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　元 年　 ６ 月 ２１ 日現在

機関番号：６２６１８
研究種目：基盤研究(A)（一般）
研究期間：2015～2018
課題番号：１５Ｈ０１８８３

研究課題名（和文）日本語歴史コーパスの多層的拡張による精密化とその活用

研究課題名（英文）Refinement and utilization of the Corpus of Historical Japanese through multilayered extension

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・言語変化研究領域・教授

研究者番号：２０３３７４８９

交付決定額（研究期間全体）：（直接経費）　　34,300,000円

研究成果の概要（和文）：国立国語研究所の「日本語歴史コーパス」のシステムを拡張し、読み下し本文とは大きく異なる原文を扱えるように、また、掛詞・洒落・臨時的な振り仮名などの多重の読みを付与できるように改善した。その上で、国語研「通時コーパス」プロジェクトと共同で、原文付き「万葉集」、ローマ字本文と和文を併記したキリシタン資料、多重の読みを付与した洒落本と人情本、掛詞情報付きの「八代集」、明治初期口語資料と「東洋学芸雑誌」のコーパスを整備し、Web上の「中納言」を通して公開した。
さらに、このコーパスを活用した日本語史研究を展開し、特に上代・中古および近世の文法、近代の語彙等の分野で研究発表を行った。

研究成果の学術的意義や社会的意義
データベースを拡張することで日本語の歴史を研究する上で重要な資料が、原文や掛詞を含む完全なコーパスとして利用可能になった。これにより国立国語研究所の「日本語歴史コーパス」が質的に大きく向上したのみならず、上代から近代までの資料を増補したことで量的にも拡大した。その結果、このコーパスは、日本語史研究の基本となる資料として学界で広く利用されるようになった。
また、構築したコーパスはインターネット上で無償で公開し、一般にも利用できるようになっている。このコーパスを古典教育等に応用するための研究や教材開発も進められており、社会的にも意義のあるものとなった。

研究成果の概要（英文）：We improved the system of the Corpus of Historical Japanese at NINJAL, to handle the original text significantly different from the transliteration text, and to annotate multiple readings such as kakekotoba, paronomasia and temporary furigana. Then, in cooperation with NINJAL "Diachronic corpus" project, "Manyoshu" with original text, Christian materials written in both Roman characters and Japanese, Sharebon and Ninjobon with multiple readings, "Hachidai-shu" with kakekotoba, corpus of "Toyo Gakugei Zasshi" and colloquial materials in the early Meiji period were prepared and made available to the public through "Chunagon" on the Web.
In addition, we developed studies of Japanese language history using this corpus, and in particular, we published the research in the fields of grammar of Old and Early Middle and Early Modern Japanese, and vocabulary of Modern Japanese.

研究分野：日本語学

キーワード：日本語史　コーパス　形態素解析　文法　語彙　表記　自然言語処理

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

１．研究開始当初の背景

国立国語研究所において日本語史研究者にとっての研究の基盤を整備すべく「日本語歴史コーパス」が構築されている。このコーパスを日本語史研究で活用できるものにするためには、従来の日本語学や古典研究の成果を十分に活かした信頼できるものにする必要がある。特に、文献学的研究に基づく精緻な資料研究や、作品・作者や修辞法に関する研究の蓄積を取り込むことができれば、コーパスは一層信頼性と価値を増す。しかしながら、本研究開始時点の「日本語歴史コーパス」は、現代語のコーパスのために作られたデータベースシステムを利用しているため、日本語史資料を扱う上で次のような点で限界があった。

・本文が、形態素解析対象となる校訂済みテキストのみで、たとえば『万葉集』における万葉仮名などの原文のテキストが取り扱えない
・掛詞、左右のふりがな、洒落のような多重化した「読み」が扱えない

