

MeCabを用いた現代韓国語の形態素解析

須賀井 義教*

1 はじめに

1.1 本稿の目的

本稿¹は、現代韓国語の自動形態素解析を行うための辞書構築について検討し、それを韓国語教育などに活用するための試みについて述べるものである。形態素の解析には、オープンソース形態素解析エンジンであるMeCab(めかぶ)を用いる。

形態素解析の技術は言語研究、言語教育だけでなく、機械翻訳や構文解析などの前処理において重要な役割を果たす。また、この技術を応用してインターネットのウェブページ上からキーワードを抽出したり、検索や入力補助に利用することも可能である。インターネットやコンピュータ技術の進展とともに、自然言語を処理するための技術が果たす役割も、さらに重要度を増していると考えられる。

韓国語の形態素解析、あるいは形態素解析済みコーパスでは、韓国・国立国語院などが進めてきた21世紀世宗計画(以下「世宗計画」とする)による一連の成果がよく知られている。また、その他さまざまな自動形態素解析器が開発されているが、これらのツールに利用者が手を加えたり、動作や出力をカスタマイズすることはできない。

そこで本稿では、特定の言語に依存しない、汎用的な形態素解析エンジンであるMeCabを利用して、ユーザーが辞書を自由に定義することができ、さらにインターネット上でも利用することができる韓国語自動形態素解析の方法について提案したい。MeCabを用いることで、利用者のニーズに合わせた出力結果のカスタマイズ、新たな語句の登録、インターネットを通じたサービスの提供などが可能となり、さらには韓国語学習で活用することも可能となる。本稿では特に、形態素解析用辞書の構築、辞書の応用事例について検討する。

1.2 先行研究

韓国語の形態素解析²について、いくつかの先行研究について紹介する。

*近畿大学 総合社会学部 . sugaiy@kindai.ac.jp

¹本稿は、2010-2011 年度科学研究費補助金(基盤研究(B))「朝鮮語CALL教材作成技法の開発と普及」(研究代表者：須賀井義教・研究分担者：油谷幸利、大名力、中西恭子、中村麻結)による研究成果の一部である。

²なお、韓国語の形態素解析に関して韓国語学の側から概説したものに (2004)、 (2006)がある。いずれも形態素解析、構文解析について説明しており、特に前者は構文解析、後者は形態素解析についてより詳しい解説が

セットによりタグ付けを行っている。後者はKAISTの「[Morphology Analyzer](#)」に用いられたタグセットを使っているようであるが、分析可能な候補を全て挙げている点が特徴である。

これらの自動形態素解析器において共通するのは、その動作、出力結果などを利用者が自由に設定できないという点である。また出力されるのは品詞タグだけであり、その他の情報を付け加えたりすることができない。さらに、新語を独自に追加することができないため、インターネットなどを中心に新たに創り出される語などに対応できない可能性がある。

韓国語の形態素解析についての日本における研究として、形態素解析のための韓国語品詞体系を提案した論考がある。平野善隆(1997)は用言の活用体系に注目し、語尾を不変として用言語幹のみを変化させ、変化する部分を「活用語尾」として記述することで、語尾に関する操作が一切不要になるとする。しかし、語幹の活用形が7種類にものぼり、辞書の記述自体が複雑になるのではないかと考えられる。また、「[가](#)(語幹)+なし(活用語尾)+[ㅁ](#)(語尾)」となるが、「[ㅁ](#)(語幹)+[ㅁ](#)(活用語尾)+[ㅁ](#)(語尾)」とするなど、語尾を単純化しようとするあまり、従来の文法記述から若干距離のある内容となってしまう。なお、形態素解析のシステムとしてはChaSen(茶筌)を用いている。

山本和英(2000)では「[가](#) + [ㅁ](#)」が「[ㅁ](#)」, 「[ㅁ](#) + [ㅁ](#)」が「[ㅁ](#)」となるような「音韻縮約」を念頭に置き、「残留文字」という概念を利用して解析処理を行うことを提案している。詳細は割愛するが、こうした処理により、「一般補助詞」の「[-](#)」と「文末語尾」の「[-](#)」とを区別することができるという(山本和英 2000:30)。

また、油谷幸利(2008, 2009)など、油谷幸利による一連の取り組みがある。朝鮮語Web辞典や用言の活用形練習問題など、独自のアルゴリズムに基づいた形態素解析を応用したものである。

なお、日本語以外の形態素解析にMeCabを用いた試みとして、古典中国語を扱った守岡知彦(2008), 『[釈譜詳節](#)』 『[法華経諺解](#)』といった15世紀のハングル文献を対象とした村田寛(2010), 須賀井義教・村田寛(2011)がある。村田寛(2010)ではMeCabを利用したハングル文献の形態素解析が可能であることを明らかにし、続く須賀井義教・村田寛(2011)では解析率の向上について考察を行った。本稿はこれらの試みの延長として、現代韓国語にその方法を適用するものである。

2 MeCabとは

2.1 MeCabの概要

本稿で利用するMeCab⁷とは、言語、辞書、コーパスに依存しない汎用的な設計を基本方針とするオープンソース形態素解析エンジンで、京都大学情報学研究科—日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクトにより開発されたものである。コマンドラインから実行するプログラムで、直接文を入力するか、入力ファイルを指定して解析

⁷現在のバージョンは0.98である。Windows, Unixなどのオペレーティングシステムで利用することができる。詳細についてはホームページ<http://mecab.sourceforge.net/>を参照。

を行う。コマンドラインからMeCabを実行し、「今日も勉強しないとね。」という文を解析すると、図1のような結果を得ることができる。

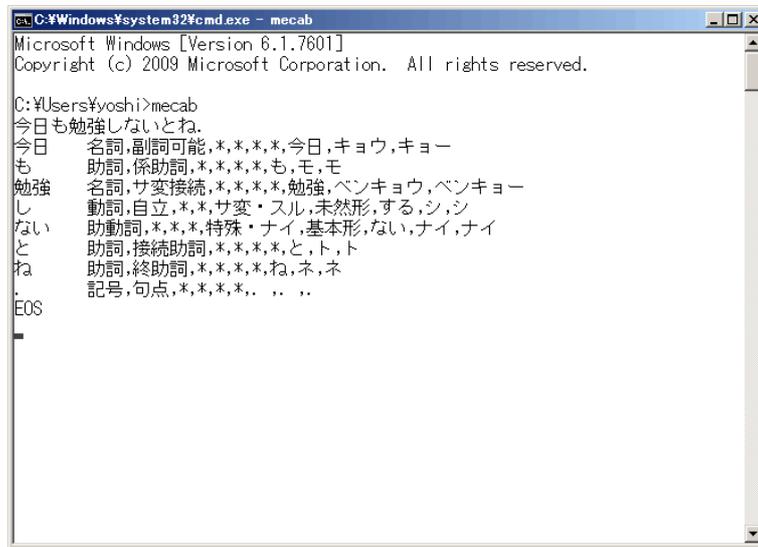


図 1: MeCabの実行結果

解析した結果は、表層形に続き「素性」をコンマで区切った形式で出力される。この素性は自由に定義することができ、いくつでも連ねることができる。Windows版バイナリに同梱されているIPA日本語辞書⁸では、以下の素性が定義されている：

