様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月　 ８ 日現在

機関番号：３４４１９
研究種目：基盤研究(C)（一般）
研究期間：2017～2020
課題番号：１７Ｋ０２７５８

研究課題名（和文）自動形態素解析を利用した15世紀朝鮮語解析済みコーパスの構築

研究課題名（英文）Development of tagged Middle Korean corpus using 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）　　2,100,000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15世紀朝鮮語を解析するための辞書を構築し，代表的な文献である『釈譜詳節』に加え，仏教経典の翻訳である『阿弥陀経諺解』『金剛経諺解』，教化書である『三綱行実図諺解』の形態素解析を行った．構築した解析用辞書は，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．
また，上記の解析済みデータを利用して，15世紀朝鮮語の計量的な分析を行った．

研究成果の学術的意義や社会的意義
本研究で構築した形態素解析用辞書をオープンソースで公開することにより，朝鮮語情報処理の質的向上に寄与することができると考える．解析用辞書の公開は，朝鮮語のみならず他の言語についても同様の試みを行うことが可能であることを示しており，様々な言語の自然言語処理技術に貢献することが見込まれる．
また形態素解析済みデータを用いた朝鮮語史の記述を試みることで，計量的な手法による朝鮮語史研究に新たな展開をもたらし，従来の知見の補充や刷新を行うことが可能となることが期待できる．

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 middle Korean language. For this research, we built a morphological analysis-use dictionary from approximately 9,000 registered items and analyzed representative documents such as "Seokbosangjeol", "Amitagyeong Eonhae", etc. The dictionary that we built is available to the public on the internet as open source software.

研究分野：言語学

キーワード：朝鮮語史　形態素解析　コーパス　計量的分析

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

15 世紀半ばのハングル創製によって，それまで固有の文字を持たず，漢字を用いて表記していた朝鮮語の姿を完全な形で表すことができるようになった．この点で，ハングル創製以後の朝鮮語文献が朝鮮語史の研究において持つ意義は大きい．15 世紀以降，現在までの朝鮮語の変遷を知る上では出発点であり，また漢字の音訓を借りて表記された，それ以前の朝鮮語の姿を類推する際の起点ともなるためである．

こうした朝鮮語の歴史的文献に関する電子データの整備状況を見てみると，テキストファイル（平文コーパス）はある程度の分量があるものの，15 世紀の朝鮮語文献について形態素解析を行い，タグ付けを施したコーパスは未だ構築が途上にある．洪允杓によれば，15 世紀資料の平文コーパスは 93 万文節程度が作成されているが，解析済みコーパスは 20 万文節程度しかなく，平文コーパスの半分にも満たないという（洪允杓 2006）．また，解析の誤りなども散見される．計量的研究への応用なども視野に入れた，形態素解析済みデータの整備が求められる．

２．研究の目的

本研究は，ハングルが創製された 15 世紀の朝鮮語文献について形態素解析を行い，形態素情報が付与された解析済みコーパスを構築し，公開することが目的である．形態素解析の際には既に須賀井義教（2013，2016）などで利用した解析エンジン「MeCab（めかぶ）」を用いる．

本研究では 15 世紀朝鮮語資料の代表的な文献である『釈譜詳節』（1447 年刊，約 3 万文節）などの形態素解析を行う．解析の際に用いた解析用辞書はアップデートを継続して行い，オープンソースソフトウェアとして公開する．

また，構築した解析済みデータなどの応用として，計量的なアプローチからの朝鮮語研究を試みる．

本研究を通じて公開するコーパス，解析用辞書などにより，朝鮮語情報処理の質的向上および朝鮮語史研究の新たな展開に寄与することを目指す．

３．研究の方法

本研究における基本的な作業は，①MeCab で形態素解析を行うための解析用辞書を構築し，②実際に文献の形態素解析を行う，という 2 点に集約される．さらに③解析済みデータを活用した朝鮮語研究が，これらの作業の応用として行われる．

①解析用辞書の構築

既に須賀井義教（2016，2017）でプロトタイプとして構築した解析用辞書を元に，新たな項目を追加する．主に既存の辞書（『李朝語辞典』，『ウリマルクンサジョン』など）から項目を選択し，辞書項目として登録する．

次に，既存の解析済みデータを学習用データとして利用し，項目を追加した辞書の学習及び辞書構築を MeCab で行う．

②15 世紀朝鮮語文献の形態素解析

①で構築した辞書を使って，新たな文献の形態素解析を行う．解析の結果は手作業で修正し，辞書に未登録の項目は新たな辞書データとして追加する．これらの作業を通じて得られた辞書データを元に，修正済みの解析結果を学習データに追加して，再度①の作業を行う．

③解析済みデータを利用した朝鮮語研究の試み

上記の作業で得られた解析済みデータを元に，計量的な記述を試みる．コーパスを用いた 15 世紀朝鮮語の記述は多く行われているが，形態素解析済みのデータを用いた計量的な研究はあまり見られない．朝鮮語の計量的な分析は，主に著者判別や文体的特性の探求といった方面で進められており（ハン・ナレ 2009，カン・ナムジュンほか 2010 など），その分析においては共通して助詞や語尾といった機能語が多く利用されている．本研究でもこうした機能語に着目して，計量的分析を試みる．

