

OpenCHJ を活用した明治・大正期国定理科教科書コーパスの公開

呉 子凡（総合研究大学院大学／国立国語研究所）

金 賢眞（総合研究大学院大学／大阪大学）

小木曾 智信（国立国語研究所／総合研究大学院大学）

Release of a Corpus of Government-Designated Science Textbooks from the Meiji and Taishō Periods through OpenCHJ

Zifan Wu (The Graduate University for Advanced Studies, SOKENDAI / National Institute for
Japanese Language and Linguistics)

Hyunjin Kim (The Graduate University for Advanced Studies, SOKENDAI / The University of
Osaka)

Toshinobu Ogiso (National Institute for Japanese Language and Linguistics / The Graduate
University for Advanced Studies, SOKENDAI)

要旨

本稿では、現在『日本語歴史コーパス』に含まれていない明治・大正期の国定理科教科書について、オンライン公開されている画像資料からテキストデータを作成し、再利用可能なコーパスとして整備した成果について報告する。対象は、文部省『尋常小学理科書』のうち 1910 年刊行の第五学年用、および 1918 年刊行の第五・第六学年用である。

データ整備にあたっては、教科書画像を対象に、Gemini を用いて OCR 処理および TEI 準拠 XML の生成を行った。その後、人手による内容確認・修正を経て、OpenCHJ の文書定義である OCX mini に準拠した形式へ変換した。さらに、形態素解析と修正を行った。完成したデータは GitHub 上でも公開し、従来は画像として閲覧するにとどまっていた資料を、検索・分析・再利用が可能なオープンソースデータとして提供した。

1. はじめに

『日本語歴史コーパス』（Corpus of Historical Japanese; CHJ）は、奈良時代から明治・大正期までの日本語資料に読み・品詞などの形態論情報を付与したコーパスであり、国立国語研究所のコーパス検索アプリケーション「中納言」から検索できる¹。「中納言」では、文字列だけでなく、形態論情報を条件とした用例検索や集計も行える。OpenCHJ は、「中納言」において、研究者がオープンライセンスのテキストに形態論情報を付与し、公開・共有するための枠組みである²。

本稿では、明治・大正期の理科教科書を対象として、OpenCHJ XML 文書型定義である OCX mini³に基づくデータの作成から、形態素解析・修正、公開準備に至るまでの工程を述べる。併せて、CHJ「明治・大正編Ⅱ教科書」の既収録資料と語種構成および語彙多様性を比較し、整備したデータの利用可能性を検討する。

¹ <https://chunagon.ninjal.ac.jp/>

² <https://openchj.github.io/>

³ <https://openchj.github.io/ocx-mini.html>

2. コーパスデータの構築

2.1 資料選定の背景

本研究で対象とした資料は、国立教育政策研究所（教育図書館・文部科学省図書館）⁴で一般公開されている明治・大正期の国定理科教科書である。具体的には、文部省著『尋常小學校理科書：兒童用 第5學年』（1910年刊）、『尋常小學校理科書：兒童用 第5學年』（1918年刊）、および『尋常小學校理科書：兒童用 第6學年』（1918年刊）の3冊である。なお、当時の国定理科教科書は第5学年以降を対象として刊行されていた。

CHJ「明治・大正編Ⅱ教科書」⁵には、国定読本を中心とする小学校国語教科書に加え、明治初期の理科教科書である『物理階梯』（1872）、『小学化学書』（1875）、『初学人身窮理』（1876）なども収録されている。今回の資料は、国定読本とは刊行時期が重なる一方、既収録の理科教科書とは教科を同じくして時期が異なる。このため、同時期の国語・理科教科書の教科間比較と、理科教科書の通時的比較の双方に用いることができる。