- (2) 品詞, 品詞細分類 1, 品詞細分類 2, 品詞細分類 3, 活用例, 活用形, 基本形, 読み, 発音

素性の中には品詞によって用いられないものも存在するが、その場合は空白にせず、図1の例のように、「*」(アスタリスク)を記述しておく。

なお、用言などの活用について、MeCabでは文法による活用の展開を行わないため、辞書に全ての活用形を記述しておく必要がある。先のIPA日本語辞書から「書く」のエントリを一部挙げれば、以下の通りである：

- (3) x書く,679,679,7325,動詞,自立,*,*,五段・カ行イ音便,基本形,書く,カク,カク
 書か,683,683,7745,動詞,自立,*,*,五段・カ行イ音便,未然形,書く,カカ,カカ
 書こ,681,681,7299,動詞,自立,*,*,五段・カ行イ音便,未然ウ接続,書く,カコ,カコ
 ...
 書きゃ,677,677,7123,動詞,自立,*,*,五段・カ行イ音便,仮定縮約1,書く,カキャ,カキャ

⁸本来のIPA日本語辞書は日本語形態素解析器ChaSen(茶筌)用に作成されたものであるが、それをMeCab用に修正したものがともに配布されている。IPA日本語辞書の詳細については <http://sourceforge.jp/projects/ipadic/> にて公開されているダウンロードファイルとドキュメントを参照のこと。なお、MeCabに同梱されているバージョン(2.7.0-20070801)の辞書には、392,126項目が登録されている。

2.2 なぜMeCabを用いるか

では、本稿でなぜMeCabを利用するか、その理由を挙げれば以下の通りである：

- 言語，辞書に依存しない汎用的な設計
- カスタマイズの自由度が高い
- 無償で配布されている

MeCabの大きな特徴として、「言語，辞書に依存しない」という点を挙げる事ができる。MeCabとともに日本語解析用の辞書が配布されているが、この辞書を別個に用意することで、他の言語にも対応することができる。辞書の文字エンコーディングとしてUTF-8を利用することができ、さまざまな言語の文字を同時に扱えるため、本稿のような目的での利用に最適であるといえよう。先に述べた15世紀のハングル文献や古典中国語に対する形態素解析が可能であるのも、MeCabの持つこうした特徴を活かしたものである。

また、MeCab用辞書の設定ファイルを編集することで、その出力内容などを自由に変更できる。さらに、MeCabは未知語の処理を字種に基づいて行うが、この字種カテゴリも自由に設定することができ、未知語の処理方法自体も変更することが可能である。このように、カスタマイズの自由度が高いということもMeCabの特徴のひとつと言えよう。カスタマイズによって、他のツールでも使用できる出力を得られるという利点がある。これまで国立国語院、世宗計画で開発され公開されてきた一連のツール群は、それらのツールでのみ利用できるような形式となっており、汎用性という点で柔軟さが見られなかった。MeCabを用いることで、この点を補うことができるだろう。

さて、韓国語自動形態素解析器の多くが無償で公開されているが、MeCabもまたフリーウェアとして無償で配布されている。さらにその解析結果を活用するためのツールも無償で公開されているため、環境さえ整えれば誰でもこれらのツールを利用することができる、というメリットがある。そうしたツールの例としてChaKi(茶器)⁹を挙げておく。ChaKiは自然言語コーパスの構築、検索、および言語要素へのタグ付けをサポートするツール群であり、ChaSenやMeCabによる形態素解析済みテキストを読み込んで、形態素タグなどを組み合わせた条件によるKWIC検索や統計処理を行うことができる。また係り受け解析済みのテキストを用いて検索などを行えるという。こうしたツールの存在も、MeCabの利用を積極的に後押ししてくれるものである。

さらに、PerlやRubyといった他のプログラム言語からMeCabを扱うためのライブラリも用意されているため、単なる形態素解析だけでなく、より複雑な処理を行うことが可能になっている。筆者はPerlとMeCabを組み合わせ、インターネット上で韓国語の形態素解析を行えるようなプログラムを作成した。本稿では、そのための辞書作成について述べ、さらにインターネットを通じて公開しているプログラムについて概要を紹介する。

⁹詳細はホームページ <http://sourceforge.jp/projects/chaki/> を参照のこと。

3 MeCab用辞書の作成と構築

さて、MeCabで形態素解析を行うための辞書を構築するには、Seed辞書、学習用データが必要である。Seed辞書と学習用データを用いて生起コスト、接続コスト¹⁰などを学習させ、解析用の辞書を構築する¹¹。

なお、村田寛(2010)、須賀井義教・村田寛(2011)と同様に、本稿でも用言活用の記述に「語基」¹²の概念を利用し、ハングルの表記にローマ字転写を用いた。MeCabは文字を単位として解析を行うため、さまざまな形態素が文字のなかにとけあってしまう韓国語のような場合、その中から形態素を抽出することが難しい¹³。処理に際してローマ字転写を用いることで、ハングルの文字単位ではとらえることのできない形態素を抽出することができる。アルファベットは内部的に用いるため、入力、出力においてはハングルで表示するようにプログラムを作成すれば、ローマ字転写の部分は目に触れずに済む。ローマ字転写の方法については本稿末に付録として掲げた。

3.1 Seed辞書の準備

3.1.1 元となるデータについて

解析用辞書の元になるのがSeed辞書である。本稿ではSeed辞書の作成に、既存の電子化データを利用した。具体的には国立国語院が公開している「[日本語基本辞書](#)」(以下「学習用語彙」と呼ぶ)を主として利用し、不足分を「[日本語基本辞書](#)」(以下「頻度調査」と呼ぶ)によって補った。前者は国立国語院が2003年に発表したものであるが、その成果物としてExcel形式、テキスト形式のデータがホームページを通じて公開されている¹⁴。本稿ではこのExcelファイルを利用して、3段階までの語彙を収録した。解析の便を考慮して、一部の項目の品詞を変更している。また「[日本語基本辞書](#)」などのように、複数の品詞にまたがって用いられる項目については、それぞれの品詞に重複して登録した。よって、「学習用語彙」の登録語彙数と完全に一致するものではないことを断っておく。なお「学習用語彙」には語尾、助詞の類が収録されていない。このため、「頻度調査」¹⁵の「[日本語基本辞書](#)」から頻度10以上の項目を、同じく「[日本語基本辞書](#)」から頻

