マイナー言語に対する言語処理基盤開発 –キリル文字モンゴル語の場合–

村脇 有吾
murawaki@nlp.kuee.kyoto-u.ac.jp
京都大学 大学院情報学研究科

概 要

モンゴル語は日本語と構造的に類似しているため、日本語における自然言語処理の成果が応用できる。また、言語処理の基盤が確立されていないので、日本語では今更取り組めないような基礎を再検討する機会を与えてくれる。本稿では、自然言語処理における解析の全体像を概観した上で、開発の初期段階で解決すべき主要な課題である品詞体系の設計と語彙の整備に取り組む。

A Groundwork for Processing Minority Languages –In the Case of Cyrillic Mongolian–

Yugo Murawaki
murawaki@nlp.kuee.kyoto-u.ac.jp
Graduate School of Informatics, Kyoto University

1 はじめに

なぜ日本人である著者が、何の縁もないモンゴル語の言語処理を行うのか。それは、モンゴル語を解析することが、日本語を知る上で役立つからである。

そもそも、計算機科学の究極的な目標が、計算機に知能を与えることとすれば、自然言語処理の目標は、計算機が人間のように言語を操れるようにすることである。自然言語処理の現状は目標の達成には程遠いが、それでも、人間をモデルとし、その機能の一部を計算機上に実装しようとしている。そのため、研究者は、人間は何を知っているから言語を正しく解析できるのかと考えたり、逆に、何も教えられていない計算機から世界がどう見えているのかと想像をめぐらせている。

計算機と人間を比べると、身体の有無が決定的に異なる。また、猿はシンボル表現としての言語を持たないが、ある程度の知能を持っている。従って、知能の本質が何かは分からないが、少なくとも、言語のみを観測して解析を行うのは奇形的と言える。言語のみしか観測できないという制約は、言語が持つ意味を言語以外と関連付けることによって表現できないということを意味する。過去にはSHRDLU[3]のように、言語を世界に関連付ける試みもあったが、自然言語処理は、現在でも言語外の情報を本格的に活用するに至っていない。

自然言語処理は言語だけで完結しているが、自然言語処理を用いたアプリケーションの目的は、言語を解析することではなく、解析結果から書かれた内容を知ることにある。そのために、知りたい情報と相関のある言語上の振る舞い、すなわち、文法的制約などの様々な特徴を総動員する。しかし、そもそも、意味を復元できるだけの情報が言語にコード化されているのだろうか。答えは否である。話し手は、聞き手の間で了解されている常識をわざわざ言語として表現しない。従って、計算機は、話し手が想定する常識を知らない限り、言語を理解できない。簡単な例として、複合名詞を挙げる。複合名詞の構成要素間の関係に強い制約はない。複合名詞の構成要素についていえることは、何らかの関係を持っているらしいということだけである。例えば、「産業革命」や「農業革命」といった用例から、「革

命」に前接する要素は、「革命」の対象 (object) と推測できるかもしれない。しかし、「IT 革命」の「IT」は対象というよりも手段である。このように、文の論理規則より指示対象が特定できるのではなく、あらかじめ指示対象を知っているからこそ理解できると考えられる。すなわち、言語を正しく解析するためには、様々な常識をあらかじめ知っていなければならない。しかし、そのような知識は、人工知能における知識表現の歴史が示すように、人手での記述が困難である。結局、常識もテキストを観測することによって獲得するほかない。複合名詞については、Nakov ら [1] が、tear gas に対する gas that brings tears のような、構成要素に関する動詞を介した用例を収集し、その関係を明らかにしている。

言語が持つ様々な特徴を総動員する以上、利用できる特徴は言語によって異なる。例えば、文章から人名を抽出するタスクは、日本語に比べて英語が圧倒的に有利である。なぜなら、英語では固有名詞を義務的に大文字で書き始めるが、日本語文にはそのような特徴が現れないからである。逆に、構文解析においては、必ず前から後ろに係るという日本語の特徴が、強力な制約として働く。そのため、前に係るか後ろに係るかわからない英語や中国語に比べて、日本語は構文解析しやすい。

特定言語に固有の特徴を利用する一方で、特定言語に依存しない普遍性の解明も言語研究の目的の一つである。人間の脳が特定言語の話者ごとに異なるということはないため、言語にも何らかの形で普遍性があると考えられる。この矛盾する両者の折衷案は、同じ特徴を共有する複数の言語を考察することである。