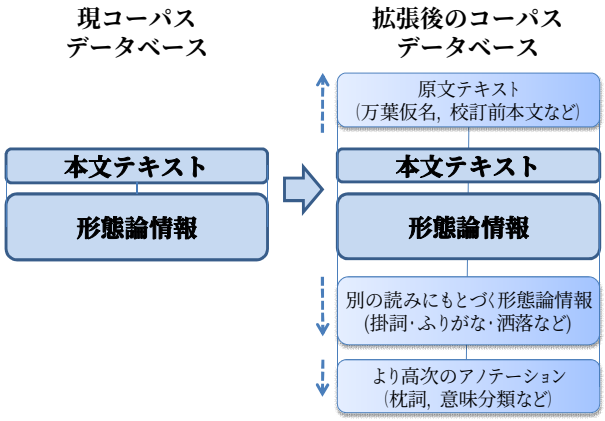
「日本語歴史コーパス」を日本語史研究に適したコーパスとするために、上記の情報を適切に扱ってコーパスの検索や集計に利用できるようにシステムを拡張する必要があった。

２．研究の目的

上記の問題に対処するため、まず右下図のようにコーパスのデータベースシステムを拡張する。これにより、「万葉集」の万葉仮名やキリシタン資料のローマ字本文等の歴史的資料の原文を利用可能にし、さらに和歌における掛詞や戯作における洒落、近世・近代の振り仮名による多重の読みなどをコーパス上で扱うことを可能にする。

そして、国立国語研究所の「通時コーパス」プロジェクトと共同で、こうした拡張を活かすことのできるコーパスを実際に構築し、完成したものを一般に公開する。対象とする資料は、万葉仮名の原文をもつ「万葉集」のほか、掛詞を付与する「八代集」、掛詞や洒落等を含む近松の世話物浄瑠璃等である。また、近代の資料については、資料性を検討し、新たに対象を選定したうえでコーパスの構築を行う。また、このシステムを用いて別途構築したキリシタン資料・洒落本・人情本などのコーパスの質的な拡張に対応する。

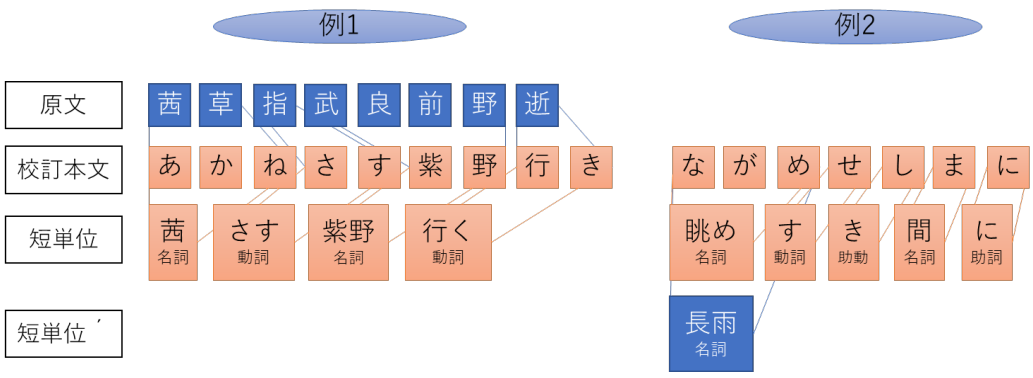
そのうえで、構築されるコーパスを用いた日本語史研究を各時代・各分野で展開する。



３．研究の方法

「万葉集」を扱う「上代グループ」（野村・鴻野）、八代集と散文中の和歌を扱う「和歌グループ」（近藤みゆき・山元・冨士池・松崎）、近松世話物浄瑠璃を中心とする「近世グループ」（村上・岡部・市村・高田）、近代雑誌コーパスを補う新資料を対象とする「近代グループ」（田中・岡部・近藤明日子）、およびコーパス整備のためのツールとデータベースを開発する「言語処理グループ」（松本・小木曽）の５つのグループを置き、時代ごとに分担してコーパスの構築と利用を行った。上代グループには、研究協力者としてオックスフォード大学で上代語コーパスを構築したフレレスビッグ教授が参加し、全体の統括は小木曽が行った。

言語処理グループでは、データベースの拡張により、下図のように原文テキスト（例 1）、掛詞を含むテキスト（例 2）を扱えるようにした。時代別の各グループでは、コーパスの構築とこれを活用した日本語研究に取り組んだ。構築側と活用側が互いにフィードバックを行い、精密な研究とコーパスの質の向上をはかった。



