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　２ 年　 ６ 月 １８ 日現在

機関番号：６２６１８
研究種目：挑戦的研究（萌芽）
研究期間：2017～2019
課題番号：１７Ｋ１８５０５

研究課題名（和文）日本語研究用オントロジーの設計と開発

研究課題名（英文）Design and Development of the Ontology for Japanase Linguistics

研究代表者
山崎　誠（Yamazaki, Makoto）

大学共同利用機関法人人間文化研究機構国立国語研究所・言語変化研究領域・教授

研究者番号：３０１８２４８９

交付決定額（研究期間全体）：（直接経費）　　4,900,000円

研究成果の概要（和文）：日本語学の文法分野の文献に付されたキーワードのうち、頻度2以上の約3900語について、上位下位関係による分類を第3階層まで行った。言語に関する二次情報のデータベース化については、(1)『二十巻本和名類聚抄』、(2)『言語地図画像データベース』に含まれる地図名、(3)大正期の『読売新聞』及び大正から昭和の『文藝春秋』における言語について書かれた記事をそれぞれデータベース化した。それらのデータベースと「日本語歴史コーパス」「現代日本語書き言葉均衡コーパス」から得た統計情報とを併せて検索できるようにした語誌情報ポータルのサイトを2019年3月に内部公開した。

研究成果の学術的意義や社会的意義
学術用語の整理は当該の分野がどのような広がりを見せているか現状を把握し、かつ、関連分野との関係を知る上で重要である。分野間融合の可能性が生じるきっかけになる可能性もある。また、言語に関する二次情報のデータベース化は、これまで分散して検索されていた二次情報を一か所で検索できるようにしたことで、研究の効率化につながり、また、コーパスからの頻度情報を可視化することで、日本語史に興味を持ってもらえるような工夫がなされた。

研究成果の概要（英文）：Of the keywords annotated to the literatures in the grammar field of Japanese language, about 3900 words with a frequency of 2 or more were classified up to the third hierarchy according to the super-subrelations. Regarding the databases of secondary information related to language, we made a database consisting of (1)entries of "Nijukanbon Wamyo Ruijusyo", (2) map names included in "Language Map Image Database", (3)articles written about language in "Yomiuri Shimbun" in the Taisho era, and "Bungei Shunju" from Taisho to Showa. In March 2019, the site of a Goshi Joho Portal (bibliographic information portal) that enables users to search both those databases and statistical information obtained from the "Japanese History Corpus" and "Balanced Corpus of Contemporary Written Japanese" was released internally.

研究分野：日本語学

キーワード：学術用語　概念体系　オントロジー　古辞書　言語地図　言語記事

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

21 世紀に入り、日本語研究において言語コーパスが盛んに利用されるようになってきた。例えば、国立国語研究所で構築・公開が進められている『日本語話し言葉コーパス』（2004 年公開）、『現代日本語書き言葉均衡コーパス』（2011 年公開）、『日本語歴史コーパス』（2014 年から順次公開）などを用いた研究をよく目にするようになっている。

これらのコーパスに格納されているデータは、実際に使用された話し言葉・書き言葉であり、これらは一次資料と位置付けられる。一方、言語研究にとって重要な資料として、論文等の研究文献や言語を対象とした調査結果等があり、これらは二次資料と位置づけられる。これらの二次資料は、一部はデータベースないしは電子資料として利用されているが、多くは単体で存在し、有機的に関連付けられておらず、利用者にとっては個別に参照しなければならない（あるいは存在がよく知られていない）という、不便な状態に置かれている。

これらの二次情報の活用のためには、日本語研究に関する二次情報の全体像を俯瞰し、その中で個々の情報の価値や相互の関連性を明らかにするような枠組みがまず必要である。この枠組みは、近年盛んになっている、オントロジーの技術を利用することで適切に整理できるのではないかという着想に至った。

２．研究の目的

本研究では、学術用語のオントロジー化とそれに関連付ける日本語研究二次情報の集約を行うことを目的とする。具体的には以下の 2 点である。

（1）文法・語彙分野の学術用語をオントロジーとして整理する。

コーパス利用でニーズの高い、文法・語彙の分野を対象とする。通常の階層関係（is-a 関係）だけでなく、類義関係・対義関係なども付与する。また、必要に応じて新たな属性を作り、整理する。文法・語彙という分野は言語研究の中核的な部分に当たり、歴史言語学、社会言語学、認知言語学、計量言語学などで幅広く使われる。また、国語教育、日本語教育、自然言語処理、社会学、心理学等の関連領域でも用いられるため、研究領域を横断的に眺める要素として適している。