日本語の場合、似た特徴を持つ言語は内陸アジアに広がっており、アルタイ諸語と総称される。モンゴル語は、このアルタイ諸語の一つである。後述するように、いくつかのアルタイ諸語の中で、モンゴル語が言語処理を通じた考察に適していると考えられる。

2 モンゴル語とその特徴

モンゴル語は、モンゴル国や中国領の内モンゴル自治区などで話されている言語である。モンゴル語には、ハルハ、チャハル、ホルチンなど、いくつかの有力方言がある。このうち、ハルハ方言が、モンゴル国全域で話されており、モンゴル国における標準語の地位を確立している。また、国際的にも、現在では、単にモンゴル語と言えばハルハ方言を指すことが多い。話者数は、モンゴル語全体で 570 万、ハルハ方言は 230 万と推定される。

2.1 文字

モンゴル語は、歴史上様々な文字で表記されてきたが、現在用いられているのは、モンゴル文字とキリル文字の 2 種類である。モンゴル文字は、古くから用いられてきた伝統的な文字で、中国領の内モンゴル自治区を中心とする地域で現在でも用いられている。モンゴル文字の綴りは古いモンゴル語に基づいており、現代のあらゆる方言の発音と乖離している。

キリル文字は、ソ連の衛星国家であったモンゴル人民共和国 (モンゴル国の前身) で導入された。キリル文字の正書法はハルハ方言に基づいている。モンゴル語を表記するキリル文字は、ロシア語に使用される文字に、母音を表記するために ө と ү の 2 (大文字を含めると 4) 文字が追加されている。モンゴル国では、民主化後の 1990 年代に、モンゴル文字の復活を目指す運動が一時盛んになったが、現在は事実上頓挫している。

図 1 にモンゴル文字の例を示す。これをラテン文字で転写すると、

čі yaGaGad iregsen ügei bui?

となる。同じ文をキリル文字で表記すると、

Чи яагаад ирсэнгүй вэ?

となり、同様にラテン文字で転写すると、

Chi yaagaad irsengüi ve?

となる。このように、モンゴル文字とキリル文字の綴りは単純には対応しない。

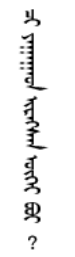


図 1: モンゴル文字

計算機上での処理を考えたとき、モンゴル文字は、その特異な性質ゆえに、現存する中で最も実装が難しい文字の一つである。モンゴル文字は、縦書きで、日本語とは反対に左から右へ改行する。また、アラビア文字のように続け書きするため、字形は語中の位置によって変化する。さらに、字母と音価の関係が多対多のマッピングとなる。例えば、o と u、t と d は同じ字形で現されるため、あらかじめ単語を知らなければ正しい発音すらままならない。このような複雑な問題は、音価を無視し、字形にのみ注目して符号化表を作れば回避できる [5]。しかし、この方式では、単に表示が行えるようになるだけで、辞書順ソートのような単純な言語処理すらできない。言語処理を行えるように音価を保存するには、特別な処理が必要となる [10][12]。

一方、本稿が対象とするキリル文字は、計算機上で問題なく取り扱える。旧来の 8 ビットの符号化方式は、ロシア語の文字コードとほとんど同じため、自

動判別が難しいという問題があった。こうした問題はUnicodeにより解決される。

2.2 言語的特徴

言語としてのモンゴル語の特徴は、日本語に近い構造を持っていることである。日本語で「京都へ」のように、内容語 (京都) に、文法機能を担う付属語 (へ) が後続する。同様に、モンゴル語も付属語が内容語に後続される。また、語順はいわゆるSOV型である。そのため、word-by-wordで日本語に翻訳することができる。上の例文は、Чи 「あなた (は)」、яагаад 「どうして」、ирсэнгүй 「来なかった」、вэ? 「か (疑問の助詞)」、と語順を入れ替えることなく翻訳できる。

モンゴル語は、構造的に日本語と類似しているものの、歴史上日本語との接触がほとんどないため、日本語と語彙が単純に対応しない。そのため、実際には、逐語訳ではなかなか自然な日本語とならない。この点、韓国語とは対照的である。韓国語は、大量の漢語を日本語と共有しているため、比較的単純な処理で実用レベルの機械翻訳が成し遂げられている。本格的なモンゴル語の機械翻訳は、日韓翻訳や日英翻訳などとは違った意味で挑戦的な課題になると思われる。

