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

2023 年 5 月 30 日　近藤明日子

1.　はじめに

この文書では、『昭和・平成書き言葉コーパス』（バージョン 2023.5）雑誌レジスターの収録資料とテキストの仕様、形態論情報、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

『日本語歴史コーパス 明治・大正編 I 雑誌』[2]に接続して利用できること目指し、収録雑誌として昭和・平成期を代表する月刊総合雑誌『中央公論』『文芸春秋』を選定した。収録刊年は、『日本語歴史コーパス 明治・大正編 I 雑誌』で 6〜8 年おきに収録していることに倣い、1933〜2013 年の間の 8 年おきとし、各年の通常号 12 冊、11 か年分、計 132 冊を収録対象とした。収録した雑誌名と刊行年・巻号は以下の表 1 のとおりである。

表 1　雑誌レジスター収録資料

雑誌名	刊行年	巻号
中央公論	1933（昭和8）	48年1号-12号
中央公論	1941（昭和16）	56年1号-12号
中央公論	1949（昭和24）	64年1号-12号
中央公論	1957（昭和32）	72年1号-3号、5号-6号、8号-14号
文芸春秋	1965（昭和40）	43巻1号-12号
文芸春秋	1973（昭和48）	51巻1号、3号-4号、6号-7号、9-10号、12-13号、15号-16号、18号
文芸春秋	1981（昭和56）	59巻1号-7号、9号-13号
文芸春秋	1989（平成元）	67巻1号-3号、5号-13号
文芸春秋	1997（平成9）	75巻1号、3号-5号、7号-14号
文芸春秋	2005（平成17）	83巻1号-3号、5号、7号-9号、11号-13号、15号-16号
文芸春秋	2013（平成25）	91巻1号-4号、6号-13号

3.　テキストの範囲とサンプル

収録対象資料の内、(1)表紙、(2)目次、(3)奥付、(4)グラビア等の写真・絵・図表を中心とする記事、(5)前号の誤植等を謝罪する記事、(6)広告、(7)付録、を除くすべての文書要素のテキストをコーパスに収録した。ただし、文書要素中の写真・絵・図表に付属するテキスト、漢文・外国語・数式からなる段落のテキストは入力対象外とした。

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、1 記事＝1 サンプルとして分割した。各サンプルを一意に

[1] https://chunagon.ninjal.ac.jp/
[2] 明治・大正期の各年代を代表する総合雑誌・女性雑誌を収録するコーパス。
https://clrd.ninjal.ac.jp/chj/meiji_taisho.html#zasshi

特定するサンプル ID は「70M 中公 1933_01001」のような 15 桁の英数字・記号からなる。その構成を表 2 に示す。

表 2　サンプル ID の構成

左からの桁数	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	M	レジスターを表す。すべて「M」で「雑誌」を表す。
4～5	中公 / 文春	雑誌名を表す。「中公」は『中央公論』、「文春」は『文芸春秋』を表す。
6～9	（4桁の数字）	雑誌の刊行年を西暦で表す。
10	_	サンプルIDの区切り記号（アンダーバー）。
11～12	（2桁の数字）	雑誌の号番号を表す。
13～15	（3桁の数字）	各号内のサンプルの通し場号を表す。

4.　テキストの仕様

4.1.　テキストに使用する文字

電子化テキストに使用した文字の範囲は、JIS X 0213（2004）の文字集合（JIS 漢字の第 4 水準までを含む）に準拠した[3]。

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)～(5)の手順で電子化した[4]。

(1)　JIS X 0213 の「6.6.3　漢字の字体の包摂規準」に基づき、文字集合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のうち、近代に特有な微細な字形差を持つ漢字については、須永・堤・近藤ほか（2013）の「追加包摂規準」に基づき文字集合内の文字に包摂する。
(3)　(2)の追加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・用法を持ち、同一の漢字部品を有する同音・同訓の文字集合内の文字で代用する。
(4)　(3)による代用が不可能な文字のうち、Unicode に収録されている漢字はそれにより電子化する。
(5)　(4)による入力が不可能な文字は、「〓」（げた記号、JIS 面区点 1-02-14、U+3013）で入力する。

