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 ２６ 日現在

機関番号：６２６１８
研究種目：挑戦的研究（開拓）
研究期間：2019～2022
課題番号：１９Ｈ０５４７７・２０Ｋ２０４１１

研究課題名（和文）日本語コーパスに対する情報付与を核としたオープンサイエンス推進環境の構築

研究課題名（英文）Development of an open science promotion environment focusing on annotation of Japanese corpora

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）　　19,800,000円

研究成果の概要（和文）：本研究課題ではコーパス検索システム「中納言」上のコーパスに対する汎用のアノテーションの共有環境を構築した。この共有環境の適用例として『日本語歴史コーパス』の形態論情報の誤り修正報告機能を実装した。機能の愛称を「みんなごん」とし、2022年度より修正報告アノテーションの収集の運用を開始し、クラウドソースによるコーパス修正環境の準備を整えた。その後、2023年2月までに集まった形態論情報誤り修正報告をデータに反映させて『日本語歴史コーパス』の更新を行った。このシステムの運用については国立国語研究所の共同研究プロジェクトに引きつぎ、今後もコーパスの修正を定期的に実施することとした。

研究成果の学術的意義や社会的意義
日本語研究の分野では研究に欠くことのできないインフラとして機能しつつあるコーパス検索アプリケーション「中納言」に、ユーザーが新たな情報を任意の場所に付加するアノテーション機能を追加し、情報を他のユーザーと共有して活用することができるシステムを構築した。その応用例として形態論情報の誤り修正報告機能（愛称「みんなごん」）を実装し、実際にこれを運用してクラウドソースによる『日本語歴史コーパス』の修正・更新を実現した。構築された共同研究環境は、テキストを中心とするオープンサイエンスを実現する基盤であり、ユーザーによるコーパス修正は今後の学術の方向性を示す先進的な事例として価値を持つものである。

研究成果の概要（英文）：In this research project, we have developed a general-purpose annotation sharing environment for corpora. As an example of application of this shared environment, we implemented an error correction reporting function for the morphological information of the "Corpus of Historical Japanese" in the public corpus search system "Chunagon". We nicknamed this reporting function "Minnagon" and began operating the collection of correction report by corpus users in FY2022. With this, the crowdsourced corpus correction environment is now ready, and the "Corpus of Historical Japanese" was updated by reflecting the morphological information correction reports collected by February 2023. We have decided to continue to operate the corpus under a joint research project of the National Institute for Japanese Language and Linguistics (NINJAL), and to periodically revise the corpus.

研究分野：日本語学

キーワード：コーパス　アノテーション　形態論情報　日本語歴史コーパス

## １．研究開始当初の背景

国立国語研究所が公開してきた『日本語歴史コーパス』は上代の「万葉集」から近世・近代の文学作品や雑誌まで、2 千万語以上の歴史的資料を収録した通時コーパスである。また、『現代日本語書き言葉均衡コーパス』は 1 億語を超えるバランスのとれた現代日本語の大規模コーパスである。これらのコーパスを利用する Web 上の検索アプリケーション「中納言」の登録ユーザーは研究開始時に 15,000 人を超えており、特に日本語研究の分野では研究に欠くことのできないインフラとして機能している。この「中納言」の機能を拡張して、利用者が新たな情報を任意の場所に付加するアノテーション機能を追加し、その情報を他のユーザーと共有・活用することができるシステムを構築することができれば、さまざまな応用が考えられる。従来、コーパスの用例を分類したデータなどの研究データは再利用されることなく消えていく状態にあったが、本研究課題によるシステムを用いることで、「中納言」ユーザーが、コーパスへのアノテーションという形でこうしたデータを共有・公開し、再利用できる環境を作ることができる。本研究課題では、こうした環境の構築とアノテーションの実践をおこなうことで、コーパスにもとづく研究データの公開とコーパス利用者間での共有を促し、将来的に書き言葉テキストを中心とした研究のプラットフォームとすることを目指すこととした。

