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 １９ 日現在

機関番号：６２６１８
研究種目：研究活動スタート支援
研究期間：2018～2019
課題番号：１８Ｈ０５６１３・１９Ｋ２０８１９

研究課題名（和文）精緻な文字表記情報を持つ近代新聞コーパスの構築による表記・文体変遷の計量的研究

研究課題名（英文）A Corpus-based Study of Diachronic Change on Notation and Style in Modern Japanese

研究代表者
間淵　洋子（MABUCHI, Yoko）

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・特任助教

研究者番号：１０４１５６１４

交付決定額（研究期間全体）：（直接経費）　　2,300,000円

研究成果の概要（和文）：　本研究では、社会と言語が大きく変化した近代（明治・大正時代）において、言語使用の実態がどのように変化したかを明らかにすることを目的として、近代の新聞のコーパス（テキストを元と研究に利用できる形で電子化したデータベース）を構築し、それに基づく言語分析を行った。
　コーパスを用いて、主に語彙・表記・文体について分析した結果、活版印刷における変体仮名利用の実態と近代を通じての変遷の一端を把握することができた。また、変体仮名を含む表記の変化と並行して生じていた、新聞の形態や内容の比重の変化、文体の変化（平易な談話体から文語体、現代語に繋がる口語文体へ）も明らかになった。

研究成果の学術的意義や社会的意義
本研究は、近代の新聞コーパスの構築を主眼とした。これまで利用できなかった形態論情報付き新聞コーパスの完成によりメディアやジャンルを考慮した近代語研究の精緻化が可能となった。
また、このコーパスでは、近代の活版印刷で多く用いられていた変体仮名に対して、Unicode変体仮名（文字コードの国際規格Unicodeに収録された変体仮名）に対応づけるためのコーディングを行っている。これにより、近代の出版物における変体仮名の使用実態を計量的に調査・分析することが可能となった。

研究成果の概要（英文）：The purpose of this study is to clarify how the actual situation of language usage changed in the modern times (Meiji and Taisho era) when society and language changed drastically. I constructed a corpus of modern newspapers and conducted linguistic analysis based on it.
As a result of quantitative analysis of notations and styles using the constructed corpus, I clarified the actual conditions and changes in the use of Hentaigana in letterpress printing. In addition, the changes in the form of newspapers, the weight of contents, and the style were also clarified.

研究分野：日本語学

キーワード：コーパス　新聞　近代語　言語変化　表記　文体　Unicode変体仮名　ルビ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年、日本語の歴史的資料のコーパス（言語研究用の情報が付けられたテキストデータベース）が続々と公開され、日本語の歴史・変化をデータに基づき計量的に分析することが可能になっている。国立国語研究所の公開する『日本語歴史コーパス』には、奈良時代から大正時代にかけて日本語の歴史を知る上で重要な資料が収録されているが、中でも近代（明治・大正時代）については、1500 万語を超える大規模なデータを有しており、言文一致による文体の変化や、西洋文明流入による語彙の変化など、日本語史上の重要トピックを論証するための有用な資料として利用されている。しかし、含まれる資料の多くは雑誌であるため、日本語においては決して小さくはないメディアやジャンルによる言語使用の差異を捉えるのに十分であるとは言い難い。

一方、現代日本語の研究においては、1990 年代から新聞各社による新聞本文テキストデータの販売が行われ、これをコーパスとして活用し計量言語学的な手法による研究が盛んになされてきた。これは、コーパスとして用いることのできるデータが新聞であったということだけでなく、現代における新聞というメディアの持つ公共性や一般性、また規範性が、「一般的な日本語」の実態把握に有用だったためでもある。その後登場した均衡コーパス（多種多様な資料をバランス良く収録したコーパス）に基づく研究では、メディアやジャンルによる言語使用の差異が数多く報告されている。

そこで、近代語の研究においても、メディア・ジャンルによる言語差を把握することを目指して、現代語研究で多く用いられている新聞を対象としたコーパスの構築とこれを用いた表記・語彙の調査・研究を行うことを企図した。

