MeCab 用韓国語形態素解析辞書の構築

須賀井 義教
近畿大学 総合社会学部
sugaiy@kindai.ac.jp

1　はじめに

現代韓国語の形態素解析器は既にいくつかが開発・公開されているが，その多くは解析器と辞書が分離されておらず，辞書に新たな項目を追加できないなどといった問題点を抱えている．そこで本研究では，オープンソース形態素解析エンジン MeCab[1] を用いて現代韓国語を形態素解析するための辞書を構築し，ユーザーがカスタマイズ可能な辞書を提供する．

また，Perl などのプログラム言語から MeCab を利用できるという利点を生かし，韓国語学習用ツールの開発についても検討する．

2　先行研究

韓国・国立国語院が国語情報化プロジェクト「21 世紀世宗計画」(以下「世宗計画」) の成果として公開している「지능형 형태소 분석기 (知能型形態素分析器)」は，EUC-KR でエンコーディングされたテキストファイル，あるいは直接入力された韓国語文を解析し，「어절 (語節)」(文節と似た区切り) ごとに品詞タグを付与したテキストを出力するものである．

その他の自動形態素解析器としては，高速な解析を目指した「MACH」(심광섭・양재형 2004)，「HAM」(現在は「KLP」)(강승식 2002;2003) などがあり，「POSTAG SEJONG/K」[2] や「KRISTAL Morphological Analyzer」[3]，「꼬꼬마 형태소 분석기」[4]，「UTagger」[5](신준철・옥철영 2012) などのように，インターネットを通じて公開されている形態素解析サービスもある．

これらの多くは辞書に新たな項目を追加することができず，ユーザーがカスタマイズできる余地がないといってよい．ブログや掲示板など，インターネットで使用される韓国語の表現にはほとんど対応できないのが現状である．

また，日本においては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, 山本和英 2000)．いずれも機械処理を優先しており，語尾の処理など従来の文法記述とやや異なる部分があるため，解析結果の言語研究への利用などに困難が少なくない．

3　解析用辞書の作成

3.1　辞書の基礎データ

本研究では，「한국어 학습용 어휘 선정 결과 보고서 (韓国語学習用語彙選定結果報告書)」(조남호 2003)(以下「学習用語彙」とする) および「현대국어 사용 빈도 조사 (現代国語使用頻度調査)」(조남호　2002)(以下「頻度調査」) の電子データを利用した[6]．「学習用語彙」についてはほぼ全てを収録し，「頻度調査」は「名詞」と「副詞」の全て，さらに「助詞」と「語尾」，「形容詞」，「動詞」の一部を収録した．

このほか，韓国人の姓，韓国の行政区域名，ソウルの地下鉄駅名，世界の国名と首都名など，インターネット上のリソースも活用し，項目を追加してある．さらに上記「世宗計画」成果物 DVD-ROM(2011 年修正版) に含まれる「형태분석 말뭉치 구축 지침 (形態分析コーパス構築指針)(Ver. 2005-1)」を参考に，体言接頭辞なども追加した．

上記のデータなどから，60,177 項目からなる解析用辞書を作成した．機関や団体といった固有名詞，話しことばで用いられる語尾など，今後も作業を進めて項目を追加していく予定である．

なお，登録項目の記述においてはハングル 1 文字を字母の単位に分解し，Unicode の「Hangul Jamo」(U+1100-11FF) を用いて表記することにする．例え

[1] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[2] http://isoft.postech.ac.kr/Research/POSTAG/sejong/postag_sejong_k.php
[3] http://www.kristalinfo.com/K-Lab/ma/
[4] http://kkma-sc.snu.ac.kr/
[5] http://nlplab.ulsan.ac.kr/
[6] いずれも韓国・国立国語院ホームページの「자료실 (資料室)」から入手可能 (http://www.korean.go.kr/).