4.2.　テキストの校訂

コーパスのテキストを形態素解析用辞書 UniDic[5]による形態素解析に適したものとするため、底本の本

[3] ただし、①JIS X 0213 附属書 7　2.1　b）に掲載される、戸籍法施行規則付則別表「人名用漢字許容字体表」（昭和 56 年法務省令 51）の漢字、及び常用漢字表（昭和 56 年内閣告示第 1 号）のかっこ書き内の漢字（「いわゆる康熙字典体」）のうち、JIS X 0208 で包摂していた漢字、②JIS X 0213:2004 において UCS との互換のために追加された 10 字、についてはこれを用いない。
[4] この電子化の手順は須永・堤・高田ほか（2011）の方法を参照した。

行のテキストに対して以下のA)～C)にあげる校訂を施し、コーパスのテキストを作成した。同様に本行に付されたルビも校訂を施してテキストを作成したが、A)の校訂は行わずB)C)の校訂のみ行った。

なお、コーパス検索アプリケーション「中納言」(6 節参照) では、本行のテキストに関して、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。ただし、ルビテキストは「中納言」では常に校訂後のテキストのみ表示され、校訂前のテキストの確認はできない。

A)　踊り字

踊り字は繰り返される文字列に置き換える (表 3)。ただし、「人々」「愈〻」等、1 短単位内部で直前の漢字 1 字を繰り返す「々」「〻」は置き換えの対象としない。

表 3　踊り字の電子テキスト化の例[6]

コーパステキスト	原文 KWIC・原本文字列
英國國旗	英國々旗
流れつつある	流れつゝある
意味をなさざるもの	意味をなさゞるもの
チチハル	チヽハル
ヒビ割れ	ヒヾ割れ
ははははは	はゝはゝゝ
いろいろ	いろ/＼
それぞれ	それ/〃＼

B)　濁点無表記

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、該当の濁音を表す濁点付き仮名に置き換える (表 4)。ただし、清濁両形がある語[7]については、置き換えの対象としない。

表 4　濁点無表記の電子テキスト化の例

コーパステキスト	原文 KWIC・原本文字列
異様な男であるのを	異様な男てあるのを
ブルータスよ	フルータスよ

C)　誤植

原文の誤植 (誤字、前後文字列の転倒、脱字、衍字) と思われる表記は訂正する (表 5)。ただし、仮名遣いの揺れや、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[8]などは、

---

[5] https://unidic.ninjal.ac.jp/

[6] 表中の電子化例ではルビテキストの表示は省略した。表 4・表 5 も同様。

[7] 当該の語に清濁両形があるかどうかの判定は、原則として『日本語国語大辞典　第二版』によるものとする。清濁両形が辞書の「見出し」にある場合のほか、「語義説明」内に「〇〇とも」「古くは〇〇」の形で異語形を示す場合は、清濁両形があるものと判断する。

[8] 語形のバリエーションかどうかの判定は、注 7 に示した清濁両形の有無の判定に準ずる。また、通用の異表

訂正の対象としない。

表 5　誤植の電子テキスト化の例

コーパステキスト	原文 KWIC・原本文字列
寄與しなかつた許りでなく	寄與しなかつた許りでたく
アメリカ人は	アメカ人は
常識では考えられない	常識では考ええられない

5.　形態論情報

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、長単位は未付与である[9]。短単位の形態論情報は、2 種類の形態素解析辞書 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。使用した UniDic の種類とそれで解析したサンプルとの対応は以下のとおりである。

- 旧仮名口語 UniDic　…1933～1957 年の全サンプル、1965～2013 年の旧仮名遣いのサンプル
- 現代書き言葉 UniDic …1965～2013 年の残りのサンプル

形態論情報の精度評価を行ったところ、発音形レベル[10]の精度（F 値）は 97.28%であった。

形態論情報の各項目については表 6 を参照のこと。

6.　「中納言」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）。

記かどうかの判定は、①『日本語国語大辞典　第二版』及び②近代語のコーパスによる出現状況によるものとする。①は、見出しの「漢字表記」のほか、「用例文」中の表記、「表記」欄の表記などを通用の異表記と見なす。①が適用できない場合、②近代語のコーパス（公開済みのもののほか、内部資料を含む）において、複数のサンプルに出現し、出現数が少なくない表記を異表記と見なす。

[9] 短単位・長単位の詳細については、小椋・小磯・冨士池ほか（2011）を参照のこと。

[10] 形態論情報の品詞・活用型・活用形・語彙素・語彙素読み・語彙素細分類・発音形のすべてを評価対象とするレベル。

3,173 件の検索結果が見つかりました。そのうち 500 件を表示しています。
検索対象語数：31,120,433　記号・補助記号・空白を除いた検索対象語数：27,399,474　検索対象サンプル数：10,003

