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[†] 宇津呂 武仁^{††} 佐藤 理史^{†††}

本論文では、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から、モンゴル語の名詞句・動詞句を生成する手法を提案する。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する。評価実験の結果において、100%近くの場合について、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。さらに、本論文では、この句生成に基づいて、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する。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手で抽出する。次に、これらの語幹に対して、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。

キーワード: モンゴル語自然言語処理, 句生成, 形態素解析

Mongolian Phrase Generation and Morphological Analysis based on Phonological and Morphological Constraints

SANDUIJAV ENKHBAYAR[†], TAKEHITO UTSURO^{††} and SATOSHI SATO^{†††}

Using currently available Mongolian linguistic resources such as the lists of stems of nouns and verbs as well as the list of suffixes, this paper proposes a method for morphologically analyzing noun/verb phrases of the Mongolian language. More specifically, we first examine phonological and morphological constraints on connecting stems of nouns/verbs and suffixes, and invent inflection/conjugation rules for nouns/verbs. We experimentally show that, almost 100 % cases, correct noun/verb phrases can be found among the candidates of noun/verb phrases generated by the proposed method. Then, we compile a table for mapping a stem-suffix pair and a phrase to be generated from the stem-suffix pair. Morphological analysis of noun/verb phrases is performed by simply consulting this mapping table. We experimentally show that, by the proposed method, correct candidate of stem-suffix pair can be obtained from the given noun/verb phrases.

KeyWords: *Mongolian Natural Language Processing, Phrase Generation, Morphological Analysis*

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などにおいては、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。ところが、現時点で利用可能なモンゴル語の言語資源としては、数千語程度の規模の単語について語幹情報が登録された電子辞書、および、ウェブ上で収集可能な新聞記事等の電子テキストが存在するにすぎない。また、モンゴル語に関して、名詞あるいは動詞の語幹と語尾の組から名詞句あるいは動詞句を生成するための言語知識や規則なども全く整備されていない。また、そのような句生成のための言語知識・規則を運用すれば、モンゴル語の句の形態素解析を行なうこともできるが、現時点では、モンゴル語文の形態素解析を実用的規模で行なうことも実現されていない。

本論文では、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から、モンゴル語の名詞句・動詞句を生成する手法を提案する。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する。評価実験の結果において、名詞句の場合は98%程度、動詞句の場合は100%という性能で、生成された句の中に正しい句候補が含まれるという結果が得られた。さらに、本論文では、この句生成に基づいて、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する。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手で抽出する。次に、これらの語幹に対して、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。

以下、まず、2節においては、モンゴル語の文法の概要について述べる。3節においては、名

† 日立ソフトウェアエンジニアリング株式会社, Hitachi Software Engineering Co., Ltd.

†† 京都大学情報学研究科知能情報学専攻, Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

††† 名古屋大学大学院工学研究科電子情報システム専攻, Department of Electrical Engineering and Computer Science, Graduate School of Engineering, Nagoya University

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
xvvxэд (子供)
- (2) 名詞語幹 + 複数語尾
xvvxdvvd (子供達)
- (3) 名詞語幹 + 複数語尾 + 格語尾
xvvxdvvdтэй (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
xvvxdvvdтэйгээ (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
xvvxdvvdтэйгээрээ (自分の子供達とは別に)

図 1 名詞の語形変化の例

詞・動詞の語幹に語尾が接続する際に、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満たされるべき接続制約について述べ、4節においては、名詞・動詞の語幹に語尾が接続する際の語形変化規則について述べる。5節においては、モンゴル語句生成の評価実験について、6節においては、モンゴル語形態素解析の評価実験について、それぞれ述べる。また、7節においては、関連研究について述べる。

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

表 2 動詞の活用語尾の一覧 (その 1)

		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

表 3 動詞の活用語尾の一覧 (その 2)

		活用の分類	語尾種類数
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			96

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

- (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

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2～3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。

