音韻論的・形態論的制約を用いたモンゴル語句生成・形態素解析

Sanduijav ENKHBAYAR　　佐 藤 理 史　　宇 津 呂 武 仁

京都大学 情報学研究科

1. はじめに

モンゴル語においては，自立語の語幹に対して格を表す語尾や動詞の活用を表す語尾・接続助詞等が結合したものが句を構成し，ヨーロッパ言語と同様に，空白で区切られた句の列により文を構成する．ここで，モンゴル語の形態素解析の問題について考えると，この問題は，モンゴル語文中の名詞句や動詞句が与えられて，それらの句を名詞あるいは動詞の語幹と語尾とに分解することであると言える．この処理を実現するためには，名詞あるいは動詞の語幹に語尾が接続する際の接続可能性や語変形の規則性を明らかにする必要がある．また，例えば，他の言語からモンゴル語への機械翻訳においては，名詞あるいは動詞の語幹および語尾が与えられると，その語幹・語尾の組に対する語変形や活用の過程を規則化し，名詞句あるいは動詞句を生成する機構を確立する必要がある．

このような状況をふまえて，我々は，文献[1]において，現時点で利用可能なモンゴル語の言語資源，特に，名詞・動詞の語幹のリスト，および，名詞・動詞に接続する語尾のリストから，モンゴル語の名詞句・動詞句を生成する手法を提案した．そこでは，具体的には，名詞・動詞の語幹に語尾が接続する際の音韻論的・形態論的制約を整備し，語幹・語尾の語形変化の規則を作成した．また，評価実験の結果において，8 割以上の場合について，生成された名詞句・動詞句の中に正しい句候補が含まれるという性能を達成した．また，文献[2]ではモンゴル語の名詞句・動詞句の形態素解析を行なう手法を提案した．具体的には，まず，既存のモンゴル語辞書から名詞語幹および動詞語幹を人手で抽出する．次に，これらの語幹に対して，文献[1]において整備したモンゴル語名詞句・動詞句生成規則を適用することにより，語幹・語尾の組から句を生成するための語形変化テーブルを作成する．そして，この語形変化テーブルを参照することにより，与えられた名詞句・動詞句を形態素解析して語幹・語尾に分離する．評価実験の結果においては，語形変化テーブルに登録されている句については，形態素解析の結果得られる語幹・語尾の組合せの候補の中に，正しい解析結果が必ず含まれることが確認できた．この結果をふまえて，本稿では，数個のバグを訂正し，文献[2]で実装されていなかった語尾変化☆ を含めて語幹に複数個 (名詞句の場合は 4 個，動詞句の場合は 8 個) の語尾を接続した句の生成を実現した．そして，句生成・形態素解析の過程を実験によって評価し，句生成・形態素解析の性能が改善したことを確認した．特に，実在しない誤った句候補も含めて，生成された句候補を全て用いて語形変化テーブルを作成し，形態素解析の評価を行った結果では，誤った句の影響による性能の低下はほとんどなかった．

2. モンゴル語の文法

現代モンゴル語で使われる文字はキリル文字である．モンゴル語では，自立語の語幹に対して格を表す語尾や動詞の活用を表す語尾・接続助詞等が結合したものが句を構成し，ヨーロッパ言語と同様に，空白で区切られた句の列により文を構成する．モンゴル語の語順は日本語と同じ SOV で，動詞が文末に位置し，その他の句の語順は比較的自由である．

通常，名詞の語幹には，数を表す語尾，格を表す語尾，再帰所属を表す語尾がこの順に接続する．名詞に接続する語尾の分類，および，各分類ごとの語尾の種類数を表 1 に示す．通常，同一の分類に対応する語尾には数種類の可能性があり，一つの名詞に接続する語尾を決定する際には，その複数の可能性の中から，3 節で述べる母音の接続制約，および，4 節で述べる語幹・語尾の接続制約を満たす語尾が選ばれる．さらに，4 節の語形変化規則により，語幹・語尾が語形変化する．名詞の語幹にこれらの語尾が接続した場合の語形変化の例を図 1 に示す．

