『昭和・平成書き言葉コーパス』（バージョン 2023.5）新聞レジスター　概説書

2023年12月5日　間淵洋子

１．はじめに

『昭和・平成書き言葉コーパス　新聞』（以下、SHC新聞）は、明治初期に創刊され今日まで続く『読売新聞』の昭和・平成期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、収録資料、テキストの仕様、コーパス検索アプリケーション「中納言」[1]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。

２．収録資料

本コーパスでは、『日本語歴史コーパス　明治・大正編Ｖ新聞』[2]（以下、CHJ明治・大正編Ｖ新聞）に接続して利用でき、また、戦前から戦後、さらに現在にかけての公共的な書き言葉の変化を捉えることができる資料として、「CHJ明治・大正編V新聞」が収録対象とする『読売新聞』を引き続き収録対象とし、「CHJ明治・大正編V新聞」がおよそ8年おきにテキストを収録するのに倣い、1933・1941・1949・1957・1965・1973・1981・1989・1997・2005・2013の各年を収録年次とした。また、1か年につき原則として5月2日・11月2日の2日分の全国版朝刊1冊のテキストを取得した「明治・大正編V新聞」を拡張し、より多くのデータ量を確保することを目的に、原則として奇数月2日（2日が休刊日の場合は3日）の計6日分の全国版朝刊1冊を対象とした。本コーパスの収録年月日と、次節に述べるサンプル単位の別、サンプル数、短単位数を表1に示す。

表1　収録年月日の一覧

収録年	収録月日	サンプル単位	サンプル数	短単位数（万）
1933(昭和 8)年	5月2日・7月2日・9月2日・11月2日	面	47	12.8
1941(昭和 16)年	1月3日・3月2日・5月2日・7月2日・9月2日・11月2日	面	43	14.6
1949(昭和 24)年	1月3日・1月4日・3月2日・5月2日・7月2日・7月3日・9月2日・11月2日	面	29	11.7
1957(昭和 32)年	5月2日・7月2日・11月2日	面	37	10.1
1965(昭和 40)年	1月3日・3月2日・5月2日・7月2日・9月2日・11月2日	記事	407	27.9
1973(昭和 48)年	1月3日・3月2日・5月2日・7月2日・9月2日・11月2日	記事	770	37.9
1981(昭和 56)年	1月3日・3月2日・5月2日・7月2日・9月2日・11月2日	記事	787	35.6
1989(昭和 64,平成元)年	1月3日・3月2日・5月2日・7月2日・9月2日・9月3日・11月2日	記事	821	31.7
1997(平成 9)年	1月3日・3月2日・5月2日・7月2日・9月2日・11月2日	記事	627	27.1
2005(平成 17)年	1月3日・3月2日・5月2日・7月2日・9月2日・11月2日	記事	563	23.7
2013(平成 25)年	1月3日・3月2日・5月2日・7月2日・9月2日・11月2日	記事	553	22.9
合計			4684	255.9

[1] https://chunagon.ninjal.ac.jp/

[2] https://clrd.ninjal.ac.jp/chj/meiji_taisho.html#shinbun

３．テキストの収録範囲とサンプル

新聞紙1冊分における本文テキストの収録範囲は、「明治・大正編V新聞」に倣い、広告記事および固有名詞や数値の羅列を中心とする記事（叙任・辞令，スポーツの試合結果・株式の取引結果等）を除外した全ての記事のうち、図表や写真・挿絵の中のテキストと、それらのキャプションに相当する文書要素を除外した全文である。

これらの対象テキストをコーパスに収録する際には、テキストを一定の範囲で分割する必要があり、その各範囲を「サンプル」と呼ぶ。新聞レジスターのサンプル単位は、発行年により、新聞の1記事に相当する範囲（「記事」）と、新聞の1ページに相当する単位（「面」）の2種に分けられている。本来は、一律に、「CHJ明治・大正編Ｖ新聞」やSHCの他のレジスターで採用する「記事」による分割とすべきところであるが、発行年が1933年から1957年の4か年分については、構築作業上の都合により「記事」を単位とする分割ができなかったため、「面」による分割になっている。ただし、その場合でも、新聞紙1ページ内に文芸作品（小説、短歌・俳句、詩など）を含む場合は、個別にその部分を「記事」として分割した。このように、新聞レジスターにおいては、サンプルの分割単位が不統一であるため、サンプルを単位とした分析が不適当となる場合がある点に、特段の注意が必要である[3]。