なお、名詞・動詞に関して、本論文の執筆段階において実装されていないものとして、派生語がある。派生語とは、名詞語幹あるいは動詞語幹に派生語語尾が接続して語形変化した結果の語であり、名詞語幹から構成される派生動詞、名詞語幹から構成される派生名詞、および、動詞語幹から構成される派生名詞がある。派生語語尾としては、数十種類のものがある。派生語の内部構造を解析するためには、派生語に対して語幹・語尾の語形変化規則を適用して形態素解析を行なう必要があるが、実用的には、既知の派生語を語幹として登録し、形態素解析を行

表 4 モンゴル語の母音

基本母音字	а, о, У, и, з, Ө, v
補助母音字 1 (子音+母音と見なされる)	я, е, ё, ю
補助母音字 2 (長母音として扱う)	ы
補助母音字 3 (基本的に単独では使われない)	й
長母音	аа, оо, УУ, ий, зз, ӨӨ, vv
二重母音	ай, ой, Уй, зй, вй

表 5 モンゴル語の子音字

母音を必ず伴う子音字	м, н, г, л, б, в, р
母音を伴わなくてもよい子音字	д, т, ч, ж, ц, з, с, ш, х
特殊子音字 (外来語に使われる)	п, ф, щ, к

なうという方式が妥当であると考えられる。また、句生成および形態素解析において、名詞と同様の扱いが可能なものとして形容詞がある。モンゴル語の形容詞には、名詞に接続する語尾のうち複数語尾を除く語尾がすべて接続可能であり、語幹・語尾の語形変化の規則についても、名詞句の語形変化で用いている規則がそのまま適用できる。

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。本章では、この語尾の接続において、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、どのような接続制約が満たされる必要があるかについて述べる。なお、本章および次章で述べる内容は、(Ганболор and Тунгалаг 2000) に基づいており、日本語での用語等は(栗林 1992)に従っている。

表 6 モンゴル語の母音と性の関係

男性母音	а, о, У, я, ё, ю, ы
女性母音	з, Ө, v, е, ю
中性母音	и

表 7 モンゴル語の母音字の接続制約

語幹の母音字	語尾の母音
а, ʏ, я	а, я, ай
о, ё	о, ё, ой
з, в, и, е, ю	з, е, зй
Ө	Ө
а, о, ʏ, я, ё, ю, ы	ʏ, ы
з, Ө, в, и, е, ю	в, ий

表 8 母音字の接続制約の例外: 語幹の末尾の文字と語尾の先頭字の接続制約

語幹の末尾の文字	語尾の先頭字
ж, ч, ш, и, г, ь	ий

3.1 モンゴル語の母音字と子音字

モンゴル語で用いる文字は全 35 文字で、母音字 13 字、子音字 20 字、記号文字 2 字から構成される。長母音、二重母音を含めたモンゴル語の母音の一覧を表 4 に、子音の一覧を表 5 に、それぞれ示す。母音字のうち補助母音字 1 は、多音字であり子音+母音とみなされる。また、補助母音字 1 に基本母音を一つつけるという形でも使われる。補助母音字 2 は、単独で長母音として扱われる。補助母音字 3 は、単独で使われることはなく、他の母音とともに使われ二重母音を構成する。

また、モンゴル語の母音は、男性・女性・中性の三つの性を持ち、その内訳は表 6 となる。モンゴル語の単語の性は、強勢が置かれる母音の性によって決まる。

3.2 母音字の接続制約

動詞・名詞の語幹と語尾の接続の際に、双方の母音字の間で満たされるべき接続制約をまとめると、表 7 となる。この表は、語幹の母音字、および、それに接続可能な語尾の母音の組の一覧となっている。なお、ここで接続可能な母音同士は、表 6 で示した性が同一のものとなっている。ただし、この場合、中性母音は女性母音として扱う。