同様に，動詞の語幹に接続する語尾は，命令・願望類，叙述類，完了・習慣等を表す類，順序関係を表す類，等に分類される．動詞の活用語尾の分類，および，各分類ごとの語尾の種類数を表 2 に示す．動詞の場合も，同一の分類に対応する語尾には数種類の可能性があり，一つの動詞に接続する語尾を決定する際には，その複数の可能性の中から，3 節で述べる母音の接続制約，および，4 節で述べる語幹・語尾の接続制約を満たす語尾が選ばれる．そして，4 節の語形変化規則により，語幹・語尾が語形変化する．動詞の語幹にこれらの語尾が接続して動詞が活用する例を図 2 に示す．

3. モンゴル語の母音字の接続制約

モンゴル語においては，名詞・動詞の語幹に語尾が接

☆ 名詞に関しては複数と否定，動詞に関しては叙述・過去，受身，使役，否定，完了である

表 3　動詞語幹＋形動詞・予定形語尾＋与位格語尾＋再帰所属語尾の語形変化の例

動詞語幹	*авар* (救う)
語形変化	*авар*＋*х*(形動詞・予定形語尾) −→ *авах* (救うこと) *авах* ＋*ыг* (与位格語尾) −→ *авахыг* (救うことを) *авахыг* ＋*аа* (再帰所属語尾) −→ *авахыгаа* (自分が救うことを)

表 1　名詞に接続する語尾の一覧

語尾の分類	語尾種類数
属格	7
対格	2
与位格	3
奪格	4
造格	4
共同格	3
再帰所属	4
複数	4
否定	1
合計	32

(1)　名詞語幹
хүүхэд (子供)
(2)　名詞語幹 + 複数語尾
хүүхдүүд (子供達)
(3)　名詞語幹 + 複数語尾 + 格語尾
хүүхдүүдтэй (子供達と一緒に)
(4)　名詞語幹 + 複数語尾 + 格語尾 + 再帰所属語尾
хүүхдүүдтэйгээ (自分の子供達と一緒に)
(5)　名詞語幹 + 複数語尾 + 否定の語尾
+ 格語尾 + 再帰所属語尾
хүүхдүүдгүйгээрээ (自分の子供達とは別に)

図 1　名詞の語形変化の例

(1)　動詞語幹
ид (食べる)
(2)　動詞語幹 + 受動態語尾
идэгд (食べられる)
(3)　動詞語幹 + 使役態語尾
идүүл (食べさせる)
(4)　動詞語幹 + 意志の語尾
идье (食べよう)
(5)　動詞語幹 + 単純過去の語尾
идсэн (食べた)
(6)　動詞語幹 + 形動詞・完了の語尾 + 否定の語尾
идсэнгүй (食べなかった)
(7)　動詞語幹 + 従属節 (限界) 語尾
идтэл (食べるまで)
(8)　動詞語幹 + 受動態語尾 + 単純過去の語尾
идэгдсэн (食べられた)

図 2　動詞の活用の例

続する際に，名詞・動詞の語幹の末尾および語尾において語形変化・活用が起こる．この語尾の接続においては，名詞・動詞に含まれる母音字と，語尾に含まれる母音字の間で，一定の接続制約が満たされる必要がある．これらの制約の詳細については，文献 [1] を参照されたい (なお，文献 [1] で述べた内容は文献 [4] に基づいており，日本語での用語等は文献 [3] に従っている)．また，文献 [2] では，語幹の全ての母音を考慮して，語尾との接続制約を満たすかどうかを調べていた．しかし，実際には，語幹の全ての母音を考慮する必要はなく，語幹の末尾の主母音のみを考慮すればよい．本稿ではこの方法を採用することにより，接続する語尾がほぼ一つに絞られている．ただし，この方法では外来語の場合に正しい句が生成できない場合がある．

表 2　動詞の活用語尾の一覧

		活用の分類	語尾種類数
1	命令・願望	1 人称意思 1	3
2		1 人称意思 2	2
3		2 人称命令	0
4		2 人称勧告	4
5		2 人称催促	4
6		2 人称懇願	2
7		1-3 人称願望	4
8		1-3 人称懸念	2
9	叙述	現在・未来	4
10		単純過去	4
11		体験過去	4
12		伝聞過去	2
13		過去	1
14	形動詞	完了	4
15		継続	4
16		予定	1
17		習慣	4
18		可能性	4
19	副動詞	連合	1
20		並列	2
21		分離	4
22		条件	8
23		継続	4
24		限界	4
25		即刻	4
26		随伴	4
27		付帯 1	2
28		付帯 2	4
29	その他	受身	1
30		使役	2
31		否定	2
32		完了	1
合計			95