表 1: 品詞分類の一覧

品詞 1	品詞 2(品詞 3)
名詞	普通，代名詞，하다語根，助数詞 (固有のみ)，不完全名詞，接尾語，数詞 (漢数詞，固有数詞)，固有名詞 (姓，名，国名，地名，都市，人名，など)
動詞	自立，非自立
形容詞	自立，非自立
指定詞	自立
存在詞	自立
副詞	一般 (名詞可能)，否定
助詞・語尾	助詞 (主格，対格，など)，語尾 (終止形，接続形，連体形，名詞形成)
先語末語尾	過去，尊敬，将然，など
冠形詞	数詞 (固有数詞)
接尾辞	名詞派生，動詞派生，形容詞派生
接頭辞	体言接頭辞
接続語	
間投詞	
記号	

ば「말씀」(おことば) は「ㅁ ㅏㄹ ㅆㅡㅁ」のように表記される．韓国語の形態素は字母の単位に対応するものがあり，ハングル 1 文字ずつを解析の対象とした場合，その中から個別の形態素を抽出することが難しい[7]．村田寛 (2010) や須賀井義教・村田寛 (2011) ではローマ字で転写を行ったが，元の入力にローマ字が含まれる場合にうまく処理できなかった．本研究では字母単位に分解することで，こうした問題を解決した．

3.2 素性の記述

本研究では，辞書の素性を以下のように設定した：

(1) 品詞 1，品詞 2，品詞 3，活用形，接続情報，辞書形，表層形，漢字，備考，学習用語彙レベル

「品詞 1」は概ね「学習用語彙」の品詞分類に従った．「品詞 2」はその細分類であり，「品詞 3」はさらにその補足情報である．それぞれの詳細については表 1 に示した．

本研究では用言の活用記述にあたり「語基」(菅野裕臣 1997) の概念を利用している．韓国の文法研究における活用体系を利用する場合，語幹末の母音によって語尾に異形態があり，その選択を記述しなければならず，処理が煩雑になる．語尾を不変とし，用言語幹が 3 つの語基を持つという方式で記述を行うことで，日本語の活用体系と同様に処理することができ，結果的に辞書の記述が容易になる．そのため，用言や接尾辞では「活用形」に何番目の語基であるかを，語尾では「接続情報」に第何語基につくかを記述する．

また，「学習用語彙」などとは異なり，用言に存在詞 (「있다」(ある) など) と指定詞 (「–이다」(…である)) などを認め，別に分類してある．

その他の素性のうち，「辞書形」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，概ね『표준국어대사전 (標準国語大辞典)』(국립국어연구원 1999) に準じた同音異義番号がつけてある．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．

漢字表記が可能な項目については，「漢字」素性にその表記を記述した．「정말」(本当に) のように，語の一部のみ漢字表記できる場合には，「正ㅁ ㅏㄹ」のようにしてある．また，「学習用語彙」「頻度調査」の「補充情報」を「備考」という素性に記述した．多くは同音異義語を区別するための情報である．

「学習用語彙レベル」は，「学習用語彙」における，A から C の学習レベル表示をそのまま記述したものである．学習補助ツールの作成の際に用いる．

3.3 学習用コーパスの作成

本研究では，学習用コーパスに「世宗計画」の「형태분석 말뭉치 (形態分析コーパス)」を利用した．データとして用いたのは，文語コーパスのうち 8 つのファイル[8]で，それぞれ章などの見出しや前書きを除いた冒頭の 100 文ずつ，総 800 文 (延べ形態素数 23,441) である．なお，学習用コーパスには 228 項目の未知語を含んでいる．

学習用コーパスの作成にあたっては，品詞分類や活用形の記述など，元の解析結果と異なる部分や解析の誤りなどがあるため，適宜修正を行った．また，「졸업하다」(卒業する) など「学習用語彙」に含まれている項目は，元の解析結果で「졸업」(卒業) と「하다」(する) に分けられていても，そのまま一つの項目として記述してある．この他に，「世宗計画」では母音終わりの名詞に一部の語尾がつき，指定詞が脱落する場合でも，解析済みコーパスでは指定詞を復元しているが，本研究ではこうした復元を行わない．

[7]例えば「갔어요.」(行きました) から過去の接尾辞を取り出す場合など．

[8]タイトルとファイル名はそれぞれ以下の通り．함께 걷는 이 길은 (BTHO0111.txt), 알기 쉬운 인권 지침 (BTHO0112.txt), 인간을 위하여 미래를 위하여 (BTHO0116.txt), 우리 학문의 길 (BTHO0124.txt), 인간과 사회—전통윤리와 현대풍조의 갈림길에서 (BTHO0131.txt), 대중화술 (BTHO0376.txt), 고객과 경쟁하라 (BTHO0390.txt), 논술의 정석 (BTHO0414.txt)

