『日本語歴史コーパス　明治・大正編Ｖ新聞』（短単位データ 0.8）概説書

2023 年 3 月 31 日　更新　髙橋雄太

１．はじめに

『日本語歴史コーパス　明治・大正編Ｖ新聞』（短単位データ 0.8）は、明治初期に創刊された『読売新聞』の明治大正期間に刊行された一部をコーパス化したものである。本文書では、コーパスの設計方法、テキストの仕様、コーパス検索アプリケーション「中納言」の検索結果に表示されるテキストおよびアノテーション（テキストに付与する付加情報）の項目について、その概要を示す。

※2023 年 3 月に、短単位データ ver.0.7 から ver.0.8 にデータ更新を行いました。一部書誌情報に誤りがあったものを修正したほか、形態論情報の精度向上を実施しました。

２．コーパスの概要

２．１　収録範囲の設計

本コーパスは、共著者である間淵洋子の科学研究費の助成を受けた個人開発の『読売新聞』の通時的コーパスの構築に始まり、国立国語研究所の通時コーパスプロジェクトが協力することで CHJ での公開に至った。コーパス設計の基本方針や考え方については、間淵（2018）を参照されたい。

「明治・大正編Ｖ新聞」は「明治・大正編Ｉ雑誌」とのメディア間比較を可能とするため、「Ｉ雑誌」と同じ 1875 年、1881 年、1887 年、1895 年、1901 年、1909 年、1917 年、1925 年の 8 か年を収録対象年次とした。「Ｉ雑誌」のコアデータ（人手修正が全編にわたって施されたデータセット、各年およそ 3~4 万語）程度と同程度の語数を各年で確保できるよう、特殊な時期（年・年度・月の始めおよび末）を避けて、5 月 2 日と 11 月 2 日と定めた（2 日が休刊日の場合、3 日を採用した）。収録対象とする範囲は各号（1 日分）全体とした。ただし、明治期は 1 号あたりの頁数（語数）が少ないため、語数が充足する程度まで 3 日以降の続く号も収録対象とし構築を進めた。本コーパスの収録年月日と、各年の記事数と短単位の概数の一覧を、表 1 に示す。

表 1　収録年月日と記事数・短単位数の一覧

収録年	収録月日	記事数	短単位数（万）
1875（明治8）年	5月2、3、4、5、7、8、9日・11月2、4、5、7、8、9、10日	314	4.4
1881（明治14）年	5月2、4、5、6日・11月2、4、5、6日	291	5.1
1887（明治20）年	5月3、4、5日・11月2、3日	243	4.6
1895（明治28）年	5月2日・11月2、3日	203	5.1
1901（明治34）年	5月2、3日・11月2日	205	5.6
1909（明治42）年	5月2日・11月2日	123	4.7
1917（大正6）年	5月2日・11月2日	161	5.7
1925（大正14）年	5月2日・11月2日	194	6.2
計		1734	41.4

２．２　　形態論情報の精度

本コーパスは全編が、人手修正は入っているが部分的に形態素解析の結果のままを残している「非コアデータ」である。短単位の形態論情報は、形態素解析辞書 UniDic を使用した形態素解析に基づき、一部を人手により修正することで付与した。「明治・大正編Ｖ新聞」の短単位バージョン 0.8 における形態論情報の精度（適合率）は、非コアデータが 98.1%となっている[1]。

３．サンプル ID

テキストをコーパスに収録する際にテキストを一定の範囲で分割する必要があるが、その各範囲をサンプルと呼ぶ。本コーパスのサンプル単位は、各作品の最も細かい構成要素（記事相当）に分割した、その各文書要素である。各サンプルを一意に特定する ID の構成を表 2 にあげる。

表 2　サンプル ID の構成

桁数	値	説明
1~2	60	時代区分を表わす。すべて「60」で、「明治・大正」を表わす。
3	P	サブコーパスのジャンル（新聞、PaperのP）を表わす。全サンプルで共通。
4~5	作品ID	「読売新聞」の「読売」を共通で表わす。
6~9	（4桁の数字）	サンプルの発行年を西暦で表わす。
10	_	サンプルIDの区切り記号（アンダーバー）。
11	（1桁の数字）	発行月を32進数で表わす（1~9月はアラビア数字で、10月以降はABC…に対応する)。
12	（2桁の数字）	発行日を32進数表わす（1~9日はアラビア数字で、10日以降はABC…に対応する)。
13~15	（3桁の数字）	記事の通し番号を表わす。