４．研究成果

国語研コーパス開発センターと共同で当初予定したデータベースシステムの拡張を行い、まず原文の取り扱いを可能にした。このシステムで、まず万葉集の原文データをデータベースに取り込み、原文と読み下し本文との対応付け（アラインメント）を行った。また、データベースに「掛詞テーブル」を新設し、これによって掛詞・洒落・振り仮名による多重の読み（多重形態論情報）を付与することを可能にした。これを用いて近世の洒落本・人情本の本文に対し、洒落や振り仮名による多重の読みの情報を付与したほか、近松の世話物浄瑠璃や八代集に掛詞情報の付与を行った。

これらの拡張に並行して、コーパス本文と形態論情報そのものの整備も行った。「万葉集」、八代集、近松の世話物浄瑠璃のほか、明治初期の啓蒙書と「安愚楽鍋」から成る「明治初期口語資料」、雑誌コーパスを拡張する「東洋学芸雑誌」は、この科研費による成果である。これらのコーパスは『日本語歴史コーパス』の一部としてコーパス検索アプリケーション「中納言」を通して一般に公開した。

2017.9　奈良時代編Ⅰ万葉集　ver.1.0　新規公開
2019.3　明治・大正編Ⅲ明治初期口語資料　　ver.0.8　新規公開
2019.3　明治・大正編Ⅰ雑誌　ver.1.2　『東洋学芸雑誌』追加公開
2019.3　和歌集編（八代集）　ver.0.8　新規公開

未公開の近松の世話物浄瑠璃と八代集の掛詞データについては、2019 年度中に公開予定である。

なお、他の研究費によって構築されたコーパスで、本科研費による成果によって原文や多重形態論情報の付与を行ったコーパスとして、次のものがある。

2018.3　室町時代編Ⅱキリシタン資料　ver.1.0　新規公開
2019.3　江戸時代編Ⅰ洒落本　ver.1.0　短単位データ更新
2019.3　江戸時代編Ⅱ人情本　ver.0.8　新規公開

いずれも『日本語歴史コーパス』の一部として『コーパス検索アプリケーション「中納言」を通して一般に公開している（https://pj.ninjal.ac.jp/corpus_center/chj/update.html）。

５．主な発表論文等

〔雑誌論文〕（計　14 件）

1．近藤明日子「語種率・品詞率からみる近代文語文の通時的変化」
日本語学論集 15　pp.(64)97-(78)83、査読なし、2019 年

2．近藤明日子「明治・大正期の文語文における一人称代名詞の通時的変化 ： 『日本語歴史コーパス 明治・大正編Ⅰ雑誌』と『東洋学芸雑誌』を用いた分析」
『国立国語研究所論集』14　pp.73-88、10.15084/00001413、査読あり、2018 年

3．鴻野知暁「上代日本語の複合動詞の項構造について―二つの内項を取る場合を中心に―」
言語・情報・テクスト（東京大学大学院総合文化研究科言語情報科学専攻紀要）25　pp.41-50、査読なし、2018 年

4．岡島昭浩「「ひいやり・ふうわり」型から「ひんやり・ふんわり」型へ」
国語語彙史の研究 36　pp.107-117、査読なし、2017 年

5．冨士池優美「中古歌合日記の品詞比率」
紀要　言語・文学・文化第 119 号（通巻第 264 号）　pp.57-67、査読なし、2017 年

6．市村太郎，村山実和子「洒落本コーパス構築の試行」
国立国語研究所論集 12　pp.29-45、10.15084/00000852、査読あり、2107 年

7．村山実和子，小木曽智信，中村壮範「形態論情報の多重化による洒落本コーパスの質的拡張」 情報処理学会研究報告 2017‐CH‐114　pp.8-18、査読なし、2017 年

8．田中牧郎「演説の文末表現の変遷―明治時代から昭和 10 年代まで」
『SP 盤演説レコードがひらく日本語研究』笠間書院　pp.248-270、査読なし、2016 年

