日本語－ウイグル語機械翻訳における単語接続関係を用いたウイグル語文の生成方法

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
名古屋大学　三重大学†

1　はじめに

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して、形態素解析が終了した段階で各単語を逐語訳するという手法が考えられる[1][2]。その中でも我々は、日本語とウイグル語を共に派生文法[3][4]で記述することにより、複雑な語形変化をする動詞句も、基本的には逐語訳によって翻訳が可能になることを示した[2]。

しかし、日本語とウイグル語の間の文法的差異から、単純な逐語訳では不自然な翻訳となる例が存在する。その問題に対して、我々は日本語形態素解析システムに与える文法を変更する形での対処方法を示した[2]。ところが、この手法は、翻訳システムとは独立であるべき日本語形態素解析部分を変更するという点で問題があった。

そこで、本稿では日本語形態素解析システムに与える文法は変更せず、単語間の接続関係を考慮し、逐語訳された単語を別の訳語に置き換えることにより、適切な翻訳文を生成する手法を提案する。

なお、本稿では音韻論的手法である派生文法に基づいて日本語を記述するため、日本語表記の一部にローマ字を用いる。その結果、日本語、ウイグル語の両者ともローマ字で表記され混同しやすいため、本文中では日本語の単語は「」、ウイグル語の単語は“”で囲んで区別する。

2　派生文法に基づく日－ウ機械翻訳

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。そのため日本語－ウイグル語機械翻訳においては、構文解析を行わず、日本語入力文の形態素解析を行っ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で、ある程度の翻訳が可能となる。図1にその例を示す。

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。そうした点に着目すると、「書カセラレタ。」のように動詞に複数の接尾辞

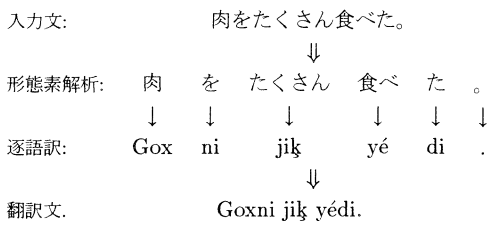

図1: 日本語－ウイグル語逐語翻訳

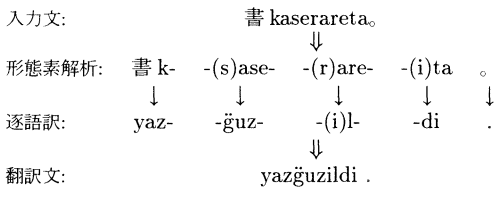

図2: 派生文法に基づく動詞句の翻訳

が接続した文も逐語訳で翻訳が可能となる。その例を図2に示す。

ここで、日本語－ウイグル語間の語形成における共通点を派生文法の用語を用いて示すと以下のようになる。

1. 語幹に接尾辞が接続することで語が形成される。
2. 語幹に子音幹と母音幹の区別がある。
3. 新たな語幹を派生する接尾辞が存在する。
4. 連結母音、連結子音の欠落規則が存在する。
5. 動詞接尾辞の順序の制約がほぼ同じである。

日本語とウイグル語では、動詞語幹に接尾辞が接続することによって動詞句が形成される。学校文法では、日本語の動詞は活用するとされ複雑な活用規則があったが、派生文法においては、動詞は活用しないとされ、いわゆる動詞の活用も語幹への接尾辞の接続として表現される。

また、動詞の語幹は音素単位で区切られるため、動詞「書ク」の語幹は「書」ではなく、「書k」となり、このように子音で終わる語幹は**子音幹**と呼ばれる。そ

れに対して、「食べル」の語幹「食 be」のように母音で終わる語幹は**母音幹**と呼ばれる。日本語の動詞の語幹は子音幹と母音幹に分類されるが、ウイグル語の動詞においても子音幹、母音幹の区別がある。たとえば、「書 k」に相当する動詞 "yaz-" の語幹は子音幹であり、「食 be」に相当する動詞 "yé-" の語幹は母音幹である。

図2 における「-(s)ase-」「-(r)are-」は学校文法の助動詞に相当する接尾辞である。派生文法ではこれらの接尾辞を、ある語幹に接続し新たな語幹を派生する接尾辞であるととらえ、**派生接尾辞**と呼ぶ。ウイグル語においても、派生接尾辞は存在し、その役割も日本語と一致する物が多い。たとえば、使役を意味する日本語の「-(s)ase-」にはウイグル語の派生接尾辞 "-ğuz-" が、受身を意味する日本語の「-(r)are-」にはウイグル語の派生接尾辞 "-(i)l-" がそれぞれ対応する。また、日本語で完了を表す「-(i)ta」や、それに対応するウイグル語の "-di" は新たな語幹を派生しない接尾辞であり、派生接尾辞に対して**統語接尾辞**と呼ばれる。

