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　５ 年　 ６ 月　 １ 日現在

機関番号：１２６０１
研究種目：若手研究
研究期間：2020～2022
課題番号：２０Ｋ１３０１２

研究課題名（和文）日本語レキシコンプロジェクトの開発と評価

研究課題名（英文）The Japanese Lexicon Project

研究代表者
大関　洋平（OSEKI, Yohei）

東京大学・大学院総合文化研究科・講師

研究者番号：１０８２１９９４

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：本研究では、日本語レキシコンに関する語彙統計・形態構造・行動実験データを統合した語彙データベース「日本語レキシコンプロジェクト」を開発し、先行研究における理論・実験形態論の研究成果に基づいて評価した。また、開発した語彙データベースは、オープンアクセスとして一般に公開する予定であり、形態論・レキシコン研究のみならず、心理言語学・応用言語学・自然言語処理・リハビリテーションなど広範な隣接分野に対する学術的・社会的インパクトが期待される。

研究成果の学術的意義や社会的意義
心理言語学：実験研究において、刺激の統制に必要な語彙データベースが20年前かつ入手困難な『日本語の語彙特性』に制限されているという現状を打破できる。応用言語学：日本語教育において、客観的な語彙統計・行動実験データに基づいた初級から上級までの語彙サイズを網羅する新たな学習教材を開発できる。自然言語処理：形態素解析において、『分類語彙表』と『形態素解析辞書UniDic』の対応表を利用することで語彙統計・形態構造データを自動付与できる。リハビリテーション：言語聴覚療法において、医師および言語聴覚師の直感・経験に基づいて実施されてきた言語検査を標準化できる。

研究成果の概要（英文）：In this study, we developed a lexical database called the "Japanese Lexicon Project" by integrating lexical statistics, morphological structures, and behavioral experimental data related to the Japanese lexicon. We evaluated the database based on the theoretical and experimental morphological research findings from previous studies. Additionally, the developed lexical database is planned to be publicly available as open access, and it is expected to have academic and societal impact on a wide range of adjacent fields, including morphology, lexicon research, psycholinguistics, applied linguistics, natural language processing, and rehabilitation.

研究分野：言語学

キーワード：形態論　レキシコン　心理言語学　自然言語処理

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

人間言語のレキシコンは、「語」という言語単位を脳内に記憶する「心的辞書」であり、過去に様々な語彙データベースが開発されてきた。例えば、英語の単語親密度・心像性など語彙統計データを収録した『MRC 心理言語データベース』、英語・ドイツ語・オランダ語の形態構造データを収録した『CELEX 語彙データベース』がある。また、語彙判断課題（呈示された文字列が当該言語の単語であるか判断する課題）における反応時間や精度など行動実験データを収録した『英語レキシコンプロジェクト』を皮切りに、フランス語・オランダ語・中国語・マレー語に拡張され、注目を集めている。

しかしながら、既存の語彙データベースでは、語彙統計・形態構造・行動実験データが分散しており、日本語に至っては、形態論・レキシコン研究の長い蓄積があるにも関わらず、ほとんど語彙データベースが存在しない。日本語レキシコンに関する唯一の語彙データベースである『日本語の語彙特性』も、語彙統計データしか収録しておらず、出版から２０年が経過しており古く入手困難である。従って、本研究課題の核心には、日本語レキシコンが対象かつ語彙統計・形態構造・行動実験データを統合した語彙データベースを開発できるか、という学術的「問い」がある。

２．研究の目的

そこで、本研究では、日本語レキシコンに関する語彙統計・形態構造・行動実験データを統合した語彙データベース「日本語レキシコンプロジェクト」(Japanese Lexicon Project)を開発し、先行研究における理論・実験形態論の研究成果に基づいて評価することを目的とする。具体的には、日本語レキシコンプロジェクトの「開発」と「評価」を対象とする２つのフェイズを通して、各フェイズ４つの課題を達成することを目指す：

◇「開発」フェイズ：日本語レキシコンプロジェクトの開発
　開発１：国立国語研究所の『分類語彙表－増補改訂版－』から、収録語彙の選定。
　開発２：国立国語研究所の『日本語話し言葉コーパス』から、語彙統計データの計算。
　開発３：大規模な形態構造アノテーションによる、形態構造データの付与。
　開発４：大規模な反復測定デザイン語彙判断実験による、行動実験データの計測。
◇「評価」フェイズ：日本語レキシコンプロジェクトの評価
　評価１：開発した語彙データベースに基づく、理論形態論の研究成果の再現。
　評価２：開発した語彙データベースに基づく、実験形態論の研究成果の再現。
　評価３：国内学会「日本言語学会」ワークショップで、語彙データベースの公開。
　評価４：国際ジャーナル Behavior Research Methods で、語彙データベースの出版。

