『日本語歴史コーパス　明治・大正編Ⅱ教科書』（短単位データ 1.0）

テキストの凡例と「中納言」表示項目について

2018 年 10 月 15 日　服部紀子　間淵洋子　近藤明日子

１．はじめに

『日本語歴史コーパス　明治・大正時代編Ⅱ教科書』は，近代の国定教科書制度のもと，小学校[1]および高等小学校で使用された国定国語科教科書を収録するコーパスである．小学校教科書のデータは『国定読本用語総覧　CD-ROM 版』（国立国語研究所編，1997）を作成する際に用いられた本文のテキストデータを，高等小学校教科書のデータは別途作成していた『高等小学読本』の形態論情報付きコーパス（近藤・小木曽・加藤，2010）のデータを基礎として，それぞれ『日本語歴史コーパス』の設計に合わせて再構築・統合したものである．

本コーパスの本文は,『日本語歴史コーパス（CHJ)』の他の時代のコーパスと整合的な形態論情報を付与するために，教科書本文に対して校訂を加えたものとなっている．また，本コーパスの本文は，収録対象教科書の全文を含んだものになっているが，検索対象サンプルは教科書内の 1 課分を単位として分割されており，各サンプル（＝各課）は個別の ID と書誌的情報（期・学年・原資料等に関する情報）を有する．

この文書では，コーパス本文の成り立ちと，検索アプリケーション「中納言」における検索結果の表示項目について，例示しながらその概要を示す．

なお，本コーパスでは，研究上必要と思われる情報を，できるだけ原文の状況に即して記述するよう努めたが，不十分・不適切な箇所が残存する可能性もある．適宜，原資料の情報を基に，原文を確認されることを推奨する．

２．テキストの凡例

２．１　テキストに使用する文字

本コーパスの本文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）に準拠している[2]．また，この文字集合に含まれない漢字については，以下の順で，異体の文字に包摂・代用した[3]．

(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本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．

---

1 本コーパスでまとめて「小学校」と呼ぶ教育課程は，実際には各時期の学校制度改革によって呼称が「尋常小学校」→「国民学校初等科」→「小学校」と変遷している．

2 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，③Unicode における CJK 統合漢字拡張 B については，これを用いない．

3 異体字の拡張包摂および代用については，須永・堤・高田（2011），須永ほか（2013）を参照のこと．

(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) (2)による代用が不可能な文字は，外字として「=」（げた記号，JIS 面区点 1-02-14，U+3013）で表す．

２．２　本文校訂

本コーパスでは，本文に対して，『日本語歴史コーパス（CHJ）』の他の時代や『現代日本語書き言葉均衡コーパス（BCCWJ）』等の他のコーパスと整合的な形態論情報を提供することを目的として，形態素解析辞書 UniDic に基づいた短単位情報を付与している．そのため，UniDic による形態素解析に適した解析用本文を用意する必要があり，原資料の教科書本文に対して，いくつかの改変を施している．

例えば，国定国語科教科書には，漢字カタカナ交じり文，踊り字，濁点無表記，棒引き仮名遣い，分かち書きなど，UniDic が対応できないような特徴的な表記が見られる．そのため，これらを現代の一般的な表記法に変換して解析用本文として用いる必要がある．そこで，漢字カタカナ交じり文は漢字ひらがな交じり文へ，踊り字は通常の文字へ，棒引仮名遣いは通常の母音表記へそれぞれ変換し，分かち書きのスペースは削除するなど，自動形態素解析や検索に適した形式に改変した．このような，近代に特有の表記に対して原資料の本文を改変することを，ここでは「校訂」と呼ぶ．

以下に，具体的な校訂の方法について，事例に基づき説明する（実例の出典をサンプル ID により示す）．なお，解析用の本文はこれらの校訂後の本文であるが，「中納言」では，校訂前の文字列を原文の状態に近い形で電子化したものを「原文文字列」として表示させることができるほか，一部で原資料の画像を参照することもできる．利用に際しては，必要に応じて原文の情報を確認されたい．

［漢字カタカナ混じり文］

漢字カタカナ混じり文は，カタカナ部分をひらがなに置き換える．その際，外来語等，現代もカタカナ表記が一般的なものは，置き換えの対象としない（例：60T 小読 1910_25B17）．