4.　モンゴル語の語幹・語尾の語形変化

通常，同一の分類に対応する語尾には数種類の可能性があり，一つの名詞あるいは動詞に接続する語尾を決定する際には，その複数の可能性の中から，まず，前節で述べた母音字の接続制約を満たす語尾が選ばれ，さらに，語幹・語尾の接続制約を満たす語尾が選ばれる．そして，以下の四種類の語形変化規則により，語幹・語尾が語形変化する．

(1)　母音字消失の規則

表 4 名詞+語尾の語形変化テーブルの例 1

語幹/語幹品詞	*боловсрол*(教育)/名詞
語尾/語尾種類	*ын*(～ の)/属格
語形変化後の語	*боловсролын*(教育の)

表 5 名詞+語尾の語形変化テーブルの例 2

語幹/語幹品詞	*боловсрол*(教育)/名詞
語尾/語尾種類	*д*(～ に)/与位格
語形変化後の語	*боловсролд*(教育に)

表 6 動詞+活用語尾の語形変化テーブルの例 1

語幹/語幹品詞	*арилга*(消す)/動詞
語尾/語尾種類	*ж*(～ して)/副動詞:並列
語形変化後の語	*арилгаж*(消して)

表 7 動詞+活用語尾の語形変化テーブルの例 2

語幹/語幹品詞	*арилга*(消す)/動詞
語尾/語尾種類	*уузай*(～ するな)/命令・願望:1-3 人称懸念
語形変化後の語	*арилгуузай*(消すな)

(2) 軟音符ьが母音字 *u* に変化する際の規則
(3) つなぎの母音字の挿入規則
(4) 母音以外のつなぎの文字の挿入規則

これらの制約・規則の詳細は文献[1]を参照されたい.

5. 名詞・動詞の語幹リストの作成

見出し語数約 7,500 語の日本語・モンゴル語対訳辞書☆のモンゴル語見出し語から，以下の手順で，名詞・動詞の語幹を抽出した．まず，名詞については，見出し語が名詞の語幹で記述されているので，1,926 語を人手で抽出した．一方，動詞については，見出し語が形動詞・予定形で記述されている．そこで，まず，動詞の形動詞・予定形 1,254 語を人手で抽出し，形動詞・予定形を動詞・語幹と予定形・活用語尾に分離する形態素解析規則を適用した．この形態素解析における語幹の候補語数は，形動詞・予定形一単語あたり，平均で 1.365 語であり，この中に正しい語幹を含む率は 100%であった．この形態素解析結果に対して，人手で正しい語幹を選択し，動詞の語幹リストを作成した．さらに，形態素解析の実験に用いる句から語幹を人手で抽出したものを追加し，合計で名詞語幹 2,048 語，動詞語幹 1,258 語のリストを得た.

6. モンゴル語句候補生成の評価

前節で作成した，名詞語幹 2,048 語，および，動詞語幹 1,258 語について，以下の手順で名詞句・動詞句の句候補生成を行ない，その性能を評価した.

(1) 与えられた名詞もしくは動詞の語幹に対して，格や活用の分類に応じた語尾の全候補をまず求める.
(2) 3 節で述べた母音の接続制約に基づいて，語尾の候補を絞り込む.
(3) 4 節で述べた語幹・語尾の語形変化の規則を用いて，名詞・動詞の句候補を生成する.

☆ 制作者の清水幹夫氏から提供して頂いた.