２．研究の目的

近代は、日本の歴史上、長い武士の時代と鎖国が終わり、劇的に社会体制が変化した時代であるが、日本語の歴史においても大きな転換期であったとされる。前述の日本語史上の重要トピックである言文一致や語彙の変化も、社会の近代化に伴う日本語の近代化と捉えるべき事象であると考えられる。本研究の大きな目的は、このように社会と言語が大きく変化した近代から現代にかけての、言語使用実態とその変遷を明らかにすること、そして、資料の形式・表記・文体が、メディアの発達や社会の変化とどのように関わりながら変化したかを明らかにすることである。そこで、特に、幕末明治初期に誕生し、形式・表記・文体すべてにおいて急速に変化・発展しながら現代へと途切れなく続く「新聞」というメディアを取り上げ、言語使用実態とその変化に迫ること、資料の形式や言語の表記についての大量かつ精緻な実態把握に基づき、言語変化の様相を明らかにすることを、より具体的な目的として設定した。

３．研究の方法

近代における言語使用の実態を把握するために、以下の二つの方法で研究を実施した。

(1) 新聞コーパスの構築

①新聞の形式，②文字・表記，③単語、のそれぞれに関する詳細な情報を付与した新聞本文のデータベース（コーパス）を作成した。収録の対象と範囲は、通時的な変遷を追うことができること、既存のコーパスとの比較が可能となることを条件とした。これに合致するものとして、収録対象資料に 1874 年の創刊から現代まで続く『読売新聞』を選定した。収録範囲は、『日本語歴史コーパス　明治・大正編 I 雑誌』に倣い、1874（明治 7）年から 6〜8 年おきに 1925（大正 14）年まで、計 8 か年分とし、1 か年につき 4 日〜8 日分の新聞（1 冊全文、広告等を除く。1874 年は 1 冊の分量が少ないため 22 日分）をコーパスとした。

本文テキストは、近代資料の文字入力の経験がある専門業者に、新聞紙面のコピーに基づき入力を依頼し作成した。研究用情報の付与については、以下の通り行った。

① 形態：判型（紙面のサイズ），ページ数，段組み，面・欄割りについて、1 日分の新聞紙ごとにリストを作成した。

② 文字・表記：漢字は、JIS X0213:2004 規格（JIS 第 1〜4 水準の漢字を含む）に基づき、字体を判別し入力した。『読売新聞』は総ルビ（難読かどうかに関わらず、数字を除くほぼ全ての漢字に振り仮名を付したもの）の形態に特徴があるため、本行に加えてルビの入力も行った。更に、近代初期の活版印刷資料に多く使われている変体仮名については、Unicode 変体仮名との対応付けを行い、Unicode 番号を付与した。

③ 単語：『日本語歴史コーパス』の形態論情報と同様に、形態素解析辞書 UniDic による情報付けを行うこととし、本文テキストを Web 上で動作する形態素解析ツール「Web 茶まめ」（https://chamame.ninjal.ac.jp）を用いて解析した。

(2) 新聞コーパスに基づく計量的言語研究

(1)により作成した新聞コーパスを用いて、文字・語・文体の実態を調査し、時代・時期による差異・変化を計量的分析手法により導き出す。特に、変体仮名活字の使用実態、言語量（語数）、文体（口語文体か文語文体か）の 3 項目について重点的に調査を実施し明らかにした。

４．研究成果

(1) 新聞コーパスの概要

本研究の成果の第一の成果は、これまで収録のメディアに偏りのあった近代日本語の形態論情報付きコーパスに、日本語における重要メディアの一つとして新聞のデータを加えた点にある。以下に、そのデータ概要と、情報付けの概要について示す。