9．小木曽智信「『日本語歴史コーパス』の現状と展望」
国語と国文学 93 巻 5 号　pp.72-85、査読なし、2016 年

10．村上謙「近世上方における二段活用の一段化とその後の展開」
国語と国文学 93 巻 5 号　pp.99-112、査読なし、2016 年

11．松本裕治「第６章　形態論と自然言語処理」
漆原朗子編『形態論』(朝倉日英対照言語学シリーズ)　pp.141-154、査読なし、2016 年

12．村上謙「近世上方語研究における研究手法について―用例収集と分析・解釈―」
近代語研究第 19 集　pp.43-60、査読なし、2016 年

13．小木曽智信「使用頻度から見た中古仮名文学作品の語彙―コーパスにもとづく分析―」
国語語彙史の研究 35　pp.15-37、査読なし、2016 年

14．小木曽智信「中古和文における文体別の特徴語」
『コーパスと日本語史研究』ひつじ書房　pp.93-117、査読なし、2015 年

〔学会発表〕(計　19 件)

1. Toshinobu OGISO“Corpus of Historical Japanese ver. 2018.9” 29th EAJRS (European Association of Japanese Resource Specialists) conference、2018 年

2. 片山久留美, 小木曽智信, 中村壮範「キリシタン資料のローマ字原文対応和文テキストの作成」 人文科学とコンピュータシンポジウム「じんもんこん 2018」、2018 年

3. スティーブン・ライト・ホーン, 鴻野知暁, アラステア・バトラー, 小木曽智信, ビャーケ・フレレスビッグ「「オックスフォード・NINJAL 上代語コーパス」の公開」日本語学会 2018 年度秋季大会、2018 年

4. 鴻野知暁「疑問詞疑問文におけるカとゾの出現位置について」
NINJAL-Oxford 通時コーパス国際シンポジウム、2018 年

5. 岡島昭浩「近代語資料の今後―発掘・賦活・保持―」第 343 回日本近代語研究会、2017 年

6. 鴻野知暁「複合動詞「Ｖ＋暮らす」と「Ｖ＋暮る」について―項構造を中心に―」
第 117 回国語語彙史研究会、2017 年

7. 鴻野知暁, 岡照晃, 小木曽智信「『日本語歴史コーパス　奈良時代編Ｉ万葉集』の公開」
日本語学会 2017 年度秋季大会、2017 年

8. 鴻野知暁「万葉集における動詞の複合のしやすさについて」
韓国日本語学会 第 36 回学術発表会、2017 年

9. Tomoaki KOUNO The design and characteristics of the Man'yoshu corpus”
The 15th EAJS (the European Association for Japanese Studies) International Conference、2017 年

10. 市村太郎, 小木曽智信「文書構造を利用した近世期洒落本の形態素解析」
言語処理学会第 22 回年次大会、2016 年

11. Teruaki OKA, Tomoaki KOUNO ”Original-Transcribed Text Alignment for Man'yosyu Written by Old Japanese Language” Language Technology Resources and Tools for Digital Humanities、2016 年

12. Toshinobu OGISO and Yuki WATANABE“Construction of the Corpus of Historical Japanese” PNC (Pacific Neighborhood Consortium) 2016 Annual Conference and Joint Meetings、2016 年

13. 小木曽 智信, 池上 尚, 渡辺 由貴, 市村 太郎, 近藤 明日子, 間淵洋子「『日本語歴史コーパス』の拡張とその課題―「通時コーパス」をめざして―」
日本語学会 2016 年度春季大会、2016 年

14. 小木曽 智信「日本語歴史コーパスの量的・質的拡張」
「通時コーパス」国際シンポジウム、2015 年

15. 堤 智昭, 小木曽 智信「歴史的資料を対象とした複数の UniDic による形態素解析支援ツール『Web 茶まめ』」人文科学とコンピュータシンポジウム 「じんもんこん 2015」、2015

年

16. 小木曽智信「『日本語歴史コーパス』にもとづく中古仮名文学作品の語彙分析」第 110 回国語語彙史研究会、2015 年

17. 間淵 洋子, 小木曽智信「異なる文体の混在するテキストに対する複数辞書切り替えによる解析手法の提案」人文科学とコンピュータシンポジウム 「じんもんこん 2015」、2015 年

