明治・大正期『読売新聞』コーパスの構築と課題

間淵洋子（明治大学大学院国際日本学研究科）
mabuchi@meiji.ac.jp

1　はじめに

興隆し続けるさまざまなコーパスの開発や，それらを用いた自然言語処理・コーパス言語学・言語教育・辞書開発等多様な分野における研究の発展を背景に，近年，日本語史研究の分野へのコーパスの応用が期待されている。従来の国語学的研究手法に，コーパス言語学の手法を加えることで，新たな知見の獲得が期待できるためである。

現在，国立国語研究所を中心に構築・公開が進められている大規模な歴史的日本語資料のコーパス『日本語歴史コーパス』（以下『CHJ』）は，上代（奈良時代）から近代（明治・大正時代）までをカバーする「通時コーパス」を志向したコーパスであり[1]，このようなニーズに応え，日本語史研究への基礎資料を提供してくれる大変貴重な言語資源である。しかし，『CHJ』は，日本語史研究において代表的かつ重要な資料を収録対象としているために，カバーする時代の広さに加え，資料・分野が幅広く多数であり，その完成までには長い期間が必要とされる。現状は開発途上にあるため，ある特定の時代の言語のありさまを捉えようとする共時的な視点においては，資料の偏りが研究上のネックになると言わざるを得ない。これは，同研究所が公開する，現代語の「代表性」を志向して母集団からの精密なサンプリングにより収録資料をバランスして構築された現代語のコーパス『現代日本語書き言葉均衡コーパス』（以下『BCCWJ』）[2]と対照的な様相である。

そこで，発表者は，今日我々が用いている書記言語（＝現代日本語書き言葉）が確立される過程における語彙の変化を捉えることを目的とした時，『CHJ』に加えるべき資料として，明治・大正期の「新聞」を定め，独自にコーパスの構築を試みた。本発表では，試作した「読売新聞コーパス」の構築とその構築過程において明らかになった課題について述べる。

2　コーパスの設計

2.1　資料の選定

現在公開されている『CHJ　明治大正編 I 雑誌』（以下『CHJ 明治大正』）約 1,250 万語は，その大部分である 8 割近く（970 万語）が総合雑誌『太陽』の本文データによって占められている[3]。

これは，「現代の書き言葉の形成と確立」という日本語史における大きなトピックを扱う際，①漢語を中心とする新しい語彙の創造と定着，言文一致による口語文の創成と普及という，語彙と文体の大きな変化を経て現代語が確立した時期＝20 世紀初期（明治後期から大正時代）が極めて重要な時期であること，②当時の日本語を把握する材料として多種多様な資料が存在する中，ジャンルが広く，著者層や読者層が厚く，分量も多い，総合雑誌としての『太陽』が，十分な規模と多様性を持つという面で当時の日本語を代表できる資料としての価値を持つこと[4]から，これを起点に国立国語研究所の近代語のコーパス開発が進んだことによる。その点で，『CHJ 明治大正』は，設計段階において当時の日本語としての「代表性」という観点が考慮されたものではあるが，ここに現れる言語実態が，『太陽』という特定の雑誌，あるいは，他の収録雑誌を含めても「雑誌」という特定のメディアの特徴である可能性を否定できない。

そこで，この時期の日本語語彙の実態として，メディアによる言語的差異がどの程度存在しているかを検討するための資料として，雑誌以外のメディアの日本語を映すコーパスを作成することにした。メディアの選定においては，雑誌に並んで，読者層の厚さ，著者の多さ，ジャンルの広さ，分量の多さなどの点で言語の多様性を有する資料として「新聞」というメディアに着目した。中でも，当時，高い漢学的素養を身につけた伝統的知識人（旧武士階級，豪農・豪商など）を読者層として政論を主体とした大新聞（『東京日日新聞』（のちの『毎日新聞』）『郵便報知新聞』（のちの『報知新聞』）等）ではなく，教育を受けない庶民層にも多く読まれ，市井の事件に関する記事を中心とした小新聞（『平仮名東京絵入新聞』（のちの『東京絵入新聞』）『仮名読新聞』等）[5]を対象とすることとし，かつ，現代語との比較を念頭に，現代まで引き続き刊行されており『BCCWJ』の収録対象にもなっている新聞として『読売新聞』（1874 年 11 月創刊）を選定した。

コーパスの原資料としては，読売新聞社が提供する有料データベース『ヨミダス歴史館』から収録対象日の記事画像を取得し，これを用いた。

2.2　収録言語量

本コーパスは発表者が個人で作成するものであるため，『CHJ 明治大正』のような大規模なものは作成できないが，『CHJ 明治大正』との比較に適した言語量が必要である。そこで，『CHJ 明治大正』に包含されるスモールセットとして設定されているコアデータを基準に言語量を定めることとした。