原文　敷キタルレールニヨリテ

校訂　敷きたるレールによりて

［棒引き仮名遣い］

棒引きによる仮名遣いは現代の仮名遣いに置き換える（例：60T 小読 1904_13A19）．

原文　たいそー、おもしろいところです。

校訂　たいそう、おもしろいところです。

［分かち書き］

分かち書きのスペースは削除する（例：60T 小読 1904_12A18）．

原文　はまぐり　を　おとしました。

校訂　はまぐりをおとしました。

なお，「中納言」では「原文文字列」と「原文 KWIC」において，分かち書きのスペースを「□」で示している（図 1）．

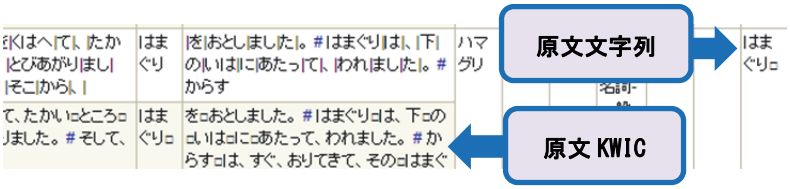

図 1　「中納言」での表示例

［踊り字］

踊り字は，繰り返される文字・語に置き換える（例：①60T 小読 1918_35B08，②60T 小読 1918_35B12）．

① **原文**　高い／＼　／　**校訂**　高い高い

② **原文**　此のまゝ　／　**校訂**　此のまま

［誤植］

原文の誤植（誤字・脱字・衍字）と思われる表記は，訂正する（例：60T 小読 1947_66C06）．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[4]などは，訂正の対象としない．

原文　どんな子だって、おかさんはひとりぎりです。

校訂　どんな子だって、おかあさんはひとりぎりです。

３．テキストの範囲とサンプル

本コーパスはに収録されるテキストは，表紙・目次・図表・刊記等を除く全ての本文である．

教科書は，複数の課を収録したものであり，それぞれの課は，個別の文章・作品であるため，本コーパスにおいては，それぞれの課を個別の文章として区別し，それぞれに書誌的・言語的な情報を付与することとした．そして，この方針に基づき，課を単位とした「サンプル」というテキストの範囲を定めた．

個々の「サンプル」は，サンプル ID という個別に認識される ID を持つ．サンプル ID は 15 桁からなり，構成は以下の通りである（網掛けは記号や数値の意味を表す）．

表１　本コーパスにおけるサンプル ID の構成

<table>
<tr><td colspan="2">1-2桁目</td><td colspan="2">3桁目</td><td colspan="2">4－5桁目</td><td>6－9桁目</td><td>10桁目</td><td>11桁目</td><td>12桁目</td><td>13桁</td><td>14-15桁目</td></tr>
<tr><td colspan="2">時代通し番号</td><td colspan="2">ジャンル</td><td colspan="2">作品ID</td><td>使用開始年</td><td>区切り記号</td><td>期</td><td>学年</td><td>上・中・下巻</td><td>出現順通し番号</td></tr>
<tr><td rowspan="2">60</td><td rowspan="2">明治・大正</td><td rowspan="2">T</td><td rowspan="2">教科書</td><td>小読</td><td>小学校国語科教科書</td><td>1904, 1910, 1918, 1933, 1941, 1947</td><td rowspan="2">－</td><td>1,2,3,4,5,6</td><td>1,2,3,4,5,6</td><td>A, B, C</td><td rowspan="2">課の連番2桁</td></tr>
<tr><td>高読</td><td>高等小学校国語科教科書</td><td>1904</td><td>1</td><td>1,2,3,4</td><td>A, B</td></tr>
</table>

このうち，ID 末尾 2 桁目の課の連番について，以下に教科書の構造との関係を示す．

本コーパスに収録した教科書本文には，個々の課ではなく，教科書そのものについての記載が含まれている（図 2 の白地部分）．これらは，教科書全体を構成する本文ではあるが，個々の課（図 2 の網掛け部分）とは性質を異にしているため，本コーパスでは，課とは別に，“教科書本体の構造要素”としてひとまとまりに扱っている．この“教科書本