（2）日本語研究の二次情報を集約し、それらをオントロジーに関連付ける。

二次情報として重要なものとして、日本語を対象とした研究文献、日本語について書かれた記事（新聞、雑誌）、日本語に関する調査資料（言語地図や各種調査結果）、古辞書等がある。これらの資料を「語」を共通のキーとして持つものと学術用語を共通のキーとして持つものに分け、「語」に学術用語を付与することにより、すべての二次情報が学術用語に関連づけられるようにする。これにより、日本語研究に関する二次情報が学術用語に結びつくことになり、トップダウン方式で目的の概念を見つけることができるようになる。

３．研究の方法

（1）国立国語研究所の「日本語研究・日本語教育文献データベース」（約 22 万件）を利用して、文法・語彙分野に関する学術用語の抽出を行い、それらに対して出現頻度、出現期間、出現分野を付与し、かつ、概念の階層化、類義関係、関連する属性などの付与を行う。

（2）日本語学、言語学の用語事典、日本語学の教科書等を参考にして、学術用語の階層関係を記述する。属性付与の作業はマニュアル化し、データと合わせて公開し、将来の拡張に備えるようにする。

（3）すでに電子化されている古辞書、言語地図、日本語についての新聞・雑誌記事（言語記事）のデータを中心にして、学術用語との関連づけを行う。

（4）データ全体をオントロジーとして整理し、検索・閲覧ができるインターフェイスにより、ウェブ上で公開する。

４．研究成果

（1）分類体系の構築

日本語学の研究文献に用いられた学術用語を分類・体系化するために、まず、全体を以下の 23 個の下位分野（第一階層）を設けた。

「L01 音声・音韻、L02 文字・表記、L03 語彙・意味・辞書学、L04 文法、L05 語用論、L06 文章・談話、L07 文体、L08 方言（言語地理学）、L09 日本語史（歴史言語学）、L10 社会言語学・言語問題・言語政策、L11 対照言語学・言語類型論・翻訳、L12 自然言語処理・計算言語学、L13 コーパス言語学・計量言語学、L14 記号論・言語哲学、L15 認知言語学、L16 心理言語学・病理言語学、L17 法言語学、L18 国語教育・日本語教育、R01 研究資料、P01 人名、P02 地名・国名、P03 組織名、P04 その他の固有名詞」

さらに、「L04 文法」分野について、以下の第 2、第 3 階層を設けた。

第 2 階層	第 3 階層
論	統語論、形態論、学校文法、規範文法、記述文法、主語廃止論

構文	文型、命題、語順、係り受け、主題、修飾、呼応、順接、逆接、引用、指示、原因、理由、主語、述語、修飾語、視点、条件、譲歩、省略、伝聞、提題、文の成分、文節、主要部、修飾部、補語、必須補語、副次補語、共起、肯否、とりたて、前提、含意、南の分類
文	疑問文、現象文、判断文、否定文、複文、単文、命令文、存在文、ウナギ文、有題文、無題文、埋め込み文、受身文、使役文、指定文、措定文
節	主節、従属節、名詞節、副詞節、引用節、並列節、補足節、連体節
語	単語、語形、形態素、異形態、語構成、語形成
品詞	動詞、名詞、形容詞、副詞、連体詞、接続詞、助動詞、助詞、慣用句、オノマトペ、擬音語、擬態語、指示詞
活用	活用の型、活用形、語幹、語尾
格	ガ格、ヲ格、ニ格、デ格、ヘ格、カラ格、マデ格、文法格、意味格、表層格、深層格、必須格、随意格、主格、属格、所有格、対格、能格、主体、動作主、対象、相手、経験者、場所、着点、起点、経過域、手段、道具、起因、根拠、時、限界、領域、目的、様態、役割、割合
テンス（時制）	過去、現在、非過去、未来、タ形、ル形、テイル形、発話時、基準時
アスペクト（完了や継続）	完了、結果、継続、完成相、継続相、結果相、進行、結果の状態、状態動詞、継続動詞、瞬間動詞、第四種の動詞
ヴォイス（態）	能動受身、受動、受身、直接受身、間接受身、使役、可能、自発、授受、やりもらい、恩恵、ベネファクティブ
モダリティ（ムード）	対事的モダリティ、対人的モダリティ、概言、確言、願望、勧誘、可能、依頼、許可、禁止、当為
待遇表現・敬語	尊敬語、謙譲語、丁寧語、美化語
文体	普通体、丁寧体、ですます体、だ体、である体
その他	（個別の語、あるいは上記に該当しない非専門的な一般名詞等）