コアデータは，形態論情報を人手により修正し高い精度を保証したデータセットである。『CHJ 明治大正』の収録資料のうち，『明六雑誌』はその全体（約 18 万語）が含まれるが，その他の資料は，文体（文語文・口語文）やジャンル（文芸・非文芸），記事長のバランスを取って各資料（『太陽』については 5 カ年の各収録年）につき約 3~4 万語程度になるようサンプリングによって選定した記事から構成されている[6]。

これらの語数に準じて，本コーパスでは収録対象年次ごとに 3〜4 万語を目安に収録することとした。

2.3　収録年次と収録対象範囲

本コーパスは，『CHJ 明治大正』とのメディア間比較を目的として作成するため，『CHJ 明治大正』の収録年次に合わせて，1874-5 年，1887 年，1895 年，1901 年，1909 年，1917 年，1925 年の 7 カ年を収録対象年次とした。

収録対象文書範囲は，多様性確保の観点から，欄などを限定せず新聞 1 日分全体を収録することとし，目標とする 1 カ年 3 万語程度を目安に各年 2〜5 日分を収録サンプル（号。1 号につき 6-8 頁）として選定した。

具体的には，特殊な時期（年・年度・月の初めや末）を避け，原則として各年 5 月 2 日と 11 月 2 日の 2 日間分を収録対象号と定めた。2 日間で語数が 3 万語に届かない場合（1 号の分量が少ない 1874-5 年および 1887 年のもの）は，語数が充足する程度まで取得基準号に続く号も収録対象とした。

2.4　収録対象

『CHJ 明治大正』に準じて，収録対象とする文書要素を定めた。具体的には，記事本文を主対象とし，『読売新聞』をはじめとする小新聞に特徴的に見られるルビ（傍訓）も漏れなく収録対象とした。

広告は新聞本文と言いがたいため，収録対象外として除き，言語表現を主体としない要素は採録対象外として，図表を除いたほか，『CHJ 明治大正』と異なる仕様として，固有名や数値主体の要素として，人名の羅列（叙任・辞令情報等）やスポーツ欄における試合結果の羅列なども除いた。

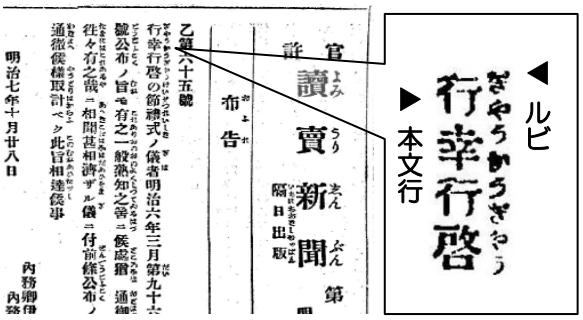

図 1：新聞紙面の例（1874 年 11 月 2 日創刊号）

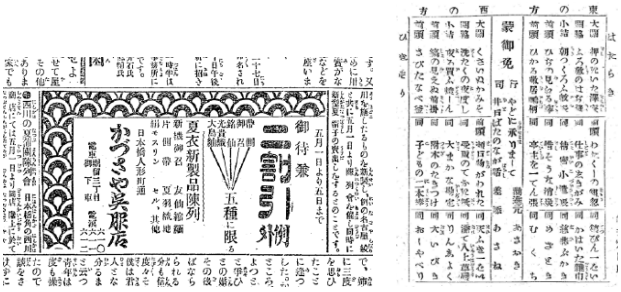

図 2：収録対象外の広告の例（左：1917 年 5 月 2 日「讀賣婦人附録」面）および，図表の例（右：1874 年 11 月 16 日「投書」欄）

3　コーパスの構築

3.1　文字化

文字化は，入力用文字セットに JISX0213:2004 を用い，この範囲で原資料を忠実に電子化することとした。

漢字は，上記 JIS 規格内の全ての文字を区別して用い，それ以外の文字は，同規格の包摂基準に則して包摂するか，それに当てはまらない場合は「=」で入力した。この点は，『CHJ 明治大正』の文字仕様[7]が，漢字において JIS 規格内の文字のうち，「康熙別掲字(「社」「褐」など 104 字)」「UCS 互換字(「叱」「嘘」など 10 字)」を用いず，常用漢字に包摂して文字化されている点と相違がある。