表 2 の基準によると、例えば、1875 年の 5 月 2 日の 1 記事目のサンプル ID は「60P 読売 1875_52001」に、1875 年 11 月 10 日の 10 記事目のサンプル ID は「60P 読売 1875_BA010」になる。

４．テキスト

４．１　テキストに使用する文字

本コーパスの電子化テキストに使用した文字の範囲は、JIS X 0213（JIS の文字コード規格）の文字集合（JIS 漢字の第 4 水準までを含む）に準拠した。

文字集合に含まれない変体仮名については文字集合内の仮名によって電子化し、文字集合に含まれない記号類は、形・用途の近い文字集合内の記号によって電子化した。また、底本の文字のかすれや破損・抹消によって判読が困難な文字・記号は、「␣」（空白記号、JIS 面区点 1-07-93、U+2423）によって表した。

文字集合に含まれない漢字については、以下の（1）～（4）の手順で電子化した（須永・堤・高田

[1] ここでいう精度（適合率）は、（調査対象とした）整備済みコーパスの語数で、そのうちの正解語数を除した値である。語形、活用型、活用形のみの誤りも含む。

2011、須永ほか 2013)。

(1) JIS X 0213 の「6.6.3 漢字の字体の包摂規準」に若干の拡張を施した近代語コーパス用の包摂規準に基づいて、JIS 内の文字に包摂する。
(2) (1) の包摂規準を適用できない字形差をもつ漢字は、類似の意味・用法を持つ同音・同訓の JIS 内の文字で代用する。
(3) JIS X 0213 中の「Unicode における CJK 統合漢字拡張 B」(サロゲートペアの文字) を文字集合に含めるほか、コーパス文字集合外の Unicode 文字に同一字体があればそれで入力する。
(4) (1) ～ (3) の手順で入力ができない場合は、外字として「=」(げた記号、JIS 面区点 1-02-14、U+3013) で表す。

４．２　テキストの校訂

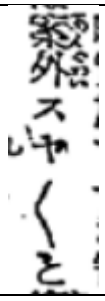
本コーパスでは、『日本語歴史コーパス』の他のサブコーパスや『現代日本語書き言葉均衡コーパス』等の国立国語研究所構築の他のコーパスと斉一な形態論情報を付与するため、形態素解析辞書 UniDic を使用した形態素解析に基づき形態論情報を付与した。そのため、コーパスのテキストを UniDic による形態素解析に適したものとするため、底本のテキストに対して以下の A) ～C) にあげる改変 (ここでは「校訂」と呼ぶ) を施し、コーパスのテキストを作成した。

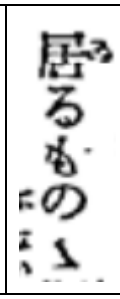
なお、「中納言」では、校訂後のコーパスのテキストと同時に、校訂前のテキストを底本の状態に近い形で電子化したものを「原文 KWIC」「原文文字列」として表示させることができるほか、底本の画像リンクから底本の字形を参照することもできる。利用に際しては、必要に応じて「原文 KWIC」「原本文字列」や底本画像を確認されたい。

A) 踊り字

踊り字は繰り返される文字列に置き換える。ただし、「国々」「人ゝ」等、1 短単位内部で直前の 1 字を繰り返す「々」「ゝ」は置き換えの対象としない。

表 3　踊り字の電子化例

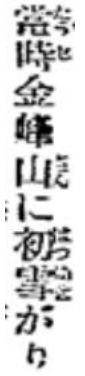
種類	例	コーパステキスト	原文 KWIC・原本文字列
ノ＼		案外スヤ**スヤ**と	案外スヤ**ノ＼**と。

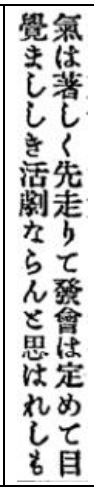
ゝ		居るもの**の**	居るもの<u>ゝ</u>

B）誤植

原文の誤植（脱字、衍字、前後文字列の転倒、誤字）と思われる表記は、訂正する。ただし、仮名遣いの誤りや、語形のバリエーション、当時通用していたと考えられる同音漢字による異表記などは、訂正の対象としない。

