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 ２８ 日現在

機関番号：６２６１８
研究種目：基盤研究(A)（一般）
研究期間：2019～2022
課題番号：１９Ｈ００５３１

研究課題名（和文）昭和・平成書き言葉コーパスによる近現代日本語の実証的研究

研究課題名（英文）An Empirical Study of Modern and Contemporary Japanese Language Using the Showa-Heisei Corpus of Written Japanese

研究代表者

小木曽　智信（Ogiso, Toshinobu）

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・研究系・教授

研究者番号：２０３３７４８９

交付決定額（研究期間全体）：（直接経費）　　34,500,000円

研究成果の概要（和文）：昭和・平成期の雑誌・ベストセラー書籍・新聞を収録した『昭和・平成書き言葉コーパス』を構築し、国立国語研究所のコーパス検索アプリケーション「中納言」を通してオンラインで公開した。このコーパスは1933年から2013年までの間を8年おきに11か年分，合計約3,340万語収録した大規模なものである。明治・大正期までの『日本語歴史コーパス』の後を承けて、現代に至るまでの日本語の成り立ちを探ることのできる資料として、日本語研究に重要な役割を果たすことが期待される。本研究課題では、このコーパスを活用した日本語史研究も実施し、研究発表会・論文を通して成果を発表した。

研究成果の学術的意義や社会的意義
本研究課題で構築･公開された『昭和・平成書き言葉コーパス』は、従来欠けていた昭和・平成期の大規模な日本語研究用資料として、今後の日本語研究において重要な役割を果たすことが期待される。本コーパスを『日本語歴史コーパス』に接続することで、上代から現代までの日本語を通時的に、実証的に研究する環境が初めて整った。また、本コーパスから得られる語彙統計情報は、国語辞典編纂・自然言語処理などの分野でも活用されることが期待される。

研究成果の概要（英文）：The Showa-Heisei Corpus of written Japanese, which contains magazines, best-selling books, and newspapers from the Showa and Heisei periods, has been constructed and made available online through corpus search application "Chunagon" by NINJAL. This corpus is a large scale one, containing 33.4 million words for 11 years from 1933 to 2013, with 8-year intervals. It is expected to play an important role in Japanese language research as a resource for exploring the history of the Japanese language up to the present day, following in the footsteps of the Historical Corpus of the Japanese Language, which covers the Meiji and Taisho periods. In this research project, we also conducted research on the history of the Japanese language using this corpus, and presented the results through research presentations and research papers.

研究分野：日本語学

キーワード：コーパス　日本語史　言語変化

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

2000 年代以降、コーパス（大規模テキストデータ）を用いた日本語の実証的研究が盛んに行われている。研究開始時点で、『現代日本語書き言葉均衡コーパス（BCCWJ）』を用いた研究論文は昨年度までに 840 件以上、『日本語歴史コーパス（CHJ）』では 310 件以上が刊行されていた。この中には明治期以降を対象に現代語がどのように形成されてきたかをデータにもとづいて実証しようとするものも多い。しかし、主要な日本語コーパスである『日本語歴史コーパス（CHJ）』は前近代から大正期までを対象としており、『現代日本語書き言葉均衡コーパス（BCCWJ）』は 1975 年以降の主として 2001～2005 年を対象としている。そのため、両者の間の時期の書き言葉のコーパスが欠けており、明治から現代までを通じた日本語の変化を実証的に研究することが困難な状況にあった。

今日我々が用いている日本語が語彙・文法・表記の各分野でどのように形成されてきたのか、近代以降どのような変化を経て現在に至っているのかを客観的・定量的に明らかにすることは、きわめて重要な問題であり、そのために昭和・平成期の書き言葉コーパスを構築することが強く求められた。

２．研究の目的

本研究課題の第一の目的は、コーパスが不足している昭和・平成期の雑誌・新聞・ベストセラー書籍を収録する「昭和・平成書き言葉コーパス」を構築し、上記の CHJ と BCCWJ とをつなぎ、明治から現代までの一貫した約 150 年の間の日本語の長期的な変化の研究を可能にすることである。