以上を踏まえ、公開画像から3冊の本文テキストを作成し、文書構造と形態論情報を付与した。以下では、その作業工程を順に述べる。

2.2 OCRとXMLファイルの作成

データの構築にあたっては、まず、一般公開されている教科書画像（図1）をPDF形式で取得し、作成したプロンプトに基づいて、Googleの生成AIモデル Gemini3 によるOCRおよびTEI形式XMLデータの生成を半自動的に行った。TEI（Text Encoding Initiative）⁶は、人文学資料の構造や内容をXMLで記述するための国際的なガイドラインである。今回は本文テキストと文書構造を対象とし、図版は構造化の対象外とした。処理方法の詳細は金・小木曾（2026）に示されている。生成結果は原画像と照合し、本文の誤認識を修正するとともに、資料間で要素の配置と記述形式を統一した。

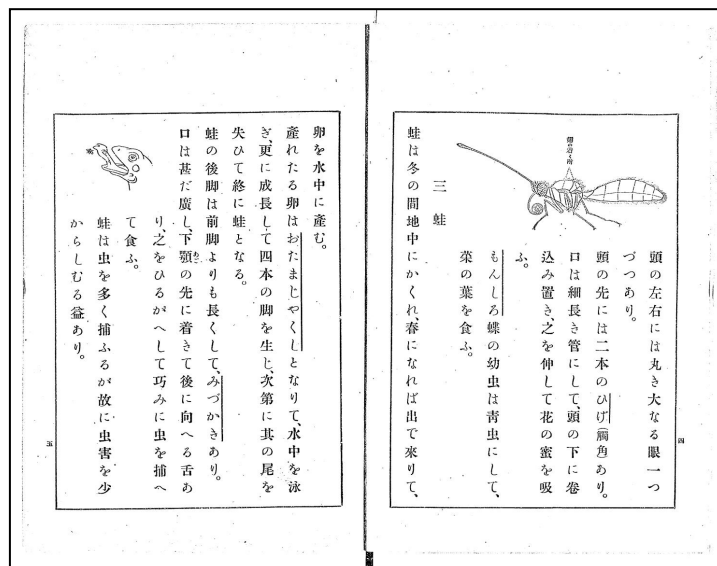


図1 教科書の画像

⁴ https://nierlib.nier.go.jp/opac/opac_search/?lang=0

⁵ https://clrd.ninjal.ac.jp/chj/meiji_taisho.html#kyokasho

⁶ <https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html>

生成された TEI 文書の全体の構造としては、<teiHeader>に書誌情報を、<text>に書籍内容を記述する形となっている。さらに、<text>は、表紙・目録にあたる<front>、本文にあたる<body>、奥付にあたる<back>に分かれている。

次に、TEI 文書を OpenCHJ XML 文書型定義である OCX mini に変換する。OCX mini は、OpenCHJ での利用を想定して設計された最小限のマークアップ方式であり、すでに構造化されている TEI 文書から比較的容易に変換できる点も特徴である（小木曾 2025）。

変換に先立ち、自動生成された TEI データの構造を手作業で整理した。とくに、<teiHeader>と<back>に記述される書誌・出版情報、および目録・奥付の構造に資料間の不統一があったため、要素の配置と記述形式を統一した。

OCX mini では、<ocx:doc>要素が形態素解析の対象範囲を示す。本研究では本文のみを解析するため、<body>内の本文部分を<ocx:doc>で囲んだ。また、@textID に資料の識別子、@corpusName にコーパス名を記述した。

```
<body>
  <ocx:doc textID="〇" corpusName="〇〇〇" xmlns:ocx="https://openchj.github.io/ns/ocx">
    ...
  </ocx:doc>
</body>
```

本文部分では、各課を<div>で区切り、課の標題を<head>で示した。さらに、形態素解析における文境界を明示するため、各文末に<ocx:eos/>を付与した。

```
<s xml:id="s080"><ocx:r rt="ゆび">趾</ocx:r>は四本づつありて、一本は後に向ふ。<ocx:eos/></s>
```

