# 『日本語歴史コーパス　明治・大正編Ⅳ近代小説』（短単位データ 1.0）概説書

2021 年 3 月 31 日　髙橋雄太

## １．はじめに

『日本語歴史コーパス　明治・大正編Ⅳ近代小説』（短単位データ 1.0）は、明治中期から大正末期にかけて刊行された主要な小説作品群をコーパス化したものである。本文書では、コーパスに収録した作品の選定や設計の方法、コーパスのテキストの仕様、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。

## ２．コーパスの概要

### ２．１　作品選定の基本方針

近代小説は、言文一致の研究をはじめ、語彙・文法・文体・表記など分野を選ばず、日本語史研究に取り上げられてきた重要な言語資料である。『CHJ』に近代小説のコーパスを追加する上では、「明治・大正編」に収録されている公開済のサブコーパスとの併用や、統計的な分析に耐えうる設計を基本の設計方針とし、著者ごとの作品数や作品ごとの言語量のバランスを考慮して構築を行った。

本コーパスに収録した資料の概要及び延べ短単位数（記号類を除く）を表 1 に示す。本コーパスに収録された作品は、明治中期（『浮雲』1887 年）から大正末期（『伊豆の踊子』1926 年）までの 21 作品（21 著者）であり、およそ 70 万語と規模になる。次節以降で、その設計について詳細に述べる。

### ２．２　作品の選定方法

本コーパスは作品の選定を、かつて国立国語研究所で計画された『日本大語誌』の準備資料の一つである、「用例採集のための主要文学作品目録」（以下「目録」） に準拠して行った。「目録」では、①現代（昭和当時）の文学全集 15 種のうち 3 種以上に収録されている 1506 作品を選出、②文学研究者、言語研究者などの有識者 10 名による用例採集に適すると考える 100 作品への投票、③票数が 4 以上集まった 139 作品を選定、という手順で、現代語の用例採集に適する作品群を選定している（国立国語研究所 1980）。ここから『CHJ　明治・大正編』がメインターゲットにしている明治時代と大正時代の作品に絞ると、74 作品となる。

74 作品のうち、基本的には票数の多いものから採用し、年代や規模のバランスを考慮して、21 作品（21 著者）の選定を行った[1]。具体的には、①一致度の高い作品を優先して採用する、②一著者につき一作品を採用する、③各時代・各年代の作品数の偏りを可能な限りなくす、④同一の一致度に複数の作品がある場合「目録」に多く作品が選定されている著者のものを優先する、の 4 点を基準として定めた。この 21 作品を採用することにより、1890 年代、1900 年代、1910 年代、1920 年代に 5 作品ずつ含まれ、加えて、明治時代に 12 作品、大正時代に 9 作品が含まれるようになり、各年代・各時代に均等に振り分けられるように設計した。

[1] 採用されなかった作品については、髙橋ほか（2019）を参照されたい。

表 1　収録作品一覧

| 作品名 | 著者 | 発表年 | 底本（出版年・出版社） | 短単位数 |
|---|---|---|---|---|
| 浮雲 | 二葉亭四迷 | 1887（明治20）年 | 『新編浮雲第一篇』（1887年・金港堂） | 22,866 |
| 舞姫 | 森鴎外 | 1890（明治23）年 | 『国民小説（第一）』（1890年・民友社） | 9,244 |
| 五重塔 | 幸田露伴 | 1891（明治24）年 | 『尾花集』（1892年・青木嵩山堂） | 32,303 |
| たけくらべ | 樋口一葉 | 1895（明治28）年 | 『一葉全集』（1897年・博文館） | 16,744 |
| 今戸心中 | 広津柳浪 | 1896（明治29）年 | 『柳浪叢書 前編』（1910年・博文館） | 20,377 |
| 武蔵野 | 国木田独歩 | 1898（明治31）年 | 『武蔵野』（1901年・民友社） | 8,889 |
| 思出の記 | 徳富蘆花 | 1900（明治33）年 | 『思出の記』（1901年・民友社） | 27,760 |
| 高野聖 | 泉鏡花 | 1900（明治33）年 | 『高野聖』（1908年・左久良書房） | 20,999 |
| 吾輩は猫である | 夏目漱石 | 1905（明治38）年 | 『吾輩ハ猫デアル』（1907年・大倉書店） | 76,669 |
| 蒲団 | 田山花袋 | 1907（明治40）年 | 『花袋集』（1908年・易風社） | 26,991 |
| 何処へ | 正宗白鳥 | 1909（明治42）年 | 『何処へ』（1910年・易風社） | 28,598 |
| 或る女 | 有島武郎 | 1911（明治44）年 | 『或女　前編』（1919年・叢文閣） | 84,822 |
| あらくれ | 徳田秋声 | 1915（大正4）年 | 『あらくれ』（1915年・新潮社） | 71,840 |
| 腕くらべ | 永井荷風 | 1916（大正5）年 | 『腕くらべ』（1917年・十里香館） | 65,536 |
| 田園の憂鬱 | 佐藤春夫 | 1918（大正7）年 | 『田園の憂鬱　或は病める薔薇』（1919年・新潮社） | 47,048 |
| 蔵の中 | 宇野浩二 | 1919（大正8）年 | 『蔵の中』（1919年・聚英閣） | 19,734 |
| 暗夜行路 | 志賀直哉 | 1921（大正10）年 | 『暗夜行路』（1930年・新潮社） | 36,836 |
| 無限抱擁 | 瀧井孝作 | 1921（大正10）年 | 『無限抱擁』（1927年・改造社） | 31,144 |
| 伸子 | 宮本百合子 | 1924（大正13）年 | 『伸子』（1928年・改造社） | 35,082 |
| 檸檬 | 梶井基次郎 | 1925（大正14）年 | 『檸檬』（1932年・武蔵野書院） | 2,908 |
| 伊豆の踊子 | 川端康成 | 1926（大正15）年 | 『伊豆の踊子』（1930年・先進社） | 10,318 |
| 計 | | | | 696,708 |