¹⁰「生起コスト」とはその形態素の出現しやすさを表し、「接続コスト」とは二つの形態素のつながりやすさを表す。MeCabはこれらのコストの和が最小になる場合に正解と判定する「接続コスト最小法」(松本裕治ほか 1997:62-63)を採っている。特に接続コストについては「日本テレビ東京で学ぶMeCabのコスト計算」(http://www.mwsoft.jp/programming/munou/mecab_nitteretou.html)が分かりやすい。

¹¹辞書構築の概要については須賀井義教・村田寛(2011:42-46)も参照されたい。

¹²語基については菅野裕臣(1997)などを参照のこと。

¹³ハングルの1文字を単位として処理すると、例えば「[日本語基本辞書](#)」から過去の接尾辞を取り出すことができなくなってしまう。

¹⁴国立国語院ホームページの「[日本語基本辞書](#)」「[日本語基本辞書](#)」より、「[日本語基本辞書](#)」(2003年6月4日付け掲示物)にて入手可能。3段階のランク付けがなされており、収録語数は1段階982語、2段階2,111語、3段階2,872語の総5,965語となっている。選定の経緯などについては[日本語基本辞書](#)(2003)を参照のこと。

¹⁵「学習用語彙」と同様に、国立国語院ホームページから入手可能である。Excelファイルは「[日本語基本辞書](#)」

度 40 以上の項目を抽出し，Seed 辞書に収録した．また「頻度調査」からは，頻度 10 以上の名詞 5,095 項目，頻度 4 以上の副詞 778 項目を追加した．実際に「学習用語彙」だけで辞書を構築し，解析を試してみたところ，名詞の辞書項目が若干不足しているようであったため，これらの項目を追加したものである．副詞については単に作業の進行上追加できたものであり，今後用言などについても追加していく予定である．

3.1.2 品詞の記述について

Seed辞書の素性は以下のように記述した：

(4) 品詞1，品詞2，品詞3，活用形，接続情報，辞書項目，表層形，漢字，その他

品詞 1 は概ね「学習用語彙」の品詞に従った．その細分類である品詞 2 も合わせて，本稿の辞書で用いた品詞体系を示せば表 1 の通りである．

表 1: 本稿での品詞体系概要

品詞1	品詞2
名詞(Noun)	普通，代名詞，語根，助数詞，不完全名詞，接尾語，数詞，固有名詞
動詞(Verb)	自立
形容詞(Adjective)	自立，非自立
指定詞(Siteisi)	非自立
存在詞(Sonzaisi)	自立
副詞(Adverb)	一般，否定
助詞・語尾(Ending)	助詞，語尾
接尾辞(Suffix)	過去，尊敬，将然，etc.
接頭辞(Prefix)	
接続語(Conjunction)	
間投詞(Interjection)	

品詞 2 として定義した内容の一部は，解析の便宜を図るために本稿で提案するものである．特にMeCabでは解析に接続コストを利用するため，前後にどのような要素が来るか，という点を重視して細分類を行った．以下，品詞 1 のカテゴリ別に詳しく説明する．

名詞 いわゆる普通名詞だけでなく，単独では使われず，名詞に後接して，全体が名詞として機能するような要素も「名詞」に入れた．例えば「接尾語」としたものがそれである．ここには複数を表す「-」や，「」「」などが含まれる．「学習用語彙」で「」とされたも

()」(2003 年 1 月 20 日付け掲示物)からダウンロードした．

のが主である．名詞に後接するものを「接尾語」，用言の連体形に後接するものを「不完全名詞」とした．これらの例を示せば以下の通りである¹⁶：

- (5) goqbu,0,0,0,Noun,普通,*,*,*, 01, ,工夫
gos,0,0,0,Noun,不完全名詞,*,*,*, 01, ,*
dal,0,0,0,Noun,助数詞,固有のみ,*,*, 05, ,*
dyl,0,0,0,Noun,接尾語,*,*,*, , ,*

また「語根」とは，「」に前接することができるが単独では名詞として用いられない要素，例えば「」の「」，「」の「」などを含める．本稿では「」で終わる用言を，全て前接要素と「」とに分けて解析するようにしたためである．「」などを一つの語として辞書に登録した場合，「」「」などのように，助詞が挿入されると解析に問題が生じる可能性がある．また一律に分離して登録することで，「」と「」の両方を辞書に登録する手間を省くことができ，結果として辞書のサイズを抑えることができる．

- (6) jo'ioq,0,0,0,Noun, 語根,*,*,*, , ,*

「数詞」には品詞3として，漢数詞と固有数詞の別を記述する．「」 「」など冠形詞として用いられるものも，「固有数詞-冠形詞」のように記述した．

- (7) pal,0,0,0,Noun,数詞,漢数詞,*,*, 03, ,八
han,0,0,0,Noun,数詞,固有数詞-冠形詞,*,*, 01, ,*

「固有名詞」は「学習用語彙」の「」に該当する．品詞3として「人名」「地名」「都市名」などを補足した：

- (8) seijoqdai'oaq,0,0,0,Noun,固有名詞,人名,*,*, , ,世宗大王
namsan,0,0,0,Noun,固有名詞,地名,*,*, , ,南山

動詞 「学習用語彙」の「」が該当する．ただし「」 「」は除外した．品詞2を「自立」としたが，補助用言として用いられる語を「非自立」として記述する予定である．現状の辞書では補助用言を区別していない．品詞3は原則として空白であるが，変格活用の用言など，一部の用言で活用のタイプを記述した．

また，素性「活用形」には語基を記述した．≡語幹の用言は，≡が脱落した形も登録してある．例えば「」の場合，以下のように記述した：

¹⁶以下の例でローマ字転写に続く3つの数字はそれぞれ「左文脈ID」「右文脈ID」「生起コスト」にあたる数値である．Seed辞書の段階ではこれらの値を0としておき，学習用コーパスを用いた学習を行うことで，例(3)に挙げたIPA日本語辞書のエントリのように，これらの数値が決定される．

- (9) 'al,0,0,0,Verb,自立,r語幹,語基1,* , ,*
 'a,0,0,0,Verb,自立,r語幹-脱落形,語基1,* , ,*
 'al,0,0,0,Verb,自立,r語幹,語基2,* , ,*
 'a,0,0,0,Verb,自立,r語幹-脱落形,語基2,* , ,*
 'al'a,0,0,0,Verb,自立,r語幹,語基3,* , ,*

形容詞 動詞と同様,「学習用語彙」の「 」が該当する.大部分は品詞2に「自立」と記述したが,「 」のみ「非自立」とした.動詞の場合と合わせ,補助用言をどのように記述するか,今後の課題である.「 」を例として挙げておく:

- (10) sip,0,0,0,Adjective,非自立,* ,語基1,* , ,*
 sip'y,0,0,0,Adjective,非自立,* ,語基2,* , ,*
 sip'e,0,0,0,Adjective,非自立,* ,語基3,* , ,*

指定詞・存在詞 「学習用語彙」では存在詞や指定詞が動詞あるいは形容詞として登録されていたが,本稿ではそれぞれ独立したカテゴリとして設定し,指定詞として「-」「 」を,存在詞として「 」「 」「 」をそれぞれ辞書に登録した.