こうして分割したサンプルには、各サンプルを一意に特定するID（「サンプルID」と呼ぶ）が付けられている。サンプルIDの構成を表2にあげる。表2の基準によると、例えば、1933年11月2日の3つめのサンプルは「70P読売1933_B2003」というサンプルIDが付けられていることになる。

表2　サンプルIDの構成

桁数	値	説明
1〜2	70 / 80	時代区分を表す。「70」は「昭和」、「80」は「平成」を表す。
3	P	レジスターを表す。すべて「P」で「新聞（Paper）」を表す。
4〜5	読売	紙名を表す。すべて「読売」で「読売新聞」を表す。
6〜9	(4 桁の数字)	サンプルの発行年を西暦で表す。
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 進数で表す(1〜9 月はアラビア数字で、10 月以降は ABC... に対応する)。
12	(1 桁の数字)	発行日を 32 進数で表す(1〜9 日はアラビア数字で、10 月以降は ABC... に対応する)。
13〜15	(3 桁の数字)	サンプルの通し番号を表す。

４．テキストの仕様

４．１　テキストに使用する文字

新聞レジスターでは、テキスト電子化の文字集合として、原則としてUnicodeを使用した。構築作業における便宜上の仕様であるが、その結果「CHJ明治・大正編Ｖ新聞」やSHCの他のレジスターと文字集合の異なっている点については、特に留意が必要である。表3に、他のレジスターで包摂対象となる漢字字体を一覧にして示す。

[3] 原則として面を単位に分割した1933年から1957年のサンプル数は156、１サンプルあたりの語数平均は3151.1語、記事を単位に分割した1965年から2013年のサンプル数は4984、１サンプルあたりの語数平均は456.6語であり、サンプルの性質が大きく異なることが推定される。

表3　留意すべき字体の一覧

新聞使用文字		包摂対象文字		備考*
文字	Unicode	文字	Unicode	
衛	U+4619	衛	U+885B	JIS 包摂
倶	U+4FF1	倶	U+5036	雑・書包摂
併	U+5002	併	U+4F75	雑・書包摂
剝	U+525D	剥	U+5265	雑・書包摂
卽	U+537D	即	U+5373	雑・書包摂
呑	U+541E	呑	U+5451	雑・書包摂
嘘	U+5653	嘘	U+5618	雑・書包摂
增	U+589E	増	U+5897	雑・書包摂
妍	U+59F8	妍	U+598D	雑・書包摂
寛	U+5BEC	寛	U+5BDB	雑・書包摂
屛	U+5C5B	屛	U+5C4F	雑・書包摂
巣	U+5DE2	巣	U+5DE3	雑・書包摂
幷	U+5E77	并	U+5E76	雑・書包摂
强	U+5F3A	強	U+5F37	JIS 包摂
徴	U+5FB5	徴	U+5FB4	雑・書包摂
德	U+5FB7	徳	U+5FB3	雑・書包摂
戻	U+623E	戻	U+623B	雑・書包摂
掲	U+63ED	掲	U+63B2	雑・書包摂
撃	U+64CA	撃	U+6483	雑・書包摂
旣	U+65E3	既	U+65E2	JIS 包摂
昻	U+663B	昂	U+6602	JIS 包摂
晩	U+665A	晩	U+6669	雑・書包摂
普	U+669C	普	U+666E	JIS 包摂
暦	U+66C6	暦	U+66A6	雑・書包摂
槪	U+69EA	概	U+6982	雑・書包摂
横	U+6A6B	横	U+6A2A	雑・書包摂
橫	U+6ACE	横	U+6A2A	JIS 包摂
欽	U+6B2B	缺	U+7F3A	JIS 包摂
歩	U+6B65	歩	U+6B69	雑・書包摂
歷	U+6B77	歴	U+6B74	雑・書包摂
毎	U+6BCF	毎	U+6BCE	雑・書包摂
渉	U+6D89	渉	U+6E09	雑・書包摂
涙	U+6DDA	涙	U+6D99	雑・書包摂
渇	U+6E34	渇	U+6E07	雑・書包摂
溫	U+6EAB	温	U+6E29	雑・書包摂
瀬	U+7028	瀬	U+702C	雑・書包摂