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	振り仮名	本文種別	話者	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-雑誌	70M中公1933_01003	8960	5750	人間|的|で|なく|、|却つて|それ|以前|の|世界|に|屈し|て|ゐる|。#馬|に|とつ|て|は|	走る	|とが|、|魚|に|とつ|て|は|泳ぐ|こと|が|、|鳥|に|とつ|て|は|飛ぶ|こと|が	ハシル	走る	ハシル	動詞-一般	五段-ラ行	連体形-一般	走る			非文芸	中央公論	1933	実践の存在論的性格	山内得立(作)		中央公論 <1933-01>	本欄 25
昭和-平成-雑誌	70M中公1933_01008	64900	42240	なら|ず|、|また|軍部|が|政党|に|反感|を|有つ|た|の|は|、|本分|を|忘れ|私利|に|	はしる	|と|いふ|過誤|に|憤慨|した|から|に|他ならない。#有産|有閑|階級|を|背景|とする	ハシル	走る	ハシル	動詞-一般	五段-ラ行	終止形-一般	はしる			非文芸	中央公論	1933	議會から見た政局の波瀾	吉野作造(作)		中央公論 <1933-01>	本欄 65
昭和-平成-雑誌	70M中公1933_01011	64030	42890	、|すでに|金|の|大半|を|失つ|て|ゐた|日本|資本|主義|は|、|急遽|金|輸出|禁止|に|	走ら	|ざる|を|得|なかつた|の|だ|。#ところ|で|、|政府|の|大量|的|な|、|資金|放出|は	ハシル	走る	ハシル	動詞-一般	五段-ラ行	未然形-一般	走ら			非文芸	中央公論	1933	非常時景氣の展望 景氣好轉、それは何處まで本當か？	ABC(作)		中央公論 <1933-01>	本欄 91

図 1　「中納言」の検索結果の表示例

次の表 6 に、「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、初期設定では表示されないが注意が必要な項目（表中*を付す）について内容を示す。

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	サブコーパス名	「キー」の含まれるコーパスとレジスター名。本レジスターはすべて「昭和・平成-雑誌」である。
	サンプル ID	「キー」の含まれるサンプルの ID（3 節参照）。「詳細な文脈情報」へのリンクを付与する。 最後の 3 桁が「000」のサンプルは、各号のテキストから記事のテキストを切り出した残りのテキストを集めたものである。複数の記事群をまとめる内容のテキスト等が収録されている。
	開始位置	「キー」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	「キー」の短単位の、サンプル内における位置を表す ID。10 きざみの連番。
形態論情報[11]	前文脈	「キー」の前方文脈。 「前文脈」中で用いられる記号「|」は短単位境界を、「＃」は文境界を示す。
	キー	検索対象の含まれる短単位[12]の出現形。

[11] 形態論情報の個々の項目の内容は、「前文脈」「キー」「後文脈」「原文 KWIC」「原文文字列」「振り仮名」を除き、UniDic の見出しに対応している。各項目の詳細については小椋ほか（2011）を参照のこと。また、一部の文語体の形態論情報については国立国語研究所コーパス開発センター（近藤明日子）（編）（2016）を参照のこと。

[12] 「キー」の表示範囲は設定により変更可能であるが、ここでは初期設定で表示される「キー」の範囲を指して言う（表中の他の項目も同様）。初期設定では、短単位検索の場合は検索対象とした 1 短単位、文字列検索の場合は検索対象文字列に含まれる最後の 1 短単位、が表示範囲となる。

情報種別	項目名	内容
	後文脈	「キー」の後方文脈。 「後文脈」中で用いられる記号「|」は短単位境界を、「＃」は文境界を示す。
	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）。「キー」に対するルビのテキストは本行テキストの上側に表示される。ただし、ルビのテキストは校訂後のテキストである。 「原文 KWIC」中で用いられる記号「＃」は文境界を示す。
	語彙素読み	「キー」の短単位の語彙素（下記項目「語彙素」参照）の読み。片仮名表記である。
	語彙素	「キー」の短単位の語彙素の表記。語彙素は、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。
	語彙素細分類*	「キー」の短単位の語彙素を語義等で更に分類するもの。「ライト」に対する「light（光）」「light（軽い）」「light（光）」「right」「write」、「引く」に対する「自動詞」「他動詞」、「たり」に対する「完了」「断定」など。細分類する必要がある場合のみ表示される。
	語形	「キー」の短単位の語形。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もの。片仮名表記である。
	品詞	「キー」の短単位の品詞。形態素解析用辞書 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。 UniDic の体系に基づかない特殊な品詞には、「**欠損**」（「␣」［4.1 節参照］を含む文字列）、「**未知語**」（形態論情報が未付与の文字列）「**correct 処理済**」（「未知語」に同じ）がある。これらの特殊な品詞の付与された短単位は、形態論情報に関する項目のうち、「前文脈」「キー」「後文脈」「原文 KWIC」「品詞」「原文文字列」「振り仮名」以外は空欄になっている。
	活用型	「キー」の短単位の活用の型。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。

