『日本語歴史コーパス　明治・大正編Ⅳ近代小説』の公開

髙橋雄太（たかはしゆうた）　服部紀子（はっとりのりこ）　小木曽智信（おぎそとしのぶ）（国立国語研究所）

1. はじめに

国立国語研究所では、上代から近代までの日本語を通時的に研究するための基礎資料として、『日本語歴史コーパス』（以下、CHJ）の構築を進めている。本発表では、その「明治・大正編」の第四弾として 2021 年 3 月に公開した『日本語歴史コーパス　明治・大正編Ⅳ近代小説』について、概要と特色を報告する。

2. コーパスの概要

2.1 概要

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 作品選定と各作品の収録範囲

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]。これにより、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、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] 詳細な作品の選定方法については、髙橋ほか（2019）、髙橋（2021）に説明を譲る。

表 1　収録作品と言語量

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、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**

2.2　底本テキストと中納言上での表示方法

本コーパスのコーパステキストの底本には、各作品の「単行本・全集などの書籍化されたもののうち最も出版年の古いもの」を採用している。底本を作品の初出ではなく書籍にした背景には、①書籍は原本（復刻版を含む）の入手が容易であること、②連載作品の初出は「タイトル」「著者名」「煽り文」などの文書要素を毎号に含みコーパステキストにする上で多くの校訂を伴うこと、③コーパスの利用者が容易に原本にアクセスできること、などの事情があった。

しかし、小説作品は多く初出の年がその作品の発表年として認知されていることがあるため、『CHJ』の検索アプリケーションツールである「中納言」上（あるいはダウンロードした EXCEL ファイル）では、初出と底本の双方の情報が表示されるようにした。具体的には、次の表 3 のように、作品の初出の情報は「サンプル ID（の 6-9 桁)」欄と「成立年」欄に発表年を、底本の情報は「底本」欄と「出版社」欄にそれぞれ書籍名と出版年、出版社を示している。

表 3　作品情報と底本情報の「中納言」上での表示

サンプル ID	作品名	成立年	底本	ページ番号	出版社
60N 舞姫 1890_11001	舞姫	1890	国民小説（第一）<1890>	83	民友社
60N 何処 1909_11003	何処へ	1909	何処へ<1910>	26	易風社
60N 腕く 1916_11001	腕くらべ	1916	腕くらべ<1917>	7	十里香館
60N 暗夜 1921_11001	暗夜行路	1921	暗夜行路<1930>	25	新潮社

底本の情報（書籍名、出版社）と「ページ番号」欄を参照することで、容易に原本の当該用例の周辺を参照することができる。

2.3　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]。

2.4　原本画像リンクと作者情報のリンク

収録された 21 作品のうち、10 作品については国立国会図書館デジタルコレクションに原本の画像が公開されているため、コーパスから当該ページへのリンクを行った。図 1 の「中納言」の検索結果から「底本リンク」欄にあるリンクをクリックすると、当該用例のあるページにアクセスすることが可能である。また、同じく図 1 の「中納言」の検索結果の作者欄にある作者名をクリックすると、国立国会図書館典拠データ検索・提供サービス（Web NDL Authorities）の作者情報にアクセスすることができる。

図 1　「中納言」上での画像リンクと作者情報のリンク
（画像の検索結果は「明治・大正編Ⅲ明治初期口語資料」のものである）

[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。

2.4　本文種別と話者情報

近代小説という資料の性質上、会話と地の文の区別がされていることが望ましいため、本コーパスでは全編にわたって引用タグを付与してこれを区別している。当該用例が会話文中のものである場合には、「中納言」の検索結果の「本文種別」列に「会話」と表示されるほか、書籍などの引用文中であれば「引用」と表示される（何も表示がない場合は地の文の用例である）。

さらに、「話者情報列」には、引用元の発話がどの人物によって行われたかの情報が表示される。小説作品では、場面によって登場人物の名称や呼称が変化する、またはその正体が作品の中途で変更になる場合があるが、本コーパスでは同一の人物について、全編にわたって統一した話者情報を付与している。話者情報は、原則的に地の文中で最も多く用いられる名称（固有名詞）を採用し、文脈中で話者の名称が明らかでない場合には、地の文から属性名（「男」「醫者」「先生」「老人」など）を付与している。名称・属性名ともに明らかでない場合には、「＊」を付与している。

３．形態論情報

本コーパスの形態論情報の付与方針は、原則的には公開済みの『CHJ　明治・大正編』ならびに『現代日本語書き言葉均衡コーパス』と同じくしており、互換性がある。

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」[3]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 4　形態論情報の多重化を行ったレコード

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	文字列（毎時）を保持したまま 「何時｜も」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）	文字列（吾夫）を保持したまま 「家｜の｜人」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

[3] 奈良時代編、江戸時代編、和歌集編の一部のコーパスでは実装済みである。

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。

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、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。

これらの処理を行った件数は、近代小説コーパス全体において 300 件程度であるが、近代の資料として特徴的な表記システムを損ねることなく、正しく形態論情報を付与できるようになった点において、大きな進歩を遂げることができたと考える。

4. おわりに

以上、「明治・大正編Ⅳ近代小説」の概要と特色について述べた。本コーパスの追加により、公開済みの「明治・大正編Ⅰ雑誌」「Ⅱ教科書」などのサブコーパスとのジャンル差の観点による比較研究が期待される。

ただし、本コーパスに収録された作品は、近代小説のごく一部に過ぎず、さらなる拡張が望まれる。将来的な課題となるが、今回収録に至らなかった作家の作品の追加や、収録範囲外となった明治前期や昭和期以降の作品の追加など、多方面の拡充の可能性を検討していきたい。

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」

国立国語研究所（2021）『日本語歴史コーパス』https://pj.ninjal.ac.jp/corpus_center/chj/

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』日本語学会、p.173-178.

髙橋雄太（2021）「『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」https://pj.ninjal.ac.jp/corpus_center/chj/doc/abstract-shosetsu-202103.pdf

付記

本コーパスおよび本発表は国立国語研究所の共同研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」の研究成果を報告したものである。