新聞使用文字		包摂対象文字		備考*
文字	Unicode	文字	Unicode	
狀	U+72C0	状	U+72B6	雑・書包摂
玆	U+7386	茲	U+5179	JIS 包摂
甁	U+7501	瓶	U+74F6	雑・書包摂
畫	U+7575	畫	U+756B	JIS 包摂
痩	U+7626	痩	U+75E9	雑・書包摂
研	U+784F	研	U+7814	雑・書包摂
綠	U+7DA0	緑	U+7DD1	雑・書包摂
緖	U+7DD6	緒	U+7DD2	雑・書包摂
緣	U+7DE3	縁	U+7E01	雑・書包摂
繫	U+7E6B	繋	U+7E4B	雑・書包摂
羨	U+7FA1	羨	U+7FA8	JIS 包摂
脅	U+810B	脅	U+8105	JIS 包摂
薰	U+85B0	薫	U+85AB	雑・書包摂
虛	U+865B	虚	U+865A	雑・書包摂
賴	U+8CF4	頼	U+983C	雑・書包摂
郎	U+90DE	郎	U+90CE	雑・書包摂
鄕	U+9115	郷	U+90F7	雑・書包摂
錄	U+9304	録	U+9332	雑・書包摂
鍊	U+934A	錬	U+932C	雑・書包摂
髙	U+9AD9	高	U+9AD8	JIS 包摂
鬪	U+9B2A	闘	U+95D8	雑・書包摂
鬭	U+9B2D	闘	U+95D8	雑・書包摂
黃	U+9EC3	黄	U+9EC4	雑・書包摂
黑	U+9ED1	黒	U+9ED2	雑・書包摂
欄	U+F91D	欄	U+6B04	雑・書包摂
蠟	U+F927	蠟	U+881F	JIS 包摂
廊	U+F928	廊	U+5ECA	雑・書包摂
朗	U+F929	朗	U+6717	雑・書包摂
虜	U+F936	虜	U+865C	雑・書包摂
殺	U+F970	殺	U+6BBA	雑・書包摂
類	U+F9D0	類	U+985E	雑・書包摂
隆	U+F9DC	隆	U+9686	雑・書包摂
塚	U+FA10	塚	U+585A	雑・書包摂
猪	U+FA16	猪	U+732A	雑・書包摂
礼	U+FA18	礼	U+793C	JIS 包摂
神	U+FA19	神	U+795E	雑・書包摂