なお本コーパスでは、特定の作品（著者）に言語量が偏ることを避けるために、作品当たりの収録の上限（10 万語）を設定している。21 作品のうち 7 作品は、表 2 のように、その中途までをコーパスに収録している。

冊で分割されている作品（『浮雲』『吾輩は猫である』『或る女』）については、その 1 巻目を採用し、1 冊に全編が収められている作品（『思出の記』『暗夜行路』『無限抱擁』『伸子』）については、3 万語以上の規模が確保できる範囲で、各作品の冒頭から、きりのよいところまでを採用した。

表 2　部分的に採用した作品の収録範囲

| 作品名 | 収録範囲 | 収録範囲外 | 作品全体の語数 |
|---|---|---|---|
| 浮雲 | 第一篇 | 第二篇、第三篇 | 約10.5万語 |
| 思出の記 | 一の巻-二の巻 | 三の巻-十の巻、巻外 | 約23.0万語 |
| 吾輩は猫である | 上巻（-第五） | 中巻、下巻 | 約21.2万語 |
| 或る女 | 前編 | 後編 | 約17.6万語 |
| 暗夜行路 | 第一部 | 第二部-第四部 | **約18.4万語** |
| 無限抱擁 | 第一部-第二部 | 第三部-第四部 | **約10.3万語** |
| 伸子 | 第一部-第二部 | 第三部-第七部 | **約15.6万語** |

### ２．３　各作品の収録範囲とコア・非コアの設計

本コーパスには、形態論情報の人手修正が一通り入った「コアデータ」と、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」の、2 種類のデータセットを含む。本コーパスでは、コアデータを「作品の冒頭 1 万語程度のきりのよいところまで」と設定しており、全 21 作品にコアデータが存在する。

このような設計にした背景には、非コアデータの形態素解析用の辞書の教師データに、人手による形態論情報の修正を施した各作品のコアデータを用いることで、非コアデータの解析精度の向上を狙ったことがある。また、コアデータの位置を作品の冒頭と定めたのは、形態素解析を行う上で誤解析を起こしやすい固有名詞（人名や地名など）が、作品の冒頭に出現しやすいため、これらの修正を行うことで非コアデータの解析精度の向上、ならびに人手による修正作業の効率化を図ったことによる。

こうした設計のもとで構築した近代小説の短単位バージョン 1.0 における形態論情報の精度（適合率）は、コアデータが 99.6%、非コアデータが 97.3%となっている[2]。

### ３．サンプル

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（章相当）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 3 にあげる。

表 3 の基準によると、例えば、『伸子』の第 2 部第 2 章のサンプル ID は「60N 伸子 1924_12002」に、『或る女』の前編の第 11 章のサンプル ID は「60N 或女 1911_11011」になる。

[2] 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。

表 3　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 | N | サブコーパスのジャンル（小説、NovelのN）を表わす。全サンプルで共通。 |
| 4~5 | 作品ID | 各作品のタイトルの冒頭2文字を表わす。 |
| 6~9 | （4桁の数字） | 各作品の発表（初出）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） | 冊レベルの通し番号を表わす。（前編や上巻ならば1、後編や下巻ならば2、など）※1 |
| 12 | （2桁の数字） | 部レベルの通し番号を表わす。※2 |
| 13~15 | （3桁の数字） | 章レベルの通し番号を表わす。 |