今回扱った 3 資料のうち、原資料にルビが付されているのは『尋常小學理科書：児童用 第 5 學年』（1910 年刊）のみである。そのため、同資料の構造化データでは、ルビを <ocx:r> 要素によってマークアップした。図 2 は変換後の本文データの一部を示している。

```
<body>
  <ocx:doc textID="1910第5学年" corpusName="明治・大正国定理科教科書" xmlns:ocx="https://openchj.github.io/ns/ocx">
    <pb n="1"/>
    <div type="chapter" n="1">
      <head>油菜</head>
      <p>
        <s xml:id="s001">油菜は根・莖・葉を具へ、花を開く。<ocx:eos/></s>
        <s xml:id="s002">根は地中より水及び養分を吸取り、また莖を支ふ。<ocx:eos/></s>
        <s xml:id="s003">葉には多くのすぢ（脈）あり。<ocx:eos/></s>
        <s xml:id="s004">花は外側にかく（萼）あり、其の内にはなびら（花瓣）あり、又其の内をしべ（雄蕊）あり、中心にめしべ（雌蕊）あり。<ocx:eos/></s>
        <s xml:id="s005">がくは四片に分れ、はなびらは四枚あり。<ocx:eos/></s>
      </p>
      <pb n="2"/>
      <p>
        <s xml:id="s006">をしべは六本ありて、各、その先にこなぶくろ（葯）を有し、其の中より花粉を出す。<ocx:eos/></s>
        <s xml:id="s007">めしべは一本あり。<ocx:eos/></s>
        <s xml:id="s008">其の下部の膨れたる所はしばう（子房）にして、中に多くの小ざき粒あり。<ocx:eos/></s>
        <s xml:id="s009">花の底には蜜を出す所あり。<ocx:eos/></s>
        <s xml:id="s010">蝶・蜂など来りて此の蜜を吸ふ。<ocx:eos/></s>
        <s xml:id="s011">此のとき花粉は虫に附着し、更にめしべの先端に附着すべし。<ocx:eos/></s>
        <s xml:id="s012">かくて花粉<pb n="3"/>を受けたるとめしべは其のしばう次第に成長して後にみ（果實）となり、しばうの中の小ざき粒はたね（種子）となる。<ocx:eos/></s>
        <s xml:id="s013">油菜のたねより種油を取り、其の搾り粕は肥料に用ふ。<ocx:eos/></s>
      </p>
    </div>
    <div type="chapter" n="2">
      <head>もんしろ蝶</head>
      <p>
        <s xml:id="s014">もんしろ蝶の體は頭と胸と腹との三部に分たる。<ocx:eos/></s>
        <s xml:id="s015">はね（翅）は四枚ありて、胸の上側より左右に二枚づつ出づ。<ocx:eos/></s>
        <s xml:id="s016">脚は六本ありて、胸の下側より左右に三本づつ出づ。<ocx:eos/></s>
        <pb n="4"/>
        <s xml:id="s017">頭の左右には丸き大なる眼一つづつあり。<ocx:eos/></s>
        <s xml:id="s018">頭の先には二本のひげ（觸角）あり。<ocx:eos/></s>
      </p>
    </div>
  </ocx:doc>
</body>
```

図 2 変換後の XML データ（本文）

2.3 形態素解析

形態素解析については、複数の UniDic を用いた形態素解析をオンラインで試行できる「Web 茶まめ」（堤・小木曾 2023）による解析結果を検討した。その結果を踏まえ、本研究では、近代の文語資料に対応した形態素解析用辞書「近代文語 UniDic」（小木曾・小町・松本 2013）と、形態素解析エンジン「MeCab 0.996」（Kudo, Yamamoto & Matsumoto 2004）を用いて、本文の形態素解析を行った。

解析にあたっては、本文テキストのみを対象とし、<ocx:eos/>タグを文境界として MeCab による形態素解析を行った。自動生成した TEI データには各種の構造化タグが付与されているが、形態素解析に直接用いたのは、解析対象範囲を示す<ocx:doc>、文境界を示す<ocx:eos/>、ルビを示す<ocx:r>である。これらの情報をもとに、その結果を文字表・短単位表・タグ表・ルビ表を生成し、各表をファイル先頭からのオフセット位置によって相互に関連付けた。解析結果は国語研の「形態論情報データベース」（小木曾・中村 2014）にインポートした。