表 4　誤植の電子化例

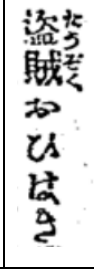
種類	例	コーパステキスト	原文 KWIC・原本文字列
脱字		當時金峰山に初雪が**あ**り	當時金峰山に初雪がり
誤字		…も商談は依然として捗取らざる爲め<u>人</u>氣は益々…	…も商談は依然として捗取らざる爲め<u>入</u>氣は益々…

衍字		…は著しく先走りて發會は定めて目覺ましき活劇ならんと思はれしも…	…は著しく先走りて發會は定めて目覺ましき活劇ならんと思はれしも…

C）濁点落ち

濁音が期待される仮名に濁点付き仮名が用いられていない場合は、濁点の無表記と判断し、該当の濁音を表す濁点付き仮名に置き換える。

表 5　濁点落ちの電子化例

種類	例	コーパステキスト	原文 KWIC・原本文字列
濁点落ち		盗賊おひは**ぎ**	盗賊おひは**き**

５．形態論情報

本コーパスでは、原則として底本の本行のテキストを主本文（主たる本文）として、それに対して形態論情報（語彙素・語彙素読み・品詞・活用型・活用形等の語に関する情報）を付与した。テキストの読みは右ルビのある場合はそれに拠った。

形態論情報は短単位のみ付与しており、長単位は未実装である。短単位の形態論情報は、形態素解析辞書 UniDic を使用した形態素解析に基づき、人手により修正することで付与した。

また、「Ⅲ明治初期口語資料」「Ⅳ近代小説」と同様に、短単位をまたがるルビ（例：有之〔コレ｜アリ〕、親父〔オ｜トッ｜サン〕）が付された、初期の『読売新聞』に多く見られる事例についても、同一文字列に複数の形態論情報を付与する機能を用いて、原文の文字列を保持しながら、読み通りの形態論情報を付与している。従来のコーパスでは、短単位の規程を外れて例外的に語彙素として認めた「之有る」を適用する、あるいはテキストに返読を施して「之有」と校訂する必要があったが、本コーパスにおいては「有之」の文字列を維持しつつ、語彙素「此れ」で検索をかけても、語彙素「有る」で検索をかけても同一のレコードが検出されるよう形態論情報が付与されてい

る。

６．「中納言」上の表示項目

本コーパスでは、テキストおよびアノテーションのデータは、コーパス検索アプリケーション「中納言」での検索結果の形で利用者に提供する（図 1）。


188 件の検索結果が見つかりました。
検索対象語数：406,663　記号・補助記号・空白を除いた検索対象語数：385,627

サンプル ID	開始位置	連番	コア	前文脈	キー	後文脈	語彙素読み	語彙素	語形	品詞
60P読売 1875_52011	2090	1290	0	の金でするのはかまはぬかぞんじませんが餬に見苦しい仕業と思ふゆゑ	新聞	紙を借りて四方の君子へうかがひます＃横濱寄留某＃白黒の何かあらそふ其はては腮に鎖で石やふるはん	シンブン	新聞	シンブン	名詞-普通名詞-一般
				でするのはかまはぬかぞんじませんが餬に見苦しい仕業と思ふゆゑ	しんぶん 新聞	紙を借りて四方の君子へうかがひます＃横濱寄留某＃白黒の何かあらそふ其はては腮に鎖で石やふるはん				
60P読売 1875_52012	2400	1540	0	が赤熊へうつり燃出して大騒ぎが有りましたといづれも銀座二丁目の繪入	新ぶん	にごく面白く畫て有りましたが此外に中島座で紅皿かけ皿の狂言に	シンブン	新聞	シンブン	名詞-普通名詞-一般
				へうつり燃出して大騒ぎが有りましたといづれも銀座二丁目の繪入	しん 新ぶん	にごく面白く畫て有りましたが此外に中島座で紅皿かけ皿の狂言に				
60P読売 1875_52012	4550	2850	0	とて人を賴んで頭を叩かせずと濟みそふなものだ＃其くらゐなら讀うり	新ぶん	へ賴んで附録にでも出して貰ツた方が早く人に知れますものを	シンブン	新聞	シンブン	名詞-普通名詞-一般
				かに名が賣りたいとて人を賴んで頭を叩かせずと濟みそふなものだ＃其くらゐなら讀うり	しん 新ぶん	へ賴んで附録にでも出して貰ツた方が早く人に知れますものを醫者				
60P読売 1875_52012	4910	3090	0	附録にでも出して貰ツた方が早く人に知れますものを醫者さまが	新ぶん	で名を賣るのも今にさがし出して投書と出かけましやう＃淺草柴井恭元	シンブン	新聞	シンブン	名詞-普通名詞-一般
				で附録にでも出して貰ツた方が早く人に知れますものを醫者さまが	しん 新ぶん	で名を賣るのも今にさがし出して投書と出かけましやう＃淺草柴井恭元				