日本語と異なる点として、付属語の独立性が低いことが挙げられる。日本語の「の」は、どのような語に後続しても形を変えないが、モンゴル語の-ын は、表1のように、5つの異形態が複雑な条件で書き分けられる。また、母音調和の影響が付属語にも及ぶ。モンゴル語における母音調和とは、母音が男性母音と女性母音 (およびそのどちらにも属さない中性母音) にグループ化され、一つの単語中には同一グループの母音しか現れないという現象である。現代語では更に円唇性の対立が加わり、4グループに分けられることもある。例えば、造格 (〜によって) の語尾-aap は、前接する語の母音型によって-aap, -ээр, -оор, -өөр という4つの異形態のいずれかをとる。

このように、付属語が様々な異形態をとるだけでなく、形態素の連結も一筋縄ではいかない。 形態素を連結する際には、必要に応じて、子音挿入、母音挿入、母音削除、母音削除+母音挿入のいずれかの操作が行われる。例えば、ах (兄)、-тай (〜と)、-аа (再帰語尾) を連結すると、ахтайгаа (自分の兄と) となり、子音г が挿入される。あるいはявх (行く)、-сан (完了の語尾)、-аар (造格、〜によって) を連結すると、явснаар (行ったことによって) となり、母音а が削除される。これらの規則には、それぞれ適用される条件が決まっている。このような規則は、モンゴル語の固有語を対象としているため、外来語の扱いが問題となる。例えば、文字を同じくするロシア語の語彙は、そのままの綴りで借用することが定められているため、綴りと発音が乖離し、連接規則についても例外的な振る舞いをする。しかし、ほとんどの語学書や文法書は、外来語に関する連接規則を明らかにしていない。

2.3 研究動向

自然言語処理分野では、モンゴル語は、マイナー言語の中では比較的盛んに研究されている。例えば、言語処理学会の年次大会では、2003年以降、毎年モンゴル語に関する研究が発表されている。しかし、日本語と異なり、基盤技術について標準が確立されていない。これは、見方を変えれば、日本語の場合には今更できないような基礎作業に取り組めるということである。現在の研究者があまり振り返ることのない自然言語処理の基礎を考え直す機会をモンゴル語は提供している。

言語学の分野でも、モンゴル語は、マイナー言語としては研究が盛んである。日本では、政治的な背景から戦前にモンゴル語の研究が開始され、戦後も廃されることなく続いている。海外でも、アメリカ、ドイツ、ロシア、中国などに研究者が多い。

2.4 コーパス

言語学にしろ、自然言語処理にしろ、言語の解析には大量の実例 (コーパス) が欠かせない。コーパスには、人間が品詞などの注釈を施したタグ付きコーパスと、注釈なしの単なるテキスト (生コーパス) がある。自然言語処理では、タグ付きコーパスは統計的なパラメータの学習に広く用いられている。日本語のタグ付きコーパスとしては、例えば、毎日新聞を元にした京都コーパスがある。モンゴル語についても、タグ付きコーパスが欲しいところだが、整備には多くの人的リソースを必要とする。タグ付きコーパスの整備は今後の課題とし、ひとまず生コーパスを収集する。満洲語などの多くのマイナー言語では生コーパスの整備すら難しいが、モンゴル語はこの要求を容易にかなえてくれる。意外なことに、ウェブ上にはモンゴル語のテキストが豊富にある。実際、Google SOAP APIとwgetを組み合わせただけの非常に単純なクローラで、約100万のモンゴル語のページを収集できた。

3 言語学の工学的応用

言語が持つ性質を自然言語処理に応用するには、そもそも言語が持つ性質を明らかにしなければならない。

表 1: 属格語尾の書き分け規則 [4]

語形	前接する形態素の条件
-ийн	子音н で終わる語を除く女性語、及びж, ч, ш, ь, и, г で終わる男性語
-ын	ж, ч, ш, ь, и, г 以外の文字で終わる男性語
-ы	н で終わる男性語
-ий	н で終わる女性語
-н	二重母音及び長母音ий で終わる語

そのために、言語学の研究成果に期待がかかる。しかし、言語学側で提案された理論のほとんどが、自然言語処理に応用されていない。実際に、言語学的な知見を自然言語処理にグラウンディングさせるのは容易ではない。

橋本 [7] は、言語理論が工学に応用されるための条件を以下の四つに整理している。

1. **現象の重要性** 理論の説明する現象が応用において重要である
2. **設計の単純性** 理論が応用システムの設計を容易にする
3. **計算の容易性** 理論の予測を導くための計算が容易である
4. **入力の利用可能性** 理論の参照する情報が容易に入手できる