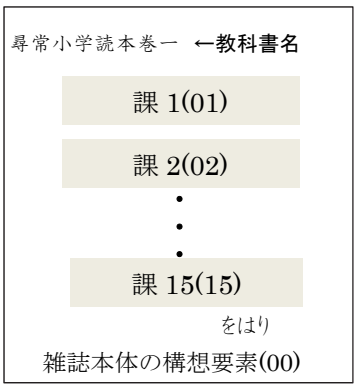

図 2　教科書の構成

[4] 語形のバリエーションかどうかの判定は，原則として『日本語国語大辞典　第二版』による．「見出し」にある場合のほか，「語義説明」内に“(「○○」とも)”と異語形を示す場合を語形のバリエーションと判断する．また，通用の異表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おけ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みなす．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，複数の記事に出現し，出現数が少なくない表記を異表記とみなす．

体の構造要素”を連番「00」とし，内部の個々の課について，出現順に「01」「02」…と ID を与える．連番「00」は，教科書全体の本文から，連番「01」以下の個々の課の本文を取り除いた，非常に特殊なテキストになっている．利用にあたっては，この点に留意されたい．

４．「中納言」における検索対象の選択方法と表示項目

４．１　検索対象の選択方法

１節において述べたように，本コーパスには 2 種の教育課程で使用された国語科教科書が収録されている．検索する際には「中納言」の「検索対象の選択」画面（図 3）において，それぞれの教育課程および期を選択し，さらに学年を選択することが可能である．

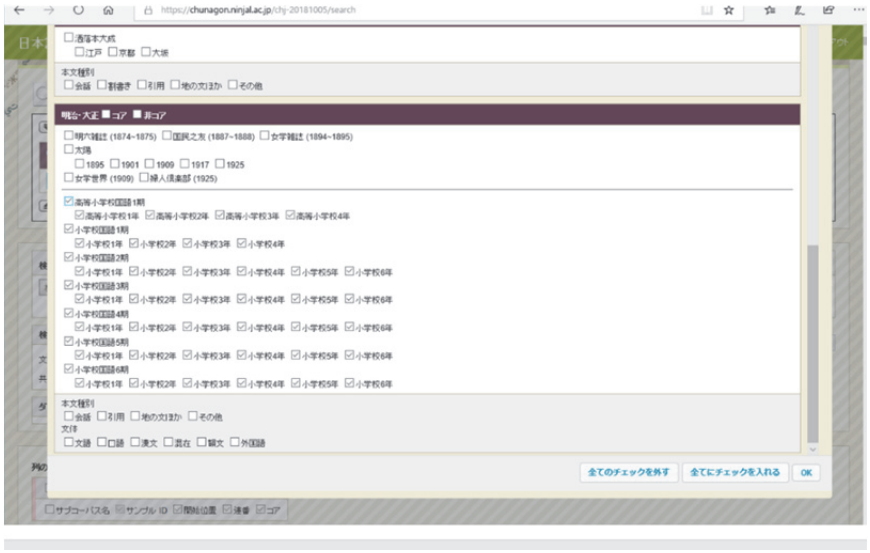

図 3　検索対象の選択画面

４．２　表示項目

本コーパスは，言語研究を目的とした利用に資する，様々な情報を有している．これらの情報は，コーパス検索アプリケーション「中納言」において参照可能な形で提供される（図 4）．

図 4　「中納言」の検索結果表示画面

表 2 に，「中納言」の検索画面で参照できる付加情報のうち，初期設定で表示される主な項目と，「明治・大正時代編Ⅱ教科書」で特に注意が必要な項目「部」「語形代表表記」「本文種別」「出版社」について，内容を示す．