図 1　「中納言」の検索結果のイメージ

「中納言」の検索結果で表示されるテキスト・アノテーションのうち、初期設定で表示される項目と、本コーパスで特に注意が必要な項目である「多重化種別」「文体」「出版社」（表中に「*」を付す）について、表 6 に内容を示す。

表 6　「中納言」検索結果の主な表示項目

情報種別	項目名	内容
コーパス情報	サンプル ID	検索対象の含まれるサンプルの ID（3 節参照）。
	開始位置	検索対象の含まれる短単位の先頭の文字の、サンプル内における位置を表す ID。10 きざみの連番。
	連番	検索対象の含まれる短単位の、サンプル内における位置を表す ID。10 きざみの連番。
	コア	検索対象の含まれるサンプルが非コアデータであることを表す。「0」が非コアを表す。
	多重化種別*	「掛詞」や「振り仮名」などの、多重化を行う要因を表す。本コーパスでは、全件が「振り仮名」である。
形態論情報	前文脈	検索対象の前方文脈。
	キー	検索対象の含まれる短単位の書字形出現形（表記形）。
	後文脈	検索対象の後方文脈。
	原文 KWIC	上記項目「前文脈」「キー」「後文脈」に対する、校訂前の底本に近い形のテキスト（4.2 節参照）。
	語彙素	検索対象の含まれる短単位の語彙素の表記。語彙素は、単語の様々なバリエーション（語形、活用形、表記形など）を統合した辞書の見出しに相当するもので、一般の和語・漢語は漢字平仮名表記、外来語・人名・地名は片仮名表記である。

情報種別	項目名	内容
形態論情報	語形	検索対象の含まれる短単位の語形。語形は、語彙素では統合される語形の別（例：語彙素「矢張り」に対する「ヤハリ」「ヤッパリ」など）や活用型の別（例：語彙素「読む」に対する「ヨム（五段-マ行)」「ヨム（文語四段-マ行)」「ヨメル（下一段-マ行；可能動詞形)」など）等を区別した語の個々の形に相当する。片仮名表記である。
	品詞	検索対象の含まれる短単位の品詞で、UniDic の体系に基づく。学校文法における「形容動詞」は、語幹が「形状詞」、活用語尾が「助動詞」に分割される点に注意が必要である。 このほか、本コーパスに含まれる、UniDic の体系に基づかない特殊な品詞には以下の種類がある。 言いよどみ…会話の中での言いよどみにあたる文字列。 例：お、親方様、ゑゝありがたうござりまする、 漢文 …漢文の文字列。 外国語 …外国語の文字列。 欠損 …原文の欠損、かすれにより判読できない文字列。コーパステキストでは「␣」で表示される。 読取不可 …原文の文字潰れにより判読できない文字列。コーパステキストでは「**=**」で表示される。 絵文字・記号等 …入力のできない絵文字や企業マークなど。コーパステキストでは「**=**」で表示される。 未知語 …形態論情報の付与を保留した文字列。
	活用型	検索対象の含まれる短単位の活用の型。活用語の場合のみ表示される。口語活用は活用の型と行で「五段-サ行」のように、文語活用は「文語」が加わり「文語四段-サ行」のように示される。検索対象の「文体」項目の値が「文語」である活用語には文語活用型を、「口語」である活用語には口語活用型を割り当てる。
	活用形	検索対象の含まれる短単位の活用形。活用語の場合のみ表示される。
	原文文字列	検索対象の含まれる短単位の、校訂前の底本に近い形のテキスト（4.2 節参照)。
	振り仮名	検索対象の含まれる短単位に付された振り仮名（右ルビ）の文字列。振り仮名の誤植は校訂したものを示す。校訂の基準はテキスト校訂における誤植の判定に準ずる。
	本文種別	検索対象の含まれる文が「地の文」以外の場合の、その種別。以下の種類がある。なお、地の文の場合、当項目は空白となる。 会話 …会話・独話・心内発話等の引用 引用 …文献等からの引用 その他 …漢文等