新聞使用文字		包摂対象文字		備考*
文字	Unicode	文字	Unicode	
祥	U+FA1A	祥	U+7965	雑・書包摂
福	U+FA1B	福	U+798F	雑・書包摂
靖	U+FA1C	靖	U+9756	JIS 包摂
諸	U+FA22	諸	U+8AF8	雑・書包摂
都	U+FA26	都	U+90FD	雑・書包摂
侮	U+FA30	侮	U+4FAE	雑・書包摂
僧	U+FA31	僧	U+50E7	雑・書包摂
免	U+FA32	免	U+514D	雑・書包摂
勉	U+FA33	勉	U+52C9	雑・書包摂
勤	U+FA34	勤	U+52E4	雑・書包摂
卑	U+FA35	卑	U+5351	雑・書包摂
喝	U+FA36	喝	U+559D	雑・書包摂
嘆	U+FA37	嘆	U+5606	雑・書包摂
器	U+FA38	器	U+5668	雑・書包摂
塀	U+FA39	塀	U+5840	雑・書包摂
墨	U+FA3A	墨	U+58A8	雑・書包摂
層	U+FA3B	層	U+5C64	雑・書包摂
悔	U+FA3D	悔	U+6094	雑・書包摂
慨	U+FA3E	慨	U+6168	雑・書包摂
憎	U+FA3F	憎	U+618E	雑・書包摂
懲	U+FA40	懲	U+61F2	雑・書包摂
敏	U+FA41	敏	U+654F	雑・書包摂
暑	U+FA43	暑	U+6691	雑・書包摂
梅	U+FA44	梅	U+6885	雑・書包摂
海	U+FA45	海	U+6D77	雑・書包摂
渚	U+FA46	渚	U+6E1A	雑・書包摂
漢	U+FA47	漢	U+6F22	雑・書包摂
煮	U+FA48	煮	U+716E	雑・書包摂
琢	U+FA4A	琢	U+7422	雑・書包摂
碑	U+FA4B	碑	U+7891	雑・書包摂
社	U+FA4C	社	U+793E	雑・書包摂
祉	U+FA4D	祉	U+7949	雑・書包摂
祈	U+FA4E	祈	U+7948	雑・書包摂
祐	U+FA4F	祐	U+7950	雑・書包摂
祖	U+FA50	祖	U+7956	雑・書包摂
祝	U+FA51	祝	U+795D	雑・書包摂
禍	U+FA52	禍	U+798D	雑・書包摂
禎	U+FA53	禎	U+798E	雑・書包摂
穀	U+FA54	穀	U+7A40	雑・書包摂
突	U+FA55	突	U+7A81	雑・書包摂
節	U+FA56	節	U+7BC0	雑・書包摂
練	U+FA57	練	U+7DF4	雑・書包摂
繁	U+FA59	繁	U+7E41	雑・書包摂
署	U+FA5A	署	U+7F72	雑・書包摂
者	U+FA5B	者	U+8005	雑・書包摂
臭	U+FA5C	臭	U+81ED	雑・書包摂
著	U+FA5F	著	U+8457	雑・書包摂
褐	U+FA60	褐	U+8910	雑・書包摂
視	U+FA61	視	U+8996	雑・書包摂
謁	U+FA62	謁	U+8B01	雑・書包摂
謹	U+FA63	謹	U+8B39	雑・書包摂
賓	U+FA64	賓	U+8CD3	雑・書包摂
贈	U+FA65	贈	U+8D08	雑・書包摂
逸	U+FA67	逸	U+9038	雑・書包摂
難	U+FA68	難	U+96E3	雑・書包摂
響	U+FA69	響	U+97FF	雑・書包摂
頻	U+FA6A	頻	U+983B	雑・書包摂
叱	U+20B9F	叱	U+53F1	雑・書包摂
吉	U+20BB7	吉	U+5409	JIS 包摂

《備考》欄の凡例

JIS包摂…JIS X0213:2004の包摂基準により包摂対象文字に包摂されるもの

雑・書包摂…雑誌・ベストセラー書籍レジスターでは須永他（2011）、須永他（2013）に準じて包摂対象文字に包摂されるもの

なお、Unicode文字集合に含まれるものの、入力・表示環境が十分ではない変体仮名については、現行の平仮名によって電子化した。また、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ほか、文字集合に含まれない漢字については、Unicodeで表現可能な異体字を有する場合はその異体字により代用し、代用が不可能な場合は「=」（げた記号、U+3013）で表した。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、U+2423）によって表した。

４．２　テキストの校訂

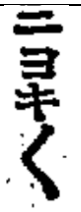
本コーパスでは、国立国語研究所が開発・公開する他のコーパスと斉一な形態論情報を付与するため、形態素解析辞書UniDic（以下UniDic）を使用した形態素解析に基づき形態論情報を付与した。そこで、コーパスのテキストをUniDicによる形態素解析に適したものとするため、底本のテキストに対して以下にあげる改変（ここでは「校訂」と呼ぶ）を施した。

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。利用に際しては、必要に応じて「原文KWIC」や「原本文字列」を確認されたい。

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」「人ゞ」等、1短単位内部で直前の1字を繰り返す「々」「ゞ」は置き換えの対象としない。

表4　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
/＼		ニヨキニヨキ	ニヨキ/＼
ゝ		ここへ味を	こゝへ味を

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。ルビの情報はデータ化していないため、語彙素認定において一切考慮していない。よって、「CHJ明治・大正編Ｖ新聞」が実装している、同一文字列に複数の形態論情報を付与する機能は用いていない。