同様に、本稿でも、次の観点から、自然言語処理に応用できる研究成果を探し、また整理しなおす。

1. **観測可能性** 知りたい情報に対応する (少なくとも相関のある) 現象がテキスト上で観測できなければ解析できない。橋本の言う「入力の利用可能性」にほぼ対応する。
2. **網羅性** 解析器にとって、一部の現象を深く解析するよりも、浅い説明でもよいので、いかなる入力に対してもある程度正しい解析結果を返す方が重要である。しかし、語学書は典型的な用法しか説明しないし、言語学の研究もあらかじめ対象となる現象を絞り込んだ議論が少なくない。

4 問題の所在

自然言語処理における解析の全体像を概観し、言語処理基盤を一から開発する際に解決すべき課題が、品詞と語彙であることを述べる。

自然言語の解析に対しては、伝統的に、形態素解析、構文解析、意味解析という段階が設定されてきた。形態素解析を字句解析と言い換えればプログラミング言語と変わらない。このうち、形態素解析と構文解析については、一定の成果を挙げてきたが、意味については何も分かっていないに等しい。実のところ、計算機に解析させる以前に、人間が意味の適切な表現方法を知らない。自然言語処理の歴史は、テキストのみを観測している限り、テキストからあまり遠く離れられないということを教えている。形態素解析と構文解析がある程度成功している理由は、観測できる表層的な情報を使って、テキストに若干のラベルを付与する問題だからである。モンゴル語についても、構文解析を視野に入れつつ、ひとまず形態素解析器を作ることが目的となる。

解析器は、一般に、入力に対して出力を返す。プログラミング言語のコンパイラであれば、出力は入力に対して原則として一意に決まる。これに対して、自然言語には曖昧性があるため、複数の出力候補が考えられる。従って、解析器が行うべきことは、出力候補の列挙と、何らかの評価基準による最善の候補の選択である。

形態素解析器の場合、入力は文字列で、出力は品詞ラベルが付与された形態素の列である。日本語の形態素解析の場合、複数の出力候補を効率よく表現するために、一般に、図 2 のように、ラティス (格子) が用いられ、文頭から文末までのすべてのパスが候補となる。最適な候補を選択するための評価基準として、品詞間の連接と単語にコストを設定し、ラティスからコストが最小のパスを選択するという手法が用いられている。

モンゴル語の形態素解析も、基本的には日本語と同じである。もっとも、モンゴル語は分かち書きされるため、日本語ほど複雑なラティスは現れない。基本的には、図 3 のように、分かち書きされた単位 (語と呼ぶ) を自立語と付属語に分割するだけである。ただし、語の候補を列挙するためには、2.2 節で述べたような複雑な正書法の規則を整理してプログラム化しなければならない。

日本語の形態素解析には長い歴史があり、研究は既に終了していると思われるかもしれないが、現在でも

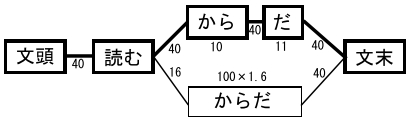


図 2: 日本語における形態素解析

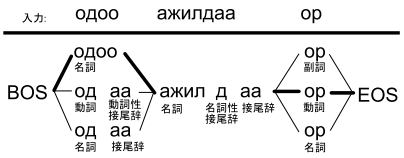


図 3: モンゴル語における形態素解析

行われている [9]。しかし、研究の焦点は、品詞間の連接や単語のコストというパラメータの機械学習による最適化にある。モンゴル語のように一から自然言語処理を始める場合、パラメータの学習に利用できるタグ付きコーパスが存在しない。複数の出力候補から最適な解を選び出す以前に、正解を含むような出力候補を列挙するところから始めなければならない。まずは、形態素に与えるべきタグ、すなわち品詞を適切に設計する必要がある。

また、語彙の整備も問題となる。形態素解析器は、辞書を引き、マッチした形態素を出力候補に加える。辞書にない未知語をその場で適切に処理するのは難しいので、出来るだけ多くの形態素を辞書に収録する必要がある。しかし、日本語の解析に用いられているような大規模な辞書は、とても著者一人で構築できるものではない。文法に関わる特殊な形態素や、ごく基本的な語彙は人手で整備するとしても、それ以外の形態素については、他の方法を模索しなければならない。