また，初期の紙面では特に変体仮名活字がバリエーション豊富に用いられているが，これらは現行の平仮名に置き換えて入力した。

図 3：現行の仮名文字に置き換えた変体仮名の例

3.2 アノテーション

コーパスには，以下（主なものを抜粋）のアノテーションを加え，XML 形式（稿末図 4 参照）で情報を格納した。

書誌情報

発行年月日，紙面記載の通号

分類　（『ヨミダス歴史館』における記事のジャンル分類。「皇室」「健康」「宗教」「情報」等）

文書構造情報

冊　（新聞 1 号分の文書範囲）

記事（内容のまとまりにより弁別される文書範囲。『ヨミダス歴史館』の記事範囲による）

文　（１文の範囲）

頁・段・行　（各要素の区切り目，頁数・段数）

テキスト属性情報

文体　（文語体・口語体の別）

書記体（漢字ひらがな混じり文・漢字カタカナ混じり文の別）

引用　（会話文・典拠引用の範囲）

文字情報

外字　（JIS 外字を「=」で電子化した文字と字形情報）

ルビ　（ルビ付与元の本文文字列と対応するルビ文字列）

変体仮名（変体仮名を現行仮名に置換した文字。字形の情報はなし）

欠損　（紙面の保存状態等により読み取りが不可能な文字）

踊り字　（踊り字を読みに則して展開した場合の本文文字列と原文の文字列）

3.3 形態素解析

『CHJ 明治大正』との比較を目的として，UniDic 体系（短単位）に基づく形態素解析を施した。解析ツールとして『Web 茶まめ』[8]を用いた。

表 1：明治大正期『読売新聞』コーパス収録語数

発行年	記事数	総計	和語	漢語	外来語	混種語	固有名詞
1874-5	279	42,129	30,606	9,513	62	505	1,443
1887	206	36,827	23,470	10,828	126	513	1,890
1895	121	32,894	18,899	11,625	72	473	1,825
1901	139	35,234	20,806	11,937	111	594	1,786
1909	135	44,932	26,981	14,666	198	744	2,343
1917	146	54,452	32,714	18,374	207	872	2,285
1925	190	54,091	33,075	17,028	352	633	3,003

『読売新聞』は，本コーパス収録年次のうち，1874-5 年及び 1925 年が口語体を主体とし，1887 年から 1917 年は文語体を主体としている。そのため，『Web 茶まめ』で用いる解析辞書にはそれぞれ，前者に「旧仮名口語 UniDic」を，後者に「近代文語 UniDic」を指定して解析を行い，一部に解析結果の修正を加えた。

解析結果に基づく，コーパスの収録語数と語種内訳を年次別に示す（表 1）。

3.4 コーパス構築における課題

文字化やアノテーションに関して，今後研究に利用する際，検討すべき課題として以下の点が挙げられる。

1) 漢字の字体包摂：『CHJ 明治大正』で常用漢字への包摂対象となっている「康熙別掲字」「UCS 互換字」については，比較に際して仕様の差が問題となるほか，当該の字体が解析辞書に登録されていないため，『Web 茶まめ』による自動解析において解析誤りの原因になる。本コーパスを用いた漢字字体の変遷調査等の研究トピックを考慮すれば，字体情報を保持した上での解析可能な常用字体への変換が妥当と思われる。
2) ルビの活用：小新聞の特徴であるルビの多様性と形態論情報付与との関連，活用方法についての検討が再重要課題である。特に，『読売新聞』のルビが，語の「読み」ではなく「意味補足」の性質が強い点を考慮したアノテーションが必要となる。
3) ジャンル情報の整理：『CHJ 明治大正』は大部分のデータに NDC によるジャンル情報が付与されている。今回ジャンル情報付与に用いた「ヨミダス歴史館」のジャンル情報を，これに則して NDC に割り当てることで，ジャンル情報を含めた分析が可能となる。
4) 変体仮名の電子化：現状の『CHJ 明治大正』との比較に留まらず，今後コーパスを表記研究に用いる可能性等を考慮し，変体仮名の電子化の必要性・方法について検討・対応が必要である。
5) 原資料の保存状態と電子化の限界：紙質や活字の小ささ等の新聞の特徴から，古いものほど読み取り困難な箇所が多くなり電子化が難しい。このような資料における電子化が，どのように実現可能かを探る必要がある。
6) データの拡充：1 カ年 3-4 万語程度の言語量は，語彙の分布，文末表現形式等の分析においては