## ２．研究の目的

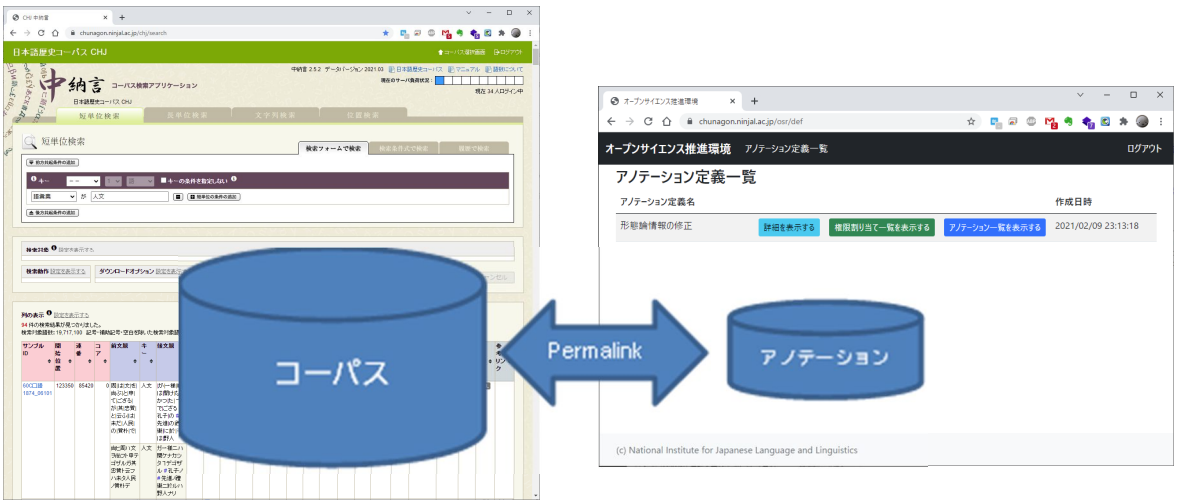
本研究の目的は、上述のとおり、「中納言」の機能を拡張して、利用者が新たな情報を任意の場所に付加するアノテーション機能を追加し、その情報を他のユーザーと共有・活用できるシステムを構築することである。この環境の上で、コーパスにもとづく研究データの公開とコーパス利用者間での共有を促し、書き言葉テキストを中心とした研究のプラットフォームとする。構築される共同研究環境は、テキストを中心とするオープンサイエンスとしての人文学研究の基盤の構築につながるものである。

このコーパスへのアノテーションは汎用的に利用できるものとして設計する。これにより、たとえば、助詞・助動詞の用法分類、係り結び、副詞と述語の呼応、格関係、代名詞や人物呼称の指示対象、話者と聞き手の情報、談話機能、枕詞・序詞などの修辞、異本注記、異なる解釈の注記、読解問題の作成などさまざまなレベルへの応用が考えられる。本研究課題は書き言葉コーパスを主たる対象とするが、このシステムは将来的に日常会話、方言、日本語教育などさまざまなコーパスに活用することも可能である。本システムが提供する研究環境は、言語研究・文学研究における新たな研究のあり方を提案するものである。

## ３．研究の方法

### （1）アノテーション用のシステムの設計

国立国語研究所の書き言葉コーパスは、作品と章段等を示すサンプル ID と開始位置の情報によって、個別の文字や語を参照することが可能になっている。この情報を用いて、利用者がコーパス中のテキストの任意の単位（字、語、文、発話、指定した範囲等）に、ラベルと任意のコメントを付与できるようにすることとして、汎用のアノテーションを実現するためのシステムを設計した。




コーパス検索アプリ「中納言」データベース（https://chunagon.ninjal.ac.jp/chj）

コーパスアノテーション環境「外記」（https://chunagon.ninjal.ac.jp/osr/def/1/annotation）

図１　アノテーションシステムの構成

単純なアノテーションとしては、利用者がメモを付箋のように記入できるようにすることが考えられるが、これ以外に、たとえば次のような情報のアノテーションが考えられる。

- 文字単位でのアノテーション例：誤字、異体字、異本注記など
- 単語単位でのアノテーション例：語の用法種別、副詞と述語との呼応、代名詞・人物呼称の指示対象、枕詞・掛詞、語釈など
- 発話単位でのアノテーション例：話者・聞き手、表現意図、談話的機能、読解問題など
- 指定した範囲へのアノテーション例：序詞、別解、場面注記、異本注記、読解問題など