- (11) a. 'i,0,0,0,Siteisi,自立,* ,語基1,* , ,*
 'i,0,0,0,Siteisi,自立,* ,語基2,* , ,*
 'i'ei,0,0,0,Siteisi,自立,後接,語基3,* , ,*
 'iei,0,0,0,Siteisi,自立,後接,語基3,* , ,*
 'ie,0,0,0,Siteisi,自立,母音体言前接,語基3,* , ,*
 'i'e,0,0,0,Siteisi,自立,子音体言前接,語基3,* , ,*
 b. 'iss,0,0,0,Sonzaisi,自立,* ,語基1,* , 01, ,*
 'iss'y,0,0,0,Sonzaisi,自立,* ,語基2,* , 01, ,*
 'iss'e,0,0,0,Sonzaisi,自立,* ,語基3,* , 01, ,*

副詞 「学習用語彙」の「 」をほぼそのまま登録した.「 _」などのように,項目自体に助詞を含むと見られるものも,分離せずにひとつの項目として登録した.「 」「 」などについてのみ,品詞2に「否定」という素性を記述した.

また「 」「가 」など,助詞の接続が可能な語については,いったん副詞に含め,品詞3に「名詞可能」と記述した.重複する表層形をできるだけ減らすためである.もちろん,名詞と副詞に重複して登録したとしても,学習によって適切に解析してくれるものと考えられるが,現時点では煩雑を避けるため,副詞にのみ登録した.

- (12) gyniaq,0,0,0,Adverb,一般,* ,* ,* , , ,*
 mos,0,0,0,Adverb,否定-不可能,* ,* ,* , 04, ,*
 nai'il,0,0,0,Adverb,一般,名詞可能,* ,* , , ,來日

なお、「」「」などといった 用言の副詞形については、全体をひとつの項目として記述した。前述のように、 用言は前接する要素と とを分離して「 語根」として記述することとしたが、副詞についてのみ、暫定的にひとつの項目として登録するという措置をとった。

(13) bunmieqhi,0,0,0,Adverb,一般,*,*,*,,,分明-

助詞・語尾 3.1.1 節で述べたとおり、「学習用語彙」には文法的項目が収録されていないため、「頻度調査」から項目を追加して登録した。品詞1としては「Ending」としてひとくりにし、品詞2で助詞と語尾を区分している。品詞3としては、助詞の場合その意味を、語尾の場合は「接続形」「連体形」などの別を記述した。語尾については、第何語基に接続するかという情報を、「接続情報」という素性に記述した。

- (14) a. ga,0,0,0,Ending,助詞,主格,*,*,가,가,*
n,0,0,0,Ending,助詞,縮約,*,*,ㄴ,ㄴ, /
b. l,0,0,0,Ending,語尾,連体形,*,2接続,ㄹ,ㄹ,*,*
da,0,0,0,Ending,語尾,終止形,*,1接続,,,*,*
se,0,0,0,Ending,語尾,接続形,*,3接続,,,*,*
dai,0,0,0,Ending,語尾,引用融合,*,1接続,,,*,-

接尾辞 こちらは初級・中級で必要と思われる「II--」「I--」「III-ㄴ-」「I--」を登録した。素性「活用形」には活用語基を記述した：

- (15) a. si,0,0,0,Suffix,尊敬,*,語基1,*,,,*
si,0,0,0,Suffix,尊敬,*,語基2,*,,,*
sie,0,0,0,Suffix,尊敬,*,語基3,*,,,*
sei,0,0,0,Suffix,尊敬,後接,語基3,*,,,*
b. geiss,0,0,0,Suffix,将然,*,語基1,*,,,*
geiss'y,0,0,0,Suffix,将然,*,語基2,*,,,*
geiss'e,0,0,0,Suffix,将然,*,語基3,*,,,*

接頭辞 主に「学習用語彙」の「」を含む。単独で用いられず、直後に名詞しか接続することができない項目が該当する。「学習用語彙」では「」「」など数詞の冠形詞形も「」とするが、既に例(7)に挙げた通り、本稿では名詞-数詞のカテゴリに含めた。「」「」「」などの指示詞も接頭辞に含まれる：

- (16) gy,0,0,0,Prefix,*,*,*,*, 01,,*
'yeny,0,0,0,Prefix,*,*,*,*, 01,,*
'iag,0,0,0,Prefix,*,*,*,*, 03,,約

接続語 「 」 「 」 などといった，文と文とをつなぐ要素が該当する．「学習用語彙」ではほとんど副詞として扱われているが，本稿では「接続語」として別のカテゴリとした．

間投詞 「 」 「 」 などが該当する．「学習用語彙」の「 」をそのまま収録した．

3.1.3 各品詞に共通する素性

例(4)に挙げた素性のうち，全ての品詞に共通して記述した素性である「辞書項目」「表層形」「その他」について述べる．

「辞書項目」は「学習用語彙」「頻度調査」の見出し語をそのままコピーしたもので，辞書形に『 』に準じた同音異義語の番号がつけてある¹⁷．同音異義の語を区別するための措置であり，今後何らかの用途に使うことを前提としている．

これとは別に「表層形」という素性を用意し，実際に表れた形式，すなわち活用しない語であれば辞書形から同音異義番号を除いた形，活用する語であれば活用形をハングルで記述した．

また，「学習用語彙」の「補充情報」を「その他」という素性に記述した．「 」のように，語の一部のみ漢字表記できる場合には，「正-」のように表記した．「その他」の多くは同音異義語を区別するための情報である．今後はデータ利用の利便性を考慮し，漢字表記のみ別の素性として記述する予定である．

3.1.4 Seed辞書の登録項目数

このようにして構成したSeed辞書の登録項目数は表 2 (12 ページ) の通りである．

なお，表 2 で用言は活用形のバリエーションを含んでいるため，実際に登録されている用言の異なり語数は，表の数字の約 3 分の 1 となる点に注意されたい．また，3.3 節で後述するように，解析率を上げるため学習用コーパスに含まれる未知語も追加した．2 つの学習用コーパスから追加した未知語は 155 項目である．よって，最終的なSeed辞書の登録項目数はこれらの合計である 14,439 項目となる．以下，本稿で解析例を挙げる際には，全てこの辞書で解析した結果を示すこととする．

