自動詞かで区別できるのではないかと考えられる．

なお，1例だけ「の」が“din”に翻訳された例があるが，これは動詞「利用する」“paydilan”がその対象を“ni”ではなく格助詞“din”で示すためであり，これも「の」が“ni”に翻訳される場合と同じ枠組みで扱うことができる．

3.1.2 「の」が名詞接尾辞に翻訳される場合

広義では，格助詞も名詞接尾辞の一種と考えられるが，ここでは，格を示すのではなく，名詞に接続して

それを形容詞化する接尾辞を取り上げる．文献 [5] を参考に説明する．

“-diki” は主に場所を示す名詞に接続して，「〜にある」の意味で形容詞化する接尾辞であり，「北京の学校 (=北京にある学校)」“Beyjindiki méktép” のように使用される．場所を示す以外の名詞に接続する例もあり，今回の実験では表 1 に示すような例があった．

“-ki” は主に時を表す名詞に接続して形容詞化するものであり，表 1 の例の他にも，「現在の」“ḫazirḳi”，「後の」“keyinki” といった物が翻訳結果に現れた．

“-lik” は，「〜のある」の意味で名詞を形容詞化する接尾辞であり，例えば，「知恵」“éḳil” に “-lik” が接続すると，「賢い」“éḳillik” という意味になる．

こうした単語の翻訳には，“-diki”，“-ki”，“-lik” が付加された形を辞書に登録することが考えられる．実際，本研究で利用する日本語–ウイグル語機械翻訳辞書の元となった「ウイグル語辞典」[6] には，こうした派生形容詞も多く収録されているので，それらをそのまま機械翻訳辞書に登録している．また，「の」の訳語候補として “ning” に加えて “-lik”，“-diki”，“-ki” も用意すれば，辞書に派生形容詞が登録されていない場合にも対応することが可能となる．

3.1.3 「の」が訳出されない場合

次に日本語「の」がウイグル語では訳出されていない場合を検討する．これには，表 1 の例の他に，今回の実験では出現しなかったが，「友達の田中君」“dostim Tanaka” のように同格を示す「の」，「桑の木」“üẓmé dérihi” における「の」などがある．ただし，こうした中には “ning” を付加できる場合もあり，個々の事例については，現在調査中である．

3.1.4 その他

最後に，「の」を逐語訳することができず，「の」を含む名詞句全体を翻訳する場合について検討する．

表 1 において「形容詞」とされているのは，「多くの」“köpligén” のように，「名詞+の」がウイグル語で 1 語の形容詞に相当する語である．3.1.2 節で述べた例と比較すると，名詞からの派生形かどうかの違いがあるが，機械翻訳における対応を考えれば，どちらも「名詞+の」を 1 語として辞書に登録すれば良いという点では同じである．ただし，その場合，翻訳辞書に「名詞」単独の訳語と「名詞+の」の訳語の両方があった場合に，訳語選択の問題が生じることになる．

一方「過度の」“ḫéddidin artuḳ” の例は，相当する単語がなかったり，慣用句の中に含まれる「の」の例であり，それを含む名詞句全体を言い換えている．“ḫéddidin artuḳ” の直訳は「限度から超えた」であり，機械翻訳の点では，こうした物は用例を収集して辞書に登録するのが望ましい．

なお，「A の B の C」のように「の」が複数現れる表現は，ウイグル語でも “A' ning B' ning C'” と表現される．ただし，個々の「の」が “ning” になるか “ni” になるかなどは，この節で述べた規則に従う．

3.2 ウイグル語の人称接尾辞

日本語では主語がしばしば省略されると言われるが，これはウイグル語にも共通する性質である．例えば，図 1 に挙げた日本語文「肉をたくさん食べた．」に相当するウイグル語文 “Goxni jiḳ yédim.” にも主語がないが，これはウイグル語でも自然な文である．よって，日英翻訳において必要とされるゼロ代名詞の推定が日ウ翻訳では不要になると期待できるが，残念ながらそうではない．

日本語文「肉をたくさん食べた」の主語が誰であるかは不明であるが，ウイグル語文 “Goxni jiḳ yédim.” では，主語は「私」であることが，動詞句「食べた」“yédim” の末尾にある人称接尾辞 “m” によって明示されている．もしも主語が「あなた」であれば「食べた」は “yéding” となり区別される．

ウイグル語では，こうした人称接尾辞が名詞にも動詞にも接続する．日本語には人称接尾辞がないため，ウイグル語への翻訳では，この人称接尾辞を補う必要がある．その際に，主体が省略されていれば，日英翻訳におけるゼロ代名詞の推定のような手法が必要となる．

現在では，翻訳対象が論説文であるため，文中に主体が明示されている場合はそれに対応する人称接尾辞を，そうでなければ三人称の人称接尾辞を付加している．

3.3 ウイグル語の限定語尾

ウイグル語では，二つの名詞が修飾・被修飾の関係で結ばれる場合，被修飾語に接尾辞 “(s)i” [2]が接続する．なお，これは前述の人称接尾辞と同時に現れることがないので，名詞に接続する三人称の人称接尾辞と考えることも可能であるが，文献 [5] では，人称語尾とは別の品詞である限定語尾と分類されている．機械