以上のような措置をとったため，元の解析済みコーパスと本研究の学習用コーパスとは若干の違いがあることを断わっておく．

4　構築した辞書による解析

以上の辞書と学習用データを用いて，MeCab 用の辞書を構築した．辞書構築と解析に用いた MeCab のバージョンは 0.994 である．ここでは，構築した辞書を用いて解析を行なった結果について述べる．

なお，構築直後の辞書では用言活用形と語尾との対応，例えば母音語幹用言の第 I 語基と第 II 語基を誤るなどといった誤りが多く見られた．そこで，辞書構築後に連接コストの一覧である matrix.def ファイルを修正し，許容されない連接のコストを 0 にして再び辞書を構築した．以下の記述では，こうした措置を取った後の辞書を用いることを断っておく．

解析精度の測定には MeCab 本体とともに配布されている評価用スクリプト mecab-system-eval を用いた．テスト用のデータには韓国・西江大学校の韓国語上級教材 (서강대학교 한국어교육원　2009a,b) から読解パート 3 課分の 95 文 (形態素数 2003，うち未知語 1 項目) と，「世宗計画」の生コーパスから 1 ファイル[9]，冒頭の 50 文 (形態素数 1372，うち未知語 18 項目) を用いた．解析の結果は表 2，3 の通りである．

表 2: 解析の精度 (韓国語教材)

Level	精度	再現率	F 値
0 (境界)	99.0514	99.0514	99.0514
1 (品詞 1)	98.5022	98.5022	98.5022
2 (品詞 2)	97.7034	97.7034	97.7034
3 (品詞 3)	97.6036	97.6036	97.6036
6 (辞書形)	96.8547	96.8547	96.8547

表 3: 解析の精度 (世宗計画)

Level	精度	再現率	F 値
0 (境界)	95.1289	96.7930	95.9538
1 (品詞 1)	94.7708	96.4286	95.5925
2 (品詞 2)	94.1977	95.8455	95.0145
3 (品詞 3)	94.1261	95.7726	94.9422
6 (辞書形)	92.3352	93.9504	93.1358

韓国語教材を解析した場合 (表 2)，一般的な内容を含んだセクションを対象としたこともあり，高い精度を見せている．一方未知語を多く含むテクスト (表 3) では品詞 1 の正解が 95%程度にとどまっており，やはり辞書の項目が不足しているためと見られる．また，誤りの部分には，助詞の縮約などが解析できていないケースが見られた．例えば「땐」(＜때「時」＋는「は」) を，直前に用言の連体形が続いているにも関わらず，動詞「때다」(火をくべる) に連体形語尾「II–ㄴ」がついた形として解析している．誤りの理由としては，硬い書きことばのデータを中心に学習用コーパスを作成したため，上記のような助詞の縮約形がほとんど表われていないことが考えられる．辞書の項目だけでなく，学習用データについても補充していく必要があると思われる．

5　読解補助ツールの作成

Perl などのプログラミング言語から MeCab を利用できるという利点を生かし，形態素解析の結果を用いた韓国語読解補助ツールの作成を試みた．ここでは Perl による CGI を用いた，インターネットで利用できるツールのプロトタイプを紹介する[10]．

図 1: メイン画面

図 1 に示すメイン画面で韓国語を入力し，「形態素解析」または「漢字表記に変換」ボタンを押すと，それぞれの結果が表示される．例えば「학교에서 수업을 들으면 유익한 정보를 많이 얻을 수 있지요.」(学校で授業を聞けば，有益な情報がたくさん得られますよ) という文を入力し，それぞれのボタンを押した結果は図 2，図 3 の通りである．

図 2 に示す色分けは，「学習用語彙」に示された語彙の学習レベルを反映させたものである．それぞれの項目にカーソルを合わせると，辞書形がツールチップで表示される．また，メイン画面では学習レベルごとにウェブ辞書[11]へのリンクを設定することができ，解

[9] タイトルとファイル名は소설 창작 강의 (BRHO0402.txt) である．「正解」データは同内容の解析済みファイル (BTHO0402.txt) を加工した．