※1 バージョン1.0では、前編や上巻のみの収録であるので、全サンプルが1となる。
※2　部構成になっていない作品は全サンプルが1となる。

## 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 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

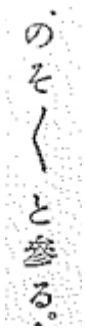
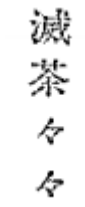
ス』等の国立国語研究所構築の他のコーパスと斉一な形態論情報を付与するため、形態素解析辞書 UniDic を使用した形態素解析に基づき形態論情報を付与した。そのため、コーパスのテキストを UniDic による形態素解析に適したものとするため、底本のテキストに対して以下の A）～D）にあげる改変（ここでは「校訂」と呼ぶ）を施し、コーパスのテキストを作成した。

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

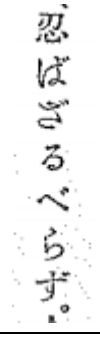
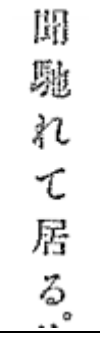
表 4　踊り字の電子化例

| 種類 | 例 | コーパステキスト | 原文 KWIC・原本文字列 |
|---|---|---|---|
| ノ＼ |  | のそ**のそ**と參る。 | のそ**ノ＼**と參る。 |
| 々々 |  | 滅茶**滅茶** | 滅茶**々々**。 |

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。

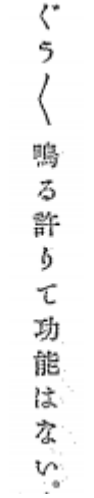
表 5　誤植の電子化例

| 種類 | 例 | コーパステキスト | 原文 KWIC・原本文字列 |
|---|---|---|---|
| 脱字 |  | 忍ばざるべ**か**らず。 | 忍ばざるべらず。 |
| 誤字 |  | 聞**馴**れて居る。 | 聞**馳**れて居る。 |

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 6　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 |  | ぐうぐう鳴る許り**で**はない。 | ぐう／＼鳴る許り**て**はない。 |

D）ハイフン・点線

本コーパスでは、「―（ハイフン）」と「…（点線）」は、原本における長さ（何文字分か）・個数によらず、それぞれ全角 1 文字分の「―」と「…」に校訂している。なお、「原文 KWIC」「原文文字列」で

も、原本における長さ・個数の再現は行わず、コーパステキストと同様に表示される。

## ５．形態論情報

### ５．１　形態論情報の概要

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。

### ５．２　形態論情報の多重化

本コーパスの形態論情報の特長としては、「明治・大正編」のコーパスとしては初めて「形態論情報の多重化」を行ったことが挙げられる。「形態論情報の多重化」は、同一の文字列に複数の形態論情報を付与する機能である（小木曽 2017）。本コーパスにおいては、短単位をまたがるルビを持つ文字列において、従来の「明治・大正編」のコーパスとは異なる処理を行っている。

表 7　形態論情報の多重化を行った短単位をまたがるルビが付されたレコード

| 例 | 従来のコーパスでの処理 | **本コーパスでの処理** | |
|---|---|---|---|
| 「毎時」（いつも） | 例外的に語彙素「いつも」を適用 | 主本文 | 文字列を保持したまま「何時｜も」を付与 |
| | | 副本文 | 「毎時」の形態論情報を付与 |
| 「吾夫」（うちのひと） | 読みを無視した形態論情報を付与（我が｜夫） | 主本文 | 文字列を保持したまま「家｜の｜人」を付与 |
| | | 副本文 | 「我が｜夫」の形態論情報を付与 |

本来であれば「いつ｜も」のように 2 短単位に分割をする場合であっても、「毎時（いつも）」のように表記上分割ができない場合には、例外的に語彙素「いつも」を適用していたが、本機能を用いることで、文字列（毎時）を保持したまま「いつ」と「も」の形態論情報を付与することが可能になった。また、「吾夫（うちのひと）」の例では、同じく短単位をまたがるルビが付されており、従来のコーパスでは読み（ルビ）を無視して文字列通りの形態論情報を付与していたが、こちらも同じく文字列（吾夫）を保持したまま、読み通りの形態論情報の付与を実現した。さらに、本コーパスでは形態論情報の多重化を行った案件に限り、副本文として、文字列の読み（表 7 の「毎時」と「我が｜夫」）の形態論情報を付与している。副本文は、「中納言」の検索ページの「検索動作」欄にある、「副本文」と記されたタブをクリックし、「副本文を検索対象に含まない（デフォルト）」を「副本文を対象に含む」に変更することで、検索対象に含まれるようになる。

また、同機能を応用することで、「不殘（のこらず）」のような返読要素にも対応することが可能となった。文字列（不→殘）と形態論情報（「残る」→「ず」）の順序が異なり、従来のコーパスでは形態論情報を順序正しく付与することが不可能であったため、文字列を返読したコーパステキスト（殘らず）を用いる必要があった。返読要素についても、短単位を超えたルビを持つレコードと同様に、次の表 8 のように、同一の文字列（不殘）に対し、「残る」と「ず」の複数の形態論情報を付与することを実現