国立国語研究所の「日本語研究・日本語教育文献データベース」（224,207 件、提供を受けた当時）から、キーワードが付けられているの 70,140 件を抜き出した。これらのキーワードの異なりは約 35,500 語であった。この中から文法の分野に現れる、頻度 2 以上のキーワード約 3,900 語について上記第 3 階層までの関係づけを行った。

（2）二次情報のデータベース化

学術用語以外の二次情報のデータベース化については、『二十巻本和名類聚抄』に含まれる 5640 か所の和訓のうち、3,284 項目に対して、形態論情報（語彙素、語彙素読み、品詞、語種、語彙素 ID)を付与して検索できるようにした。これらの中には、「紅藍（くれのあい)」のような複合語を「くれ」「の」「あい」に分割したものを含む。言語地図については、「国立国語研究所言語地図データベース」のうち、GIS 画像を含む「言語地図画像データベース」から、19 種の地図集に含まれる地図名に対して形態論情報を付与した。項目数は 1,654 である（複合語を分割した場合を含む)。言語記事情報は、言葉について言及された新聞・冊子の記事のメタ情報であるが、新聞記事は、大正期の『読売新聞』（大正 2（1913）年、大正 6（1917））の記事に現れたもの 552 記事に付された 119 のキーワードに対して形態論情報を付与した。雑誌記事は『文藝春秋』の大正 12（1923）年 1 月（創刊号）～昭和 20（1945）年に含まれる 849 の記事に付けられた 471 のキーワードに対して形態論情報を付与した。

（3）文献データの公開

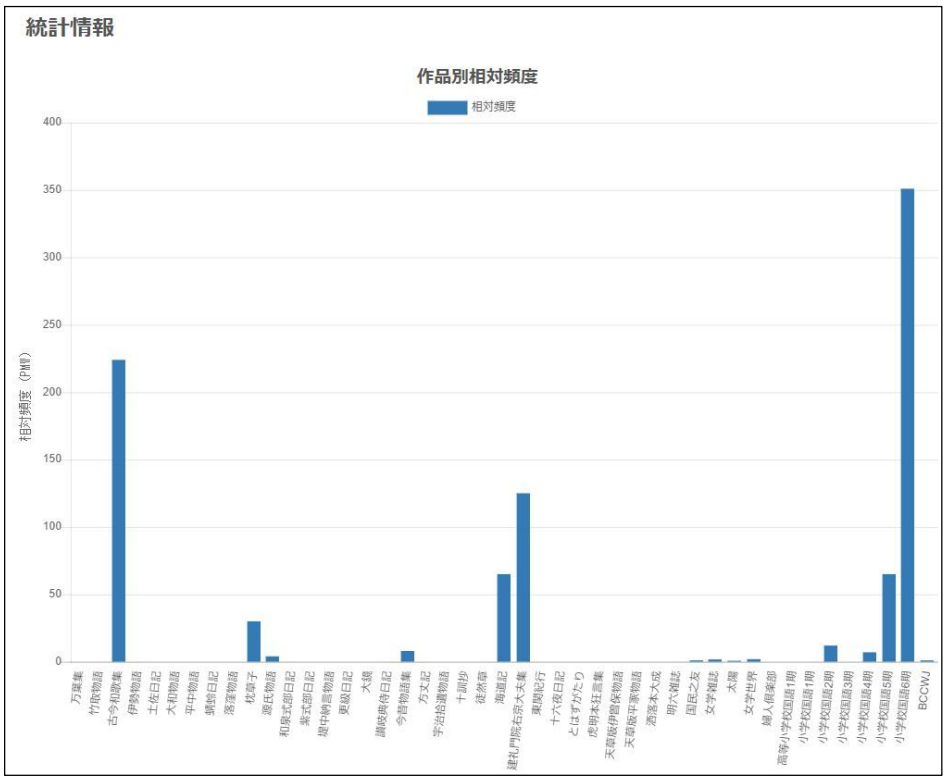
佐藤喜代治編『語彙研究文献語別目録』(1983 年、明治書院刊）を著作権継承者の承諾を得て、全文テキストとして公開した（https://textdb01.ninjal.ac.jp/gobetumokuroku/)。項目数は 17,300 である。さらに、データベース化のため、12,325 項目について形態論情報を付与した。

（4）語誌情報ポータルの構築

（2）で整備した各データと「日本語歴史コーパス（CHJ)」「現代日本語書き言葉均衡コーパス（BCCWJ)」から得た統計情報も検索できるようにした「語誌情報ポータル」のサイトを 2019 年 3 月に内部公開した。(https://goshidb.ninjal.ac.jp/goshidb/)

