キリル文字モンゴル語形態素解析器の開発
—マイナー言語に対する言語処理基盤開発—

1. 背景

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。話者は、全体で 570 万、本プロジェクトが対象とするハルハ方言は 230 万程度と推定される。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）が後続するのと同様、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。しかし、歴史上に日本語との接触がほとんどないため、翻訳にあたっては訳語が単純に対応しない。

モンゴル語のように構造的に日本語と類似した言語は内陸アジアに広く分布している。このような言語に対しては、日本語における自然言語処理の成果が応用できるが、実際にはそうした取り組みは盛んではない。従来の自然言語処理の開発は、大量の人的リソースや言語資源を前提としていたため、マイナー言語は無視されがちであった。

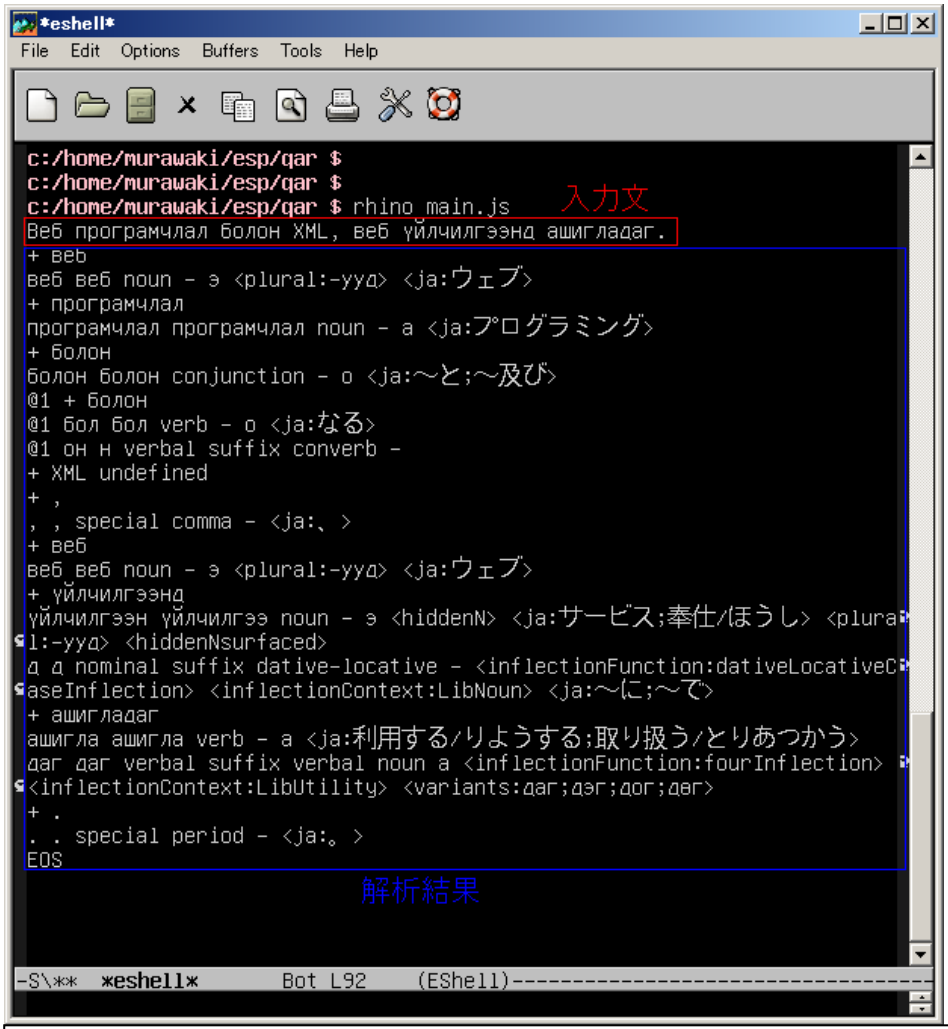

図 1 形態素解析の実行例

2. 目的

形態素解析器をキリル文字モンゴル語に対して開発する。形態素解析器とは、入力として文を受け取り、それを最小の単位である形態素列に分解して出力するプログラムである。計算機は、人が読み書きする言語、自然言語のデータをただの文字の並びとしか認識していない。計算機に言語を扱わせるためには、文字の並びが構造を持つことを教えなければならない。形態素解析は、そのもっとも基礎的な処理である。