3.2 学習用コーパスの準備

Seed辞書の段階では生起コストや接続コストが計算されていない．MeCabの出力と同じ形式で記述した学習用コーパスを用いて学習を行い，これらのパラメータを決定する．

学習用コーパスの量は多ければ多いほどよいようであるが，日本語の場合と異なり，すぐに利用できる大規模のデータが存在するわけではない¹⁸．大量の解析済みデータといえば，もちろん

¹⁷なお，『 』にない同音異義語は，80 以降の番号がふられている(2002:15)．

¹⁸MeCab用IPA辞書の学習には，4 万文程度のデータを使っているという．「気まぐれ日記：Yahoo!の形態素解析をMeCabで無理やり再現してみる」(<http://chasen.org/~taku/blog/archives/2007/06/yahoomecab.html>)を参照．

表 2: 本稿で作成したSeed辞書の項目数

品詞	項目数	(うち追加数)
名詞	9,112	(5,095)
動詞	3,027	
形容詞	605	
指定詞	14	
存在詞	11	
副詞	1,154	(778)
助詞	87	(87)
語尾	141	(141)
接尾辞	13	
接頭辞	46	
接続語	19	
間投詞	44	
記号類	11	
合計	14,284	

世宗計画の形態素解析済みコーパスがあるが、品詞体系が完全に一致せず、変換して使うにも労力がかかるため、本稿では小規模のデータを用意して、学習用コーパスを構成した。学習のデータ量が少ないと、要素の現れやすさを表す生起コストの学習に問題が生じる可能性があるが、改めて大規模なデータを利用することとしたい。

さて、本稿で学習に用いたのは、世宗計画の「 」のうち「 」(2BH9301.txt)、「 」(BRH00414.txt)の二つである。これらのデータから、それぞれ冒頭の100文を学習用データとして利用した。

これらのデータの選定は、全く主観的に行っている。未知語ができるだけ少ないもの、縮約や などの口語的要素ができるだけ現れないものを中心に選んだ。上述の通り、MeCab用に解析された大規模データがあるわけではなく、学習用コーパスの作成を手作業で行うための措置である¹⁹。学習用コーパスの分量、作成については今後の課題としたい。

3.3 MeCab用辞書の構築と接続コストの修正

上で準備したSeed辞書と学習用コーパスを用いて、MeCab用の辞書を構築する。辞書の構築にはMeCabに同梱されているプログラムを利用した。構築した辞書はそのままで利用することができるが、解析率をより向上させるため、構築後に接続コストの修正を行った。

¹⁹実際の作業は、ごく少量のデータで学習を行なったプロトタイプの辞書を作成し、その辞書を利用して第一のデータを解析、修正したのち、それを学習用コーパスとしてさらに辞書を構築、第二のデータを解析...、というように行った。

学習によって生起コスト，接続コストが適切に設定されるが，ありえない接続が完全に排除されるわけではない．大量の学習用データを用いれば，こうした誤差が少なくなると思われるが，本稿では辞書の構築後に，直接接続コストを修正した．具体的には，用言と語尾・接尾辞の接続コスト，接尾辞と語尾の接続コストで，不可能な接続をゼロに設定した．構築した辞書をそのまま利用すると，母音語幹用言の第Ⅰ語基と第Ⅱ語基とを誤って解析する例が多くみられた．上記のような措置をとることで，例えば用言の第Ⅱ語基と「1 接続」の素性を持つ語尾の接続が排除されることになる．構築後の辞書では母音語幹用言の第Ⅰ語基と第Ⅱ語基を誤って解析する結果が多く見られたが，接続コストの修正により，こうした誤りはほぼなくなった．

3.4 解析率の評価

MeCabには，構築した辞書の解析率を評価するプログラムが用意されている²⁰．試みに，(1993)の4級第1課「」(58文，形態素数 1939)をテストデータとして，解析率を算出してみた．接続コストの修正前と後，さらにテストデータに現れる未知語 30 語の追加前と追加後とで解析率を比較すれば，以下の通りである：

表 3: テストデータの解析率比較(全素性についての正解率)

	コスト修正前	コスト修正後
未知語追加なし	87.3717	90.6242
未知語追加あり	91.2987	94.7831

構築したそのままの状態では約 87% の解析率となっているが，接続コストの修正や未知語の追加を行うことで，解析率が向上している．

なお，表 3 の結果は品詞 1 から「その他」まで全ての素性を比較した場合の数値であるが，品詞 1 のみの解析率を見ると，未知語を登録せず，接続コストも修正しない場合で 94.7125% であった．これは形態素の切り出しとその品詞判定が 94% 以上正解していることになる．さらに未

²⁰MeCabに同梱されている評価用のスクリプトmecab-system-evalは，解析結果の出力ファイルと正解ファイルを引数にとり，以下のような結果を出力する：

	precision	recall	F
LEVEL 0:	99.7070(2382/2389)	99.7905(2382/2387)	99.7487
LEVEL ALL:	98.7861(2360/2389)	98.8689(2360/2387)	98.8275

ここで「LEVEL」は比較に用いた素性のレベルを表し，最初の素性が「0」となる．「LEVEL ALL」は全ての素性を比較した値である．()内の数値は分母が形態素全体の数，分子は出力と正解とで一致する形態素の数で，「precision」の列は解析した出力ファイルについて，「recall」の列は正解ファイルについてのデータを示している．「F」列は左 2 列のパーセンテージの平均である．上の例では解析結果の形態素が 2389，正解ファイルの形態素が 2387 であり，両ファイルで最初の素性が一致する形態素が 2382，全ての素性が一致する形態素は 2360，ということになる．本研究では解析率として，全素性を比較した場合の平均値，即ち上の表で右下に示される数値を用いることとする．

なお，須賀井義教・村田寛(2011:53)でもこの評価について触れたが，「precision」「recall」の内容について逆に記述してしまった．ただし，最終的な解析率の記述には誤りのないことを断っておく．

知語を登録し、接続コストを修正した辞書では品詞1の解析率が98.5021%と、かなり向上が見られた。須賀井義教・村田寛(2011)でも辞書の登録項目数と解析率との関連について述べているが、接続コストの修正に関しては課題となっていた。もちろん、大量の学習用データを用いてコスト計算を行い、辞書を構築するのが望ましいが、学習用コーパスが少ない状況では、本稿で行ったようなコストの修正によって、解析率を上げることができると言える。

