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　６ 年　 ６ 月 １２ 日現在

機関番号：３４４１９
研究種目：基盤研究(C)（一般）
研究期間：2020～2023
課題番号：２０Ｋ００５５８

研究課題名（和文）計量的分析のための15世紀朝鮮語形態素解析済みコーパス構築：仏教諺解を対象に

研究課題名（英文）Building a Morphologically Analyzed Corpus of 15th Century Korean for Quantitative Analysis

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『月印釈譜」の形態素解析を行った．構築した解析用辞書は約1万項目が収録されており，この辞書データ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た．さらに，15世紀朝鮮語文献の電子データ化をTEIガイドラインを用いて行った．
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
本研究で構築したデータをオープンソースソフトウェアとして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期待される．解析用辞書構築の手法については，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実践することで，計量的な手法による15世紀朝鮮語研究の実例を示すこととなり，併せて従来の知見の補充や刷新を行うことが可能となる．

研究成果の概要（英文）：For this research project, a dictionary for analyzing 15th century Korean was constructed using the open source morphological analysis engine "MeCab", and morphological analysis of "Wolin-seokbo" was conducted. The lexicon contains approximately 10,000 entries, and was released on the Internet as open source software. In addition, the 15th century Korean documents were digitized using the TEI guidelines.
Using the above analyzed data, a quantitative analysis of the 15th century Korean language was conducted.

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　TEIガイドライン

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．

本研究の研究代表者は上記形態素解析済みデータを構築すべく，科研費を取得して作業を行った．本研究では，未解析の文献について作業に着手し，データの構築を行うものである．

２．研究の目的

本研究は 15 世紀の朝鮮語文献，特に仏教諺解について形態素解析を行い，コーパスとして構築することを目的とする．解析済みコーパス構築においては，形態素解析エンジン「MeCab」（めかぶ）を用いた自動形態素解析を行う．本研究では 15 世紀朝鮮語文献のうち，仏教諺解の『月印釈譜』（1459 年刊，約 15 万文節）および『法華経諺解』（1463 年刊，約 10 万文節）をコーパス化の対象とする．この過程で作成されるコーパスデータや解析用辞書については，広く今後の朝鮮語史研究に資するべく，インターネットを通じて公開する．本研究を通じて公開されるデータや検索ツールなどによって，朝鮮語情報処理の質的向上ならびに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

本研究では，①MeCab で形態素解析を行うための解析用辞書を構築し，②その辞書を用いて実際に文献の形態素解析を行う．また，③解析結果を修正したデータをコーパスとして朝鮮語研究に活用する．

①MeCab で解析を行うための解析用辞書構築

既に須賀井義教（2016，2017）でプロトタイプを構築しており，本研究ではさらに辞書項目と学習用データを拡充する．

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，新たな文献について形態素解析を行い，その解析結果をチェックして，誤りを修正し，未登録の項目は辞書に追加し，修正した解析結果を学習用データとして追加して，再び①の作業＝解析用辞書を構築する，というサイクルで行う．結果として辞書が拡充され，解析精度も上がることが期待される．

③解析済みデータを利用した朝鮮語研究への応用

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解析用辞書の性能評価

本研究では，ここまで蓄積した学習用データ 3080 文を用いて，10023 項目を含む 15 世紀朝鮮語の形態素解析用辞書を構築した．学習用データには含めていない『金剛経諺解』（1464 年刊）解義部分の冒頭 50 文を用いて性能評価を行ったところ，表層形（LEVEL6）までの解析率（F 値）は約 94.46%であった．また，形態素境界の判定（LEVEL 0）については 99.27%，品詞 3 までの判定（LEVEL 3）については 98.46%と，形態素を抽出して品詞付与を行う，といった作業については，非常に高い性能を見せた．

解析誤りの傾向としては，同音異義語の判定ミス，母音語幹の用言や〈尊敬〉を表す接尾辞「Ⅱ-si-」などの活用語基（菅野裕臣 1997）における判定ミス，等が挙げられる．こうした解析の誤りについて，どのように解決するか，今後検討する予定である．

なお，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 年刊）：巻 20，21 の本文および注釈

・『月印釈譜』（1459 年刊）：巻 1，2，7 の本文，巻 7 の注釈

これらの解析結果については，コーパスとして今後公開する予定である．

なお，当初研究の計画に含めていた『法華経諺解』については，新型コロナウイルス感染症の流行などにより作業者を確保することが難しく，作業を進めることができなかった．今後の課題としたい．

③朝鮮語研究への応用

上記の解析結果を活用して，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．

須賀井義教（2022）では，仏教経典の翻訳である『釈譜詳節』のうち，釈迦の一代記である『釈迦譜』を翻訳した巻・部分と，『法華経』などの仏典を翻訳した巻・部分とについて，その文体的特徴を抽出すべく，接続形語尾の出現頻度を利用して，クラスター分析および対応分析を行った．志部昭平（1990）によれば，15 世紀の朝鮮語文献である『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」として，「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，とも述べている．

このような指摘を参考に，『釈譜詳節』に加え，「説話体」の例として『三綱行実図諺解』を，「諺解体」の例として『金剛経諺解』も追加して分析を行った．クラスター分析の結果，『釈迦譜』を翻訳した『釈譜詳節』巻 6・23・24 と『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 及び『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．また，対応分析の結果としては，文献ごとに接続形語尾の分布に特徴が見られ，特に一部の語尾については志部昭平（1990）の指摘を指示する結果となった．

④TEI ガイドラインを用いた文献の電子データ化

本研究では，解析に用いる文献の電子データを整備する中で，一定の規格に沿った電子化の必要性を痛感し，TEI（Text Encoding Initiative）ガイドラインを用いた電子データの作成も行った．TEI ガイドラインを用いることで，本文と注釈の区別や，文献情報，原典の情報などといったメタデータを盛り込むことができるため，データ抽出の際の便宜を向上させることができる．本研究では段落構造などの情報までしか入力することができなかったが，今後人物名などをマークすることで，電子データの新たな応用が可能になると期待できる．

本研究では，最終的に『釈譜詳節』原刊本の全て，『月印釈譜』巻 1・2・7・8・9・10，『阿弥陀経諺解』（1464 年刊），『金剛経諺解』，『三綱行実図諺解』のデータ整備が完了した．

〔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(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.

須賀井義教(2022)「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―」，『朝鮮語研究』9，東京：朝鮮語研究会，pp.175-207.

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1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 須賀井 義教	４．巻 9
２．論文標題 中期朝鮮語の計量的分析の試み	５．発行年 2022年
３．雑誌名 朝鮮語研究	６．最初と最後の頁 175～207
掲載論文のDOI（デジタルオブジェクト識別子） 10.50986/koreanlinguistics.9.0_175	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

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 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〕

MeCab用形態素解析辞書MkHanDic
https://ja.osdn.net/pkg/handic/mkhandic-mecab
MeCabで韓国語
https://porocise.sakura.ne.jp/wiki/korean/mecab
MeCab用形態素解析辞書MkHanDic
https://ja.osdn.net/pkg/handic/mkhandic-mecab
MeCabで韓国語
https://porocise.sakura.ne.jp/wiki/korean/mecab
MeCab用形態素解析辞書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

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