名詞と動詞の語幹にそれぞれ，表 1 と表 2 中の語尾を一つだけ接続した句を生成する過程を評価した．評価実験の結果では，名詞については，句候補の平均数が 1.60，正しい句を含む率は 97.78%であった．動詞については，二種類の変化を除いて一意に生成できて (その二種類については句候補の平均数は 1.15)，句候補の平均数が 1.01，正しい句を含む率は 100.00%であった．ただし，句候補の平均数は全ての語幹を対象として算出したが，句生成の精度については，動詞語幹および名詞語幹を 100 語ずつ無作為に選び，それらに語尾を一つだけ接続した句を対象として算出した．動詞句の場合は生成された句は全て正しい．一方，名詞句の場合は，誤った句が生成されており，1 語幹につき 0.6 語の誤った句が生成されている．誤った句は，特に，複数，与位格，属格，奪格の語尾が接続する場合に多い.

7. 語幹・語尾の語形変化テーブルの作成

5 節で述べた名詞語幹 2,048 語，および，動詞語幹 1,258 語について，以下の手順で語幹・語尾の語形変化テーブルを作成した．文献[2]では正しい句を人手で選択していたが，本稿の実験では全ての句を登録した．名詞語幹および動詞語幹について，それぞれ，表 1 と表 2 中の全語尾を文法上の順番で接続して句を生成した．次に，語幹・語尾，および，語形変化後の句の情報を用いて語形変化テーブルを作成した．語形変化テーブルは以下の情報から構成した.

- 語幹，もしくは，語幹にいくつかの語尾が接続して語形変化した語．および，語幹の品詞.
- 新たに接続する語尾の種類，および，語尾.
- 語形変化後の語.

語形変化テーブルの例を表 4～7 に示す．名詞語幹 2,048 語，および，動詞語幹 1,258 語に対して，語形変化テーブルの数は，それぞれ，226,541 個，および，2,703,462 個となった．これらのテーブル中における句の重複数は，名詞句が 7,603 個 (名詞句の 3.36%)，動詞句が 126,945 個 (動詞句の 4.70%)，名詞・動詞の両方にわたって重複する句は 3548 個 (全体の 0.12%) であった．語形変化テーブル全体の中では，正しく生成された句と誤って生成された句が混在するが，誤って生成された句と正しく生成された句が偶然一致することはまれである．従って，正しく生成された句がテーブル中で重複する場合は，その句に対して得られる語幹・語尾の組合わせはいずれも文法的に正しく，意味・文脈情報を参照してその曖昧性を解消すべきものである.

語形変化テーブルを用いて句の形態素解析を行なった結果，語幹・語尾の組合せとして複数の候補が得られる場合の例を表 8 に示す．一つ目の例においては，“*ор*(ベッド，代わり)” あるいは “*орон*(国)” という，異なる二つの名詞語幹に対して，“*нын*(～ の)” あるいは “*ын*(～ の)” という異なる属格語尾が接続して語形変化した結果の句が同

表 8 形態素解析において複数の解析結果が得られる例

句	形態素解析結果 (語幹/語幹品詞+語尾/語尾種類)
орны	*ор*(ベッド，代わり)/名詞 + *ны*(～ の)/属格 *орон*(国)/名詞 + *ы*(～ の)/属格
хазсан	*хаз* (噛む)/動詞 + *сан* (～ した)/叙述・単純過去　(噛んだ (文末)) *хаз* (噛む)/動詞 + *сан* (～ した (連体修飾))/形動詞・完了　(噛んだ (犬))

表 9 コーパス中の句の内訳 (% (個数))

語形変化テーブル中の語幹・句の有無	名詞	動詞	重複	その他	合計
語幹・句とも存在する	93.5 (260)	93.5 (331)	100.0 (4)	0 (0)	86.3 (587)
語幹のみ存在する	6.5 (18)	6.5 (23)	0 (0)	0 (0)	6.0 (41)
語幹が存在しない	0.0 (0)	0.0 (0)	0.0 (0)	7.6 (52)	7.6 (52)
合計	40.9 (278)	52.1 (354)	0.6 (4)	7.6 (52)	100 (680)

じ表記になっている．この例の場合は，文の意味を考慮して形態素解析の曖昧性を解消する必要がある．一方，二つ目の例においては，“*хаз* (噛む)” という動詞語幹に対して，叙述・単純過去形語尾あるいは形動詞・完了形語尾が接続しているが，この二つの語尾が同じ表記となっており，語形変化した結果の句も同じ表記となっている．叙述・単純過去形の場合は，文末等に現れる過去形となり，形動詞・完了形の場合は，連体修飾用法となる．この例の場合は，この句の直後が名詞句かどうかによって，形態素解析の曖昧性を解消することができる．