5 品詞

5.1 品詞体系の設計

品詞は、形態素を似た性質によって分類したもので、すべての形態素はいずれかの品詞に属す。品詞は、自然言語処理の解析において、入力文に最初に与えるラベルとして広く用いられている。新たな品詞体系の設計は、品詞にどのような情報を含ませるかという基本方針と、具体的な品詞の分類基準を明らかにする作業となる。

自然言語処理で用いられる品詞は、英語と日本語で割り当て方が異なる。英語では、多くの語が複数の品詞となりえる。例えば、dog には名詞と他動詞の可能性がある。英語における解析 (part-of-speech tagging) は、テキスト中の用語がどの用法で使われているかを明らかにするタスクである。一方、日本語では、原則的に、一つの形態素には一つの品詞が与えられている。例えば、「実行」のようなサ変名詞は、名詞として働くだけでなく、「する」が後続して動詞として振舞う。日本語の形態素解析では、テキスト中での用法にかかわらず、「実行」にサ変名詞とタグ付けする。テキスト中での用法の特定は、構文解析の段階で行われる。モンゴル語の品詞は、基本的に日本語と同じ方針で設計する。

日本語の処理に用いられる品詞は、一般にツリー状の階層構造をとる。例えば、日本語辞書である ipadic[8] に採用されている品詞体系は非常に詳細で、「名詞-固有名詞-人名-姓」のような深い階層が設定されている。この詳細な品詞体系により、形態素解析の段階で必要な情報がすべて品詞に詰め込まれている。見方を変えれば、解析器が個々の形態素について知っているのは、ある品詞に属すということだけである。

こうした方針と異なり、提案する品詞体系では、すべての情報を品詞に詰め込まない。例えば、固有名詞か否かのような文法的に本質的でない情報を品詞の下位分類としない。その結果、品詞体系は 2 階層の単純なものとなる。細かな情報は、品詞に代えて、任意の数の素性という形で各形態素に与える。特にモンゴル語の解析は、開発の初期段階にあり、形態素に与えるべき情報の全貌を把握できていないため、品詞のようなかっちりした体系よりも、素性のように柔軟な構造の方が都合がよい。

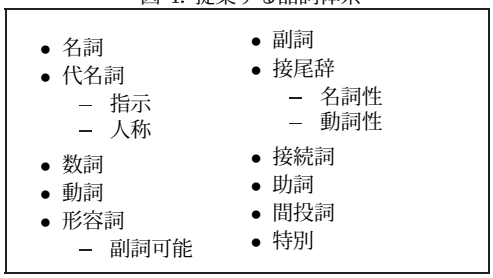
また、2.2 節で述べたように、モンゴル語における自立語に対する付属語の連接は、複雑な文字列操作を必要とする。連接に関する規則は、統語的な振る舞いとは独立である。そのため、連接に関する情報を品詞に組み込むと、統語的品詞との組み合わせとなり、品詞の総数が増大してしまう [13]。提案手法では、品詞とは独立に母音型を設定し、各形態素に記述する。

5.2 分類方針

個々の形態素がどのような基準で分類されているかはそれ程明らかではなく、様々な判断基準が渾然一体となっている。しかし、計算機で扱うためには分類基準を明確化しなければならない。

本稿では、品詞の分類基準を形態、統語、意味という三つのレベルに分類することを提案する。この分類は、形態素解析、構文解析、意味解析という自然言語処理における解析の段階にあわせたもので、形態論や統語論といった言語学の下位分野とはかならずしも一

図 4: 提案する品詞体系



致しない。このうち意味レベルは、形態レベル、統語レベルのいずれにも属さない基準で、現在の計算機には扱えない。例えば、「形容詞はおもに物事の性質や状態を表す」と説明されても、計算機はお手上げである。形態レベルとは、形態素同士の連接に関する制約である。構文解析は、日本語やモンゴル語の場合、係り受け解析を考える。そのため統語レベルは、どの品詞がどの品詞に係るか、また係らないかに関する制約となる。計算機に処理させるためには、品詞の分類基準は、形態レベルが統語レベルかのいずれかにより説明する必要がある。

一般的な議論はこれぐらいにして、モンゴル語の場合どうなるかを考えてみる。結論から言えば、提案する品詞体系は図 4 の通りである。設計の際に検討した事項をすべて説明にするには紙面が足りないので、本稿では具体例として形容詞と後置詞について述べる。