理論言語学では、マサチューセッツ工科大学の Ken Hale 氏による「MIT レキシコンプロジェクト」を先駆けとして、日本では国立国語研究所の影山太郎氏による「関西レキシコンプロジェクト」および「形態論・レキシコン研究会」において、理論的な研究成果が蓄積されてきた。一方、心理言語学では、大規模な行動実験データを収録した『レキシコンプロジェクト』を中心として、実験的な研究成果が発展してきた。本研究は、理論的な研究成果を、心理言語学で開発されてきた語彙データベースに収録することによって、理論言語学と心理言語学の融合を試みている点で、学術的独自性を備えている。また、本研究は、『レキシコンプロジェクト』を日本語に拡張することで、日本語における語彙データベース不足を解消するだけで無く、既存の語彙データベースでは散逸していた語彙統計・形態構造・行動実験データを集積した、世界に類を見ない語彙データベースの開発を目指している点で、極めて創造性が高い。加えて、既存の語彙データベースは、欧米言語における単音節・単形態素語彙に偏重していたが、本研究では、形態論・レキシコン研究に資するため、アジア言語における多音節・多形態素語彙に着目している点も独創的である。

また、開発した語彙データベースは、オープンアクセスとして一般に公開する予定であり、形態論・レキシコン研究のみならず、心理言語学・応用言語学・自然言語処理・リハビリテーションなど広範な隣接分野に対する学術的・社会的インパクトが期待される：

◇心理言語学：実験研究において、刺激の統制に必要な語彙データベースが 20 年前かつ入手困難な『日本語の語彙特性』に制限されているという現状を打破できる。
◇応用言語学：日本語教育において、客観的な語彙統計・行動実験データに基づいた初級から上級までの語彙サイズを網羅する新たな学習教材を開発できる。
◇自然言語処理：形態素解析において、『分類語彙表』と『形態素解析辞書 UniDic』の対応表を利用することで語彙統計・形態構造データを自動付与できる。
◇リハビリテーション：言語聴覚療法において、医師および言語聴覚師の直感・経験に基づいて実施されてきた言語検査を標準化できる。

３．研究の方法

以上の研究目的を達成するため、本研究は、３カ年の研究計画とし、令和２-３年度の「開発」フェイズと令和４年度の「評価」フェイズを設定する。尚、各フェイズにおいて浅原正幸氏（国立国語研究所）および伊藤たかね氏（東京大学）から助言を受ける。

◇令和２-３年度の計画：「開発」フェイズ

【開発１：収録語彙の選定】国立国語研究所の『分類語彙表－増補改訂版－』から、日本語レキシコンプロジェクトに収録する約４万語を選定する。既存の語彙データベースでは、高頻度・単形態素の語彙に偏っていたが、低頻度・多形態素の語彙を含めつつ、「体の類」・「用の類」・「相の類」から名詞・動詞・形容詞・副詞を約１万語ずつ収録する。

【開発２：語彙統計データの計算】NTT コミュニケーション科学基礎研究所の『日本語の語彙特性』から、文字・音節・形態素数、書記・単語頻度、形態・音韻類似度など語彙統計データを計算する。単語頻度は、国立国語研究所の『日本語話し言葉コーパス』および『現代日本語書き言葉均衡コーパス』、形態・音韻類似度は Coltheart 数および Levenshtein 距離に基づいて計算する。更に、親密度（浅原，2019)と心像性（浅原，準備中)も統合する。

【開発３：形態構造データの付与】関西レキシコンプロジェクトの『意味と構文』シリーズから、形態構造・統語範疇など形態構造データを収集する。形態構造は、『統語・意味解析コーパス』のアノテーション方針に則り、「体の類」・「用の類」・「相の類」に対応した『動詞の意味と構文』(影山，2001)・『名詞の意味と構文』(影山，2011)・『形容詞・副詞の意味と構文』(影山，2009)を参照しつつ、大学院生のリサーチアシスタント数名が実施する。

【開発４：行動実験データの計測】反復測定デザインの大規模な語彙判断実験を実施し、反応時間・精度など行動実験データを計測する。非単語刺激は、擬似単語生成ソフトウェア「Wuggy」(Keuleers et al.，2010)を利用し、対応する単語刺激から作成する。日本語の母語話者８０人を実験参加者とし、約８万語の語彙判断に対する報酬として２万円を支払う。

◇令和４年度の計画：「評価」フェイズ

【評価１・２：理論・実験形態論の再現】開発した語彙データベースに基づき、先行研究における理論・実験形態論の研究成果を再現する。理論的には、影山太郎氏による『構文と意味シリーズ』を、実験的には、玉岡賀津雄氏による語彙判断実験を対象とする。

【評価３・４：語彙データベースの公開・出版】再現結果および日本語レキシコンプロジェクトの概要を周知・宣伝するため、国内学会「日本言語学会」第 165 回秋季大会で、ワークショップを実施する。また、様々な言語の「レキシコンプロジェクト」が出版されている国際ジャーナル Behavior Research Methods に論文を投稿し、研究成果を出版する。