4 応用の事例

ここまで、MeCabで韓国語を解析するための辞書構築について述べてきた。以下ではその辞書をどのように活用することができるか、いくつかの試みを紹介する。

4.1 解析結果を利用する

本稿で構築した辞書を利用し、MeCabで韓国語の形態素解析を行うと、以下のような結果が出力される²¹：

```
(17) hangug'e      Noun,普通,*,*,*,      ,      ,韓國語
      goqbu        Noun,普通,*,*,*,      01,      ,工夫
      nyn          Ending,助詞,題目,*,*,      ,      ,*
      jeqmal       Adverb,一般,名詞可能,*,*,      01,      ,正-
      jaimi'iss'e  Adjective,自立,*,語基3,*,      ,      ,*
      'io          Ending,語尾,終止形,*,3接続,      ,      ,*
      .           Symbol,ピリオド,*,*,*,*,*,*,*
      EOS
```

解析された形態素と素性の間にはタブが入っている。この出力結果をテキストファイルに保存し、テキストエディタなどを使ってさまざまに活用することができる。

また、2.2節で紹介したように、コーパス検索ツールであるChaKiを利用することもできる。図2はChaKiを使って、連体形語尾に不完全名詞が続く例を検索した時の画面である。

ChaKiを利用すると、文字列を指定して検索するだけでなく、品詞や活用語基といった素性を基に、検索を行うことができる。ここでは連体形語尾に不完全名詞が続く例を検索してみた。図2ではローマ字転写のままデータを利用しているが、MeCabの出力形式を変更することで、表層形を表示させることもできる。また、ChaKiで共起頻度などに関するデータも集計することが可能であり、今後韓国語研究への活用が期待できる。

²¹入力は一文字のローマ字転写による。韓国語をそのまま入力としたい場合には、UTF-8でエンコーディングしたテキストファイルをPerlなどでローマ字転写に変換し、その結果を入力としてMeCabに渡す必要がある。

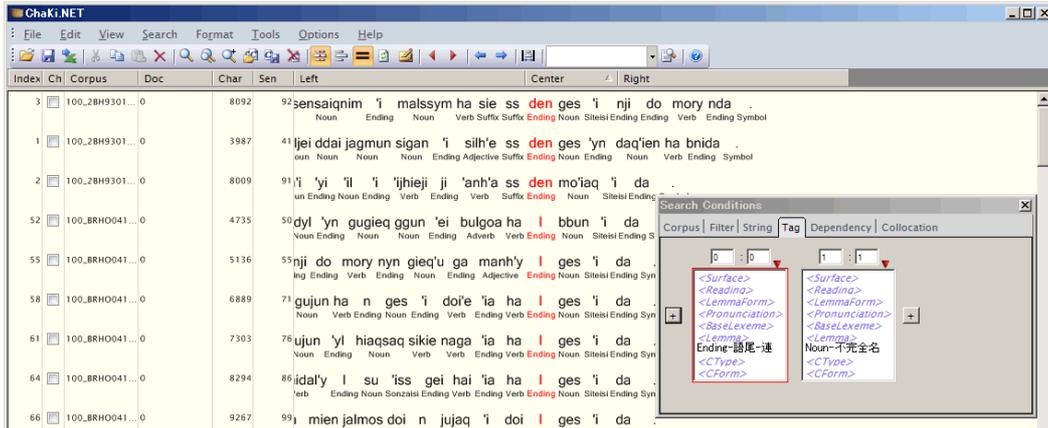


図 2: ChaKiを利用したコーパスの検索

4.2 インターネットで利用する

MeCabはPerlなどのプログラミング言語からも利用することができる．本稿ではPerlによるCGIを利用して，インターネットを通じて形態素解析を行うためのページを作成した²²．

MeCabによる韓国語の形態素解析

オープンソース形態素解析エンジンMeCabを用いた現代韓国語の形態素解析を行います．韓国語の文を入力して、「解析」ボタンを押すかEnterキーを押してください．解析結果は別ウィンドウで表示されます．(Dictionary Version: 1.14 / 14437 Entry)

함께 서울 시내 구경도 할 수 있어서 기쁩니다. 全素性を表として出力

現代朝鮮語の形態素解析 - 結果 - Google Chrome

porocise.sakura.ne.jp/cgi-bin/analyzer.cgi

元のテキスト:

함께 서울 시내 구경도 할 수 있어서 기쁩니다.

解析結果:

表層形	表層形alpha	品詞1	品詞2	品詞3	活用形	接続情報	辞書項目	その他
함께	hamggei	Adverb	一股				함께	
서울	se'ul	Noun	固有名詞	都市			서울	
시내	sinai	Noun	普通				시내03	市内
구경	gugieq	Noun	普通				구경01	
도	do	Endina	助詞	とりたて			도	

図 3: Perl/CGIを利用した，ウェブでの形態素解析ページ

このページでは解析を行うだけでなく，解析結果の出力方式をいくつか選べるようにした．「品詞のみルビとして出力」を選択した場合，品詞1にあたる素性，すなわち品詞の大分類のみ，ルビとして表示する²³．この形式では漢字表記などが記述された素性「その他」を()に入れて表示するようにしている(図 4)．

²²<http://porocise.sakura.ne.jp/korean/morph/analyzer.html> にて公開している．なお，筆者が利用しているプロバイダ「さくらインターネット」では，サーバーにMeCabが既にインストールされていたが，バージョンは0.97であった．また，Perl用モジュールとしてText::MeCab(0.20011)を利用した．

²³表示にはHTMLの<ruby>タグを利用しているため，一部のブラウザではルビではなく，()の中に表示される．

Noun Ending Noun Ending Noun Verb Ending Symbol
 학교(學校) 에서 한국어(韓國語) 만 공부(工夫) 해 요 .

학교 韓國語만 공부
 學校에서 韓國語만 工夫해요 .

図 4: 品詞をルビとして出力

図 5: 漢字表記に置き換えて出力

また「全素性を表として出力」は、図 3 に示したような表の形式で全ての素性を表示する。品詞の大分類だけを表示すれば適切に解析されているようでも、全ての素性を見ると同音異義語の判定に誤りがある場合がある。こうした誤りをチェックするためには、全ての素性を表示する必要があると考えられる。

「名詞のみ漢字表記を出力」オプションを選択すると、名詞についてのみ、「その他」素性で置き換えて出力し、ハングルをそのルビとして表示する(図 5)。現状では「その他」に漢字表記以外の内容が含まれるため、予期しない結果となることがある。今後「その他」の記述を漢字表記とそれ以外の内容とに分けて、辞書を構築する予定である。

本稿ではいくつかの試みだけを紹介したが、韓国語教育という観点からどのように活用することができるか、今後検討していく必要があるだろう。構築した辞書については、さらに項目を拡充した上で、インターネットを通じて公開する予定である。

4.3 ユーザーによる単語の追加

構築した辞書に登録されていない語を、ユーザーが自由に追加することができる²⁴。ここでは例として「가 .」という入力文を、「soniesidai,66,53,5493,Noun,固有名詞,グループ名,*,*, , ,少女時代」というエントリを持つユーザー辞書を追加して解析した結果(18a)と、ユーザー辞書を利用しなかった場合の解析結果(18b)とを挙げておく：