5.3 形容詞

数多くの品詞のうち、名詞と動詞については、どの言語にも普遍的に存在すると見られるが、その他については議論が分かれる。モンゴル語の場合、伝統的に形容詞が独立した品詞として扱われており、提案する品詞体系でもそれに従うが、仔細を検討するとさまざまな問題がある。

日本語の形容詞が歴史的に語尾を発達させてきたのに対し、モンゴル語の形容詞は基本的に語幹の形のまま用いられる。例えば、шинэ ном は「新しい 本」、Энэ ном шинэ биш. は「この本は 新しく ない。」となる。

形容詞について、まず問題となるのは副詞との区別である。モンゴル語の副詞は、形容詞と同様に、基本的には活用せず、語幹のまま動詞や形容詞を修飾する。形容詞と副詞のなかには、もっぱら形容詞として用いられるもの、もっぱら副詞として用いられるもの、形容詞としても副詞としても用いられるものの 3 種類がある。例えば、шинэ は、形容詞としてのみ用いられるが、хурдан は、形容詞 (速い) としても副詞 (速く) としても用いられる。提案する品詞体系では、形容詞としても副詞としても用いられるもののために、形容詞に副詞可能という細分類を設定した。

もう一つの問題は、名詞と形容詞の区別である。形容詞には、名詞性の接尾辞を取って名詞のように働く用法がある。例えば、шинийг хийх は、「新しいことをする」となる。一方、名詞も語幹のまま別の名詞を修飾することができる。例えば、хүрэл медаль は、「銅メダル」となる。

山越 [6] は、名詞と形容詞の違いを意味的性質に注目して説明している。しかし、意味レベルの違いは計算機には観測できないので、形態レベルあるいは統語レベルで対応する振る舞いの違いが求められる。このうち、形態レベルでは明確な区別が観測できないが、統語レベルで名詞と形容詞を識別するテストが知られている。形容詞は、強意の副詞маш (とても) を前置できるが、名詞はできないというものである。例えば、маш шинэ ном (とても新しい本) は文法的だが、*маш хүрэл медаль (?とても銀メダル) は不適格である。しかし、すべての形容詞が、このテストに通るのかは明らかでなく、さらなる分析が求められる。

実際の登録作業では、形態素にいずれの品詞を割り当てるか判断に迷うことが少なくない。例えば、сонин は、「面白い」という意味の形容詞だが、「新聞」という意味で名詞的にも働く。この場合は、意味が分化したものとみなし、二つの形態素として処理することにする。その他にも、特に「世代に関する語」と「色彩形容詞」を中心に、形容詞から名詞への意味的拡張が見られる [6]。例えば、залуу は「若い」という意味だが、単独で「若者」を指す用法もある。このような語は、現在は形容詞とみなしているが、今後方針を変えるかもしれない。

5.4 後置詞

提案する品詞体系では、多くの文法が独立した品詞とみなす後置詞を立てない。

後置詞の主な機能は、名詞類の後に置かれ、述語などを修飾する句を形成することである。例えば、наад は、「こちらに」、「〜以内に」といった意味を持ち、後置詞として働く際には、直前の名詞は「〜から」を意味する奪格-аас を取る。сараас наад ирэхгүй は、「一月 以内に 来ない」(直訳すれば、「月からこちらに来ない」) となる。

後置詞という品詞には、形態素ごとに振る舞いが大きく異なるという問題がある。実際、名詞類の後に置

かれて句を形成するという以外に、後置詞に共通の性質を見つけることは難しい [2]。例えば、直前の名詞がとる格は、нааш のように奪格のものもあれば、主格をとるもの、属格をとるもの、主格か属格かによって意味が変化するものなど様々である。

後置詞は、名詞や副詞といった別の品詞の形態素が、特殊な振る舞いをするようになったものである。仮に元の品詞としての働きが失われていれば、後置詞という品詞を設定しても問題ないかもしれないが、実際にはそうではない。例えば、нааш の元の品詞は副詞であり、наашаа суу! 「こっちに座れ!」のように、名詞を取らずに単独で出現できる。

提案する品詞体系では、後置詞は、名詞や副詞などの品詞として登録し、後置詞としての用法は素性として与える。例えば、нааш は副詞として登録し、「後置詞としては奪格を支配する」という素性を与える。

6 語彙

一般に、日本語は語彙が豊富と言われる。日本語の形態素解析器に使用されている辞書を見ると、ipadic[8] は 28 万語、Juman[11] は語彙を削減する傾向にあるが、依然として基本語彙を 3 万語収録している。モンゴル語の場合、日本語では 1 形態素の漢語とされる語が構成的に表現されるといった理由で、日本語ほどの語彙数が必要でないかもしれない。とはいえ、1 万語程度は必要だろうと推測される。