①収録データ概要

本研究で構築した新聞コーパスの収録データ概要は、表１の通りである。

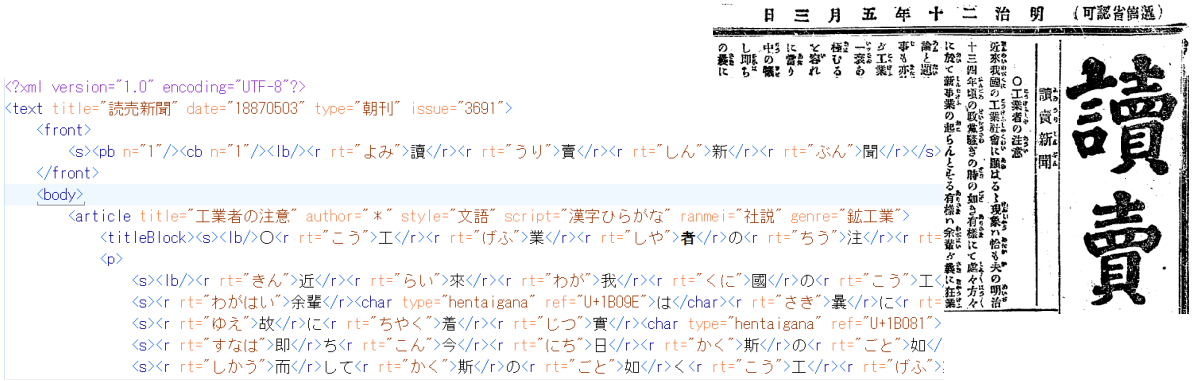
当初目標として、新聞における時期による使用語彙の偏りなど（歳時的な行事の影響等）に考慮し、各年でサンプルテキストの取得月日を揃える方法を取った上で、1 ヵ年につき 5 万語程度を収録することを計画した。収録対象期間に紙面構成が大きく変化しており、1 日分の新聞紙から取得される文字数に大きな差があるため、年次による収録語数に大きな開きが生じているが、最も語数の少ない 1881 年でも、当初目標の 5 万語は取得できた。

表１ 「新聞コーパス」の収録データ概要

発行年	冊数	内訳	文字数（千字）	短単位数（千語）	形態
1874	22	(1874 年)11 月 2・10・14・16 日/ (1875 年) 5 月 2・3・4・5・7 日/ 11 月 2・4・5・7・8 日/ 7 月 2・3・4・5 日/9 月 2・3・4・5 日	106	72	30 字 2 段 2 頁
1881	8	5 月 3・4 日/7 月 2・3 日/ 9 月 2・3 日/11 月 2・4 日	76	51	23 字 4 段 4 頁
1887	8	5 月 3・4 日/7 月 2・3 日/ 9 月 2・3 日/11 月 2・3 日	115	75	23 字 5 段 6 頁
1895	4	5 月 2 日/7 月 2 日/9 月 2 日/11 月 2 日	93	61	22 字 6 段 6 頁
1901	4	5 月 2 日/7 月 2 日/9 月 2 日/11 月 2 日	119	76	19 字 8 段 6 頁
1909	4	5 月 2 日/7 月 2 日/9 月 2 日/11 月 2 日	150	95	18 字 8 段 6 頁
1917	4	5 月 2 日/7 月 2 日/9 月 2 日/11 月 2 日	160	101	17 字 9 段 8 頁
1925	4	5 月 2 日/7 月 2 日/9 月 2 日/11 月 2 日	192	119	14 字 12 段 8 頁
合計	58		1,011	650	

②コーパスアノテーション

収録本文には、書誌情報（発行年月日、号数、記事分類、著者等）、文書構造情報（記事、見出し、文、ページ・段・行等）、テキスト属性情報（文体・書記体・引用等）、文字情報（ルビ・外字・踊り字・変体仮名等）を付与し、XML 形式のデータとして作成した（図 1）。さらに、既存の公開コーパス（国立国語研究所『日本語歴史コーパス』『現代日本語書き言葉均衡コーパス』など）と共通の枠組みによる形態論情報（単語の情報）付与を試みた。




```
<?xml version="1.0" encoding="UTF-8"?>
<text title="読売新聞" date="18870503" type="朝刊" issue="3691">
    <front>
        <s><pb n="1"/><cb n="1"/><lb/><r rt="よみ">讀</r><r rt="うり">賣</r><r rt="しん">新</r><r rt="ぶん">聞</r></s>
    </front>
    <body>
        <article title="工業者の注意" author="＊" style="文語" script="漢字ひらがな" ranmei="社説" genre="鉱工業">
            <titleBlock><s><lb/>〇<r rt="こう">工</r><r rt="げふ">業</r><r rt="しや">者</r>の<r rt="ちう">注</r><r rt=
            <p>
                <s><lb/><r rt="きん">近</r><r rt="らい">來</r><r rt="わが">我</r><r rt="くに">國</r>の<r rt="こう">工
                <s><r rt="わがはい">余輩</r><char type="hentaigana" ref="U+1B09E">は</char><r rt="さき">曩</r>に<r rt=
                <s><r rt="ゆえ">故</r>に<r rt="ちやく">着</r><r rt="じつ">實</r><char type="hentaigana" ref="U+1B081">
                <s><r rt="すなは">即</r>ち<r rt="こん">今</r><r rt="にち">日</r><r rt="かく">斯</r>の<r rt="ごと">如</
                <s><r rt="しかう">而</r>して<r rt="かく">斯</r>の<r rt="ごと">如</r>く<r rt="こう">工</r><r rt="げふ">
```