- (18) a. soniesidai Noun,固有名詞,グループ名,*,*, , ,少女時代
 ga Ending,助詞,主格,*,*,가,가,*
 joh'a Adjective,自立,*,語基3,*, 01, ,*
 'io Ending,語尾,終止形,*,3接続, , ,*
 . Symbol,ピリオド,*,*,*,...,*
 EOS
- b. sonie Noun,普通,*,*,*, 02, ,少女
 sidai Noun,普通,*,*,*, 02, ,時代
 ga Ending,助詞,主格,*,*,가,가,*
 joh'a Adjective,自立,*,語基3,*, 01, ,*
 'io Ending,語尾,終止形,*,3接続, , ,*
 . Symbol,ピリオド,*,*,*,...,*
 EOS

²⁴詳細はMeCabホームページ「単語の追加方法」(<http://mecab.sourceforge.net/dic.html>)を参照。

5 問題点・課題

最後に、本稿で構築した辞書をMeCabで利用する際の問題点、課題を挙げておく：

1. 未知語の処理
2. 同音異義語の判定
3. 学習用データの量

5.1 未知語の処理

文脈に応じた未知語の処理は、MeCabが持つ特徴のひとつである。MeCabホームページによればその精度は正確ではないというが、前後の接続に応じて、辞書にない語でも、その品詞を適当に判定してくれる。

通常日本語の形態素解析では字種に基づく未知語処理が行われるという(工藤拓・山本薫・松本裕治 2004:90)。韓国語の場合、表記にはほとんどハングルのみ使用されるため、日本語の未知語処理と同様に考えることはできないと考えられる。本稿では、漢字についてはひとつの字種カテゴリとして設定し、その連続を一般名詞あるいは人名、組織名として扱うよう、暫定的に未知語処理の定義を行った。この定義により、例えば以下の例(19)のように、「須賀井 가 .」(a)と「須賀井 .」(b)とで異なった結果を出力してくれる。もちろん漢字表記の「須賀井」は辞書に登録されていない項目である：

- (19) a. 須賀井 Noun, 固有名詞, 姓, *, *, *, *, *, *
ssi Noun, 接尾語, *, *, *, * 07, , 氏
ga Ending, 助詞, 主格, *, *, 가, 가, *
ga Verb, 自立, *, 語基3, *, 가 01, 가, *
ss'e Suffix, 過去, *, 語基3, *, ㄷ, ㄷ , *
'io Ending, 語尾, 終止形, *, 3接続, , , *
. Symbol, ピリオド, *, *, *, . . . , *
EOS
- b. 須賀井 Noun, 普通, *, *, *, *, *, *
'ei Ending, 助詞, 処格, *, *, , , *
ga Verb, 自立, *, 語基3, *, 가 01, 가, *
ss'e Suffix, 過去, *, 語基3, *, ㄷ, ㄷ , *
'io Ending, 語尾, 終止形, *, 3接続, , , *
. Symbol, ピリオド, *, *, *, . . . , *
EOS

漢字表記の語については、以上の設定に加えて「語根」として処理するような設定を追加すれば、概ねカバーできると考えられる。しかし、漢字表記は名詞にだけ現れるわけではなく、

「決」「真」のように、副詞の一部などにも現れる。字種によって未知語の判定を行う場合、これらの語から漢字部分を除いた「」「」をどのように扱うか、検討しなければならない。もちろん、「決」自体を辞書に登録する、というのもひとつの方策である。

また、ハングルの連なりが辞書に登録されていない場合、やはり未知語処理が行われるが、本稿では内部でローマ字転写を用いており、その連続をひとまとまりとして扱うため、いったん解析を誤ると、語節の残りの部分を全て未知語として処理してしまう。例えば「

」(進学率が高い学校)を解析してみると、「」は辞書に登録されているにもかかわらず、以下のような誤解析を起こす：

(20) jin Noun, 語根,*,*,*, 06, ,津-
 ha Verb,自立,*,語基2,*, 01, ,*
 griul'i Noun,普通,*,*,*,*,*,*
 nop'y Adjective,自立,*,語基2,*, , ,*
 n Ending,語尾,連体形,*,2接続,⊥,⊥,*
 haggio Noun,普通,*,*,*, , ,學校
 EOS

これは「(率)」が辞書に登録されていないためと見られるが、問題は「griul'i」の部分である。助詞の「i」()だけでも分節してくれればよいが、「gr」以下解析ができなかった部分をひとまとまりにして未知語として処理している。

こうした問題は解析しようとする文自体にローマ字が含まれている場合にも生じる。「MS Windows が .」という文の解析結果は以下の通りである：

(21) MS Noun,固有名詞,人名,*,*,*,*,*
 W Noun,普通,*,*,*,*,*,*
 i Siteisi,自立,融合,語基2,*, , | ,*
 n Ending,語尾,連体形,*,2接続,⊥,⊥,*
 do Ending,助詞,とりたて,*,*, , ,*
 wsnyn Noun,普通,*,*,*,*,*,*
 munjei Noun,普通,*,*,*, 06, ,問題
 ga Ending,助詞,主格,*,*,가,가,*
 manh'a Adjective,自立,*,語基3,*, , ,*
 'io Ending,語尾,終止形,*,3接続, , ,*
 . Symbol,ピリオド,*,*,*,...,*
 EOS

上の例の5行目で、「wsnyn」の部分から助詞を抽出できていない。アルファベットの大文字はひとつの字種カテゴリにまとめているため、「MS」「W」の部分のみ抽出され、未知語として処理されている。

未知語の解析におけるこうした誤りをなくすためには、もちろん辞書のエントリを増やして事実上未知語をなくすという方法があるが、それも限りがある。とはいえ、解析にローマ字転写を用いている以上、こうした問題は避けて通れないともいえよう。今後も引き続き検討すべき課題である。

5.2 同音異義語の判定

本稿で構築した辞書を利用して解析を行う場合、同音異義語を適切に判定することができない。「`mal`」を解析すると、以下のような結果となる：

```
(22) gyren Prefix,*,*,*,*, 01, ,*
mal Noun,普通,*,*,*, 05, ,
'yl Ending,助詞,対格,*,*, , ,*
dyl'e Verb,自立,r語幹,語基3,*, 01, , -
ss'e Suffix,過去,*,語基3,*,从,从 ,*
'io Ending,語尾,終止形,*,3接続, , ,*
. Symbol,ピリオド,*,*,*,...,*
EOS
```