[10] サーバーにインストールされている MeCab のバージョンは 0.97 であり，Perl 用モジュールは Text::MeCab-0.20013 をインストールした．なお，現在公開中のツールでは以前のバージョンの辞書を使用しており，出力などが若干異なることを断っておく．`http://porocise.sakura.ne.jp/korean/mecab/main.html`

[11] 現在は NAVER 日本語辞書を参照することにしている．`http://jpdic.naver.com/`

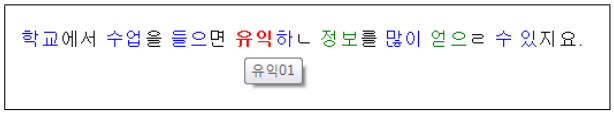

図 2: 形態素解析の結果を表示

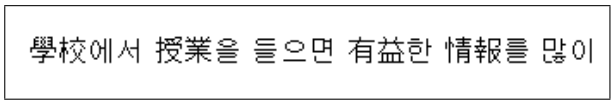

図 3: 漢字表記への変換結果を表示 (一部)

析結果で該当する項目をクリックすると，ウィンドウが開いてウェブ辞書を参照することができる．

また，日本語と韓国語はいずれも中国語由来の漢語 (韓国語では「漢字語」) を多く共有しており，図 3 のように漢字表記に変換して提示することで，文脈の把握が容易になると考えられる．

このように，形態素解析の結果を生かして辞書形や学習レベル，漢字表記を表示し，またウェブ辞書へのリンクを生成することで，学習者にとって有益な情報を提供することができる．

現在のところ上述の機能のみを提供しているが，今後学習者へのアンケートなどを通じてニーズを調査し，機能を改善，追加していく予定である．

6 おわりに

本研究では現代韓国語の形態素解析を行うために，MeCab 用の辞書を構築し，あわせて読解補助ツールの作成も試みた．辞書の項目数が未だ少ないものの，未知語がそれほど多くなければ，実用に耐えうる解析が可能であると言える．

現状では学習用コーパスが書きことば中心であること，またその量が少ないことなどが課題として挙げられる．今後は話しことばのデータや新聞記事など，様々なジャンルのテクストを追加していく．

もちろん，辞書の登録項目も増やさなければならない．特に新聞記事などは固有名詞が多く，また「大統領選挙」が「大選」となるなど縮約が多く見られ，現状の辞書では解析の精度がそれほど高くないと思われる．インターネット上のリソースなどを利用して，項目を追加していきたい．

また，Perl などのプログラム言語から MeCab を利用し，解析結果を加工して，韓国語の学習に役立てられる可能性についても確認した．今後も改良が必要であるが，信頼できる情報を提供するためには，やはり解析の精度が高くなければならない．特に漢字表記に変換する場合，同音異義語を表示しては学習の役に立たない．今後学習者のニーズも調査しつつ，機能の追加と辞書の補充を行っていく予定である．

謝辞

本研究は 2010–2012 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」(課題番号 22320115) による研究成果の一部である．

参考文献

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

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)

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

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

강승식 (2002;2003) “한국어 형태소 분석과 정보 검색”(수정판), 홍릉과학출판사

국립국어연구원 (1999) “표준국어대사전”, 두산동아

서강대학교 한국어교육원 (2009a) “서강한국어 Student’s Book 5A 읽기・말하기”, 서강대학교 국제문화교육원 출판부

서강대학교 한국어교육원 (2009b) “서강한국어 Student’s Book 5B 읽기・말하기”, 서강대학교 국제문화교육원 출판부

신준철・옥철영 (2012) ‘기분석 부분 어절 사전을 활용한 한국어 형태소 분석기’, “정보과학회논문지: 소프트웨어 및 응용”, Vol.39 No.5, 한국정보과학회, pp.415–424

심광섭・양재형 (2004) ‘인접 조건 검사에 의한 초고속 한국어 형태소 분석’, “정보과학회논문지: 소프트웨어 및 응용”, Vol.31 No.1, 한국정보과학회, pp.89–99

조남호 (2002) “현대 국어 사용 빈도 조사: 한국어 학습용 어휘 선정을 위한 기초 조사”(국립국어연구원 2002-1-17), 국립국어연구원

조남호 (2003) “한국어 학습용 어휘 선정 결과 보고서”(국립국어연구원 2003-1-4), 국립국어연구원