４．研究成果

2020 年度は、日本語レキシコンプロジェクトの開発に向けて、基礎研究を実施した。まず、国立国語研究所の「複合動詞レキシコン」と「使役交替言語地図」を Python プログラムで解析し、Morphology & Lexicon Forum (MLF) 2020 で発表「Compound verbs in transitivity harmony and alternation」を行った。また、 Morphology & Lexicon Forum (MLF) 2019 で発表した日本語の他動性交替に関する研究を論文「分散形態論と日本語の他動性交替」としてまとめ、『レキシコン 研究の現代的課題』(くろしお出版)に出版した。更に、研究室の大学院生 6 名も研究協力者として各自の専門領域で基礎研究を実施し、研究代表者の論文「形 態論の計算認知神経科学に向けて」と併せて、計 7 本の論文を『言語研究の楽しさと楽しみー伊藤たかね先生退職記念論文集ー』(開拓社)に出版した。加え て、日本英語学会第 36 回大会で企画したシンポジウム「言語理論における形態論の「分散」をめぐる諸問題」に基づき、論文集『形態論と統語論のインターフェ イス(仮)』(開拓社)を出版する予定である。最後に、David Crystal (2019) The Cambridge Encyclopedia of the English Language のレキシコンに関する 章「第 12 章 レキシコンの諸側面」を翻訳し、『ケンブリッジ英語百科事典』(朝倉書店)として出版する予定である。

2021 年度は、2020 年度に実施した基礎研究に基づき、日本語レキシコンプロジェクトの開発に着手した。まず、国立国語研究所の「分類語彙表」を Python プログ ラムで解析し、日本語レキシコンプロジェクトの元データを作成した。また、研究室の大学院生をリサーチ・アシスタントとして雇用し、各自の専門領域に応じ て「動詞班」、「形容詞班」、「名詞班」に割り振った上で、影山太郎氏の『意味と構文』シリーズ(大修館書店)をデータ化した。加えて、日本英語学会第 36 回大会で企画したシンポジウム「言語理論における形態論の「分散」をめぐる諸問題」に基づく論文集『形態論と言語理論(仮)』(開拓社)および David Crystal (2019) The Cambridge Encyclopedia of the English Language のレキシコンに関する章「第 12 章 レキシコンの諸側面」を翻訳した『ケンブリッジ英語 百科事典』(朝倉書店)を出版する予定である。

2022 年度は、2021 年度に引き続き、日本語レキシコンプロジェクトの開発を実施した。まず、国立国語研究所の「分類語彙表」を Python プログラムで解析し、日 本語レキシコンプロジェクトの元データを作成した。また、研究室の大学院生をリサーチ・アシスタントとして雇用し、各自の専門領域に応じて「動詞班」、 「形容詞班」、「名詞班」に割り振った上で、影山太郎氏の『意

味と構文』シリーズ(大修館書店)をデータ化した。加えて、2020 年度に実施した形容詞に関する基礎研究の成果を国際会議 Japanese/Korean Linguistics Conference 30 で発表し、日本英語学会第 36 回大会で企画したシンポジウム「言語理論における形態 論の「分散」をめぐる諸問題」に基づく論文集『生成形態論の新展開(仮)』(開拓社)および David Crystal (2019) The Cambridge Encyclopedia of the English Language のレキシコンに関する章「第 12 章 レキシコンの諸側面」を翻訳した『ケンブリッジ英語百科事典』(朝倉書店)を出版する予定である。

以上

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 大関洋平
２．発表標題 Compound verbs in transitivity harmony and alternation
３．学会等名 Morphology & Lexicon Forum (MLF)
４．発表年 2020年

１．発表者名 Morine Kondo, Takane Ito, Yohei Oseki
２．発表標題 Morphological Structures of Japanese Adjectival Compounds
３．学会等名 Japanese/Korean Linguistics Conference 30（国際学会）
４．発表年 2023年

〔図書〕　計4件

１．著者名 大関　洋平、漆原　朗子	４．発行年 2023年
２．出版社 開拓社	５．総ページ数 250
３．書名 生成形態論の新展開	

１．著者名 中島　平三、田子内　健介	４．発行年 2023年
２．出版社 朝倉書店	５．総ページ数 600
３．書名 ケンブリッジ英語百科事典	

1．著者名 岸本 秀樹	4．発行年 2021年
2．出版社 くろしお出版	5．総ページ数 240
3．書名 レキシコン研究の現代的課題	

1．著者名 岡部 玲子、矢島 純、窪田 悠介、磯野 達也	4．発行年 2021年
2．出版社 開拓社	5．総ページ数 536
3．書名 言語研究の楽しさと楽しみ	

〔産業財産権〕

〔その他〕

―

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	伊藤　たかね (ITO Takane)		
研究協力者	浅原　正幸 (ASAHARA Masayuki)		

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