本稿では、まず整備対象の語彙を生産性のある品詞と生産性のない品詞に分類する。接尾辞にように生産性のない品詞は自分で登録する。名詞や動詞のような生産性のある品詞については、重要な基本語彙は自分で登録する。しかし、残りの一般的な語は、人手を介さずに登録できないかと考えている。そこで、自動獲得を予備的に検討してみた。

6.1 人手による整備

辞書整備のために利用できる資料には、複数の紙の辞書、電子辞書 1 種類、及び生の言語データがある。このうち、電子辞書である『電子日蒙索引』を基礎として形態素辞書を整備した。『電子日蒙索引』は、名前の通り日蒙辞典であり、約 7,500 の日本語の見出しに対して、モンゴル語の訳語が与えられている。辞書登録に必要なのは、日本語ではなくモンゴル語の見出し語なので、見出し語と訳語を逆転させた上で利用した。モンゴル語の訳語には、複合語や説明的なフレーズが少なからず含まれるが、1 形態素と見なしうるもののみを登録対象とした。

図 5: 用例リストの抜粋

ярилц	307	ярилцаачээ	3
ярилцаа	10	ярилцав	1538
ярилцаад	10	ярилцавч	10
ярилцаагй	82	ярилцага	20
ярилцаагүй	152	ярилцагад	10
ярилцаад	2144	ярилцагийг	7
ярилцаарай	96	ярилцагч	190
ярилцаасай	27	ярилцагчаа	19
ярилцаач	128	ярилцагчаас	4

日蒙辞典には、形態素辞書に必要な品詞の情報が欠けている。紙の辞書についても、品詞の分類はあまり厳密に行われていない。また、辞書に載っているモンゴル語の語彙について予備的に検討した結果、現在はあまり使われていない幽霊語が含まれているのではないかという印象を得た。そのため、ウェブサイトから収集した用例データを適宜参照しながら辞書登録を行った

収集したウェブページからは、分かち書きされた単位 (語) を抽出した。以下では、辞書順ソートされた語にその出現回数を加えたリストを用例リストと呼ぶ。語は、ごく少数の例外を除き、名詞や動詞などの自立語に 0 個以上の付属語が後続している。また、モンゴル語には「お」、「御」のような接頭辞は存在しない。そのため、語を辞書順にソートすると、同じ自立語が近くに並ぶ。例えば、図 5 を見ると、ярилц に対して、-аад, -в など、いくつかの動詞性接尾辞が付加されているのがわかる。このことから、ярилц という語幹があり、それが動詞で母音型が a であるとわかる。このように、形態レベルの性質は、用例リストを眺めるだけで判断できる。

このようにして、動詞は約 1,000、名詞は約 2,000 語を人手で登録した。また、用例リストを出現回数順に並び替えたリストを検討した結果、高頻度語にも漏れが少なくないことが判明した。

6.2 自動獲得

辞書への形態素の登録作業にはルーチンワークが多くを占める。出来る範囲で計算機に代行させることによって、人間の手間を省きたい。そこで、将来的な辞書の大規模拡張を視野に入れて、語彙の自動獲得実験を行った。

用例データからの形態素の自動獲得には、品詞の分類基準に関する議論が応用できる。品詞の分類基準が明確であれば、新たな形態素がその基準を満たすかを

テストすることよって、品詞の割り当てが行える。

自動獲得に利用できる品詞の分類基準は、形態素解析に取り組み始めた段階では、基本的に形態レベルに限られる。形態レベルの分類基準とは、具体的には接尾辞の連接に関する制約である。動詞を例とると、希求形 (〜しましょう) の接尾辞-я は特徴的なので、この語尾の付いた語から動詞語幹の候補を復元できる。例えば、байгуулья から動詞語幹байгуул (組織する) が復元される。しかし、機械的にこの手法を適用すると、誤りが混入する。例えば、名詞гуя (太もも) から、誤った動詞語幹*гу が作られてしまう。そこで、抽出された動詞候補からいくつかの活用形を生成し、その語幹候補が本当に動詞か否かをテストする。例えば、байгуул の場合、байгуулах, байгуулсан, байгууллаг, байгуулна といった動詞の活用形が実際に用例データに現れる。しかし、誤った候補である *гу については、гух, гусан, гудаг, гуна という活用形は、出現しないか、ごく少数にとどまる。このことから、байгуул が動詞であり、*гу がノイズであるとわかる。名詞についても、やや複雑であるが、基本的には動詞と同じである。

