『日本語歴史コーパス　明治・大正編Ⅱ教科書』（短単位データ 1.2）
ハワイ日本語教科書第１期　概説書

2024 年 3 月 31 日　ヤロシュ島田むつみ　竹内綾乃

１．はじめに

本文書は、2024 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）に追加された、布哇教育会編纂『日本語読本』第１期（以下、ハワイ日本語教科書第１期とする）のコーパスについて、その概要を説明するものである。本コーパスは、筑波大学附属図書館蔵本『日本語読本尋常科用』（布哇教育会第１期６巻分）を底本とし、全文テキストデータとして国立国語研究所日本語史研究用テキストデータ集で公開したもの[1]を基礎として、『日本語歴史コーパス』の設計に合わせて再構築したものである。

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）について解説する。

２．コーパスの概要

本コーパスの短単位数と課数を一覧に示すと、表１のようになる。

表１　収録課数と短単位数の一覧
（空白、補助記号削除）

学年	課数	短単位数（万）
１年生	26	0.3
２年生	52	0.8
３年生	54	1.2
４年生	56	1.6
５年生	68	2.4
６年生	68	2.9
合計	**324**	**9.2**

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。なお、短単位のみの実装で、長単位は未実装である。

[1] https://www2.ninjal.ac.jp/textdb_dataset/hn17/

３．テキストの範囲とサンプル ID

本コーパスには、表紙・目次・図表・新出漢字表・奥付・英文テキストを除く、全ての本文テキストが含まれている。

本コーパスでは教科書の各課を最小の文書構造として 1 サンプルに定め、それぞれにサンプル ID を付与している。サンプル ID は 15 桁からなり、その構成は表 2 の通りである（なお、網掛け部分は記号や数字の意味内容を表す）。

表 2　サンプル ID の構成

1-2桁目		3桁目		4-5桁目		6-9桁目	10桁目	11桁目	12桁目	13桁目	14－15桁目
時代通し番号		ジャンル		作品ID		使用開始年数	区切り記号	期	学年	巻数	課数
60	明治・大正	T	教科書	哇読	日本語読本	1917	―	1	1～6	1～6	課の連番(2桁)

1－2 桁目は時代区分を表しており、「明治・大正編」のサブコーパスはすべて 60 である。3 桁目はサブコーパスのジャンルであり、教科書は Textbook の T で表す。4－5 桁目は作品 ID であり、ハワイ（布哇）で使用されていた日本語読本（教科書）であることから、布哇の「哇」の字を使用し「哇読」としている。6－9 桁目はサンプルの成立年を西暦で表す。10 桁目はサンプル ID の区切り記号を表す。11 桁目はハワイ日本語教科書が発行された期を表し、本コーパスは第 1 期のみを収録しているので、全サンプル共通で「１」を付与している。12 桁目は学年を表す。13 桁目は巻数であるが、ハワイ日本語教科書第 1 期では学年と巻数が同じである（6 年生が 6 巻目となる）。14－15 桁目は課の番号を 2 桁で表す。

表 2 によると、第 1 期 3 年生 23 課のサンプル ID は「60T 哇読 1917_13323」である。第 1 期 6 年生 5 課のサンプル ID は「60T 哇読 1917_16605」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）に準拠している。 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）～（3）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。

４．２　テキストの校訂

本コーパスでは、CHJ の他の時代や『現代日本語書き言葉均衡コーパス（BCCWJ)』等、他のコーパスと整合的な形態論情報を提供する目的のため、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。形態素解析に適した本文にするため、教科書の本文テキストに対していくつかの改変を施した。例えば、ハワイ日本語科教科書には、漢字カタカナ交じり文、踊り字、濁点無表記、分かち書きのほか、テキスト内部に段落の境界を示す終了括弧（」）や、一つの語が二行にまたがって書かれる語が、一語であることを示す連結記号（‖）の表記などがある。これらを形態素解析や検索に適したテキストに改変した。

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。なお、校訂前の文字列は、「中納言」の検索結果においては「原文文字列」列と「原文 kwic」欄に表示される。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。

A）漢字カタカナ交じり文

漢字カタカナ交じり文は、カタカナ部分をひらがなに置き換える。その際、外来語など、現代もカタカナ表記が一般的なものは、置き換えの対象とはしなかった。

なお、ハワイ日本語教科書では、表４の例のように漢字カタカナ交じり文の中の外来語をひらがなで表記している場合がある。このようにひらがなで表記されている外来語部分はカタカナに置き換えた。

表３　漢字カタカナ交じり文電子化例

テキストの種類	例
コーパステキスト（校訂）	ボールがとなりのいへのまどにあたつてガラスの戸がこはれました。
原本文字列・原文kwic	ボール　ガ　トナリ　ノ　イヘ　ノ　マド　ニ　アタツテ　ガラス　ノ　戸　ガ　コハレマシタ。