18. Taro ICHIMURA, Yuki WATANABE, Tomoaki KOUNO, Toshinobu OGISO ”Construction of the Corpus of Toraakira-bon Kyogen” DH2015: Annual Conference of Digital Humanities、2015 年

19. Tomoaki Kouno, Toshinobu Ogiso ”Improving an Electronic Dictionary for Morphological Analysis of Japanese: Use of historical period information” The 9th International Conference of ASIALEX (ASIALEX2015)、2015 年

〔その他〕
『日本語歴史コーパス』概要
https://pj.ninjal.ac.jp/corpus_center/chj/
『日本語歴史コーパス　奈良時代編Ⅰ万葉集』
https://pj.ninjal.ac.jp/corpus_center/chj/nara.html
『日本語歴史コーパス　明治・大正編Ⅲ明治初期口語資料』
https://pj.ninjal.ac.jp/corpus_center/chj/meiji_taisho.html#shokikogo
『日本語歴史コーパス　明治・大正編Ⅰ雑誌』
https://pj.ninjal.ac.jp/corpus_center/chj/meiji_taisho.html#zasshi
『日本語歴史コーパス　和歌集編』
https://pj.ninjal.ac.jp/corpus_center/chj/wakashu.html

６．研究組織

(1)研究分担者
研究分担者氏名：松本　裕治
ローマ字氏名：MATSUMOTO Yuji
所属研究機関名：奈良先端科学技術大学院大学
部局名：先端科学技術研究科
職名：教授
研究者番号（8 桁）： 10211575

研究分担者氏名：村上　謙
ローマ字氏名：MURAKAMI Ken
所属研究機関名：関西学院大学
部局名：文学部
職名：教授
研究者番号（8 桁）： 20431728

研究分担者氏名：冨士池　優美
ローマ字氏名：FUJIIKE Yumi
所属研究機関名：玉川大学
部局名：文学部
職名：准教授
研究者番号（8 桁）：20510572

研究分担者氏名：鴻野　知暁
ローマ字氏名：KOUNO Tomoaki
所属研究機関名：東京大学
部局名：大学院総合文化研究科
職名：助教
研究者番号（8 桁）：30751515

研究分担者氏名：岡島　昭浩
ローマ字氏名：OKAJIMA Akihiro

所属研究機関名：大阪大学
部局名：文学研究科
職名：教授
研究者番号（8 桁）：50194345

研究分担者氏名：市村 太郎
ローマ字氏名：ICHIMURA Taro
所属研究機関名：常葉大学
部局名：教育学部
職名：講師
研究者番号（8 桁）：10701352

研究分担者氏名：田中 牧郎
ローマ字氏名：TANAKA Mariko
所属研究機関名：明治大学
部局名：国際日本学部
職名：専任教授
研究者番号（8 桁）：90217076

研究分担者氏名：高田 智和
ローマ字氏名：TAKADA Tomokazu
所属研究機関名：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
部局名：言語変化研究領域
職名：准教授
研究者番号（8 桁）：90415612

研究分担者氏名：松崎 安子
ローマ字氏名：MATSUZAKI Yasuko
所属研究機関名：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
部局名：言語変化研究領域
職名：プロジェクト PD フェロー
研究者番号（8 桁）：50581724

研究分担者氏名：近藤 明日子
ローマ字氏名：KONDO Asuko
所属研究機関名：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
部局名：コーパス開発センター
職名：プロジェクト非常勤研究員
研究者番号（8 桁）：30425722

(2)研究協力者
研究協力者氏名：岡部 嘉幸（千葉大学）　※連携研究者
ローマ字氏名：OKABE Yoshiyuki

研究協力者氏名：野村 剛史（東京大学）　※連携研究者
ローマ字氏名：NOMURA Takashi

研究協力者氏名：近藤 みゆき（実践女子大学）　※連携研究者
ローマ字氏名：KONDO Miyuki

研究協力者氏名：山元 啓史（東京工業大学）　※連携研究者
ローマ字氏名：YAMAMOTO Hilofumi

研究協力者氏名：ビャーケ・フレレスビッグ（オックスフォード大学）
ローマ字氏名：Bjarke FRELLESVIG