このようにして自動獲得した結果を検討すると、頻出語についてはほぼ正しく抽出できており、誤りは、*a のように極端に短い語幹について、偶然綴りが一致する形態素と混同する場合に限られる。ただし、ウェブページというコーパスの性格を反映してか、ь をи と表記 (発音は同じ) するといった、正書法上誤った綴りが目立つ。今回は、人手でのチェックにより、こうした俗表記を取り除いた。

この手法では、形態レベルしか観測していないので、統語レベル以上が必要な情報は獲得できない。例えば、5.3 節で述べたとおり、名詞と形容詞は形態レベルでは同じ振る舞いをするため、区別がつかない。また、基本的に活用しない副詞も獲得できないという問題がある。これに対しては、本格的に統語的制約を実装する方法や、単語 N-gram によって近似的に統語レベルの情報の利用する手法などが考えられる。

7 おわりに

本稿では、自然言語処理における解析の全体像を概観した上で、その第一段階で行うべき基本設計について述べた。特に品詞体系の設計は、言語の振る舞いを直接計算機に教える作業であり、言語学的な知見に期待がかかる。しかし、言語学の理論はなかなか素直に自然言語処理に応用できない。実際の言語を調べると、しばしば理論に説明のない現象に出会う。自然言語処理の立場からは、言語学に、非典型的な用法を含めてすべての現象を包括的に扱うことと、厳密な形式化を求めたい。

モンゴル語は、日本語と構造的に類似しているため、解析器の開発は、今後も基本的に日本語における開発成果を応用しながら行うことになる。もっとも、語彙の自動獲得は、日本語ではそれ程研究されていない。理由としては、日本語には既に大規模な辞書が整備されていることが考えられる。しかし、多種多様なウェブページなどの解析するためには、日本語でも自動獲得が必要と考えている。今後は自動獲得の高精度化により、獲得した語彙は、人手を介在さず、直接解析器にフィードバックできるようにしたい。

謝辞

本研究の一部は未踏ソフトウェア創造事業未踏ユースによりご支援いただきました。ご支援いただきました IPA および竹内郁雄教授に感謝いたします。また、『電子日蒙索引』を提供いただいた清水幹夫氏に感謝いたします。

参考文献

[1] Preslav Nakov and Marti A. Hearst. Using verbs to characterize noun-noun relations. In *AIMSA*, pp. 233–244, 2006.

[2] John C. Street. *Khalkha Structure*, pp. 209–213. Curzon, 1997.

[3] Terry Winograd. *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language*. PhD dissertation, M.I.T., 1971.

[4] フフバートル. 続モンゴル語基礎文法. インターブックス, 1997.

[5] 三上喜貴. 文字符号の歴史 アジア編, pp. 264–268. 共立出版, 2002.

[6] 山越康裕. 現代モンゴル語における「名詞」と「形容詞」について. 日本モンゴル学会紀要, No. 37, pp. 97–108, 2006.

[7] 橋田浩一. 言語への情報科学的アプローチ. 大津由紀雄, 郡司隆男, 田窪行則, 長尾真, 橋田浩一, 益岡隆志, 松本裕治（編）, 言語の科学入門, 第 3 章. 岩波書店, 1997.

[8] 浅原正幸, 松本裕治. ipadic version 2.7.0 ユーザーズマニュアル, 2003.

[9] 工藤拓, 山本薫, 松本裕治. Conditional random fields を用いた日本語形態素解析. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2004, No. 47, pp. 89–96, 2004.

[10] 中里致元. モンゴル語 電子化計画. http://texa.human.is.tohoku.ac.jp/~chigen/md_cnt_j.htm.

[11] 黒橋禎夫, 河原大輔. 日本語形態素解析システム JUMAN version 6.0, 9 2007. http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman/juman-6.0.tar.gz.

[12] 満都拉, 藤井敦, 石川徹也. 伝統的モンゴル語と現代モンゴル語の双方向的な翻字手法. 言語処理学会第 11 回年次大会発表論文集, pp. 360–363, 2005.

[13] 江原暉将, 早田清冷, 木村展幸. 茶筌を用いたモンゴル語の形態素解析. 言語処理学会第 10 回年次大会発表論文集, pp. 709–712, 2004.