本研究の第二の目的は、構築される近代以降約 150 年の日本語を対象とする書き言葉コーパスを用いて、語彙・文法・表記上の問題について、現代日本語がどのように形成されてきたのかをコーパスにもとづいて客観的・定量的に、検証可能な形で明らかにすることである。従来は資料の制約から断片的なデータにのみ依拠したり、主観的に論述したりするしかなかった課題をコーパスにもとづいて探求する。そのために、各分野で優れた実績を持つ日本語の研究者 4 名が研究分担者としてコーパスを応用した研究に取り組むこととした。

３．研究の方法

昭和・平成期の書き言葉のコーパスがこれまで構築できなかった最大の原因は、著作権処理の困難さにある。従来の著作権法ではコーパスに収録するテキストの著作者の許諾が必要であったため、BCCWJ の構築にあたってはこの点に大変な労力と費用を費やしたうえに、収録できなかった資料も多い。しかし、2018 年の著作権法改正（「著作権法の一部を改正する法律」（平成 30 年法律第 30 号）平成 31 年 1 月 1 日施行）により「デジタル化・ネットワーク化の進展に対応した柔軟な権利制限規定の整備」が行われ「著作物の市場に悪影響を及ぼさないビッグデータを活用したサービス等のための著作物の利用について、許諾なく行えるようにする」ことが盛り込まれた。これにより、国立国語研究所の「中納言」のようにコーパスをウェブ上の検索アプリケーションを通して提供するサービスや、コーパスを解析した統計情報にもとづく言語データの提供であれば、著作権者の許諾なく行うことが可能になった。本研究課題ではこの改正著作権法にもとづいて、著作権者に不利益を与えないように配慮しつつ、著作者の許諾を得ることなしにコーパスを構築し公開を行うことで、低コストでのコーパスの構築・公開を実現する。これはコーパス構築に関わるきわめて先進的な取り組みであり、コーパスの設計とともに創造性を有するものである。

このコーパスに収録する資料は、①広く読まれて社会的な影響が大きい、②明治から平成まで継続的に刊行されてきた、③CHJ・BCCWJ に収録されておりコーパスを接続可能である、という点から、新聞・雑誌・書籍（ベストセラー）とする。ただし、全ての年にわたってデータを作成することは現実的でないため、長期的な変化に主眼を置いて『太陽コーパス』で行われた（ほぼ）8 年おきに収録する方法を採る。これにより 1925 年から 2013 年までの期間を 8 年おきにカバーし（対象年：1933・1941・1949・1957・1965・1973・1981・1989・1997・2005・2013）、CHJ と BCCWJ の不足期間を埋め両者を結合することとした。コーパス構築にあたっては、BCCWJ と CHJ の構築時のノウハウと既存のデータベースシステム、形態素解析用の辞書等を活かし、仕様の共通化と低コスト化を図り、また将来的により大規模なコーパスとして拡張することも視野にいれた設計を行った。

雑誌は、CHJ 収録済の雑誌コーパスを引き継いで 8 年おきに総合雑誌のデータを作成した。『太陽』は 1928 年に終刊となるため、1933 年からは当時最も有力であった総合雑誌『中央公論』を対象とし、1965 年からは、最大の刊行部数を誇り社会的な影響力が大きい『文藝春秋』を収録対象とした。新聞は、BCCWJ および CHJ「明治・大正編Ⅴ 新聞」でも収録対象となっている『読売新聞』の奇数月 2 日（朝刊）の紙面を各年収録した。書籍（ベストセラー）は各年のベストセラー上位 20 位までからコーパス化に適さないものを除外したのち、ランダムサンプリングによるテキスト採集を行った。

４．研究成果

第一の目的としたコーパスの構築は『昭和・平成書き言葉コーパス』として完成し、国立国語研究所のコーパス検索アプリケーション「中納言」を通して、2023 年 3 月に試験公開、5 月に正式版の公開を行った。収録語数は表 1 の通りである。

