派生文法に基づく日本語からモンゴル語への文節翻訳

百順，長谷部紀元，石川徹也
筑波大学大学院図書館情報メディア研究科

1 はじめに

本研究では，まだ存在していない日本語からモンゴル語への機械翻訳[1]システムの実現を目指す.

モンゴル語は日本語と同じく，類型的には膠着語に属していて，文法体系がよく似ている．特に文節単位での両言語の語順はほとんど同じである．語幹形成においても，語幹に接尾辞が膠着して新しい語幹を作る点が同じである.

このために日本語からモンゴル語への翻訳においては，日本語文の形態素解析結果を直接翻訳することで，ある程度のモンゴル語訳文を実現できる．しかし，両言語間の音韻変化と語幹形成には文法的に異なる点が多い．また，語幹と接尾辞に対して適切な訳語選択が必要となる.

本稿の段階では，文節単位翻訳の高正訳率の実現を目的にする．そのために日本語文を，日本語の膠着語的性格を基礎とする派生文法[1]に従ってとらえる．文節を派生文法により解析し，語幹と接尾辞列を個別的にモンゴル語に変換し,更にモンゴル語の音韻規則処理を適用し,文節生成を行う文節単位の翻訳手法を提案する.

以下 2 でシステム構成の方針，3 でシステムの概要を紹介し，4 で評価実験，5 で結論と今後の課題について述べる.

2 システム構成の方針

2.1 構成方針

現段階の翻訳システムの実現に当たっては，以下の方針を採用する．(1) 日本語文の文節区切りに，伝統的文法に基づく既存の解析システムを利用する．(2) 文節毎に，伝統文法の自立語と付属語の列を，派生文法によって語幹と接尾辞の列に解析し直す．これによって，より強力な文法的情報を抽出することを目指す．(3) 解析結果の語幹と接尾辞を 1 対 1 にモンゴル語へ変換する．(4) モンゴル語の語幹と接尾辞の列に音韻規則を適用して，モンゴル語の文節を生成する.

[1] 以下では，日-モ機械翻訳と表記する.

2.2 派生文法による日本語文の解析

派生文法[1][3]は日本語の膠着語としての性質に着目する．派生文法では伝統的日本語文法における用言の活用を，語幹への接尾辞の膠着として解釈する．結果として派生文法では活用の概念が存在しない．派生文法の特徴をまとめると以下のようになる.

(1) 4 種の語幹

派生文法では，伝統的文法の自立語に当たるものの主要部分を動作動詞語幹，形状動詞語幹，実名詞語幹，形状名詞語幹の 4 種に分ける．これらはそれぞれ，伝統的文法の品詞である動詞，形容詞，名詞，形容動詞に対応する.

(2) 接尾辞膠着による文節形成

これら自立語相当のものは，文節の先頭に一次語幹として現れる．文節は一次語幹に接尾辞の列が膠着することによって形成される.

(3) 派生接尾辞と統語接尾辞

派生文法の接尾辞には派生接尾辞と統語接尾辞の 2 種がある．派生文法においては，「書カセル」を kak-ase-ru と解析する．語幹 kak に接尾辞 -ase が膠着して二次語幹「書カセ」 kak-ase を派生する．つまり派生文法では，語幹に接尾辞が順次に膠着することによって，二次的な語幹が派生する．この種の接尾辞を派生接尾辞と呼ぶ.

一方 -ru のような接尾辞は，新たな語幹を派生することがなく，膠着によって文節を完成する．この種の接尾辞を統語接尾辞と呼ぶ．統語接尾辞は文節間の関係を統制する.

(4) 動作動詞の子音語幹と母音語幹

ここでは文中で特に重要な動作動詞の語幹を考える．動作動詞語幹を末尾音によって，子音語幹と母音語幹に分ける．語幹末音によって，次項で述べるように接尾辞膠着の振る舞いが変わる.

伝統的文法の五段活用動詞「話ス」の語幹は hanas である．このような子音で終わる動詞語幹を子音語幹と呼ぶ．一方，一段活用動詞「起きる」「食べる」の語幹は「oki」「tabe」である．このような母音で終わる動詞語幹を母音語幹と呼ぶ.

(5) 連結子音と連結母音

動詞語幹に接尾辞が膠着するときには，次の 2 個の規則が適用される.

規則 1: 子音語幹に子音で始まる接尾辞が膠着するときには，接尾辞先頭の子音が欠落する.
規則 2: 母音語幹に母音で始まる接尾辞が膠着するときには，接尾辞先頭の母音が欠落する.