利用可能であるものの，特定の語の分布や変化を扱う研究等においては活用が困難である。量的な拡充が必須である。

4　まとめ

以上，『CHJ 明治大正』とメディア比較が可能な補完データとして開発した明治大正期の『読売新聞』のコーパスの構築とその課題について述べた。

本コーパスは，7 カ年，各 3-4 万語程度，総計約 30 万語からなるコーパスである。表記研究への利用可能性を考慮し，漢字表記・仮名表記・ルビといった表記に関する情報を豊富に有したコーパスとなっている点が特徴的であり，構築における課題もこの点に関して多かった。特に，既存コーパスとの連携や形態素解析上の問題，情報の活用等について更に検討する必要性が明らかになった。

発表者は，近代語コーパスを資料とした漢語語彙の使用度・表記・語法の変化を研究対象としており，特に表記については，現代語においてもメディア（レジスタ）の差が大きい言語事象であるため，近代語におけるメディア比較を目的として本コーパスの構築を試みた。本コーパスを研究に適用した結果としては，観察対象とする個々の漢語の用例が充分に確保できないという問題点があったが，構築過程で，メディアの特性や構築における課題などが浮き彫りになった。

今後は，コーパスの拡充を進めながら，個々のメディアに特徴的な言語事象を適切に映すコーパスの仕様や構築方法について，更に検討を重ねたい。

付記

本研究は，JSPS 科研費 16J08872「コーパスを利用した近現代漢語の表記・語法の多様性に関する計量的・通時的研究」(代表：間淵洋子，研究期間：2016-04-01～2018-03-31)による補助を受けた。

文献

[1] 小木曽智信(2016)「『日本語歴史コーパス』の現状と展望」『国語と国文学』93(5): 72-85.

[2] 前川喜久雄(2008)「KOTONOHA 『現代日本語書き言葉均衡コーパス』 の開発 (〈特集〉資料研究の現在)」『日本語の研究』4(1): 82-95.

[3] 日本語歴史コーパス（CHJ）語彙統計：バージョン 2017.3（2018 年 1 月 15 日確認：http://pj.ninjal.ac.jp/corpus_center/chj/201703.html）

[4] 田中牧郎(2005)「言語資料としての雑誌『太陽』の考察と『太陽コーパス』の設計」国立国語研究所編『雑誌『太陽』による確立期現代語の研究:『太陽コーパス』研究論文集』: 1-48.博文館新社.

[5] 山本武利(1981)『近代日本の新聞読者層（叢書・現代の社会科学）』法政大学出版局.

[6] 近藤明日子・間淵洋子(2016)「「明治・大正編 I 雑誌」の構築と課題（ワークショップ 『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 年度春季大会予稿集』:251-254

[7] 須永哲矢(2012)「近代語文献を電子化するための異体字処理」国立国語研究所『近代語コーパス設計のための文献言語研究　成果報告書（国立国語研究所共同研究報告 12 - 03)』:65-82.

[8] 堤智昭・小木曽智信(2015)「歴史的資料を対象とした複数の UniDic 辞書による形態素解析支援ツール『Web 茶まめ』」『じんもんこん 2015 論文集』:179-184.

```
<?xml version="1.0" encoding="UTF-8"?>
<text title="読売新聞" date="18870503" type="朝刊" issue="3691">
    <titleBlock>
        <s><pb n="1"/><cb n="1"/><lb/><ruby rubyText="よみ">讀</ruby><ruby rubyText="うり">賣</ruby><rub
    </titleBlock>
    <article title="［社説］工業者の注意" author="＊" style="文語" script="漢字ひらがな" genre="鉱工業">
        <s><lb/><ruby rubyText="こう">工</ruby><ruby rubyText="げふ">業</ruby><ruby rubyText="しや">者</
        <s><lb/><ruby rubyText="きん">近</ruby><ruby rubyText="らい">來</ruby><ruby rubyText="わが">我</
        <s><ruby rubyText="わがはい">余輩</ruby><hentaiKana>は</hentaiKana><ruby rubyText="さき">曩</rub
        <s><ruby rubyText="ゆえ">故</ruby>に<ruby rubyText="ちやく">着</ruby><ruby rubyText="じつ">實</r
        <s><ruby rubyText="すなは">即</ruby>ち<ruby rubyText="こん">今</ruby><ruby rubyText="にち">日</r
        <s><ruby rubyText="しかう">而</ruby>して<ruby rubyText="かく">斯</ruby>の<ruby rubyText="ごと">如
        <s><ruby rubyText="また">又</ruby><ruby rubyText="あつ">厚</ruby>く<ruby rubyText="こ">之</ruby>
        <s><lb/><ruby rubyText="また">又</ruby><ruby rubyText="たま">偶</ruby>に<ruby rubyText="き">技</
```

図 4 : XML ファイルによるアノテーション例