表 1　『昭和・平成書き言葉コーパス』収録語数

年	雑誌	ベストセラー書籍	新聞	合計
1933	3,291,739	178,895	128,016	3,598,650
1941	2,460,554	229,811	146,149	2,836,514
1949	1,016,658	272,594	116,868	1,406,120
1957	3,134,703	457,663	100,545	3,692,911
1965	2,025,871	373,622	279,162	2,678,655
1973	2,323,584	343,542	379,015	3,046,141
1981	2,658,012	297,230	355,731	3,310,973
1989	2,744,385	315,787	316,743	3,376,915
1997	2,541,480	306,927	271,339	3,119,746
2005	2,523,450	315,032	236,960	3,075,442
2013	2,679,038	355,067	228,549	3,262,654
合計	27,399,474	3,446,170	2,559,077	33,404,721

CHJ の明治・大正編の後をうけ、現代までの日本語の変遷をたどることのできるコーパスとして、3340 万語という規模のコーパスとなった。

このコーパスの特長の一つは、全文に UniDic による形態素解析を施して CHJ や BCCWJ と互換性のある短単位による形態論情報を付与したことである。解析精度は、語彙素認定の F 値で、雑誌が 0.9758，ベストセラー書籍が 0.9855，新聞が 0.9730　となっており、十分に実用に問題のないものとなった。

図 1　『昭和・平成書き言葉コーパス』ウェブページ https://clrd.ninjal.ac.jp/shc/

コーパスの公開はウェブ上の検索アプリケーション「中納言」による検索システムでの公開と、語彙統計情報の公開とした。検索システムでの公開は、弁護士とも相談の上で、表示される文脈長は前後 30 語までに制限して、著作権法の「デジタル化・ネットワーク化の進展に対応した柔軟な権利制限規定」（第 30 条の 4，第 47 条の 4,5）にもとづき、著作権者の承諾を得ないで利用することができる「軽微な利用」の範囲内で適法に公開している。

