様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　５　月　２６　日現在

機関番号：　３４４１９
研究種目：　若手研究(B)
研究期間：　2013～2014
課題番号：　２５７７０１６４

研究課題名（和文）現代韓国語自動形態素解析のための辞書構築とその応用

研究課題名（英文）Development and application of dictionary for Korean morphological analyzer

研究代表者
須賀井　義教（SUGAI, Yoshinori）

近畿大学・総合社会学部・准教授

研究者番号：６０４５４６４１

交付決定額（研究期間全体）：（直接経費）　　1,300,000円

研究成果の概要（和文）：本研究では，オープンソース形態素解析エンジン「MeCab（めかぶ）」を利用して現代韓国語を解析するための辞書を構築し，それを研究・教育に活かすためのツール開発および公開を行った．本研究では，約10万5千の登録項目からなる形態素解析用辞書を構築した．構築した辞書は，オンラインのツールを通じて利用することができ，またオープンソースソフトウェアとしてインターネットで公開している．

研究成果の概要（英文）：For this research project, we used the open source morphological analysis engine, MeCab, to build a dictionary for analyzing the modern Korean language, and to develop and release tools for use in research and education. For this research, we built a morphological analysis-use dictionary from approximately 105,000 registered items. The dictionary that we built is able to be used through online tools and is available to the public on the internet as open source software.

研究分野：　言語学

キーワード：　韓国語学　韓国語教育　言語情報処理

１．研究開始当初の背景

（１）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．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 世紀世宗計画による一連の成果がよく知られている．例えば世宗計画の成果として公開している「知能型形態素解析器」は，テキストファイル，あるいは直接入力された内容を解析し，語節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．また，韓国でその他様々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したりすることはできない．日本国内においても韓国語の形態素解析に関する研究，特に解析のための品詞分類の提案などが行われているが，解析の便を優先してお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っている．

（２）このような状況を踏まえて，研究代表者は科研費を得てオープンソースの形態素解析エンジン「MeCab（めかぶ）」を用いた現代韓国語の形態素解析を試みた．MeCab は解析用辞書と解析器が分離されているため，韓国語の解析用辞書を準備すれば，現代韓国語の形態素解析を行うことができる．また，出力の形式をカスタマイズすることができ，さらに Perl などのプログラミング言語から利用できるという利点がある．

（３）先の科研費で構築した解析用辞書は，約 6 万の項目を含んでおり，上級韓国語教材の読解セクションを解析した際の解析率は平均 92.5%であった．固有名詞などを多く含む文章では解析率がやや落ちる．辞書に含まれる項目もそれほど多いわけではなく，また学習用データも 715 文と，補充が必要であるという見通しを得た．

２．研究の目的

（１）本研究では，「MeCab（めかぶ）」を利用して現代韓国語を解析するための辞書を構築し，それを研究・教育に活かすためのツールを開発，公開することを目的とする．本研究で構築したカスタマイズ可能な形態素解析用辞書の公開や，解析技術を利用したツールの開発により，韓国語情報処理技術の質的な向上と，韓国語の教育方法の新たな可能性を提示することを目指す．

（２）本研究の遂行により，日本語以外の言語でも，MeCab を用いて実用的な解析が可能であるということ，さらにそれを利用したツールの開発が可能であるということを示すことができる．また，ツールの公開によって同種のツール開発を促進し，新たな教育方法を提示することで，韓国語教育の向上に貢献することができると考える．

３．研究の方法

（１）本研究では，先の科研費によって構築した辞書をベースとして辞書開発を行うが，品詞体系を見直し，各項目の記述内容（「素性」と呼ぶ）を全面的に改訂した．具体的には，21 世紀世宗計画で構築，公開している「形態分析コーパス」の品詞体系を元にした．また，固有名詞などを含めて，10 万項目程度を目標として辞書の登録項目を増やす．さらに，新聞など多様なジャンルの文を用いて学習用データの増加に努める．

（２）既に公開している読解補助ツールのプロトタイプについて，学習者のニーズ調査などを行って機能の拡充を図る．

４．研究成果

（１）解析用辞書について

① 品詞体系の全面的な改訂：先に構築した辞書では，一部 21 世紀世宗計画の品詞体系に沿う形であったが，今回はこれを見直し，ほぼ全面的に 21 世紀世宗計画の品詞体系に合わせることとした．理由としては，21 世紀世宗計画の「形態分析コーパス」との互換性を考慮したことが挙げられる．このコーパスは韓国国内でもよく利用されており，この体系に合わせることで，韓国の利用者にも問題なく利用してもらえると考えられる．

　また，素性に「形態分析コーパス」のタグを追加した．出力の書式をカスタマイズできるという MeCab の特徴を利用すれば，「形態分析コーパス」とほぼ同じ書式，形式で出力が可能になる．

② 辞書の規模：計画の 1 年目終了時点で 76,395 項目，計画の 2 年目終了時点（現在）で 105,905 項目の規模となっている．主に地名や動詞，形容詞と，語根の項目補充を行った．地名については韓国・国立国語院が公開している外国地名のハングル転写資料などを利用して，日本や中国の地名なども追加した．

③ 学習用データ：従来の 715 文から倍以上に量を増やし，2,200 文のコーパスを作成した．当初の目標では 3,000～4,000 文程度としていたが，作業の都合により現状にとどまっている．少ない学習用データでも効率よく学習できるということが MeCab の特徴であるが，やはりデータが多いことに越したことはない．今後も作業を進める予定である．

　なお，データの内 2 割程度は新聞記事を含めている．名詞で文が終わるなど，従来学習用データとして利用してきた評論文とは文体が異なるためである．

　また，本研究では品詞体系の改訂に合わせて，学習用データも再構成した．特に大きな変更としては，学習用データの

作成において「形態分析コーパス」を参照した点がある. 従来の辞書構築時には, 定義した品詞体系に合わせて素性を記述していたが，本研究では,「形態分析コーパス」を正解データとして参照することで, より客観性の高い学習用データを構築することができた. 素性記述の際,「形態分析コーパス」の錯誤については適宜修正を加えた.