情報種別	項目名	内容
	活用形	「キー」の短単位の活用形。活用語の場合のみ表示される。学校文法では「未然形」と助動詞「う・よう」に分割される形態は、結合して「意志推量形」とする点に注意が必要である。 文法的に特定の活用形が期待される箇所（単語同士の接続関係や文末等）で、それとは異なる形態が用いられている場合は、語の形態に即して活用形を割り当てる。 例）性格が見出さるる。（文末だが活用形は「連体形-一般」）
	原文文字列	「キー」の短単位の出現形の、校訂前の底本に近い形（4.2 節参照）。
	振り仮名	「キー」の短単位に付されたルビの校訂後のテキスト。
本文情報	本文種別	「キー」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。 **会話**　…会話・独話・心内発話等の引用 **引用**　…文献等からの引用、記事に対する雑誌記者・編集者の説明・解説・注釈等 なお、ジャンル「非文芸」のサンプルでは、地の文と会話・引用を区別せず、すべて「地の文」として扱っている。
	話者	上記項目「本文種別」が「会話」の場合の話者名、「引用」の場合の典拠文献名や著者名。 本レジスターでは情報付与しておらず空欄である。
作品情報	ジャンル	「キー」の含まれるサンプルの、文章内容に基づく分類。以下の種類がある。 **文芸/小説**　…小説 **文芸/戯曲**　…戯曲 **文芸/詩歌**　…和歌・俳句・詩 **非文芸**　…上記以外
	作品名	「キー」の含まれるサンプルが収録された雑誌名。
	成立年	「キー」の含まれるサンプルが収録された雑誌の刊行年。
	巻名等	「キー」の含まれるサンプルのタイトル。原則、底本の記載に基づく。ただし、連載記事のタイトル統一等のため、底本の記載から改変した場合もある。
作者情報	作者	「キー」の含まれるサンプルの著者名・訳者名。著者名は後ろに「(作)」を、訳者名は後ろに「(訳)」を付けて示す。複数人を併記する場合は、「/」で区切る。 著者名の認定は底本の記載に基づく。ただし、著者名が団体名で記載されている場合や、著者名の記載がなく不明な場合は「＊」で示す。 訳者名の認定は底本の記載に基づく。ただし、訳者名が団体名で記載されている場合は「＊」で示す。

情報種別	項目名	内容
	生年	上記項目「作者」の生年。 本レジスターでは情報付与しておらず空欄である。
	性別*	上記項目「作者」の性別。 本レジスターでは情報付与しておらず空欄である。
底本情報	底本	「キー」の含まれるサンプルの収録された底本（原資料）。「中央公論〈1933-01〉」のように「雑誌名〈刊行年-号〉」の形式で示す。
	ページ番号	「キー」の底本における出現ページ番号。底本での記載に基づく。底本にないページ番号を補った場合は〔　〕で括って示す。

謝辞

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」および国立国語研究所共同研究プロジェクト「開かれた共同構築環境による通時コーパスの拡張」による成果の一部である。

参考文献

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011）『『現代日本語書き言葉均衡コーパス』形態論情報規定集第 4 版（上）（下）』特定領域研究「日本語コーパス」平成 22 年度研究成果報告書，国立国語研究所，http://doi.org/10.15084/00002855，http://doi.org/10.15084/00002856

国立国語研究所（2019）『日本語歴史コーパス 明治・大正編 I 雑誌』https://clrd.ninjal.ac.jp/chj/meiji_taisho.html#zasshi

国立国語研究所コーパス開発センター（近藤明日子）編（2016）『近代文語 UniDic 短単位規程集 Ver.1.1』国立国語研究所コーパス開発センター，https://clrd.ninjal.ac.jp/chj/doc/unidic-MLJ_rulebook_v1_1.pdf

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

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』2011(8)，pp.381-388