この例では形態素の抽出、品詞の判定に問題はないが、「`mal`」「`yl`」のエントリが期待したものになっていない。もちろん辞書には「ことば」を意味する「`mal` 01」,「聞く」を意味する「`yl` 01」が登録されている。

このうち「`yl`」については、「ことば」を指す語は「学習用語彙」の「1段階」に、動物を指す語は同じく「2段階」に登録されており、さらに接尾語としての「末(`yl` 11)」が「2段階」に登録されている。このうち「ことば」「馬」は、本稿の辞書で生起コストがいずれも 3191 となっているが、単語の重要度を基準として生起コストを変更すれば、例 (22) のような誤りはある程度回避できると考えられる。

しかし、本稿の辞書に登録されている「`mal`」は「(`mal`) 01」「(`yl`) 04」「(`yl`) 01」の第 III 語基で、いずれも「学習用語彙」の「1段階」に含まれる語である。このような場合にどのような措置をとるべきか、現在のところ明確な解決策を得られていない。「`mal`」「`yl`」などその他の自動形態素解析器は品詞だけを出力するため、こうした同音異義語の問題がゼロではないものの、より発生しにくいと考えられる。

5.3 学習用データの量

前述の通り、本稿では辞書構築の学習に用いたデータが極端に少ない。MeCabは学習用コーパスによる学習に基づいてコスト計算を行うため、学習用データの量は解析率向上において重要なファクターといえる。もちろん形態素解析の基となる辞書のエントリ量も重要である。学習用データと辞書エントリをいかにして増やすか、今後作業を続けていく必要があるだろう。

6 おわりに

ここまで、韓国語の形態素解析にMeCabを用いるための辞書構築、実際の運用例、問題点について述べてきた。形態素解析エンジンとしてMeCabを用いることで、辞書や出力、未知語処理の動作などをユーザーが自由に定義することができ、さらにインターネットを通じて解析処理を行うことができるようになる。本稿では特に辞書の構築について紹介し、さらにインターネットでの利用について簡単に触れた。

本稿で構築した辞書は未だ登録項目数が 14,000 余りであり、実際にさまざまな入力文を処理するためには今後辞書の項目を増やしていく必要がある。現状の辞書でもテストデータで 90% 近い解析率となっているが、辞書の項目、さらに学習用データを充実させることで、解析率をさらに向上させることができると思われる。

同音異義語の処理やハングルのローマ字転写に起因する未知語の処理などといった課題が残っているが、韓国語の形態素解析において、MeCabは充分実用に耐えうると考えられる。活用の方法なども含め、今後さらに検討を重ねていきたい。

参考論著

- 菅野裕臣(1997)「朝鮮語の語基について」『日本語と外国語との対照研究 IV 日本語と朝鮮語 下巻 研究論文編』, くろしお出版, pp.1-21
- 工藤拓・山本薫・松本裕治(2004)「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]』2004-NL-161, 情報処理学会, pp.89-96
- 須賀井義教・村田寛(2011)「15世紀朝鮮語の形態素解析について」『教養・外国語教育センター紀要』第1巻第2号, 近畿大学教養・外国語教育センター, pp.41-56
- 平野善隆(1997)『用言の活用を考慮した韓国語品詞体系の提案とそれを用いた韓国語形態素分析』, 奈良先端科学技術大学院大学情報科学研究科情報処理学専攻修士論文(NAIST-IS-MT9551092)
- 福井玲(1989)「中期朝鮮語文献の電子計算機による処理」『明海大学外国語学部論集』2, 明海大学, pp.17-29
- 松本裕治ほか(1997)『単語と辞書』岩波講座言語の科学 3, 岩波書店
- 村田寛(2010)「15世紀朝鮮語の形態素解析の試み—MeCabを利用して—」『福岡大学研究部論集A: 人文科学編』Vol.10 No.3, 福岡大学, pp.17-28
- 守岡知彦(2008)「MeCabを用いた古典中国語の形態素解析の試み」『情報処理学会研究報告[人文科学とコンピュータ]』2008-CH-73, 情報処理学会, pp.17-22
- 山本和英(2000)「計算機処理のための韓国語言語体系と形態素処理」『自然言語処理』Vol. 7 No. 4, 言語処理学会, pp.25-62
- 油谷幸利(2008)「朝鮮語Web辞典の設計について」『朝鮮学報』第206輯, 朝鮮学会, pp.(1)-(37)
- 油谷幸利(2009)「朝鮮語Web辞典について—用例辞典から学習辞典へ—」『朝鮮学報』第211輯, 朝鮮学会, pp.(1)-(40)
- (2002;2003) “ ”(),
- (2005) “ ”(2005-1-33), (国立国語院ホームページより入手可能)
- ・ (2004) ‘ ’ “ ” :
- ”, Vol. 31 No. 1, , pp.89-99
- (1993) “ ” (3 ・ 4),
- (2004) “ ”,
- (2002) “ — ”(2002-1-17), (国立国語院ホームページより入手可能)
- (2003) “ ”(2003-1-4), (国立国語院ホームページより入手可能)
- (2006) “ ”,
- Lafferty, J., A. McCallum and F. Pereira(2001) *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. In Proc. of ICML-01, pp.282-289

付録：本稿で用いたローマ字転写

本稿で用いたローマ字転写の方法は、概ね福井玲 (1989) に従ったものであるが、終声のㄹのみ「l」で表記することとした。

表 1: 初声字母

ハングル	ㄱ	ㄲ	ㄴ	ㄷ	ㄸ	ㄹ	ㅁ	ㅂ	ㅃ	
ローマ字	g	gg	n	d	dd	r	m	b	bb	
ハングル	ㅅ	ㅆ	ㅇ	ㅈ	ㅉ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ
ローマ字	s	ss	'	j	jj	c	k	t	p	h

表 2: 中声字母

ハングル	ㅏ	ㅑ	ㅓ	ㅕ	ㅗ	ㅛ	ㅜ	ㅠ	ㅡ	ㅚ	ㅜ	ㅝ
ローマ字	a	ai	ia	iai	e	ei	ie	iei	o	oa	oai	
ハングル	ㅟ	ㅡ	ㅣ	ㅥ	ㅦ	ㅧ	ㅨ	ㅩ	ㅪ	ㅫ	ㅬ	
ローマ字	oi	io	u	ue	uei	ui	iu	y	yi	i		

表 3: 終声字母

ハングル	(なし)	ㄱ	ㄲ	ㄴ	ㄷ	ㄸ	ㄹ	ㅁ	ㅂ	ㅃ	ㅄ	ㅅ	ㅆ	
ローマ字	(空白)	g	gg	gs	n	nj	nh	d	l	lg	lm	lb	ls	lt
ハングル	ㅈ	ㅉ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ	ㅇ	ㅈ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ
ローマ字	lp	lh	m	b	bs	s	ss	q	j	c	k	t	p	h