また、表 7 の母音字の接続制約に対する例外として、名詞・動詞の語幹の末尾の文字と語尾の先頭字の間においては、表 8 に示す接続制約が満たされなければならない。ここで、表 8 に示す接続制約は、名詞・動詞の語幹の末尾の文字が表 8 左側の文字となる場合で、しかも、語尾の候補として、表 8 右側の文字を先頭字として持つものが含まれる場合には、必ずその語尾が選ばなければならない、と解釈される。そして、このときには、表 7 の母音の接続制約は必ずしも満たされる必要はない。

表 9 語幹の末尾と語尾の接続制約

語幹の末尾	語尾
н	г
с, х	т, ч
母音	л
母音を必ず伴う子音字 + 軟音符 ь	母音を伴わなくてもよい子音字
子音字 + 軟音符 ь	я, ё
子音字 + 軟音符 ь	г в й
母音	е, я, ё

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、前節で述べた母音の接続制約を満たす語尾が選ばれ、さらに、4.1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、4.2節の語形変化規則により、語幹・語尾が語形変化する。

4.1 語幹・語尾接続制約

動詞・名詞の語幹の末尾と語尾の接続において満たされるべき接続制約をまとめると、表 9 となる。この表は、語幹の末尾、および、それに接続可能な語尾の組の一覧となっている。

4.2 語幹・語尾の語形変化

語幹・語尾の接続における語形変化の規則は、以下の四種類である。

- (1) 母音字消失の規則 (表 10)
- (2) 軟音符 ь が母音字 и に変化する際の規則 (表 11)
- (3) つなぎの母音字の挿入規則 (表 12)
- (4) 母音以外のつなぎの文字の挿入規則 (表 13)

語幹・語尾の接続における語形変化の際に、語形変化後の語のアクセントが変化する場合がある。この場合、特に、アクセントが変化して、母音が発音されなくなることがあり、この発音されなくなった母音字が消失する。その際の規則性を記述したものが「母音字消失の規則」である。ただし、以下の場合には、必ずしも母音字が消失しなくてもよい。

- (1) 「母音を必ず伴う子音字」に伴っている母音が消失する場合
- (2) н, г の直後の母音字

表 10 母音字消失の規則

	語の末尾	語尾の先頭	語形変化後
(i)	и 以外の母音	長母音	語の最後の母音が消失
(ii)	母音字 + 子音字	長母音	子音字 + 長母音
(iii)	母音字 ₁ + 子音字 ₁	子音字 ₂	子音字 ₁ + 母音字 ₂ + 子音字 ₂

表 11 軟音符 ь が母音字 и に変化する際の規則

	語幹の末尾	語尾の先頭	語形変化後
(i)	子音字 + 軟音符 ь	長母音 (xx)	子音字 + иx
(ii)	子音字 + 軟音符 ь	母音を必ず伴う子音字	子音字 + и + 母音を必ず伴う子音字
(iii)	母音を伴わなくてもよい子音字 ₁ + 軟音符 ь	母音を伴わなくてもよい子音字 ₂	母音を伴わなくてもよい子音字 ₁ + и + 母音を伴わなくてもよい子音字 ₂
(iv)	子音字 + 軟音符 ь	x (形動詞予定形活用語尾)	子音字 + иx

- (3) ж, ч, ш 以外の子音字の直後の и
 (4) 固有名詞の母音字
 (5) 形動詞予定形の母音字

語幹・語尾の接続における語形変化の際に、軟音符 ь が母音字 и に変化することがある。その際の規則性を記述したものが「軟音符 ь が母音字 и に変化する際の規則」である。また、語幹・語尾の接続において子音が連続する場合は、つながりの母音字を挿入する。その際の規則性を記述したものが「つながりの母音字の挿入規則」である。その他の場合で、語幹・語尾の接続において、母音以外のつながりの文字を挿入する場合もある。その際の規則性を記述したものが「母音以外のつながりの文字の挿入規則」である。