情報種別	項目名	内容
形態論情報	話者	上記項目「本文種別」が「引用」の場合の典拠文献名や著者名、「会話」の場合の話者名や属性名（男、先生など）を表す。不明の場合は「＊」で示す。
	文体*	検索対象の含まれる文の文体。以下の種類がある。 文語…文語体。文末辞が「なり」「たり」「つ」「ぬ」「き」「けり」のもの。 口語…口語体。文末辞が「だ」「ぢや」「である」「です」「ます」のもの。 漢文…漢文。 外国語…漢文以外の外国語の文。 なお、一つのサンプル内で複数の文体が混在しているものは、地の文および引用範囲ごとにそれぞれ文体を付与する。
本文情報	ジャンル	検索対象の含まれるサンプルの文章内容に基づく分類。小説や詩歌には「文芸」、それ以外のサンプルには「非文芸」が表示される。
	作品名	検索対象の含まれるサンプルが収録された資料名。全て「読売新聞」と表示される。
	成立年	検索対象の含まれるサンプルが収録された年。
	巻名等	検索対象の含まれるサンプルが収録された資料の編名・巻名、およびサンプルのタイトル。本コーパスでは記事タイトルを表示する（記事タイトルのないものは無表示である）。
作者情報	作者	検索対象の含まれるサンプルの著者名。著者名の認定は、底本テキストの記載に基づく。ただし、現在一般的に知られている呼称に変えた場合がある。 「国立国会図書館典拠データ検索・提供サービス（Web NDL Authorities）」のウェブページでの著者情報へのリンクを付与している。
	生年	検索対象の含まれるサンプルの著者の生年。西暦 4 桁で示す。
底本情報	底本	検索対象の底本（原資料）。「読売新聞<1875-05-02-第 90 号>」のように、<>内に年、月、日、号数を示す。
	ページ番号	検索対象の底本におけるページ（紙面）番号。
	出版社*	底本の出版社を示す。本コーパスでは、1917 年 11 月までのサンプルには「日就社」、1925 年のサンプルには「読売新聞社」が表示される（1917 年 12 月に屋号を「日就社」から「読売新聞社」に変更）。
その他	底本リンク	検索対象の底本画像へのリンク。ヨミダス歴史館へのリンクをヨミダスボタンで示す。
	参照リンク	検索対象の底本以外の参照本画像へのリンク。本コーパスでは該当画像がないため空欄である。

付記

本コーパスは、国立国語研究所共同研究プロジェクト「通時コーパスの構築と日本語誌研究の新展開」（2016-2021）、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）、JSPS 科研費 20K13060「新語彙定着期の言語変化―コーパスに基づく通時的語彙研究の実践」（代表：間淵洋子）の研究成果を報告したものである。

参考文献

須永哲矢・堤智昭・高田智和（2011）「明治前期雑誌の異体漢字と文字コード―『明六雑誌』を事例として―」『じんもんこん 2011 論文集』pp.381-388.

須永哲矢・堤智昭・近藤明日子・木川あづさ・服部紀子（2013）「明治中期雑誌の異体漢字と JIS 漢字―『国民之友』を事例として―」『じんもんこん 2013 論文集』2013(4)、pp.201-208.

間淵洋子（2018）「明治・大正期『読売新聞』コーパスの構築と課題」『言語処理学会　第 24 回年次大会発表論文集』pp.500-503 https://anlp.jp/proceedings/annual_meeting/2018/pdf_dir/P4-4.pdf

関連 URL

UniDic https://unidic.ninjal.ac.jp/

コーパス検索アプリケーション「中納言」 https://chunagon.ninjal.ac.jp/

『日本語歴史コーパス』 https://ccd.ninjal.ac.jp/chj/

国立国会図書館典拠データ検索・提供サービス（Web NDL Authorities） http://id.ndl.go.jp/auth/ndla/

読売新聞社「ヨミダス歴史館」https://database.yomiuri.co.jp/about/rekishikan/