図１　データアノテーションの例

(2) 新聞コーパスに基づく計量的言語分析

次に、構築した新聞コーパスを用いて行った、近代新聞の語彙・変体仮名・ルビについての調査分析を報告する。

近代新聞の語彙

まず、新聞の語彙の全体像を把握するために、コーパス収録語全体（短単位語彙素の異なり語

数）の語種分布を調査した。調査には、データに付与した形態論情報のうち、「語種」情報を用いた。メディア間比較のために、『日本語歴史コーパス　明治大正編 I 雑誌』収録雑誌『明六雑誌』(1874-5 年)、『太陽』(1925 年）の語種比率と対照させて図 2 に示す。また参考として、現代語における新聞と雑誌の語種分布を『現代日本語書き言葉均衡コーパス』生産実態サブコーパスの新聞・雑誌データの調査結果に基づき掲載する。

収録期間初年の 1874-5 年では、雑誌と比較して漢語比率が 20%近く低いが、収録期間最終年の 1925 年ではその差を 10%弱まで詰めており、現代の新聞における漢語比率にも近接している。これは、読売新聞』に代表される当時の大衆新聞が、その誌面構成を一般大衆の事件中心から社会・国『際・政治の報道中心へとシフトさせたことや、それと並行して「談話体」という読者に向けて話し掛けるような文体から、漢文訓読文体にルーツを持つ「近代文語文」を経て、言文一致による口語文へと、文体を変化させながら、報道を中心とする公共メディアとしての新聞の位置を確立していったことに伴う、語彙の変化と見ることができた。

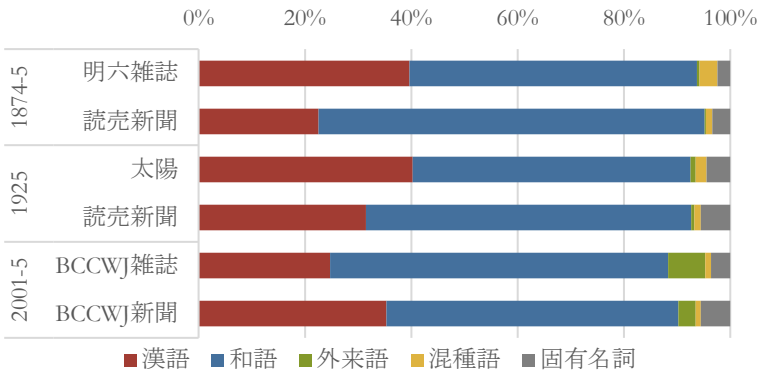


図 2　新聞コーパスと『日本語歴史コーパス　明治・大正編 I 雑誌』との語種分布

近代新聞における変体仮名

新聞における変体仮名の使用について調査を行った。計量的な全体把握に加えて、文脈による字体選択の背景分析も行った。調査には、データに付与した変体仮名 Unicode 情報を用いた。

収録期間初年の 1874-5 年では、同じ新聞記事の同一語の表記に、複数の仮名（現行ひらがなと変体仮名、あるいは同一音価の変体仮名間のバリエーション。動詞「する」連用形「し」に対する「し」と「志」(U+1B048)、助詞「が」に対する「ゝ」(U+1B019) と「う」(U+1B01A) など）が用いられており、用法の分化がなく、揺れの大きな状況が見られたが、仮名字体の統一を示した 1900 年の小学校令が出された翌年の 1901 年では、一部で揺れも残るものの、変体仮名は付属語に偏って使用されている傾向が見られた。その後、1909 年には消失していることも明らかになった。