表 12 つなぎの母音字の挿入規則

	語幹の末尾	語尾	語形変化後
(i)	母音を伴わなくてもよい子音字	子音字	母音を伴わなくてもよい子音字 + 母音字 + 子音字
(ii)	母音を必ず伴う子音字 ₁	母音を必ず伴う子音字 ₂	母音を必ず伴う子音字 ₁ + 母音字 + 母音を必ず伴う子音字 ₂
(iii)	c, ш	л	c, ш + л + 母音字

表 13 母音以外のつなぎの文字の挿入規則

	語幹の末尾	語尾の先頭	語形変化後
(i)	(女性語) 子音字	e	子音字 + 軟音符 ь + e
(ii)	(男性語) 子音字	я, ё	子音字 + 硬音符 ь + я, ё
(iii)	長母音 ₁	長母音 ₂	長母音 ₁ + г + 長母音 ₂
(iv)	長母音 ₁	長母音 ₂	長母音 ₁ + нх + 長母音 ₂

5 モンゴル語句生成

5.1 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。

5.2 モンゴル語の句が生成されるパターン

モンゴル語の句が語幹からどの順番で生成されるかを図で示す。名詞句の場合を図3に、動詞句の場合を図4に示す。

図3 名詞語幹に語尾が接続する順序

図4 動詞語幹に語尾が接続する順序

表 14 名詞+属格の語形変化の例

語幹	боловсрол
語尾候補	ны, ын, ы, ний, ий, ийн (属格の全語尾候補) → ны, ын, ы (男性名詞に接続可能な語尾候補) → ны, ын (属格固有の接続制約: ыは語の末尾がн以外には接続不可)
語形変化	(語幹+ны)→ боловсролны (語幹+ын) → боловсролын (語幹+ын) → боловсролын → (母音字消失規則(ii)) → боловсрлын
人手による検証	боловсролын

5.3 例

表 14～18 に、名詞+属格の語形変化の例、名詞+与位格の語形変化の例、動詞の活用「副動詞:並列」の例、動詞の活用「命令・願望:1-3 人称懸念」、および、「動詞語幹+形動詞・予定形語尾+与位格語尾+再帰所属語尾」の例をそれぞれ示す。

表 14 の名詞+属格の語形変化の例においては、「боловсрол」(教育) という名詞に、「～の」という意味の属格語尾が接続した場合の語形変化の様子を示す。まず、属格語尾の全候補として六種類の語尾が得られるが、このうち、男性名詞「боловсрол」に接続可能な語尾は三種類に絞られる。さらに、属格固有の語幹・語尾接続制約により、接続可能な語尾は二種類に絞られる。この二種類の語尾が語幹に接続すると、語形変化を伴わず語幹に語尾がそのまま接続した形の句候補が二種類、および、母音字消失規則(ii)が適用された形の句候補が一種類、合計三種類の句候補が生成される。今回の評価実験では行っていないが、人手でこれらの句候補の検証を行なった場合は、一種類の句候補のみが得られる。

表 15 の名詞+与位格の語形変化の例においては、「боловсрол」(教育) という名詞に、「～に」という意味の与位格語尾が接続した場合の語形変化の様子を示す。この場合、与位格語尾の全候補二種類がそのまま接続した形の句候補二種類の他に、与位格固有のつなぎの子音字の挿入、および、つなぎの母音字の挿入が適用され、さらに二種類の句候補が生成される。今回の評価実験では行っていないが、人手でこれらの句候補の検証を行なった場合は、一種類の句候補のみが得られる。

表 16 の動詞の活用「副動詞:並列」の例においては、「ир」(来る) という動詞の活用の様子を示す。また、表 17 の動詞の活用「叙述:伝聞過去」の例においては、「гар」(出る) という動詞の活用の様子を示す。表 16 および表 17 のどちらの場合も、全語尾候補二種類がそのまま接続

