『日本語歴史コーパス　明治・大正編Ⅱ教科書』明治初期理科教科書コーパス
概説書

2023 年 3 月 31 日　髙橋雄太・田中牧郎

１．はじめに

本文書は、2023 年 3 月に『日本語歴史コーパス　明治・大正編Ⅱ教科書』（以下、『日本語歴史コーパス』を CHJ と称する）に追加公開された、明治初期理科教科書の『物理階梯』『小学化学書』『初学人身窮理』のコーパスについて、その概要を説明するものである。上記の 3 つの理科教科書のコーパス作成は、明治大学大学院国際日本学研究科田中牧郎研究室で進められていたものを、国立国語研究所の第 4 期プロジェクトである「開かれた共同構築環境による通時コーパスの拡張」における OpenCHJ の枠組みで公開できるコーパスとして整備し直したものである。OpenCHJ では、コーパスの開発・構築を国立国語研究所外で、コーパスの開発協力や公開を国立国語研究所でそれぞれ受け持つ共同構築環境を展開することとしており、本コーパスをそのモデルケースの一つに位置づけるものである。

各教科書の概要、コーパスの構築方法や活用例については田中・髙橋（2023）に説明を譲り、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）の項目について解説する。

２．コーパスの概要

本コーパスの収録作品と短単位の概数の一覧を、表 1 に示す。

表 1　収録作品と短単位数の一覧

収録作品	収録年	巻	サンプル（章・課）数	短単位数（万）
物理階梯	1872（明治5）年	上巻、中巻、下巻	41	4.1
小学化学書	1875（明治8）年	一巻、二巻、三巻	24	2.7
初学人身窮理	1876（明治9）年	上巻、下巻	18	2.4
計			83	9.2

※サンプル数は総論や付録を含む　※短単位数は記号類を抜いた数

本コーパスは全編が、人手修正の入った「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、人手により修正することで付与した。

なお、CHJ は長単位と短単位で解析されているが、「明治・大正編」は短単位のみであるため、本コーパスも短単位情報のみを付与した。

３．サンプル ID

テキストをコーパスに収録する際に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・課相当）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・大正編」を表わす。
3	T	サブコーパスのジャンル（教科書、TextbookのT）を表わす。全サンプルで共通。
4-5	作品ID	それぞれの教科書の名称の一部からとる。
6-9	（4桁の数字）	サンプルの成立年を西暦で表わす。
10	_	サンプルIDの区切り記号（アンダーバー）。
11-12	（2桁の数字）	巻数を表わす。
13-15	（3桁の数字）	章や課などのサンプルごとの通し番号を表わす。

表 2 の基準によると、『物理階梯』の中巻の第 15 課のサンプル ID は「60T 物理 1872_02015」に、『小学化学書』の第 3 巻の第 21 章のサンプル ID は「60T 化学 1874_03021」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字形の漢字や仮名については文字集合内の漢字や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。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、CHJ 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）漢字平仮名交じりへの変換

CHJ では、形態素解析や形態論情報の整備の効率化のため、外来語などの一部を除き、漢字片仮名交じりテキストを漢字平仮名テキストに変換してコーパステキストを作成している。

「中納言」での検索は、コーパステキストに基づいて行われる。

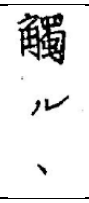
表 3　コーパステキストと原文 KWIC・原文文字列欄の表示の例

テキストの種類	例
コーパステキスト	人の五官に觸るるもの、之を「ナチュール」と云ふ、
原文 KWIC・ 原本文字列	人ノ五官ニ觸ルヽモノ、之ヲ「ナチュール」ト云フ、

B）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 字の漢字を繰り返す「々」「ゝ」は置き換えの対象としない。

表 4　踊り字の電子化例

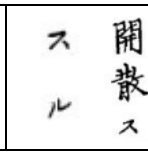
種類	例	コーパステキスト	原文 KWIC・原本文字列
ヽ		觸る**る**	觸ル**ヽ**

C）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

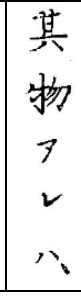
種類	例	コーパステキスト	原文 KWIC・原本文字列
誤字		**網**膜	**綱**膜