モンゴル語の一般のウェブページを正しく解析できるレベルを目標とする。もっとも、実際問題として、一般のページを正しく解析するには、大規模な辞書が必要となる。そのため、使用頻度の高い基本的な語彙を開発期間内に可能な限り登録し、残りは今後の課題とした。

3. 開発の内容

図 1 に形態素解析の実行例を示す。

形態素解析器の開発は、ルールの実装とデータの整備からなる。ルールとは、文法など、言語が持つ規則と、そうした規則を使った解析実行時の処理であり、データは主に形態素解析用の辞書を指す。

キリル文字モンゴル語の形態素解析において、必要なルールは以下の通りである。

- 形態素連接規則
- 形態素屈折変化規則
- 品詞体系
- 生成規則
- 形態素解析

こうしたルールの実装は、主に言語学の研究書や語学の教科書を参考にして行った。もっとも、このような参考資料からプログラムに落とし込む規則を抽出するのは容易ではない。理由は二つある。第一に、人間向けに記述されているため、規則が計算機で扱えるほど厳密でない。第二に、言語の振る舞いを網羅的に説明していない。プログラムを書くということは、事前にすべてのパターンを把握しておくということである。しかし、語学の教科書は、典型的な振る舞い説明していても、例外を列挙しない。また、品詞の体系的な扱いについても、文法の古典的名著は、やや通時的で、現代語を扱う上で適当でない。かといって、現代語に関する論文は、特定の現象の説明に終始している。したがって、ルールの実装は、大げさに言えば、自分で一つの体系を構築するような作業となった。

解析用辞書の整備には、人間用の既存辞書[1]を取っ掛かりとして利用した。しかし、既存辞書だけでは形態素解析に必要な情報が足りない。語彙が不足して

---

[1] 本プロジェクトでは、清水幹夫氏が作成した『電子日蒙索引』を利用させていただきました。心から感謝いたします。

いるだけでなく、品詞と活用の情報が欠けている。そこで、不足している情報を補うために、モンゴル語のウェブページを収集した。収集したウェブページから、形態論的特徴を利用して形態素を抽出する実験を行った。形態論的特徴とは、この場合、後続する付属語に関する振る舞いを指す。日本語で言えば、「遊ばない」「遊び」「遊ぶ」「遊べば」などの用例から「遊ぶ」が動詞だと分かるといった具合である。自動獲得した形態素候補に人手で修正を加えて形態素辞書に追加した。

4. 従来の技術（または機能）との相違

一般公開されているキリル文字モンゴル語の形態素解析器は、管見の限り存在しない。

5. 期待される効果

形態素解析は、自然言語処理の第一歩となる基盤処理である。形態素解析結果そのものを直接活用することはあまりなく、次の処理に利用するのが一般的である。ただし、本プロジェクトで整備した辞書は、各形態素に日本語の訳語を付加してあるので、形態素解析結果を眺めるだけで、モンゴル語学習者にとっては、辞書引きの手間を省く程度の効果があり、モンゴル語を知らない日本語話者も、書かれた内容をある程度推測できるようになる。

形態素解析結果を使ってできることは、日本語の場合と基本的には変わらないが、比較的簡単な応用として、検索エンジン用のインデックス作成を紹介する。インデックス作成に形態素解析が必要なのは、モンゴル語は活用変化を起こすからである。例えば、情報技術はモンゴル語で мэдээллийн технологи だが、ここで мэдээллийн は мэдээлэл（情報）に属格「～の」が付いた形である。このとき、мэдээллийн という活用形のままインデックスを作ると、мэдээлэл で検索した際にヒットしない。現在の Google は活用形をそのまま登録しているので、モンゴル語の検索ではヒット漏れが多く不便である。また、単語ベースではなく、文字列ベースでインデックスを作るとしても、活用時に母音 э の削除が行われているので、やはりヒットしない。これに対し、形態素解析を行い、原形を復元することによって、クエリ мэдээлэл に対して мэдээллийн технологи をヒットさせることができる。

6. 普及（または活用）の見通し

本プログラムは、修正を加えたうえでの一般公開を予定している。

形態素解析を実行する上で問題となるのは、語彙の整備である。語彙はオープンな問題であり、その整備には終わりがない。しかし、辞書の整備は専門知

識なしには難しい。そこで、計算機の支援により生の言語データを見て辞書を作る枠組みを作り、不特定多数の一般利用者が辞書整備に参加できるようにしたいと考えている。

7. 開発者名（所属）

村脇　有吾（京都大学大学院情報学研究科）