した。

表 8　返読要素のあるレコードの処理

| 例 | 従来のコーパスでの処理 | **本コーパスでの処理** | |
|---|---|---|---|
| 「不殘」<br>（のこらず） | 文字列を返読し、「殘ず」にテキストを加工 | 主本文 | 文字列を保持したまま「残る｜ず」を付与 |
| | | 副本文 | 「対象語無し」の形態論情報を付与 |
| 「失禮乍」<br>（しつれいながら） | 文字列を返読し、「失禮乍」にテキストを加工 | 主本文 | 文字列を保持したまま「失礼｜ながら」を付与 |
| | | 副本文 | 「対象語無し」の形態論情報を付与 |

## 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）。

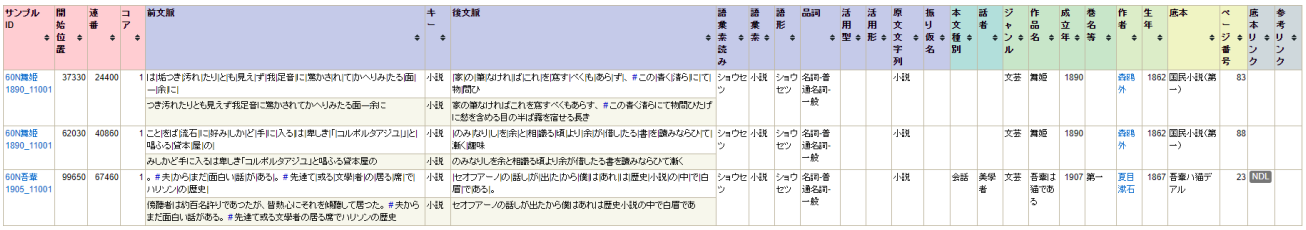



図 1　「中納言」の検索結果のイメージ

「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「主本文」「多重化種別」「文体」「出版社」（表中に「*」を付す）について、表 9 に内容を示す。

表 9　「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コア 9 データであることを表す。「1」がコア、「0」が非コアを表す（2.3 項参照）。 |
| | 主本文* | 主本文（主たる読み、本コーパスではルビ通りの読み）と副本文（副たる読み、文字列の読み）の区別を表す。「1」が主本文、「0」が副本文を表す（5.2 節参照）。 |

| | | |
|---|---|---|
| | 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（5.2 節参照）。 |
| 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。 |
| | 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。<br>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。<br>言いよどみ…会話の中での言いよどみにあたる文字列。<br>例：お、親方様、ゑゝありがたうござりまする、<br>漢文 …漢文の文字列。<br>外国語 …外国語の文字列。<br>未知語 …形態論情報の付与を保留した文字列。 |
| 情報種別 | 項目名 | 内容 |
| 形態論情報 | 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。<br>会話 …会話・独話・心内発話等の引用<br>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 |
| | 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。<br>文語…文語体。文末辞が「なり」「たり」「つ」「ぬ」「き」「けり」のもの。<br>口語…口語体。文末辞が「だ」「ぢや」「である」「です」「ます」のもの。<br>漢文…漢文。<br>外国語…漢文以外の外国語の文。<br>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。本コーパスでは全て「文芸」である。 |
| | 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。 |
| | 成立年 | 検索対象の含まれるサンプルが収録された資料の初出の年。コーパスのテキストの底本となる初出本とは異なるので注意されたい。 |
| | 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。底本テキスト中にタイトルがない場合は内容を表す名付けを〔 〕に括って示す。序・跋等以外の主要本文のタイトルは「〔本文〕」と示す。 |
| 情報種別 | 項目名 | 内容 |
| 作者情報 | 作者 | 検索対象の含まれるサンプルの著者名。後ろに「(作)」を付けて示す。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。<br>「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)。「『一葉全集』<1897>」のように、書籍名を『』で、発行年を<>で示す。 |
| | ページ番号 | 検索対象の底本における出現ページ番号。 |
| | 出版社* | 底本の出版社を示す。 |

| | | |
|---|---|---|
| その他 | 底本リンク | 検索対象の底本画像へのリンク。国立国会図書館所蔵本画像へのリンクを NDL 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-）による研究成果である。

## 参考文献

小木曽智信（2017）「多重の読みを持つテキストのコーパス化」『言語資源活用ワークショップ 2016 発表論文集』国立国語研究所、p.159-162.

国立国語研究所（1980）「用例採集のための主要文学作品目録」

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.

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.

髙橋雄太、服部紀子、小木曽智信（2019）「明治・大正期の文学作品コーパスの設計とその課題」『日本語学会 2019 年度秋季大会予稿集』

## 参考 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/