表 15 名詞+与位格の語形変化の例

語幹	боловсрол
語尾候補	д, т (与位格の全語尾候補)
語形変化	(語幹+д)→ боловсролд (語幹+д)→ боловсролд → (与位格固有のつながの子音字 н の挿入) → боловсролдн → (つながの母音字の挿入規則 (ii)) → боловсролдонд (語幹+т)→ боловсролт (語幹+т)→ боловсролт → (与位格固有のつながの子音字 н の挿入) → боловсролнт → (つながの母音字の挿入規則 (ii)) → боловсролонт
人手による検証	боловсролд

表 16 動詞の活用「副動詞:並列」の例

語幹	ир
語尾候補	ж, ч (副動詞:並列の全語尾候補)
活用	(語幹+ж)→ ирж (語幹+ч)→ ирч
人手による検証	ирж

した形の句候補二種類が生成される。いずれの場合も、人手でこれらの句候補の検証を行なった結果においては、文法的に正規の句候補が一種類だけ得られる。ただし、実際には、表 17 の「rap」の活用「叙述:伝聞過去」の場合は、慣習的に表記の揺れが起こっており、誤った句候補の方も用いられることがある。また、活用「副動詞:並列」の場合も、動詞によっては表記の揺れが起こる場合があり、現在の句生成規則は、そのような場合を考慮した設計となっている。表 16 の「ир」の活用「副動詞:並列」において、誤った句候補が生成されるのは、このことが原因である。

表 18 に形動詞・予定形の変化を示す。モンゴル語の動詞の形動詞の中に名詞として変化する一部がある。表 18 にその一例を示す。こういう語形変化済み句を更に語形変化すると一つ前の変化の結果によって、接続される語尾に限られる。本論文では語形変化を複数回すると、一つ前の語尾を考慮している。そのため、表 18 に示す句は一意に生成されている。

表 17 動詞の活用「叙述:伝聞過去」の例

語幹	гар
語尾候補	жээ, чээ (叙述:伝聞過去の全語尾候補)
活用	(語幹+жээ)→ гаржээ (語幹+чээ)→ гарчээ
人手による検証	гаржээ

表 18 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) → аврах (救うこと) аврах + ыг (与位格語尾) → аврахыг (救うことを) аврахыг + аа (再帰所属語尾) → аврахыгаа (自分が救うことを)

5.4 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。

名詞と動詞の語幹にそれぞれ、表 1 と表 2~3 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。評価の詳細な結果は、名詞の結果は、表 19 に、動詞の結果は、表 20~21 に示す。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。ここで、正しい句を含まない場合の多くを占めるのは外来語であり、外来語が語幹の場合の句生成は、本論文で用いているモンゴル語の句生成規則に従わないことが多いと言える。例えば、оптерон (Opteron マシン) という外来語名詞語幹に造格語尾 оор (~で) が接続して語形変化をする場合、句生成規則に従えば、оптеронの下線部の母音 о が消失し、句候補 оптерноор が生成される。しかし、実際には、消失された母音も発音されるため、正しい句は、оптероноор となる。このような外来語の句生成に対処する方法としては、外来語に特化した句生成規則を用意することが考えられる。また、動詞については、二種類の変化 (形態素解析において表記の揺れに対処するために、二通りの句を生成するように規則が設定されている副動詞:並列 (表 16) および叙述:伝聞過去 (表 17)) を除いて一意に生成できて (その二種類

表 19 名詞の句候補手順の評価

語尾の分類	語尾種類数	(一名詞あたり)	
		句候補数の平均	正しい句候補を含む率 (%)
属格	7	2.000	97.0
対格	2	2.803	99.5
与位格	3	1.107	97.0
奪格	4	2.00	97.0
造格	4	1.107	97.0
共同格	3	1.000	99.0
再帰所属	4	1.107	97.0
複数	4	2.237	97.0
否定	1	1.000	100.0
合計/平均	32	1.596	97.78