衍字		開散する	…開散ス<u>ス</u>ル…

D）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、該当の濁音を表わ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		其物あれ<u>ば</u>、	其物アレ<u>ハ</u>、

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。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。

本コーパスの文体は明治期に広く見られる文語文に近いものであったため、CHJ「明治・大正編」の文語テキストの規程である近藤（2016）に準拠して形態論情報の整備を行った。

短単位をまたがるルビ（例：眼底〔メ｜ノ｜ソコ〕、早晩〔イツ｜カ〕）が付された事例についても、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。

なお、理科教科書では図 1 のように、随所に左ルビが付されている。本コーパスでは右ルビを読みとして採用し、通常の検索では図 1 の例でいえば「千態（センタイ）｜万状（バンジョウ）」のみが検索対象となる。ただし、本コーパスでは主本文とは別に、副本文（従となる読み）として左ルビの形態論情報を付与しており、次の図 2 のように、「中納言」の「検索動作」の「副本文」の項目のプルダウンから、「副本文を検索対象に含む」を選択した状態で検索を行うと、左ルビの形態論情報を参照することができる。この動作を行うことにより、左ルビの形態論情報である「様々（サマザマ）｜の（ノ）｜形（カタチ）」が検索対象に含まれるようになる。

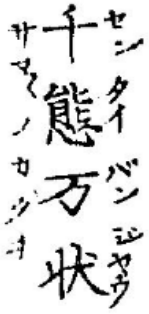


図1　理科教科書の左ルビの例

図 2　「中納言」の「検索動作」の変更手順

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 3）。


4 件の検索結果が見つかりました。
検索対象語数：95,826　記号・補助記号・空白を除いた検索対象語数：90,176　検索対象サンプル数：83

サンプル ID	前文脈	キー	後文脈
60T物理 1872_01002	以|て|、|其|至極|する|所|を|知る|ヿ|能|ま|ず|、|例へば|、|一|厘|七|毛|の|	金	|(の)||如き|、|金匠|之|を|鎚延|して|、|五|寸|八|分|三|厘|平方|を|蓋ふ|べき
	、人智ヲ以テ、其至極スル所ヲ知ルヿ能ハズ、例ヘハ、一厘七毛ノ	金	ノ如キ、金匠之ヲ鎚延シテ、五寸八分三厘平方ヲ蓋フヘキ薄葉トナ
60T物理 1872_01002	、|二百|箇|に|分つ|、＃故|に|此|二百|分|の|一||は、|即ち|一|厘|七|毛|	金	|(の)||大約|二百|万|分|の|一||に|當り|、|其|微細|なる|、|斯|の|如き||に|至る|と
	平方ヲ、二百箇ノ小片トナシテ、再ビ此一小片ヲ、二百箇ニ分ツ、＃故ニ此二百分ノ一ハ、即チ一厘七毛	金	ノ大約二百万分ノ一ニ當リ、其微細ナル、斯ノ如キ二至ルト雖モ、
60T化学 1874_03021	其|一||に|は|硝酸|を|加へ|一||に|は|塩化|水素|酸|を|加ふる|も|二|管|共||に|	金	|(の)||溶くる|を|見|ず＃然る||に|今|二|管|の|液|を|混同|すれ|ば|其|金|忽|消失
	管二入レ其一ニハ硝酸ヲ加ヘ一ニハ塩化水素酸ヲ加フルモ二管共ニ	金	ノ溶クルヲ見ス＃然ルニ今二管ノ液ヲ混同スレハ其金忽消失ス
60T人身 1876_01003	此|食物|の|プロッシ|にて|除き去り|難き|もの||は|象牙|の|細|楊枝|を|用ひ|て|之|を|掃り出す|べし＃	金	|(の)||楊枝|にて||は|彼の|イネメル|を|損ふ|ヿ|あれ||ば|決して|之|を|用ふ|可ら|ず＃朝夕||には
	シニテ除キ去リ難キモノハ象牙ノ細楊枝ヲ用ヒテ之ヲ掃リ出スベシ＃	カネ 金	ノ楊枝ニテハ彼ノイネメルヲ損フヿアレバ決シテ之ヲ用フ可ラズ＃朝夕ニハ必ス一回ソハ温湯ニテ口内ヲ嗽キプロッシヲ用ヒテ歯ノ前