[2]-(s)i は異形態を含めた表記であり，末尾が母音の名詞に接続する場合は si に，末尾が子音の名詞に接続する場合は () 内の s が消失した i になる．

翻訳の立場から見た場合も，人称接尾辞とは別の問題があるため，本稿でも区別して考える．

機械翻訳において問題となるのは，限定接尾辞 "(s)i" が名詞に接続するかどうかの判定である．例えば，「環境問題」 "muḫit mésilési" (「環境」 "muḫit"+「問題」 "mésilé") のように名詞が連続する場合や，表1の例にもある「家の庭」 "öyning ḫoylasi"(「家」 "öy"+「庭」 "ḫoyla") のように格助詞「の」 "ning" が入る場合などは，名詞が名詞を修飾しており，"(s)i" が後接する．

一方，名詞が形容詞によって修飾される場合は「広い庭」 "kéng ḫoyla"(「広い」 "kéng"+「庭」 "ḫoyla") のように，限定語尾は付加されない．この形容詞には，3.1.2節で述べた，"diki"，"lik"，"ki" が後接して派生した形容詞も含まれる．

機械翻訳においては，この規則を元に名詞が名詞を修飾している場合に "(s)i" を付加することになるが，その際に問題となるのは，修飾語の品詞が名詞か形容詞か判定できない場合である．ウイグル語では，名詞にも形容詞にもなる単語が存在する．例えば，"ḳara" は「黒い」を意味する形容詞であるが，「黒」を意味する名詞でもある．よって，例えば「黒雲」を翻訳する場合に，名詞「黒」 "ḳara"+名詞「雲」 "bulut" と解析し逐語訳すると，名詞が連続していることから，"(s)i" を付加してしまう．しかし実際には「黒雲」は「黒い雲」のことであり，そのウイグル語訳 "ḳara bulut" には "(s)i" が付加されない．

また，3.1.4節で述べた，「A の」を "*A′* ning" と翻訳するか，一つの形容詞として翻訳するかの選択に失敗すると，それに伴なって限定語尾の付加に関しても失敗することになる．

3.4 複合名詞

日本語では名詞が連続して複合名詞を形成することができる．ウイグル語でも先述の「環境問題」 "muḫit mésilési" の例のように可能である．しかし，日本語では連続しているが，ウイグル語訳では名詞間に適当な語を補う必要がある場合も存在する．例えば，「自然破壊」 "tébi'étning wéyran boluxḳa" の例では，「自然」 "tébi'ét" と「破壊」 "wéyran boluxḳa" の間に「の」に相当する格助詞 "ning" が必要となる．こうした事例は，複合名詞を辞書に登録すれば解決するが，数が多いため文献 [9] などで提案されている複合名詞の対訳獲得手法が必要になる．

3.5 ウイグル語の複数接尾辞

ウイグル語の名詞は単数を表す場合でも複数を表す場合でも同じ形であり，その意味では日本語と同様，数に関して中立である．さらに，日本語で「たち」「々」などを付加して複数形を明示するように，ウイグル語でも接尾辞 "lér" を付加することで複数形であることを示す場合がある．例えば，「国」 "dölét" に "lér" を付加すれば「国々」 "dölétlér" という意味になる．

ここで，日本語とウイグル語では複数形語尾に用法の違いがあり，日本語では複数形を明示しない場合でも，ウイグル語では必要になる場合がある．例えば「多くの国」 "köpligén dölétlér" のように「多く」「各種の」などの形容詞で修飾される場合には，複数形語尾が必要になる．現在，こうした複数形語尾を必要とする表現について，どのような種類があるか調査中である．

4 おわりに

日本語–ウイグル語間での機械翻訳実験において両者の言語的性質に依存した失敗例について収集し検討した．現段階では，収集事例が少ないため，より多くの事例を収集するとともに，今回検討した手法を実現して機械翻訳の精度向上を図りたい．

謝辞 本研究は，文部科学省科学研究費補助金 (課題番号 13780223) および堀情報科学振興財団からの補助を受けて行われています．

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] 竹内和夫: 現代ウイグル語四週間, 大学書林 (1991).

[6] 飯沼英二: ウイグル語辞典, 穂高書店 (1992).

[7] 島津明, 内藤昭三, 野村浩郷 : 助詞「の」が結ぶ名詞の意味関係の subcategorization, 情報処理学会研究会報告, NL53-1, pp. 1–8 (1986).

[8] 池原 悟, 村上 仁一, 宮本 健司 :「A の B」型名詞句の日英翻訳規則について, 情報処理学会論文誌, Vol. 43, No. 7, pp.2300–2308 (2002).

[9] イラム・シャハザド, 大竹清敬, 増山繁, 山本和英 : 非対訳コーパスを用いた日本語複合名詞の英訳語推定, 情報処理学会研究会報告, NL133-2, pp. 7–12 (1999).