また、国立国語研究所の書き言葉コーパスは、原則として単語情報（形態論情報）のみが付与されていて、上記のような情報は付与されていないが、単語情報とアノテーション情報を組み合わせることで、利用者の研究目的に即した高度な利用が可能になる。

（2）「形態論情報の誤り修正報告機能」への応用

アノテーションのデータベース・システムとしては上述のような汎用性を考慮したものとして設計・実装を行ったが、実際に応用するためのユーザーインターフェイスの実装については、より限定的で実用性の高い機能に限定して実装を行った。コロナ禍下で共同研究員との情報交換に限界があり、他の分野での応用可能性について幅広く情報を集めることが困難であったことから、コーパスの利用がもっとも進んでいる日本語学分野に範囲を絞り、中でももっとも要望の強かったコーパスの形態論情報の誤りを修正する（誤り部分の修正情報を報告する）機能を実装して運用することとした。

この「形態論情報の誤り修正報告機能」を実現するために、コーパスが依拠する形態素解析用の辞書「UniDic」の見出し語データベースを参照可能にし、コーパス中の任意の語について、表層形の辞書引きを行って正しい情報を付け直す機能を実装した（図 2）。


図 2　形態論情報の誤り修正報告（みんなごん）アノテーション画面

## ４．研究成果

主要な研究成果として、上述のアノテーションシステムと形態論情報の誤り修正報告について情報処理学会・人文科学とコンピュータシンポジウム「じんもんこん 2021」で発表した下記の発表・論文が挙げられる。本研究は、じんもんこん 2021 ベストポスター賞を受賞した。

小木曽智信・八木豊（2021）「『日本語歴史コーパス』の誤り修正プラットフォームの開発」『じんもんこん 2021 論文集』pp.206-211

実装した形態論情報の誤り修正報告機能について「みんなごん」という愛称を与えて、2022 年度より修正報告アノテーションの収集の運用を開始した。これにより、クラウドソースによるコーパス修正環境の準備が整ったことになる。

このシステムを実際に運用するために日本語学会で行った下記のワークショップも重要な研究成果の一つである。

小木曽智信・竹内綾乃・松崎安子「みんなで直す『日本語歴史コーパス』－中納言+みんなごん－」日本語学会 2022 年度秋季大会ワークショップ（2022 年 10 月 30 日）

このワークショップで学会会員に広く「みんなごん」の使い方の広報を行ったのち、『日本語歴史コーパス』の形態論情報について修正報告の収集を行った。そして、2023 年 2 月までに集まった「平安時代編」の形態論情報誤り修正報告をコーパスに反映させて公開した。

このようにコーパスユーザーからの情報をシステマティックに収集することでコーパスの誤りの修正に活かすのは、国立国語研究所としても初めての実績であった。

本科研の終了後、運用については国立国語研究所の共同研究プロジェクト「開かれた共同構築環境による通時コーパスの拡張」に引きついで、今後も、同様の方法でコーパスの修正機能を継続的に運用する予定である。また、2023 年度に公開予定の『昭和・平成書き言葉コーパス』用の「中納言」にも同機能を追加して、コーパスの誤り修正を実施することとした。

５．主な発表論文等

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

| | |
|---|---|
| １．著者名<br>小木曽 智信 , 八木 豊 | ４．巻<br>(2021) |
| ２．論文標題<br>『日本語歴史コーパス』の誤り修正プラットフォームの開発 | ５．発行年<br>2021年 |
| ３．雑誌名<br>じんもんこん2021論文集 | ６．最初と最後の頁<br>206-211 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>なし | 査読の有無<br>無 |
| オープンアクセス<br>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<br>－ |