派生接尾辞「-(s)ase」や統語接尾辞「-(i)ta」の先頭の括弧内の音素は**連結子音**もしくは**連結母音**と呼ばれる音素である。語幹と接尾辞の接続においては、「末尾が子音 (母音) の語幹に接続する場合、連結子音 (連結母音) が欠落する」という規則がある。たとえば、子音幹「書 k」に接尾辞「-(s)ase-」が接続する場合、連結子音 "(s)" が欠落し「書 kase-」となる。そのような連結子音と連結母音はウイグル語にも存在し、たとえば受身の派生接尾辞 "-(i)l-" の "(i)" は連結母音であり、母音幹 "yé-" に接続する場合には欠落して "yél-" となる。

日本語は語順が比較的自由な言語と言われるが、動詞語幹に接続する接尾辞の順序には明らかに制約がある。この接尾辞の順序も図2 に見られるように、日本語とウイグル語でほぼ同じである。

3 日 – ウ機械翻訳における問題点

日本語とウイグル語は類似性が高いが、異なる部分もある。そのため、日本語単語とウイグル語訳語を1対1に対応付けできない場合がある。その場合、単純な逐語訳では不自然な翻訳となる。本節では、そのような問題点を例を挙げて説明する。

3.1 終止形と連体形

日本語では終止形と連体形の形が同じであり、たとえば完了を表す統語接尾辞「-(i)ta」は終止形と連体形の両方の役割を果たす。しかし、次の例のように、ウイグル語では終止形と連体形を表す統語接尾辞はそれぞれ別の単語である。

終止形: 彼が書いた。　　U yazdi.
連体形: 彼が書いた本。　U yazğan kitap.

したがって、「-(i)ta」の翻訳においては、訳語として "-di" であるか "-ğan" であるかの選択が必要となる。

3.2 派生語幹の不一致

否定の派生接尾辞は日本語では「-(a)na-」であり、ウイグル語では "-ma-" である。「書 kanakatta」を単純に逐語訳すると、以下のようになる。

書 k-	-ana-	-katta
↓	↓	↓
動詞	派生接尾辞	形容詞接続 統語接尾辞
↓	↓	↓
yaz-	-ma-	-kén

ここで、 "-kén" は形容詞の語幹に接続して完了の意味を表すウイグル語であり、日本語の "-katta" に相当する。しかし、実際のウイグル語では、 "-ma-" の後に接続する単語は "-kén" ではなく、日本語の動詞接尾辞「-(i)ta」に相当する "-di" である。これは日本語の「-(a)na-」が形容詞語幹を派生するのに対して、ウイグル語の "-ma-" は動詞語幹を派生するからである。このように派生接尾辞「-(a)na-」と "-ma-" では派生する語幹が異なるため、単純な逐語訳では不自然な翻訳となる。

3.3 サ変動詞

日本語のサ変動詞とは、その単語の基本となる形が名詞であるが、接尾辞「スル」が接続することによって動詞化する単語のことである。たとえば「開発」「登録」がその例であり、日本語には数多く存在する。なお、「サ変」と呼ばれるのは、接続する動詞化接尾辞「スル」がサ行変格活用をするからである。

ウイグル語には、この「スル」に相当する単語として "qilmaq" がある。たとえば、「開発」に相当するウイグル語名詞は "kéxip" であるが、これが動詞化して「開発スル」となる場合、ウイグル語では "kéxip qilmaq" となる。よって、「スル」の訳語として "qilmaq" を登録すれば、逐語訳による翻訳が可能となる。

しかし、この手法では自然な翻訳ができない例が存在する。たとえば、「登録」に相当するウイグル語は "tizimlax" であるが、「登録スル」に相当するウイグル語は "tizimlax qilmaq" ではなく、 "tizimlamaq" である。ここで、 "tizimla-maq" の "-maq" は動詞の辞書見出しをつくる接尾辞であり、語幹は "tizimla-" である。つまり、 "tizimlax" は、動詞語幹 "tizimla-" に名詞化接尾辞 "-(i)x" が接続することによって形成される名詞であり、 "tizimlax qilmaq" は日本語で「登録スルコトヲスル」に相当する冗長な表現なのである。そのため、「登録スル」の翻訳に際しては、逐語訳 "tizimlax qilmaq" ではなく、 "tizimlamaq" と翻訳するのが望ましい。