8. モンゴル語形態素解析の評価実験

モンゴル語形態素解析の評価実験を行なうために，まず，モンゴル語コーパスを収集した．本稿で用いたモンゴル語コーパスは，ウェブ上のモンゴル語新聞一年半分のテキストを収集してコーパスとしたもの (延べ語数 206 万，異なり語数 11 万 5 千，30MBytes) である．このコーパスから，無作為に 680 語を収集し，前節で用意した動詞・名詞の語形変化テーブルを用いて各語の形態素解析を行なった．語形変化テーブルに誤った句が登録されていて，入力された句がそれと一致すると誤った解析結果が得られる．本実験では派生語が入力されて，それが誤った名詞句と一致し，誤った解析となった事例が一つあった．また，名詞句について複数の解析結果が得られたのは 11 語で，動詞句については 41 語であった．そして，名詞句かつ動詞句とした重複解析結果が 4 語であった．名詞句の 11 語のうち 6 語が，” 対格＋再帰” と” 再帰” との間の曖昧性の事例であった．動詞の 41 語のうち 32 語が表 8 の二つ目の例と同様の動詞・叙述・単純過去形と動詞・形動詞・完了形の間の曖昧性であった☆．

次に，前節で用意した動詞・名詞の語形変化テーブルが，コーパス中のどの程度の範囲の語に対応しているかのカバレージを評価するために，まず，680 語を，名詞，動詞，その他の単語に分類し，それぞれのクラスについて，語形変化テーブルに含まれるかどうかを判別し，以下に分類し，表 9 に結果を示した．名詞かつ動詞として重複した解析結果が得られた句については「重複」という欄に示した．

- 「語幹・句とも存在する」
- 「語幹のみ存在する」
- 「語幹が存在しない」

「語幹・句とも存在する」は，コーパス中の出現形が，そのままの形で，語形変化後の語として語形変化テーブルに含まれるものである．「語幹のみ存在する」は，コーパス中の出現形から判別した語幹は語形変化テーブルに含まれるが，コーパス中の出現形が，そのままの形で，語形変化後の語として語形変化テーブルに含まれてはいない，というものである．これらの語については，2 節で述べた語形変化以外の語形変化 (具体的には，派生語を生成する語形変化) を実装することにより，形態素解析が可能となる．「語幹が存在しない」は，コーパス中の出現形から判別した語幹が語形変化テーブルに含まれない，というものである．

9. おわりに

本稿では，現時点で利用可能なモンゴル語の言語資源，特に，名詞・動詞の語幹のリスト，および，名詞・動詞に接続する語尾のリストを用いて，モンゴル語の名詞句・動詞句の句生成・形態素解析を行なう手法を提案した．特に，誤った句候補も含めて，生成された句候補を全て用いて語形変化テーブルを作成し，形態素解析の評価を行った結果では，誤った句の影響による性能の低下はほとんどなかった．

謝辞: 本研究の一部は，次の研究費による: 21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」．

参 考 文 献

1) Sanduijav ENKHBAYAR, 宇津呂武仁, 佐藤理史. 音韻論的・形態論的制約を用いたモンゴル語句生成. 情報処理学会研究報告, Vol. 2004, No. (2004–NL–162), pp. 87–93, 2004.
2) Sanduijav ENKHBAYAR, 宇津呂武仁, 佐藤理史. 音韻論的・形態論的制約を用いたモンゴル語形態素解析. 情報処理学会研究報告, Vol. 2004, No. (2004–NL–164), pp. 41–46, 2004.
3) 栗林均. モンゴル語. 亀井孝, 河野六郎, 千野栄一 (編), 言語学大辞典，第 4 巻，世界言語編 (下–2), pp. 501–517. 三省堂, 1992.
4) С.Ганболор, Л.Тунгалаг. *Зөв бичих дүрмийн тулгуур дохио*. Улаанбаатар, 2000.

☆ この曖昧性は文献 2) の評価実験においても含まれていた．