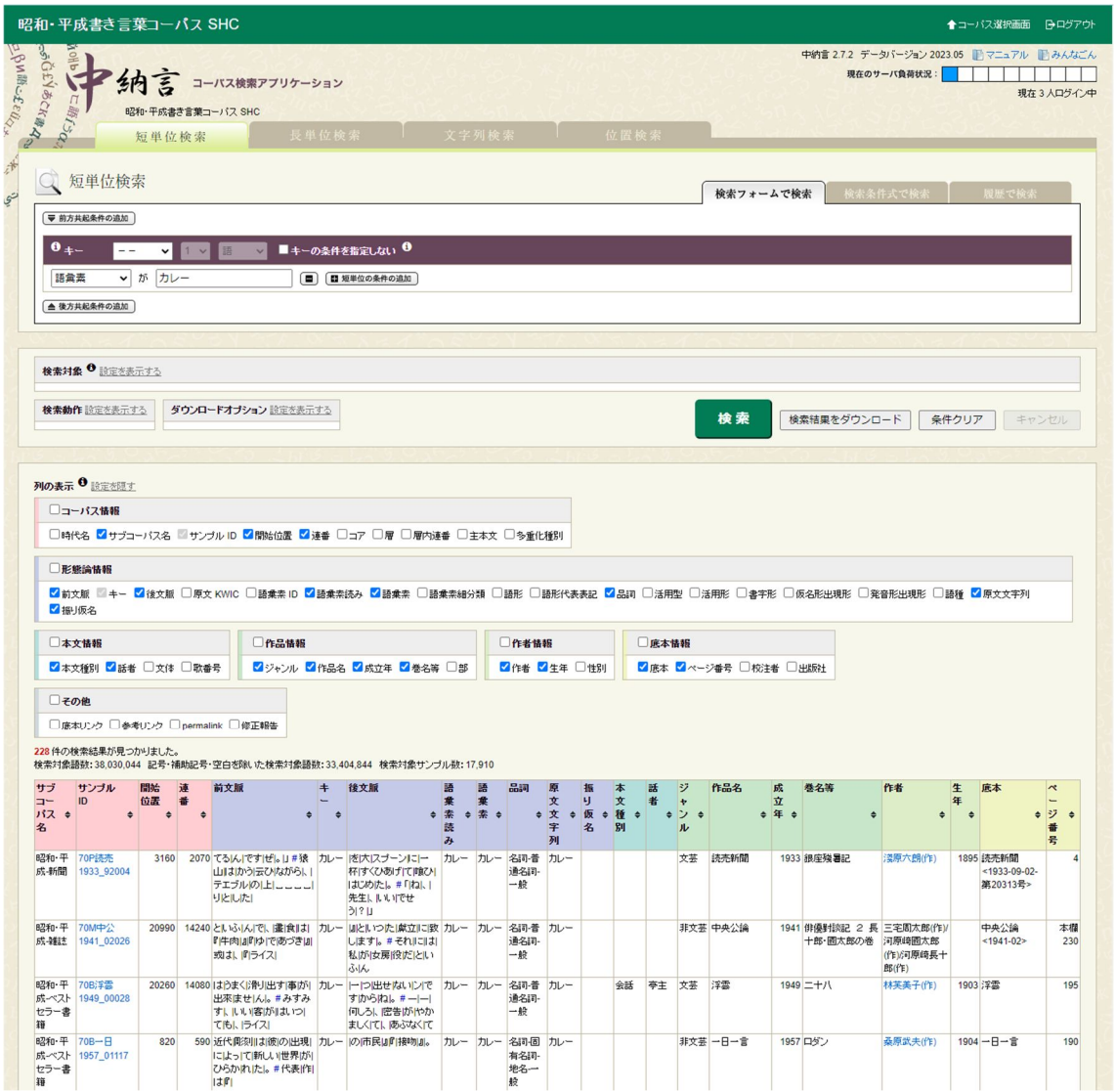

図 2　「中納言」による『昭和・平成書き言葉コーパス』検索画面
https://chunagon.ninjal.ac.jp/shc/search

このコーパスを用いた語彙・文法・表記上の問題についての研究会は、コーパスの構築途上から概ね年 3 回のペースで研究会を開催し、研究報告と議論を行って、研究成果は各自論文として発表してきた。各年 3 月には国立国語研究所「通時コーパス」プロジェクトとともに「通時コーパス」シンポジウムを共催した。この研究会は今後も「通時コーパス」プロジェクトの一部として継続し、成果物を研究書にまとめる予定である。

５．主な発表論文等

〔雑誌論文〕　計11件（うち査読付論文　8件／うち国際共著　0件／うちオープンアクセス　4件）

１．著者名 金愛蘭	４．巻 -
２．論文標題 第４章　語種	５．発行年 2020年
３．雑誌名 コーパスで学ぶ日本語学「日本語の語彙・表記」	６．最初と最後の頁 63-82
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 金愛蘭	４．巻 -
２．論文標題 新語や慣用表現から見た変化、「外来語の氾濫・乱用と叙述語化」	５．発行年 2020年
３．雑誌名 日本語の乱れか変化か―逸脱表現や新語の発生と許容―	６．最初と最後の頁 153-172
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 永澤済	４．巻 21-1
２．論文標題 「Xノタメニ」受身文の残存と衰退：近現代コーパスからみる	５．発行年 2021年
３．雑誌名 日本語文法	６．最初と最後の頁 21-37
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 小椋秀樹	４．巻 13
２．論文標題 近代における字音接頭辞「非・不・未・無」	５．発行年 2020年
３．雑誌名 立命館白川靜記念東洋文字文化研究紀要	６．最初と最後の頁 85-98
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 オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 金愛蘭	４．巻 38
２．論文標題 新聞における外来語「ルール」の叙述基本語化	５．発行年 2019年
３．雑誌名 国語語彙史研究会編『国語語彙史の研究』	６．最初と最後の頁 362-378
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 松田謙次郎	４．巻 21