4 従来の解決方法

我々が [2] で提案した日本語 – ウイグル語機械翻訳は、派生文法に基づく形態素解析システム **MAJO** [5] を利用している。MAJO は派生文法に基づいて日本語の形態素を解析するシステムであり、辞書に各単語の情報が ⟨ 日本語単語, 品詞, 意味 ⟩ の 3 項組の形で登録されている。[2] では MAJO の辞書を ⟨ 日本語単語, 品詞, ウイグル語訳 ⟩ の 3 項組で表される日本語 – ウイグル語辞書に置き換えて使用した。翻訳結果には連結子音、連結母音が含まれるため、それらの処理をウイグル語整形システムで行っている。

さらに、3 節で述べた問題に対しては、MAJO に与える形態素文法を以下のように変更することにより対処している (図 3)。

まず、終止形と連体形の区別については、元々の MAJO の形態素文法で「終止・連体形」で表現されていた属性を「終止形」と「連体形」に分割して対処している。

次に、否定の派生接尾辞「-(a)na-」 "-ma-" については、否定の派生接尾辞と後接する接尾辞を合成し、1 語の単語として辞書に登録し、また、その合成語に特別な品詞を付与することで対処してきた。たとえば「書カナカッタ」を翻訳するために、⟨-anakatta, 子体辞, -madi⟩ などの単語を辞書に登録してきた。

なお、サ変動詞の問題に関しては、[2] では対処できていなかった。

[2] の手法は、できるかぎり逐語訳で翻訳を行う点を重視している。しかし、各モジュールの独立性を考慮した場合、日本語形態素解析を行うシステムである MAJO に翻訳上の問題を処理させる方法は望ましくない。特に、この手法を日本語 – 韓国語などの他の膠着語間の機械翻訳に応用する場合には、MAJO に与える形態素文法をそれらの言語の特徴に合わせて変更することになる。また、⟨-anakatta, 子体辞,-madi⟩ のように、登録する単語の単位が不自然なものとなり、日本語の 1 単語とウイグル語の 1 単語が 1 対 1 に対応するという直観に合わなくなる。

そこで本稿では、形態素解析システム MAJO に与える文法を変更せずに、これらの問題を解決する方法として、単語の接続関係を検査し、それに基づいて訳語置換を行う手法を提案する。

5 訳語置換表

[2] の手法では、1 つの単語を翻訳するときに、その単語の情報しか利用していなかった。しかし、実際の翻訳においては、逐語訳する単語の前後には他の単語が存在しており、その情報を利用することが可能である。特に、3 節で挙げた例は、いずれもウイグル語における単語間の接続関係が原因となっている。そこで、逐語訳する単語とその前後のウイグル語との接続関係

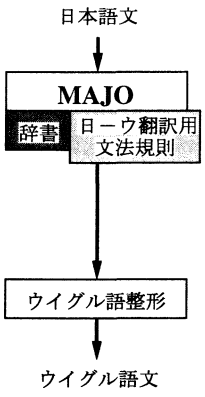

図 3: 従来の翻訳システム

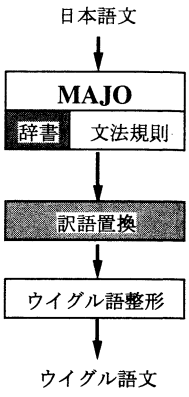

図 4: 提案するシステム

を利用して、自然な訳語を選択できるようにする。

ウイグル語の単語間の接続関係を利用するためには、訳語選択の前にすべての入力語をあらかじめウイグル語に翻訳しておく必要がある。そこで、MAJO が使用する日本語 – ウイグル語辞書に基本訳語を登録しておき、まず入力文をその基本訳語を用いて逐語訳する。その後、文の先頭から順に、各訳語とその前後の訳語を調べ、他の訳語が適切である場合は、その訳語を置き換える。この置換規則を記した表を訳語置換表 (表 1) と呼ぶ。

したがって、今回提案するシステムでは、図 4 のように、MAJO とウイグル語整形システムとの間に訳語置換システムを設け、MAJO に与える文法規則を変更することなく翻訳が可能となる。

表 1では、一番左の列に日本語が記述してあるが、これは表の理解を助けるためのものであり、実際のシステムが用いる訳語置換表では省略される。次の列のウイグル語が、その日本語の基本訳語である。前接ウイグル語および後接ウイグル語は、基本訳語を置換する場合の条件を示しており、それらのウイグル語が前後にあった場合、基本訳語を新訳語で置換する。前接ウイグル語および後接ウイグル語の欄には、基本的にウイグル語単語を記述するが、規則の記述を簡潔にするため、単語の代わりに品詞を記述することも可能とする。なお、前接ウイグル語または後接ウイグル語に依存しない置換規則の場合は条件の欄に * (don't care) を記述しておく。また、今回の手法では、訳語置換を行うかどうかを文の先頭から順に検査するため、訳語を置換した場合には、置換した訳語の品詞が必要になる。そこで、それを新品詞の欄に記述しておく。ただし、基本訳語と新訳語の品詞が同じ場合には空欄とする。