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, 二種類の変化を除いて, 生成された句は全て正しい。句候補の平均数が 1.15 となる二種類の変化についても, 形態素解析において表記の揺れに対処するために, 二通りの句を生成するように規則が設定されているためであり, これらの変化において句候補の平均数を 1.00 とすることは容易である。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。

6 モンゴル語形態素解析

6.1 語幹・語尾の語形変化テーブルの作成

5.1 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。表 1 と表 2~3 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 22~25 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

表 20 動詞の句候補生成手順の評価 (その 1)

		活用の分類	語尾種類数	(一動詞あたり)	
				句候補数の平均	正しい句候補を含む率 (%)
1	命令・願望	1 人称意思 1	3	1.000	100.0
2		1 人称意思 2	2	1.000	100.0
3		2 人称命令	0	1.000	100.0
4		2 人称勧告	4	1.000	100.0
5		2 人称催促	4	1.000	100.0
6		2 人称懇願	2	1.000	100.0
7		1-3 人称願望	4	1.000	100.0
8		1-3 人称懸念	2	1.000	100.0
9	叙述	現在・未来	4	1.000	100.0
10		単純過去	4	1.000	100.0
11		体験過去	4	1.000	100.0
12		伝聞過去	2	1.154	100.0
13		過去	1	1.000	100.0
14	形動詞	完了	4	1.000	100.0
15		継続	4	1.000	100.0
16		予定	1	1.000	100.0
17		習慣	4	1.000	100.0
18		可能性	4	1.000	100.0

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 3,548 個 (全体の 0.12%) であった。

なお、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がある。これらの例を表 26 に示す。一つ目の例においては、“op (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、“ны (~の)”あるいは“ы (~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

表 21 動詞の句候補生成手順の評価 (その 2)

		活用の分類	語尾種類数	(一動詞あたり)	
				句候補数の平均	正しい句候補を含む率 (%)
19	副動詞	連合	1	1.000	100.0
20		並列	2	1.154	100.0
21		分離	4	1.000	100.0
22		条件	8	1.000	100.0
23		継続	4	1.000	100.0
24		限界	4	1.000	100.0
25		即刻	4	1.000	100.0
26		随伴	4	1.000	100.0
27		付帯 1	2	1.000	100.0
28		付帯 2	4	1.000	100.0
29	その他	受身	1	1.000	100.0
30		使役	2	1.000	100.0
31		否定	2	1.000	100.0
32		完了	1	1.000	100.0
合計/平均			96	1.010	100.0

表 22 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	ны (~ の)/属格
語形変化後の語	боловсролны (教育の)

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。

表 23 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 24 動詞+活用語尾の語形変化テーブルの例 1

語幹/語幹品詞	арилиа (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	арилиаж(消して)

表 25 動詞+活用語尾の語形変化テーブルの例 2

語幹/語幹品詞	арилиа (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	арилиУУзай (消すな)

6.2 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, 30 MBytes)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。680 語のうち、語形変化後の語が語形変化テーブルに含まれる語については、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ていた。

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 26 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ッ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 27 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」

表 26 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 27 コーパス中の句の内訳 (% (個数))

語形変化テーブル中 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)

「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。

7 関連研究

(江原, 早田, 木村 2004) においては、日本語形態素解析システム茶釜を処理系として、モンゴル語文の形態素解析を行なうための文法体系の構築を試みている。茶釜の処理系は、基本的には、活用語・非活用語の品詞体系、および、活用語の活用語尾の語形変化の体系を定義する機能を持つ。また、各形態素、および、二個もしくは三個程度の形態素の接続に対してコストを