２．論文標題 新漢字と旧漢字が混在したテキストからの短単位形態素の抽出について	５．発行年 2021年
３．雑誌名 国立国語研究所論集	６．最初と最後の頁 123-132
掲載論文のDOI（デジタルオブジェクト識別子） 10.15084/00003440	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 小椋秀樹	４．巻 40
２．論文標題 明治・大正期における否定の字音接頭辞 ―「非」を中心に―	５．発行年 2021年
３．雑誌名 国語語彙史の研究	６．最初と最後の頁 208-189
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 髙橋雄太	４．巻 2
２．論文標題 近現代における副詞の仮名表記化	５．発行年 2022年
３．雑誌名 論究日本近代語	６．最初と最後の頁 221-234
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 小椋秀樹	４．巻 42
２．論文標題 明治期から平成期における接頭辞「非－」の変遷	５．発行年 2023年
３．雑誌名 国語語彙史の研究	６．最初と最後の頁 左97-左116
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 相田太一，小町守，小木曽智信，高村大也，持橋大地	４．巻 30
２．論文標題 異なる時期での意味の違いを捉える単語分散表現の結合学習	５．発行年 2023年
３．雑誌名 自然言語処理	６．最初と最後の頁 275-303
掲載論文のDOI（デジタルオブジェクト識別子） 10.5715/jnlp.30.275	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 小林千真，相田太一，岡照晃，小町守	４．巻 30
２．論文標題 BERTを用いた日本語の意味変化の分析	５．発行年 2023年
３．雑誌名 自然言語処理	６．最初と最後の頁 713-747
掲載論文のDOI（デジタルオブジェクト識別子） 10.5715/jnlp.30.713	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計23件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 永澤　済
２．発表標題 昭和・平成書き言葉コーパスにみる「ために （為） 」行為者受身文の残存
３．学会等名 研究発表会 「昭和・平成書き言葉コーパスによる近現代日本語の実証的研究」
４．発表年 2020年