なお、一つの訳語を置き換えた場合、すぐに新訳語を出力するのではなく、その新訳語が更に訳語置換表の別の条件を満たしていないかを検査する。これは、

表 1: 訳語置換表

日本語	基本訳語	前接ウイグル語	後接ウイグル語	新訳語	新品詞
-(r)u	-[y]diğan	*	文末	-ydu	終止接尾辞
	-[y]diğan	*	句読点	-ydu	終止接尾辞
	-[y]diğan	*	終助辞	-ydu	終止接尾辞
-(i)ta	-ğan	*	文末	-di	終止接尾辞
	-ğan	*	句読点	-di	終止接尾辞
	-ğan	*	終助辞	-di	終止接尾辞
-katta	-kén	-ma-	*	-ğan	
登録	tizimlax	*	şil-	tizimla-	サ変動詞
減少	azayix	*	şil-	azay-	サ変動詞
si-,su-,se-	şil-	サ変動詞	*	-	

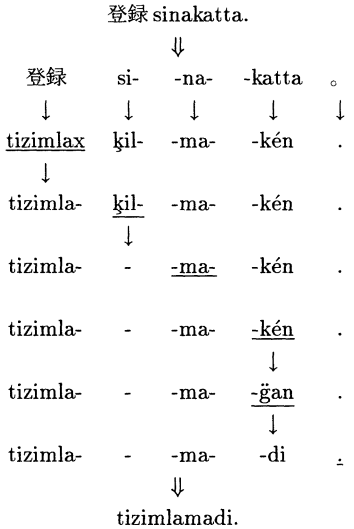

図 5: 提案手法における翻訳の例

いったん置き換えた訳語が別の置換条件を満たす場合があるからである。

この手法による例文「登録シナカッタ。」の翻訳過程を図 5 に示す。

「登録」の基本訳語は “tizimlax” であるが、日本語のサ変動詞「si-」の訳語 “şil-” が後接するため、訳語置換表の規則から “tizimla-” に置き換えられる。次の “şil-” は、前の単語が “tizimla-” に置き換えられたため、前接単語がサ変動詞であるという条件を満たす。したがって、何も語を訳出しないことを意味する “-” に置き換えられる。次に、「-katta」の基本訳語 “-kén” は前接のウイグル語単語が “-ma-” であるため、訳語置換表の条件に合致し “-ğan” に置き換えられる。また、この置き換えられた “-ğan” を訳語置換表で再検査すると、後接のウイグル語が句読点であるという条件を満たすため、終止形を表す “-di” に置き換えられる。この結果、最終的に入力文「登録シナカッタ。」に対する自然な翻訳文 “tizimlamadi.” が得られる。

6 おわりに

本稿では、派生文法に基づく日本語 – ウイグル語逐語翻訳において、不自然な翻訳を避けるための訳語置換表を提案した。その結果、動詞句の翻訳においてより自然な翻訳文を生成することが可能となった。また、今回の提案では訳語置換の条件として、ウイグル語の単語および品詞の情報に限定しているが、意味情報などを扱えるように拡張することにより、訳語の多義性の解消などにも応用が可能と考えられる。

今後は、このシステムを使用した翻訳実験を進めると共に、実用的な日本語 – ウイグル語翻訳システムの実現を目指す。

参考文献

[1] ムフタル・マフスット, 外山勝彦, 稲垣康善: 日本語 – ウイグル語機械翻訳における助動詞のパラメータ化による処理, 電子情報通信学会技術研究報告, NLC 94-13, pp. 47–53 (1994).

[2] 小川泰弘, ムフタル・マフスット, 外山勝彦, 稲垣康善: 派生文法に基づく日本語 – ウイグル語機械翻訳 — 動詞接尾辞の変換 —, 情報処理学会研究会報告, NL 120-1, pp. 1–6 (1997).

[3] 清瀬義三郎則府: 日本語学とアルタイ語学, 明治書院 (1991).

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

[5] 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 (1999) 掲載予定.

[6] 戸部実之: ウイグル語入門 – 文法と会話 –, 泰流社 (1986).

[7] 竹内和夫: 現代ウイグル語四週間, 大学書林 (1991).

[8] 金泰錫, 浦昭二: 日韓機械翻訳における意味接続関係を用いた韓国語の生成方法, 情報処理学会論文誌, Vol. 33, pp. 1578–1588 (1992).

[9] 金政仁, 権泰光, 大駒誠一: 日韓機械翻訳における活用語処理のための拡張テーブルの改善, 言語処理学会 第 2 回年次大会発表論文集, pp. 9–12 (1996).