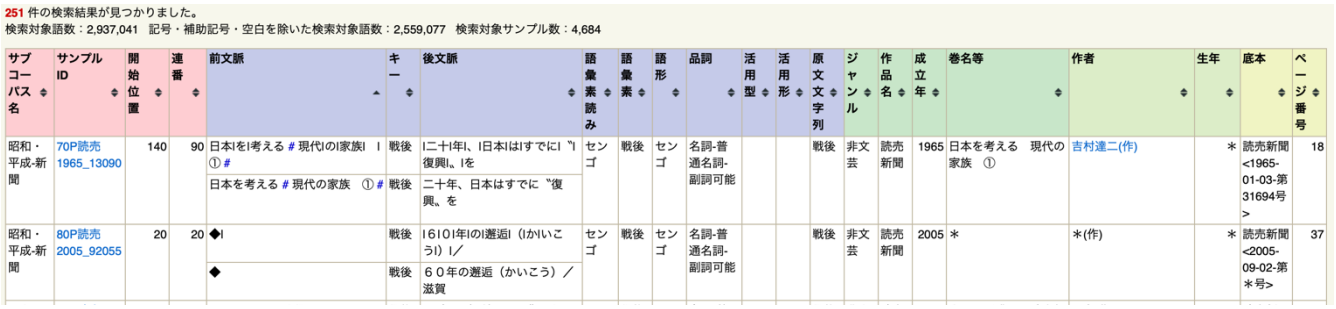
形態論情報は短単位のみ付与しており、長単位は未実装である。短単位の形態論情報は、UniDic[4]を使用した形態素解析に基づき、ごく一部について人手による修正を施した。新聞レジスター（バージョン

[4] 1933年から1941年の全サンプルに「旧仮名口語UniDic」を、1949年から2013年の全サンプルに「現代書き言葉UniDic」を使用した。

2023.5）における形態論情報（語彙素レベル）の精度（適合率）は、97.98%となっている[5]。

６．「中納言」上の表示項目

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）。


251 件の検索結果が見つかりました。
検索対象語数：2,937,041　記号・補助記号・空白を除いた検索対象語数：2,559,077　検索対象サンプル数：4,684

サブコーパス名	サンプルID	開始位置	連番	前文脈	キー	後文脈	語彙素読み	語彙素	語形	品詞	活用型	活用形	原文文字列	ジャンル	作品名	成立年	巻名等	作者	生年	底本	ページ番号
昭和・平成-新聞	70P読売1965_13090	140	90	日本を考える # 現代の家族　① #	戦後	二十年、日本はすでに〝復興〟を	センゴ	戦後	センゴ	名詞-普通名詞-副詞可能			戦後	非文芸	読売新聞	1965	日本を考える　現代の家族　①	吉村達二(作)	*	読売新聞<1965-01-03-第31694号>	18
昭和・平成-新聞	80P読売2005_92055	20	20	◆	戦後	１６０年の邂逅（かいこう）／滋賀	センゴ	戦後	センゴ	名詞-普通名詞-副詞可能			戦後	非文芸	読売新聞	2005	*	*(作)	*	読売新聞<2005-09-02-第＊号>	37

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について、表5に内容を示す。

表5　「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。

[5] 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。

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。 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。 **本レジスターでは不使用のため空欄である。**
	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。 **本レジスターでは不使用のため空欄である。**
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。 **本レジスターでは不使用のため空欄である。**
	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」「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。

情報種別	項目名	内容
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。本レジスターでは、ページをサンプル単位とした1933から1957年は面数を表示し（例：〈1面〉）、記事をサンプル単位とした1965年から2013年は、記事タイトルを表示する（ただし、1997年以降は未整備につき「＊」と表示されている)。
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（記名のないものなど著者が判明しないものは「＊」と表示されている）。 対応付けが可能な範囲で、「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。 不明な場合は空欄である。
底本情報	底本	検索対象の底本（原資料)。「読売新聞<1989-11-02-第40737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。本レジスターでは「読売新聞社」が表示される。
その他	底本リンク	検索対象の底本画像へのリンク。 **本レジスターでは未整備につき空欄である。**
	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。 **本レジスターでは該当画像がないため空欄である。**

付記

本レジスターを含む『昭和・平成書き言葉コーパス』は、JSPS 科研費 19H00531「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」（代表：小木曽智信）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

コーパス検索アプリケーション「中納言」 https://chunagon.ninjal.ac.jp/
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/
日本語歴史コーパス https://ccd.ninjal.ac.jp/chj/
日本語歴史コーパス 明治・大正編V新聞 https://clrd.ninjal.ac.jp/chj/meiji_taisho.html#shinbun
読売新聞社「ヨミダス歴史館」https://database.yomiuri.co.jp/about/rekishikan/

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