上記規則 1 に従って欠落する子音を連結子音と言う. 例えば，子音語幹 hanas に接尾辞 sase を膠着するとき，接尾辞の先頭の子音 s が欠落して hanas-ase となる. 規則 2 に従って欠落する母音を連結母音と言う. 例えば，母音語幹 tabe に接尾辞 ita を膠着するとき，接尾辞の先頭の母音 i が欠落して tabe-ta となる.

連結子音あるいは連結母音を持つ接尾辞を，欠落の可能性がある音を括弧に入れて，(s)ase，(i)ta のように表記する. これは sase と ase，ita と ta がそれぞれ，同一接尾辞の異なる表現型であることを表している.

以上のように派生文法では，文法的構造の表現のために，ローマ字による音素単位の表記を必要とする. 日本語表記で習慣的に用いられる仮名による音節単位の表記では不十分である.

(6) 正則文法

派生文法における語幹への接尾辞膠着による語幹種の変化を，膠着する接尾辞を入力とする有限状態オートマトンの状態遷移と見なすことができる. 従って派生文法を，この状態遷移を指定する接尾辞の定義表によって規定することができる. 言い換えると，文節の形成は正則文法に従う. 図 1 に，一次語幹の動作動詞語幹の出現と接尾辞膠着による状態遷移の例を示す.

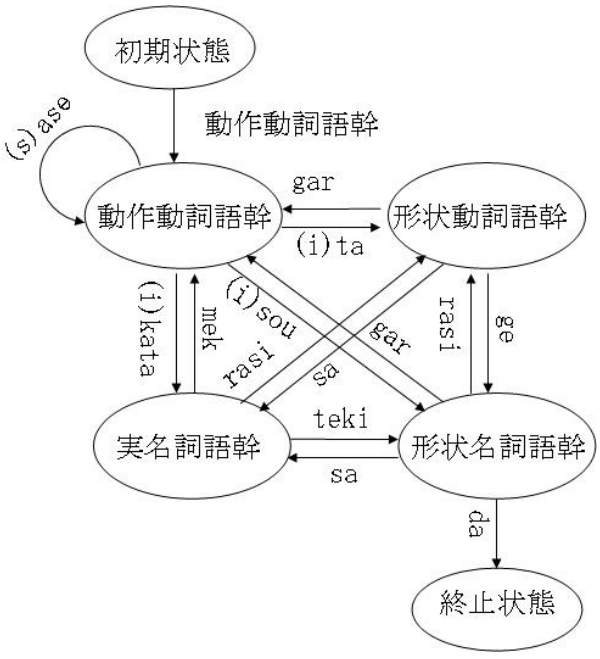

図 1 派生文法の有限状態オートマトン・モデル

3 システムの概要

3.1 システムの構成

我々が派生文法に基づいて構築した文節翻訳システムは 4 個のモジュールからなる(図 2). この中で“語幹・接尾辞の解析”が派生文法による解析を行う. 以下に各モジュールについて説明する.

形態素と構文の解析には，伝統的な日本語文法に基づいている JUMAN と KNP を利用する. JUMAN/KNP に日本語文を与えると，自立語と付属語の列からなる文節の木がえられる. 以降のモジュールは文節を単位に処理を行う.

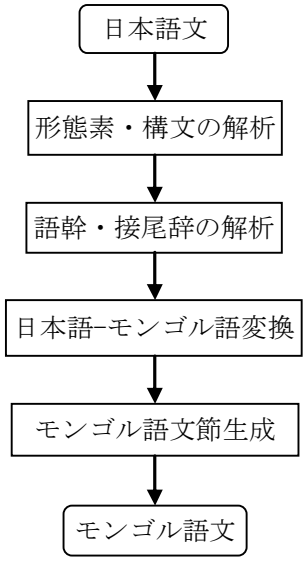

図 2 システム構成図

3.2 語幹・接尾辞の解析

このモジュールは JUMAN/KNP によって解析された文節に対し，派生文法に基づく接尾辞解析を行う. これは以下の 4 個のサブモジュールから成る(図3). この中で “接尾辞解析”が派生文法適用の中心である.

3.2.1 語幹整理

派生文法の文節では 1 個だけの語幹に接尾辞列が膠着する. JUMAN/KNP の切り出す文節が 1 個だけの自立語を含む場合には，それを派生文法の語幹に対応させればよい.

語幹整理では，文節中に複数の自立語がある場合の処理を行う. 具体的には以下の規則で処理する.

(1) JUMAN/KNP 解析結果文節中の自立語の列を合成して一つの語幹に対応させる. 例えば,「一括処理する」という文節には「一括」,「処理」,「する」の 3 個の自立語が含まれている. それを「一括処理する」という 1 個の自立語に合成し，結果から語幹を作る.