〔学会発表〕　計14件（うち招待講演　3件／うち国際学会　2件）

| |
|---|
| １．発表者名<br>小木曽智信 |
| ２．発表標題<br>『日本語歴史コーパス』の新しい語彙表とその応用例 |
| ３．学会等名<br>日本語学会2021年度秋季大会 |
| ４．発表年<br>2021年 |

| |
|---|
| １．発表者名<br>Toshinobu Ogiso |
| ２．発表標題<br>A Brief Introduction to the Corpus of Historical Japanese Edo period series (Panel: Studies of Early Modern Japanese Based on the Corpus of Historical Japanese) |
| ３．学会等名<br>EAJS2021: 16th International Conference of the European Association of Japanese Studies（国際学会） |
| ４．発表年<br>2021年 |

| |
|---|
| １．発表者名<br>近藤 明日子, 相田 太一, 小木曽 智信 |
| ２．発表標題<br>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<br>言語処理学会第28回年次大会(NLP2022) |
| ４．発表年<br>2022年 |

１．発表者名
小木曽 智信, 八木 豊

２．発表標題
『日本語歴史コーパス』の誤り修正プラットフォームの開発

３．学会等名
じんもんこん2021

４．発表年
2021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.2020.3 通時コーパス構築進捗報告

３．学会等名
「通時コーパスシンポジウム」2020オンライン

４．発表年
2020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奈良時代編Ｉ万葉集』から『オックスフォード・NINJAL 上代日本語コーパス』『万葉集校本データベース』へのリンクについて

３．学会等名
「通時コーパスシンポジウム」2020オンライン

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
小木曽智信

２．発表標題
『日本語歴史コーパス』ver.2021.3 通時コーパス構築進捗報告

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
小木曽智信

２．発表標題
国立国語研究所の言語資源とオープンデータ・オープンサイエンス

３．学会等名
第1回 SPARC Japan セミナー（招待講演）

４．発表年
2019年

１．発表者名
小木曽智信

２．発表標題
大規模研究資源の構築・整備の評価：国語研のコーパス を例に

３．学会等名
Wiley Research Seminar Japan 2019（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
小木曽智信

２．発表標題
オープンサイエンスとしてのコーパス日本語学の可能性

３．学会等名
名古屋大学大学院人文学研究科日本語学分野公開講演会（招待講演）

４．発表年
2019年

１．発表者名
小木曽智信・竹内綾乃・松崎安子

２．発表標題
みんなで直す『日本語歴史コーパス』－中納言+みんなごん－

３．学会等名
日本語学会2022年度秋季大会

４．発表年
2022年

１．発表者名
小木曽智信・竹内綾乃

２．発表標題
『日本語歴史コーパス』の活用－語彙表を用いた集計と分析－

３．学会等名
日本語学会2022年度春季大会

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

〔図書〕　計0件

〔産業財産権〕

〔その他〕
じんもんこん2021ベストポスター賞
PD2-D1：『日本語歴史コーパス』の誤り修正プラットフォームの開発
小木曽 智信(国立国語研究所), 八木 豊(株式会社ピコラボ)
http://jinmoncom.jp/sympo2021/

6．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（機関番号） | 備考 |
|---|---|---|---|
| 研究分担者 | 河内　昭浩<br>(Kawauchi Akihiro)<br>(10625172) | 群馬大学・教育学部・准教授<br>(12301) | |
| 研究分担者 | 橋本　雄太<br>(Hashimoto Yuta)<br>(10802712) | 国立歴史民俗博物館・大学共同利用機関等の部局等・准教授<br>(62501) | |
| 研究分担者 | 永崎　研宣<br>(Nagazaki Kiyonori)<br>(30343429) | 一般財団法人人文情報学研究所・人文情報学研究部門・主席研究員<br>(82683) | |
| 研究分担者 | 鴻野　知暁<br>(Kono Tomoaki)<br>(30751515) | 東京大学・大学院人文社会系研究科（文学部）・助教<br>(12601) | |
| 研究分担者 | 海野　圭介<br>(Unno Keisuke)<br>(80346155) | 国文学研究資料館・研究部・教授<br>(62608) | |
| 研究分担者 | 後藤　真<br>(Goto Makoto)<br>(90507138) | 国立歴史民俗博物館・大学共同利用機関等の部局等・准教授<br>(62501) | |

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 |
|---|---|