（60T 哇読 1917_11112）

表４　漢字カタカナ交じり文（外来語ひらがな表記）電子化例

テキストの種類	例
コーパステキスト（校訂）	リンカーンは父に従つてオハイオ州に移り、後またイリノイス州に移りしが、
原本文字列・原文kwic	りんかーんハ父ニ従ツテおはいお州ニ移リ、後マタいりのいす州ニ移リシガ、

（60T 哇読 1917_16626）

B）踊り字

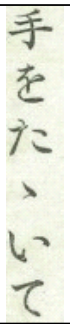
踊り字は繰り返される文字に置き換える。ただし、「国々」など、１短単位内部での直前の１字の漢字を繰り返す踊り字は置き換えの対象とはしなかった。

表５　踊り字例１

例	原文文字列 原文kwic	校訂
	なか／＼	なかなか

（60T 哇読 1917_16628）

表６　踊り字例２

例	原文文字列 原文kwic	校訂
	手をたゝいて	手をたたいて

（60T 哇読 1917_13301）

C）濁点無表記

濁点が期待される仮名に濁点が用いられていない場合は、濁点無表記とし、濁点付き仮名に置き換える。

表７　濁音無表記例

テキストの種類	例
コーパステキスト（校訂）	もちにする米とごはんの米はどう**ちがひ**ます　か。
原文文字列・原文kwic	もち　に　する　米　と　ごはん　の　米　は　どう　**ちかひ**ます　か。

（60T 哇読 1917_12240）

D）分かち書き

分かち書きの空白は削除する。

なお、各課のタイトルにおける分かち書きの空白は削除していない。

表８　分かち書き電子化例

テキストの種類	例
コーパステキスト（校訂）	ポイはじようぶんの多いたべものです。
原文文字列・原文kwic	ポイ　は　じようぶん　の　多い　たべ＝もの　です。

（60T 哇読 1917_12233）

E） 段落の境界を示す終了括弧（」）及び一語であることを示す連結記号（‖）

段落の境界を示すために付された終了括弧（」）及び、一語が二行にまたがる際に使用される連結記号（‖）はどちらも削除する。

表 9　終了括弧（」）例

例	原文文字列 原文kwic	校訂
	野へ出マス。」	野へ出ます。

（60T 哇読 1917_14430）

表 10　連結記号（‖）例

例	原文文字列 原文kwic	校訂
	サウ ウレシ‖	うれしさう

（60T 哇読 1917_11115）

左：段落の境界を示す（」）は削除する。

右：一語が二行にまたがる際に使用される連結記号（‖）は削除する。

F） その他

原文の誤植（誤字・脱字・衍字）と思われる表記は訂正する。

表 11　誤字の電子化例

テキストの種類	例
コーパステキスト（校訂）	衣川の柵は此の近所にあつ**たの**です。
原文文字列・原文kwic	衣川の柵は此の近所にあつ**のた**です。

（60T 哇読 1917_15512）

5．形態論情報

本コーパスでは、底本の本行テキストを主たる本文とし、口語のテキストは『『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版（下）』（小椋ほか 2011）に、文語のテキストは『近代文語 UniDic 短単位規程集 Ver.1.1』（近藤 2016）に基づいて形態論情報（語彙素・語彙素読み・品詞・活用型・活用形などに関する情報）を付与している。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、人手により修正している。テキストの読みは、右ルビがある場合はそれに従った。なお、ハワイ日本語教科書第 1 期（短単位バージョン 1.2）の形態論情報の精度（適合率）は、コアデータが 99.75%となっている[2]。

[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。

ハワイ日本語教科書第1期には国定教科書と共通する課が多く、これらのサンプルの形態論情報は、国定教科書における処理に基づいて付与している。ただし、部分的に国定教科書と処理を変えている箇所があり、具体的には次のように、短単位を超えるルビが付与されている文字列に対する処理を変えている。

表12　国定教科書と布哇日本語教科書で処理を変えている例

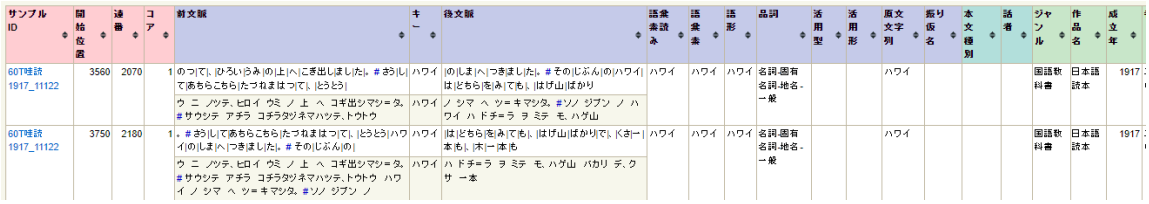
文字列	天叢雲剣	一
ルビ	あめのむらくものつるぎ	ひとつ
短単位規程に基づく単位境界	アメ｜ノ｜ムラクモ｜ノ｜ツルギ	ヒト｜ツ
国定教科書における処理	アメノ（語彙素「天」）｜ムラクモノ（語彙素「群雲」）｜ツルギ（語彙素「剣」）（3短単位）	ヒトツ（語彙素「一つ」）（1短単位）
布哇日本語教科書における処理	アメ｜ノ｜ムラクモ｜ノ（語彙素「天」「の」「群雲」「の」）｜ツルギ（語彙素「剣」）（5短単位）	ヒト｜ツ（語彙素「一」「つ」）（2短単位）