(2) 伝統文法にある助数詞のような造語接尾辞も合成対象に含める.

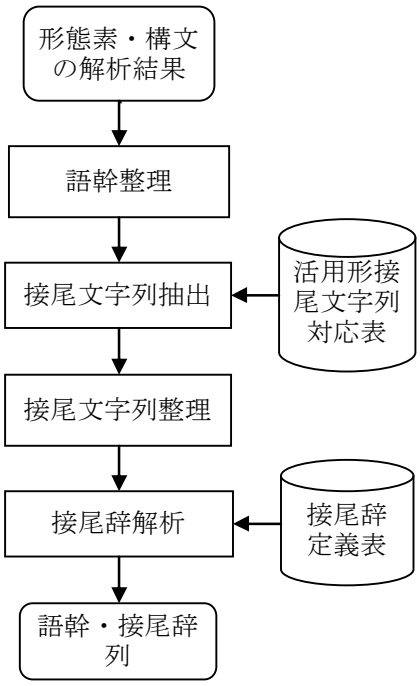

図 3 語幹･接尾辞の解析

(3) 合成語幹の種類を自立語列の最後の品詞から決定する．上例では，合成自立語「一括処理する」に対応する語幹種が，「する」に対応する動作動詞になる．

3.2.2 接尾文字列抽出

接尾文字列抽出は文節を，語幹と，それに接尾する文字列に分割する．

派生文法では伝統的文法の用言の各種活用形を，語幹に各種の接尾辞が膠着して形成されるものと考える．但し活用形によっては，活用語尾が膠着した接尾辞の先頭部分だけを含む場合がある．未然形の a，基本連用形の i がそれであって，これは多数の接尾辞で共通の連結母音である．子音動詞「書く」の活用形を形成する接尾文字列の例を表1に示す．

表1 活用形の接尾文字列の例

活用形	伝統文法	派生文法	接尾文字列
未然形	書か	kak - a	(a)
基本連用形	書き	kak - i	(i)
基本形	書く	kak - u	u
基本条件形	書けば	kak - eba	eba
意志形	書こう	kak - ou	ou
命令形	書けば	kak - e	e

抽出方法としては，先ず JUMAN/KNP 結果の用言の活用形の種類から，語幹に膠着している接尾辞またはその先頭部分を接尾辞文字列として決定する．次に，用言に後続する付属語の文字列を連結して，文節としての接尾文字列の全体とする．

語幹に膠着する接尾文字列は，活用形だけでなく用言の品詞と活用種にも依存する．活用形接尾文字列対応表はその情報を整理したものである．

3.2.3 接尾文字列整理

2.2 (4) で見たように，派生文法では動作動詞などに膠着する接尾辞をローマ字で扱う必要がある．それで必要な仮名ローマ字変換を行って，接尾文字列を整理する．

3.2.4 接尾辞解析

本サブモジュールは語幹に膠着している接尾文字列を解析して接尾辞の列に変換する．

2.2 (6) で見たとおり，語幹と接尾辞の列は正則文法に従う．それで，可能な語幹と接尾辞の列を正規表現で記述することができる．接尾辞解析処理では，この正規表現を接尾辞定義表から生成し，対象文節の接尾文字列とパターン･マッチする．接尾辞定義表は，派生文法のすべての接尾辞と，JUMAN/KNP が接尾辞として扱うものを含むように作成した．

派生文法による接尾辞は粒度が小さく，伝統文法の活用形や付属語に比較して，意味が明確である．翻訳の目的言語がモンゴル語のように膠着語である場合には，特に効果が大きい．

3.3 日本語–モンゴル語変換

日本語の語幹と接尾辞をそれぞれ，変換辞書によってモンゴル語に変換する（図 4)．

日–モ接尾辞変換表は 3.2.4 で定義した日本語接尾辞を対象としている．日本語の接尾辞に対応するモンゴル語の接尾辞が複数存在する場合には，最もよく使われるものを選んで固定的に変換している．

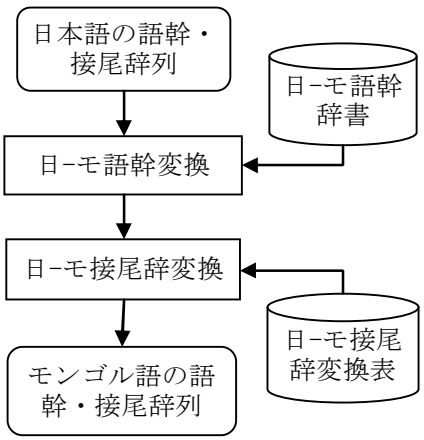

図4日本語–モンゴル語変換

3.4 モンゴル語文節生成

モンゴル語の語幹と接尾辞を膠着して，モンゴル語

文節を生成する (図5).