１．発表者名
間淵 洋子

２．発表標題
近現代日本語通史のための新聞コーパスの設計と構築

３．学会等名
研究発表会 「昭和・平成書き言葉コーパスによる近現代日本語の実証的研究」

４．発表年
2020年

１．発表者名
相田太一，小町 守，小木曽智信，持橋大地

２．発表標題
単語分散表現を用いた近現代日本語の意味変化の抽出

３．学会等名
研究発表会 「昭和・平成書き言葉コーパスによる近現代日本語の実証的研究」

４．発表年
2020年

１．発表者名
永澤　済

２．発表標題
利用者からみる通時コーパス資料 ―離縁・離婚の他動詞用法を例に

３．学会等名
「通時コーパス」シンポジウム2021

４．発表年
2021年

１．発表者名
間淵洋子，小木曽智信

２．発表標題
近現代日本語の意味変化分析のための単語データセット構築の試み

３．学会等名
言語処理学会第27回年次大会

４．発表年
2021年

１．発表者名
相田太一，小町守，小木曽智信，高村大也，持橋大地

２．発表標題
通時的な単語の意味変化を捉える単語分散表現の同時学習

３．学会等名
言語処理学会第27回年次大会

４．発表年
2021年

１．発表者名
近藤明日子，小木曽智信，髙橋雄太，田中牧郎，間淵洋子

２．発表標題
「昭和・平成書き言葉コーパス」の設計

３．学会等名
日本語学会2020年度秋季大会

４．発表年
2020年

１．発表者名
小木曽智信

２．発表標題
「昭和・平成書き言葉コーパス」の構築と活用に向けて

３．学会等名
研究発表会 「昭和・平成書き言葉コーパスによる近現代日本語の実証的研究」

４．発表年
2020年

１．発表者名
小椋秀樹

２．発表標題
近代における字音否定接頭辞 「非・不・未・無」の使用実態

３．学会等名
研究発表会 「昭和・平成書き言葉コーパスによる近現代日本語の実証的研究」

４．発表年
2020年

１．発表者名
髙橋雄太

２．発表標題
近現代における副詞の仮名表記化

３．学会等名
研究発表会 「昭和・平成書き言葉コーパスによる近現代日本語の実証的研究」

４．発表年
2020年

１．発表者名
相田太一，小町守，小木曽智信，高村大也，坂田綾香，小山慎介，持橋大地

２．発表標題
単語分散表現の結合学習による単語の意味の通時的変化の分析

３．学会等名
言語処理学会 第26回年次大会

４．発表年
2020年

１．発表者名
相田 太一，小町 守，小木曽 智信，高村 大也，持橋 大地

２．発表標題
単語ベクトルの結合学習を用いた近現代語の意味変化の分析

３．学会等名
じんもんこん2021

４．発表年
2021年

１．発表者名
小椋秀樹

２．発表標題
明治期から平成期における接頭辞「非ー」の変遷 ―『日本語歴史コーパス』『昭和・平成書き言葉コーパス』を資料として―

３．学会等名
「通時コーパス」シンポジウム2022

４．発表年
2022年

１．発表者名
髙橋雄太

２．発表標題
近現代における形容詞ムズカシイの意味と表記

３．学会等名
「通時コーパス」シンポジウム2022

４．発表年
2022年

１．発表者名
近藤 明日子，相田 太一，小木曽 智信

２．発表標題
近現代雑誌通時コーパスの語彙統計情報の公開

３．学会等名
言語処理学会第28回年次大会

４．発表年
2022年

１．発表者名
田中牧郎

２．発表標題
日本語史における和語

３．学会等名
第62回 語彙・辞書研究会

４．発表年
2022年

１．発表者名
田中牧郎

２．発表標題
日本語語彙の近代化における外来要素の受容と調整

３．学会等名
日本歴史言語学会

４．発表年
2022年

１．発表者名
Seiichi Inoue, Mamoru Komachi, Toshinobu Ogiso, Hiroya Takamura and Daichi Mochihashi

２．発表標題
Infinite SCAN: An Infinite Model of Diachronic Semantic Change

３．学会等名
The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)（国際学会）

４．発表年
2022年

１．発表者名
小木曽智信

２．発表標題
『日本語歴史コーパス』ver.2023.3通時コーパス拡張進捗報告

３．学会等名
「通時コーパス」シンポジウム2023

４．発表年
2023年

１．発表者名
井上誠一

２．発表標題
トピックモデルを用いた単語の通時的な意味変化のモデル化とその応用

３．学会等名
「通時コーパス」シンポジウム2023

４．発表年
2023年

１．発表者名
小木曽智信

２．発表標題
『昭和・平成書き言葉コーパス』と権利処理

３．学会等名
「通時コーパス」シンポジウム2023

４．発表年
2023年

１．発表者名
小木曽智信

２．発表標題
開かれた言語資源の共同構築と活用に向けて―新しい「通時コーパス」プロジェクトと「語彙資源」プロジェクト―

３．学会等名
NINJALシンポジウム「言語資源学の創成 ： 開かれた言語資源による日本語研究」

４．発表年
2022年

１．発表者名
小木曽智信，近藤明日子，髙橋雄太，間淵洋子

２．発表標題
ワークショップ『昭和・平成書き言葉コーパス』の構築と公開

３．学会等名
日本語学会2023年度春季大会

４．発表年
2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
『昭和・平成書き言葉コーパス』
https://clrd.ninjal.ac.jp/shc/
コーパス検索アプリケーション「中納言」昭和・平成書き言葉コーパス
https://chunagon.ninjal.ac.jp/shc

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	小椋　秀樹 (Ogura Hideki) (00321547)	立命館大学・文学部・教授 (34315)	

6．研究組織（つづき）

	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	間淵　洋子 (Mabuchi Yoko) (10415614)	和洋女子大学・人文学部・准教授 (32507)	
研究分担者	高橋　雄太 (Takahashi Yuta) (20840193)	明治大学・国際日本学部・助教 (32682)	
研究分担者	近藤　明日子 (Kondo Asuko) (30425722)	東京大学・大学院人文社会系研究科（文学部）・助教 (12601)	
研究分担者	松田　謙次郎 (Matsuda Kenjiro) (40263636)	神戸松蔭女子学院大学・文学部・教授 (34513)	
研究分担者	永澤　済 (Nagasawa Itsuki) (50613882)	上智大学・言語教育研究センター・准教授 (32621)	
研究分担者	持橋　大地 (Mochihashi Daichi) (80418508)	統計数理研究所・数理・推論研究系・准教授 (62603)	
研究分担者	田中　牧郎 (Tanaka Makiro) (90217076)	明治大学・国際日本学部・専任教授 (32682)	
研究分担者	金　愛蘭 (Kim Eran) (90466227)	日本大学・文理学部・准教授 (32665)	

６．研究組織（つづき）

	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	小町　守 (Komachi Mamoru) (60581329)	東京都立大学・システムデザイン研究科・教授 (22604)	

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