図 3　「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について、表 7 に内容を示す。

表 7　「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わす 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データであることを表わす。「1」がコアを表わす。
	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わす。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）。

情報種別	項目名	内容
形態論情報	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。 全て片仮名表記である。
	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。 絵文字・記号等 …入力のできない絵文字。コーパステキストでは「**=**」で表示される。
	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。 本コーパスの収録対象の教科書はすべて文語体であるため、基本的には文語活用型を適用するが、一部の口語活用型にしかない語形（一段活用の連体形や終止形など）には口語活用型を適用している。
	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。 引用 …文献等からの引用
	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が表示される。
	文体	検索対象の含まれる文の文体。本コーパスでは全て「文語」が表示される。

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。一律で「理科教科書」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。各教科書名が表示される。
	成立年	各教科書の初版が刊行された年が表示される。
	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。本コーパスでは各教科書の章や課のタイトルが表示される。
作者情報	作者	検索対象の含まれるサンプルの著者名。 「国立国会図書館典拠データ検索・提供サービス（Web NDL Authorities）」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）。
	ページ番号	検索対象の底本におけるページ。和綴じ本では、見開きの左のページと、次の見開きの右のページに同じ番号が振られているため、便宜的に、前者には「5 表」、後者には「5 裏」といったように、「表」と「裏」で区別できるようにしている。
	出版社	底本の出版社が表示される。
その他	底本リンク	検索対象の底本画像へのリンク。本コーパスでは該当画像がないため空欄である。
	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。「NDL」または「Waseda」のアイコンをクリックすると、当該用例のある原本画像のページにアクセスすることができる。

参考文献

近藤明日子編（2016）「近代文語 UniDic 短単位規程集　Ver.1.1」https://clrd.ninjal.ac.jp/chj/doc/unidic-MLJ_rulebook_v1_1.pdf.

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.

田中牧郎・髙橋雄太（2023）「明治初期理科教科書コーパスの構築と活用―『物理階梯』『小学化学書』『初学人身窮理』を対象として―」『明治大学国際日本学研究』第 15 巻 1 号、pp.1-25. https://www.meiji.ac.jp/nippon/6t5h7p00000ifucc-att/mkmht0000002odr2.pdf

関連 URL

国立国語研究所「開かれた共同構築環境による通時コーパスの拡張」https://www.ninjal.ac.jp/research/cr-project/project-4/diachronic-corpus/

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://clrd.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/

画像リンク先

・『物理階梯』: 国立国会図書館デジタルコレクション

「物理階梯. 上巻」1874 年文部省蔵版 https://dl.ndl.go.jp/info:ndljp/pid/830260

「物理階梯. 中巻」1874 年文部省蔵版 https://dl.ndl.go.jp/info:ndljp/pid/830261

「物理階梯. 下巻」1874 年文部省蔵版 https://dl.ndl.go.jp/info:ndljp/pid/830262

・『小学化学書』: 国立国会図書館デジタルコレクション

「小学化学書. 一」1874 年文部省版 https://dl.ndl.go.jp/info:ndljp/pid/830814

「小学化学書. 二」1874 年文部省版 https://dl.ndl.go.jp/info:ndljp/pid/830815

「小学化学書. 三」1874 年文部省版 https://dl.ndl.go.jp/info:ndljp/pid/830816

・『初学人身窮理』: 早稲田大学図書館

「初学人身窮理. 巻之上」1878 年版

https://archive.wul.waseda.ac.jp/kosho/ya03/ya03_01070/ya03_01070_0001/ya03_01070_0001.pdf

「初学人身窮理. 巻之下」1878 年版

https://archive.wul.waseda.ac.jp/kosho/ya03/ya03_01070/ya03_01070_0002/ya03_01070_0002.pdf

明治初期理科教科書コーパス　―開発スタッフ―

開発主任

・髙橋雄太（明治大学国際日本学部助教／国立国語研究所プロジェクト非常勤研究員）

・田中牧郎（明治大学国際日本学部教授）

開発協力者

・淺野萌花、神田脩一郎、小松寛子、土屋葵、仲村怜、深田芽生、間淵洋子、ヤロシュ島田むつみ（明治大学大学院国際日本学研究科大学院生）

※肩書は開発当時のものである