文節生成に当たっては，モンゴル語の語幹と接尾辞の列に対して音韻規則を適用し，正しいモンゴル語文節を生成する．特に，接尾辞の表現形を最大 3 個の可能性の中から選択する．他の音韻規則に母音調和規則，母音と子音の結合規則，子音調和規則，連接母音挿入規則がある[2].

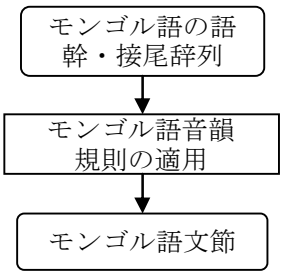

図5 モンゴル語文節生成

4 評価実験

本システムの翻訳能力の検証のために，文節単位で翻訳結果を評価した.

4.1 対象データ

「CD-毎日新聞，2002」[4]の中から農林水産関連の 50 記事を対象にした (表 2).

表2 対象データ

項目	個数	項目	個数
総文数	246	記事当たり文数	4.9
総文節数	2807	文当たり文節数	11.4
異なり文節数	2178	文当たり異なり文節数	8.8

4.2 評価方法

ここで正訳とはモンゴル語として正しい表記をいう．なお 3.2.1 の語幹整理による合成語の未知語を含む場合も，正訳に含めた．誤訳とは文法または意味が間違っている表記をいう.

4.3 結果と考察

50 記事の異なり文節数 2,178 に対して正訳文節 2,077，正訳率 95.36% の結果を得た(表 3).

誤訳の原因の内訳を表 4 に示す.

(1) 誤訳の原因のうち，最も多いのが接尾辞解析失敗の 86 個であり，全体の 85% を占める.

動作動詞語幹に関するものでは，伝統的文法のサ行変格活用動詞スルに関わるものが，合成語を含めて 21 個にのぼる．原因はこの動作動詞を，固定的に母音語幹を持つものとしたことである．活用形に応じて子音語幹に変わるとすることで，改善が期待できる.

実名詞語幹に関するものでは，語幹整理での日本語文中の記号の取り扱いの不十分さに起因するものが 13 個ある.

形状名詞語幹と形状動詞語幹に関する誤訳原因は雑多である．JUMAN/KNP 独特の付属語を派生文法的に解析するために接尾辞解析を強化するなど，各種対策の必要がある.

(2) 多義語の処理は現在行っていない.

(3) 訳語不適切表現は，モンゴル語として文法的または，意味的に正しくない文節である．原因には，「これ」などの指示詞と接尾辞の関係が両言語の間で異なることなどがある.

表3 正訳率

異なり文節数	正訳数(未知語を含む)	正訳率
2178	2077(116)	95.36%

表 4 誤訳原因の内訳

誤訳原因	文節種	個数	誤訳率
1)接尾辞解析失敗	実名詞語幹	18	0.83%
	動作動詞語幹	53	2.43%
	形状名詞語幹	5	0.23%
	形状動詞語幹	10	0.46%
2)多義語		5	0.23%
3)訳語表現不適切		10	0.46%
合計		101	4.64%

5 結論と課題

本稿では，日本語からモンゴル語への機械翻訳システムの開発を目指し，文節を派生文法に基づいて翻訳する方式を提案し実現した．評価実験の結果，文節単位での高精度翻訳が可能であること確認した.

今後の課題として，文節単位の正訳率向上のためには接尾辞解析を，JUMAN/KNP の特性や現代語で重要な補助動詞に対応するよう，強化することがある．更に文節単位翻訳の成果を元に，次の段階である文単位翻訳システムの実現に進む．文中の文節間の関係を考慮することによって，文節末の「と」の格助詞と接続助詞の混同などの解決が可能になる．文節間関係を利用しての意味処理によって初めて，重要な多義語問題の解決が可能になる．また語幹整理によって発生する合成語の未知語に関しては，合成過程を元に自動的にモンゴル語訳を生成する方法を検討したいと考えている

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] チンゲルタイ:蒙古語語法,内蒙古人民出版社 (1991).

[3] 小川泰弘，ムフタル・マフスット，杉野花津江，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,vo1.40, No.3, pp.1081-1090 (1999).

[4] CD - 毎日新聞 2002，毎日新聞社.