以下は「きりぎりす」を検索した場合の結果の一部である。まず、どの資料に何例現れるかが示される。

以下はコーパスにおける頻度情報である（各作品の相対頻度が示される）。



以下は、古辞書、言語記事、言語地図における検索結果である。古辞書を言語地図は、画像データへのリンクが示され、より具体的な情報が得られるようになっている。言語記事については、記事の簡単な内容が示される。

古辞書情報

辞書名	辞書見出し	所在
和名類聚抄	蟋蟀	巻19・虫豸部第31・虫豸類第240・19丁裏5行目

言語記事情報

資料名	発行年月(日)	掲載位置	欄名・連載名等	執筆者	題名	内容	コード	キーワード	備考
文藝春秋	1942年9月	146-152		大町文衛	蟋蟀を尋ねて	「古文書と蟋蟀」古来、キリギリス、コオロギ、イトドの三つの名が様々な虫について使われてきたこと、マツムシとスズムシの指すものが二度入れ替わったことを挙げ、昔の人の実証精神の欠如が原因とする。（150-151）	動植物名 語源・語誌	キリギリス	

言語地図情報

地図ID	地図名	地図名よみがな	質問文	分野	品詞	分類	地図集ID	地図集書名
9848	きりぎりす	キリギリス	夏から秋にかけて，野原の草むらなどにいる虫で，長い後ろ足でよくとびはねます。触角が体の長さより長い。　共通語形：キリギリス　予想語形：ギッチョン	語彙	名詞	虫	187	『神奈川県言語地図』
9854	きりぎりす（螽斯）	キリギリス（キリギリス）	（絵）秋の虫です。緑色をしていてギーッチョというようになきます。野の虫で家の中へははいって来ません。この虫を何と言いますか。	語彙	名詞	虫	201	『上伊那の方言』

５．主な発表論文等

〔雑誌論文〕　計3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 山崎誠	４．巻 39
２．論文標題 語誌情報ポータルについて	５．発行年 2020年
３．雑誌名 日本語学	６．最初と最後の頁 124-129
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
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 －

１．著者名 山崎誠	４．巻 2018
２．論文標題 日本語史研究に関する二次的情報の集約―語誌データベースの構築―	５．発行年 2018年
３．雑誌名 人文科学とコンピュータシンポジウム論文集	６．最初と最後の頁 141-146
掲載論文のDOI（デジタルオブジェクト識別子） なし	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 －

１．著者名 新野直哉	４．巻 20
２．論文標題 大正期『文藝春秋』の記事に見られる言語規範意識	５．発行年 2018年
３．雑誌名 近代語研究	６．最初と最後の頁 155-175
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
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6件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 山崎誠，桂祐成
２．発表標題 語誌データベースの構築と設計上の問題点
３．学会等名 「通時コーパス」シンポジウム2019
４．発表年 2019年

１．発表者名
山崎誠，相澤正夫，大拓一郎，柏野和佳子，高田智和，新野直哉，間淵洋子，桂祐成

２．発表標題
語誌データベースの試験公開

３．学会等名
「通時コーパス」シンポジウム2019

４．発表年
2019年

１．発表者名
山崎誠，相澤正夫，大西拓一郎，柏野和佳子，高田智和，新野直哉，藤本灯

２．発表標題
語誌データベースの設計とその活用（2）

３．学会等名
「通時コーパス」シンポジウム2018

４．発表年
2018年

１．発表者名
新野直哉

２．発表標題
「新聞記事データベース」について ―概要と活用例―

３．学会等名
「通時コーパス」シンポジウム2018

４．発表年
2018年

１．発表者名
大西拓一郎

２．発表標題
方言地図データベースについて

３．学会等名
「通時コーパス」シンポジウム2018

４．発表年
2018年

１．発表者名
大西拓一郎

２．発表標題
方言分布・言語地図データベース―時空間情報を持つ言語データ―

３．学会等名
第23回公開シンポジウム　人文科学とデータベース

４．発表年
2018年

〔図書〕　計0件

〔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
研究分担者	藤本　灯 (Fujimoto Akari) (20733017)	京都府立大学・文学部・講師 (24302)	
研究分担者	大西　拓一郎 (Onishi Takuichiro) (30213797)	大学共同利用機関法人人間文化研究機構国立国語研究所・言語変化研究領域・教授 (62618)	
研究分担者	新野　直哉 (Niino Naoya) (30218086)	大学共同利用機関法人人間文化研究機構国立国語研究所・言語変化研究領域・准教授 (62618)	
研究分担者	高田　智和 (Takada Tomokazu) (90415612)	大学共同利用機関法人人間文化研究機構国立国語研究所・言語変化研究領域・准教授 (62618)	