４．研究成果

(1)MeCab 用解析辞書

本研究では，学習用データ 2653 文を用いて 9185 項目を含む 15 世紀朝鮮語の MeCab 用形態素解析辞書を構築した．『金剛経諺解』（1464 年刊）解義部分の冒頭 50 文をテストデータとして用い，性能評価を行ったところ，表層型（LEVEL 6）までの解析率（F 値）は 97.65%であった．また形態素境界の判定（LEVEL 0）に限って言えば 98.05%と，非常に高い性能を見せている．テストデータに含まれている形態素項目を全て辞書に含んでいるということが理由の一つであろうが，学習用データには当該部分を含めていない．少ない学習用データでも効率よく学習を行う，解析エンジン MeCab のメリットがよく生かされた結果と言える．

形態素解析用辞書については，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を行った．

(2)形態素解析済みデータ

上記の辞書を用いて，以下の文献の解析および解析結果の修正を行った：

- 『釈譜詳節』（1447 年）：巻 6，9，13，19，23，24
- 『阿弥陀経諺解』（1464 年）
- 『金剛経諺解』（1464 年）：本文と割注のみ
- 『三綱行実図諺解』（1481 年？）

いずれも解析結果は公開する予定であるが, どのような方法がよいか, 現在も検討中である．今後の課題である．

(3)朝鮮語研究への応用

上記の解析済みデータの一部を用いて，計量的な分析を試みた．特に接続形語尾の分布を元に，多変量解析の手法を用いて文献の分類を試みた．

15 世紀朝鮮語資料の多くは仏教経典の翻訳であるが,『釈譜詳節』は釈迦の一代記である『釈迦譜』を翻訳した巻・部分と,『法華経』などを翻訳した巻・部分から構成されている．また, 志部昭平(1990)は『三綱行実図諺解』の文体的特徴を「いわば『説話体』とも言うべきもの」とし,「漢文の直訳である『諺解体』」と比べると，特定の接続形語尾が多く用いられると指摘している．その上で,『三綱行実図諺解』の「説話体」的な特徴は『釈譜詳節』の文体，特に仏教説話を翻訳した部分に似ているようだ，と述べている．

このような指摘は,『釈譜詳節』の内部でも文体的特徴が異なる可能性を示唆している．そこで,『釈譜詳節』・『三綱行実図諺解』と,「諺解体」の例として『金剛経諺解』も加え，接続形語尾の出現頻度を利用して，クラスター分析を行った．その結果,『釈迦譜』を翻訳した『釈譜詳節』巻 6・23・24 および『三綱行実図諺解』が一つのクラスターを成し，仏典の翻訳である『釈譜詳節』巻 9・13・19 および『金剛経諺解』がもう一つのクラスターを成していることが確認された．各クラスターでは接続形語尾の出現頻度の平均に違いが見られ，特に一部の語尾については志部昭平(1990)の指摘を支持する結果となった．こうした試みを通じて，15 世紀朝鮮語資料について計量的な分析が可能であることが示された．

〔引用文献〕

志部昭平(1990)『諺解三綱行実図研究』，東京：汲古書院

須賀井義教(2013)「MeCab（めかぶ）を用いた現代韓国語の形態素解析」，朝鮮語研究会編『朝鮮語研究』5，pp.283-312.

須賀井義教(2016)「中期朝鮮語形態素解析用辞書の開発」（口頭発表），朝鮮語教育学会・朝鮮語研究会　第 5 回合同大会，2016 年 9 月 10 日，東京大学駒場キャンパス.

須賀井義教(2017)「中期朝鮮語形態素解析用辞書の開発」，須川英徳編『韓国・朝鮮史への新たな視座』，東京：勉誠出版，pp.315-333.

カン・ナムジュンほか(2010)「『独立新聞』論説の形態注釈コーパスを活用した論説著者の判別研究—語尾の使用頻度分析を中心に—」,『韓国辞書学』第 15 号, 韓国辞書学会, pp.73-101（原文は韓国語）.

ハン・ナレ(2009)「頻度情報を利用した韓国語の著者判別」,『認知科学』第 20 巻第 2 号，韓国認知科学会，pp.225-241（原文は韓国語）.

洪允杓(2006)「国語史研究のための電子資料構築の現況等課題」,「国語史研究，どこまで来ているか」，太学社（原文は韓国語）.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 須賀井　義教
２．発表標題 中世韓国語文献の自動形態素解析と計量的研究：分析結果の活用について（原題は韓国語）
３．学会等名 第一届多元文化研究与跨文化教育国際研討会（国際学会）
４．発表年 2019年

１．発表者名 須賀井　義教
２．発表標題 中期朝鮮語の計量的分析の試み―クラスター分析による『釈譜詳節』各巻の分類―
３．学会等名 第266回朝鮮語研究会
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

韓国語形態素解析用辞書　HanDic（パッケージ：MkHanDic） https://ja.osdn.net/pkg/handic/mkhandic-mecab MeCabで韓国語 https://porocise.sakura.ne.jp/wiki/korean/mecab

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	中島　仁 (NAKAJIMA Hitoshi) (40439708)	東海大学・国際教育センター・准教授	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関