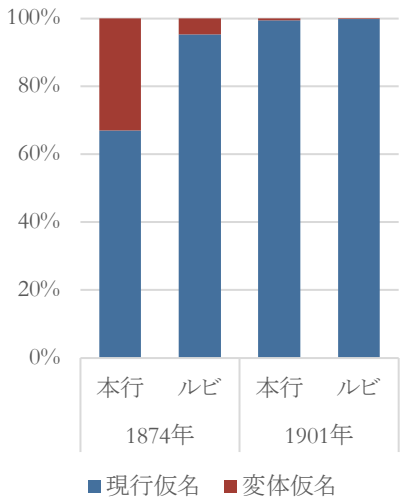


図 3 ひらがな「し」の表記分布

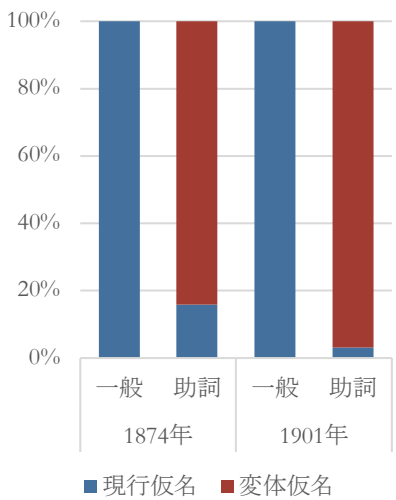


図 4 ひらがな「は」の表記分布

近代新聞におけるルビ

新聞におけるルビ（振り仮名）の使用について調査を行った。主に、二字以上の漢字から構成される語に付与されたルビ（熟字ルビ）を対象として、漢字表記語の標準的な読みを示したルビを「読みルビ」、漢字表記語に対して同義・類義の別語で言い換え意味を補足するルビを「意味ルビ」として分類し、総ルビが特徴でもある近代初期の大衆紙において、ルビがどのように用いられてきたかを、分析した。その結果、創刊当初 1874–5 年では、「政府：おかみ」「応接：つきあい」「説論：ゐけん」と言った意味ルビが、新聞の全体に渡って多く用いられているが、1887 年には、意味ルビが激減しており、意味ルビが用いられている例に関しては、半数が新聞小説内に現れる文学的効果としての修辞的なルビであることも明らかになった。

(3) 今後の展開

本研究は、日本語において重要なメディアである新聞を対象として、既に多くの蓄積のある国立国語研究所のコーパス群と共通の枠組みによる、形態論情報付きのコーパスを構築した点に特色がある。なお、本研究で構築済みの新聞コーパスは、今後『日本語歴史コーパス　明治・大正時代編』に組み入れる形で調整を行なっており、これが実現すれば、既存の雑誌、教科書等のコーパスとシームレスに比較対照することが可能となる。これまでコーパスを利用した近代語研究において問題となっていた資料の偏りを解消することに寄与する重要な成果と考える。

さらに、本研究は、変体仮名活字への Unicode コーディングの実践例としても、まだ手付かずの新しい試みであり、学術情報交換における変体仮名 Unicode の実用性や、コーディングに係る問題点の把握、検証に有用な知見を与えるものと考える。また、埋もれた人文知の活用のための取り組みとして、昨今特に興隆の見られる、「くずし字 AI」や「くずし字 OCR」といった、くずし字の自動認識システムに、テストデータ・正解データとして提供することで、システムの精度向上に寄与する可能性がある。

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　0件）

１．発表者名 間淵 洋子
２．発表標題 近代の新聞・雑誌に見られるルビの実態 ： 形態論情報アノテーションとの関わり
３．学会等名 「通時コーパス」シンポジウム2020
４．発表年 2020年

１．発表者名 間淵 洋子
２．発表標題 読売新聞に見る近代の語彙・表記
３．学会等名 国立国語研究所オープンハウス2019
４．発表年 2019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