定義することにより、形態素解析の結果得られる複数の解析結果を絞り込む機能を持つ。(江原他 2004)においては、茶釜の処理系が持つ機能のうち、活用語の活用語尾の語形変化の体系を定義する機能を利用することにより、名詞・動詞の詳細な活用型・活用形を定義している。また、語尾については、活用語とはせず、変化形をすべて別形態素として登録している。この方式と比較すると、本稿のアプローチは、茶釜辞書のような明示的な文法体系を立てるのではなく、できる限り抽象化したレベルで音韻論的・形態論的特性を整備し、この制約を用いて語幹・語尾の接続制約・語形変化規則を記述するというアプローチであると言える。本稿のアプローチでは、形態素として辞書に登録されるのは、(名詞・動詞の)語幹およびそれらの語幹に接続する語尾(の基本形)のみとなり、語幹や語尾の変化形を別途登録することはしない。そのかわりに、語幹・語尾が語形変化して生成される句については、その全ての可能性を語形変化テーブルに登録することとなる。ここで、日本語文の形態素解析と、モンゴル語における句の形態素解析を比較すると、モンゴル語においては、句が空白により分かち書きされる点が特徴的である。したがって、モンゴル語の句の形態素解析においても、実用的には、語幹に対して高々数個の語尾が連続して接続する可能性を考慮すれば十分である。本稿では、モンゴル語におけるこの特性を考慮して、句の候補をすべてテーブルに登録するアプローチを採用している。

(Cucerzan and Yarowsky 2002)においては、スペイン語およびルーマニア語について、時制・人称・数の活用変化形の生成規則を手で記述しておき、コーパス中で実際に観測される不規則変化形との間の類似度を計算して、各々の不規則変化形に対して、最も近い規則変化形の時制・人称・数を割り当てるという方法により、各言語での不規則変化形の形態素解析規則を獲得している。

8 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・動詞句の形態素解析を行なう手法を提案した。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。そして、この規則を用いて、語幹・語尾の組とそこから生成される句を対応させる語形変化テーブルを作成し、このテーブルを参照することにより、名詞句・動詞句の形態素解析を行なう手法を提案した。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。そして、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

日本語・モンゴル語対訳辞書を提供して頂いた清水幹夫氏に感謝する。また、(江原他 2004)の辞書データを提供して頂いた諏訪東京理科大学 江原暉将先生に感謝する。

参考文献

- Cucerzan, S. and Yarowsky, D. (2002). “Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day.” In *Proceedings of the 6th Conference on Natural Language Learning*, pp. 132–138.
- 江原暉将, 早田清冷, 木村展幸 (2004). “茶釜を用いたモンゴル語の形態素解析.” 言語処理学会 第10回年次大会論文集, pp. 709–712. 言語処理学会.
- Ганболор, С. and ТУНгалаг, Л. (2000). *Зөв бичих дүрмийн тулгуур дохио*. 栗林均 (1992). “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.

略歴

- Sanduijav Enkhbayar:** 2003年神戸大学工学部情報知能工学科卒業。2005年京都大学情報学研究科修士課程知能情報学専攻修了。現在、日立ソフトウェアエンジニアリング株式会社に勤務。在学中はモンゴル語の自然言語処理に関する研究に従事。
- 宇津呂 武仁:** 1989年京都大学工学部電気工学第二学科卒業。1994年同大学大学院工学研究科 博士課程電気工学第二専攻修了。京都大学博士(工学)。奈良先端科学技術大学院大学情報科学研究科助手、豊橋技術科学大学工学部情報工学系講師を経て、2003年より京都大学情報学研究科知能情報学専攻講師。自然言語処理の研究に従事。
- 佐藤 理史:** 1983年京都大学工学部電気工学第二学科卒業。1988年同大学院博士課程研究指導認定退学。京都大学工学部助手、北陸先端科学技術大学院大学情報科学研究科助教授、京都大学情報学研究科助教授を経て、2005年より名古屋大学大学院工学研究科教授。工学博士。自然言語処理、情報の自動編集等の研究に従事。

(2005年3月11日 受付)

(2005年5月2日 再受付)

(2005年5月8日 採録)