「天叢雲剣」の例のように、ルビにのみ「の」が現われる文字列の場合、国定教科書では語彙素「天」（アメ）や「群雲」（ムラクモ）に語形「アメノ」や「ムラクモノ」を立てて、短単位規程から外れた処理を行っていたが、国定教科書の構築・公開後に「中納言」に実装された、同一文字列に多重の形態論情報を付与する機能を用いて、布哇日本語教科書では短単位規程に基づく単位境界のままの形態論情報を付与している。同様に、「一」の例のように、国定教科書では表記上分割が不可能な場合にのみ適用する例外的な語彙素「一つ」を使用していたが、布哇日本語教科書では「一」の表記を維持したまま、「ヒト|ツ」の形態論情報を付与している。

なお、布哇日本語教科書に独自にみられる短単位を超えるルビとして、「中学校（ハイスクール)」などの外来語ルビが漢語文字列などに付与されているケースがあるが、これらも同様に、多重に形態論情報を付与する機能を用いて、「中学校」の文字列に「ハイ｜スクール」（2短単位）の形態論情報を付与している。

6．「中納言」における表示項目

本コーパスは言語研究を目的とした利用に対応する様々な情報を提供している。これらの情報は、コーパス検索アプリケーション「中納言」を通して利用者に提供される。



サンプルID	開始位置	連番	コア	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
60T唾読1917_11122	3560	2070	1	のつ|て|、|ひろい|さみ|の|上|へ|にぎ出しました|。|＃さう|し|て|あちらこちら|たづねまはつ|て|、|とうとう|	ハワイ	|の|しま|へ|つきました|。|＃その|じぶん|の|は|どちら|を|み|て|も|、|はげ山|ばかり	ハワイ	ハワイ	ハワイ	名詞-固有名詞-地名-一般			ハワイ			国語教科書	日本語読本	1917
				ウ ニ ノソテ、ヒロイ ウミ ノ 上 ヘ ニギ出シマシ＝タ。＃サウシテ アチラ コチラタヅネマハシテ、トウトウ	ハワイ	ノ シマ ヘ ツ＝キマシタ。＃ソノ ジブン ノ ハワイ ハ ドチ＝ラ ヲ ミテ モ、ハゲ山													
60T唾読1917_11122	3750	2180	1	。＃さう|し|てあちらこちら|たづねまはつ|て|、|とうとう|ハワイ|の|しま|へ|つきました|。＃その|じぶん|の|	ハワイ	|は|どちら|を|み|て|も|、|はげ山|ばかり|で|、|くさ|ー|本|も|、|木|ー|本|も	ハワイ	ハワイ	ハワイ	名詞-固有名詞-地名-一般			ハワイ			国語教科書	日本語読本	1917
				ウ ニ ノソテ、ヒロイ ウミ ノ 上 ヘ ニギ出シマシ＝タ。＃サウシテ アチラ コチラタヅネマハシテ、トウトウ ハワイ ノ シマ ヘ ツ＝キマシタ。＃ソノ ジブン ノ	ハワイ	ハ ドチ＝ラ ヲ ミテ モ、ハゲ山 バカリ デ、クサ ー本													

図 1　「中納言」の検索結果表示画面

表 12 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について、内容を示す。

表 12　「中納言」の検索結果表示項目

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章参照）
	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データまたは非コアデータのいずれであるかを表す。本コーパスは全編がコアデータであり、「1」が表示される。
	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。
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。
	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。 全てカタカナ表記である。
	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。

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。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。
	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章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。 会話…会話・独話・心内発話等の引用 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などが表示される。「会話」の場合はサンプル内で表示されている話者名（「おはな」など）や呼称（「おじいさん」など）が表示される。不明の場合は「＊」で表す。
	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」「けり」のもの。 口語…口語体。文末辞が「だ」「である」「です」「ます」のもの。 なお、一つのサンプル内で文体が混在しているものは、個々に文体情報を付与する。
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。一律で「国語教科書」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。ハワイ日本語教科書の場合、「日本語読本」が表示される。
	成立年	教科書の初版が刊行された年である「1917」が表示される。
	巻名等	検索対象の含まれるサンプル（課）のタイトル。
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「布哇教育会」が表示される。
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。

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』形態論情報規程集第 4 版（下）』（特別領域研究「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〈https://repository.ninjal.ac.jp/records/2872〉

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・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.