表 2　「中納言」の検索結果表示項目（「*」をつけたものはオプション項目）

情報種別	項目名	内容
コーパス情報	サンプル ID	検索対象が含まれるサンプルの ID（→3 節）.
	連番	検索対象の，サンプル内における短単位の連番.
形態論情報[5]	前文脈	上段に校訂後の前方文脈，下段に校訂前の前方文脈を示す.
	キー	検索対象の書字形出現形（表記形）.
	後文脈	上段に校訂後の後方文脈，下段に校訂前の後方文脈を示す.
	語彙素読み	検索対象の語彙素（下記項目「語彙素」参照）の読み．カタカナ表記である.
	語彙素	検索対象の語彙素の表記．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・漢語は漢字ひらがな表記，外来語・人名・地名はカタカナ表記である.
	語形	検索対象の語形．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で，カタカナ表記である.
	語形代表表記*	検索対象の語形に対する代表的な表記形で，語彙素に準じた表記である.
	品詞	検索対象の品詞で，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.
	活用型	検索対象活用語の活用の型．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．検索対象の本文情報「文体」項目（下記項目「文体」参照）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．ただし，口語文体内にあっても口語活用型が存在しない語や活用形（文語形容詞「ごとし」，文語助動詞「き」，文語二段活用の連体形語形など）については文語活用型を用い，同様に，文語文体内にあっても文語活用型が存在しない語や活用形（口語形容詞「ない」，口語動詞「ある」等の終止形など）については，口語活用型を用いた(例：60T 高読 1904_14B04)． 例）攻め來らば、諸子之を如何せんと<u>する</u>。（サ行変格/終止形-一般）

[5] 形態論情報の個々の項目の意味は，「原文文字列」「振り仮名」を除き，UniDic の見出しに対応している．詳細は小椋ほか(2011)，国立国語研究所コーパス開発センター（近藤明日子）編(2016)を参照されたい．

ただし，漢文・外国語等，一部の文字列で UniDic ベースの形態論情報を付与しないものがある．これらは，形態論情報に関する項目のうち，「前文脈」「キー」「後文脈」「品詞」「原文文字列」「振り仮名」以外の項目が空欄となっている．

情報種別	項目名	内容
	活用形	検索対象活用語の活用形．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：60T 小読 1904_14A20)． 例）軍艦などは、なにをめあてに、航路を**きむる**。(文語下二段-マ行/連体形-一般)
	原文文字列	検索対象の校訂前本文（原文）の文字列（→2.2 節）．
	振り仮名	検索対象に付された振り仮名の文字列．原資料における振り仮名の誤植は訂正したものを示す．訂正の基準は本文校訂における誤植の判定（→2.2 節［誤植］）に準ずる．
本文情報	本文種別	検索対象が「地の文」以外の場合の，その種別．文献等からの引用等は「引用」，会話・独話・心内発話等の引用は「会話」と示す．また，会話の話者を表示する部分は「その他-話者」と示す．
	話者	検索対象の本文種別が「会話」である場合の話者名，「引用」である場合の典拠文献名や著者名，用例の引用である場合は，「用例」と示す．
	文体*	検索対象が含まれる文の文体．「文語」「口語」の別を示す． なお，一つのサンプル内で，引用文に異なる文体が用いられているなど，複数の文体が混在しているものは，引用範囲について個々に文体情報を付与する．
作品情報	ジャンル	検索対象が含まれるサンプルのジャンル．すべて「国語科教科書」とする．
	作品名	検索対象の含まれるサンプルが収録された教科書が使用された「教育課程」「教科」「期」を示す．
	成立年	検索対象の含まれるサンプルが収録された教科書の使用開始年．教科書の刊行年ではないことに注意が必要である．
	巻名等	検索対象の含まれるサンプル（課）のタイトル．サンプル ID 下 2 桁（記事連番）が「00」のサンプルは空欄とする．
	部*	検索対象の含まれる教科書が使用された「教育課程」「学年」を示す．
作者情報	作者	検索対象が含まれるサンプルの作者名，すべて「文部省」とする．
	生年	検索対象が含まれるサンプルの作者の生年．すべて空欄で示す．
底本情報	底本	検索対象が含まれる原資料名．
	ページ番号	検索対象の原資料における出現ページ番号．
	出版社*	検索対象が含まれる原資料の出版社名．すべて「文部省」とする．

情報種別	項目名	内容
その他	外部リンク	検索対象の原資料画像へのリンク．小学校教科書（第 1 期）と高等小学校教科書は国立国語研究所蔵本画像へのリンクを「 Ninjal 」ボタンで示す．ボタンをクリックすると検索対象の出現するページの画像がブラウザで表示されるなお，小学校教科書の第 2 期～第 6 期については原資料への画像リンクを提供していない．

参考文献

小椋秀樹ほか（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（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.

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（編）（1997）『国定読本用語総覧　CD-ROM 版』三省堂.

近藤明日子・小木曽智信・加藤文明子（2010）「『高等小学読本』の形態論情報付きコーパス」『じんもんこん 2010 論文集』2010:15, pp.189-194.

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.

須永哲矢ほか（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.