④ 解析用辞書の性能：先の科研費の場合と同じく, やはり上級韓国語教材を用いて測定した．71 個の文を解析した結果, 形態素境界の判定で約 98.7%, 同音異義語の判定まで含めると約 93.3%の解析率であった．同時に，評論文 50 文のデータでも評価を行ったが, その場合は形態素境界の判定で約 93.5%, 同音異義語の判定まで含めて約 88.9%という結果であった. 評論文のケースで解析率が低い理由としては, データとして利用した文章に外国人名が頻出すること, また国立国語院が公開している標準語辞書である『標準国語大辞典』にも登録されていない語などが見られたためと考えられる. どのような文章でも対応できることが望ましいが, 現状の性能でも十分に実用に耐えうると考える.

⑤ 解析用辞書の公開：本研究で構築した解析用辞書は，名称を「HanDic」とし, インターネットを通じて, オープンソースソフトウェアとして公開する. ドキュメント整備などの関係で, 未だ辞書本体を公開できていない状態であるが, SourceForge.JP にプロジェクト登録を行い, 公開のためのサイトをオープンした. 今後準備が整い次第, 早急に解析用辞書を公開する予定である.

（2）解析ツールの開発について

既に公開しているプロトタイプの読解補助ツールについて，本研究では辞書項目の素性改訂にともなう再設計を行った．インターフェイス，その他の機能拡充については今後の課題とする.

（3）国内外における本研究のインパクト

本研究の進捗は，学会での発表や学会誌投稿，あるいは研究代表者のホームページなどを通じて公開してきた．こうした情報から, 韓国語解析用辞書についての問い合わせが, 国内外の企業・研究所，大学の研究者から既に数件あった．この点だけでも，本研究の持つ社会的な意義を垣間見ることができるだろう．特に，MeCab というポピュラーなプログラムを利用している点で，利用価値があると見られたものと考える.

また，本研究を参考に，韓国国内でも MeCab 用の解析辞書構築が行われるようになった（例えば, http://eunjeon.blogspot.kr/）. 本研究の辞書とは規模や品詞体系が異なるものの，今後様々な解析用辞書が構築され, 利用者のニーズに合わせて選択できるような形になるのが望ましいと思われる.

（4）今後の展望

今後も引き続き辞書の項目を追加し，学習データを増やして辞書の構築を行う．本研究で開発したツールでの利用だけでなく，韓国語の計量的研究などを行う際には，何をおいても解析率が高くなければならないと考える．研究代表者のこれまでの取り組みの中で, 辞書に登録された項目の数が解析率に影響を及ぼすことが分かっている．この点も踏まえ，辞書の規模を拡大していくことが重要である.

また，規模の拡大という点では，インターネットのブログや SNS などで用いられている表現や語彙を追加することも必要であろう．研究代表者は既にインターネットの表現について研究を行ったことがある．こうした蓄積も活かし，また日本語の解析などに関する業績も参考にして，どのように項目を追加していくか，改めて検討したい.

なお，現代語以外への応用も検討中である. 既に 15 世紀の韓国語を解析するための辞書について構築を試みており，その上に本研究での知見を活かすことができるだろう．以前の取り組みで改善すべき点など，改めて検討し，韓国語の通時的研究に利用することができる辞書の構築に取り組む予定である.

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

①須賀井義教,「自動形態素解析技術を利用した韓国語読解補助ツールの開発―日本語母語話者のための機能を中心に―」（原文韓国語）,『韓国語教育』第 24 巻 3 号，ソウル：国際韓国語教育学会, pp.139-159, 2013 年, 査読有

〔学会発表〕（計１件）

①須賀井義教,「韓国語自動形態素解析辞書の開発」（原文・発表言語は韓国語），韓国辞書学会・第 26 次全国学術大会，2015 年 2 月 25 日，NAVER Green Factory（大韓民国京畿道城南市）にて発表

〔その他〕

ホームページ等

①「MeCab で韓国語」
http://porocise.sakura.ne.jp/wiki/korean/mecab

②「MeCab を利用した韓国語自動形態素解析」（原文韓国語）
http://porocise.sakura.ne.jp/wiki/korean/mecab.ko

③「MeCab を利用した韓国語読解補助ツー

ル」
http://porocise.sakura.ne.jp/korean/mecab/main.html
④「HanDic プロジェクト日本語トップページ」
http://sourceforge.jp/projects/handic/

６．研究組織

(1)研究代表者

須賀井　義教（SUGAI Yoshinori）
近畿大学・総合社会学部・准教授
研究者番号：60456461