その結果、各資料の短単位数は、1910 年刊『尋常小學理科書：児童用 第 5 學年』が 5018 語、1918 年刊『尋常小學理科書：児童用 第 5 學年』が 8537 語、1918 年刊『尋常小學理科書：児童用 第 6 學年』が 9381 語となった。なお、これらの数値には記号など語彙項目以外の短単位も含まれるため、後述する語種別集計における合計語数とは一致しない。

2.4 形態論情報の修正

形態論情報データベースに登録した解析結果は、修正用インターフェース「大納言」で原画像と照合しながら人手で確認・修正した（図 3）。修正に際しては、『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 2013）と『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 上・下』（小椋ほか 2011）を参照した。

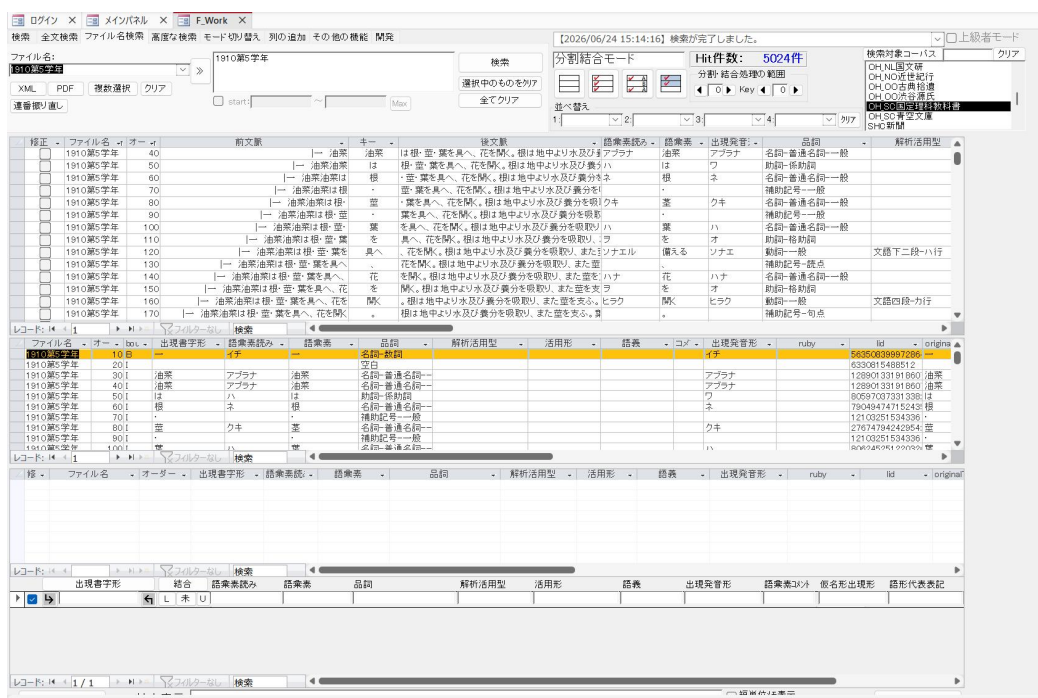


図 3 「大納言」での修正画面

3. データの公開

XML データは、書誌情報、本文、文書構造、ルビ情報を確認できる形で GitHub に公開した⁷。形態論情報付きデータは、最終確認を経たうえで、2026 年秋に「中納言」の OpenCHJ で公開する予定である。

4. データの応用

本データの応用可能性を示す一例として、CHJ「明治・大正編Ⅱ教科書」の既収録資料と語彙的特徴を比較した。比較対象は、今回の国定理科教科書と刊行時期が重なる 1910 年・1918 年の小学校国語教科書、および、同じ理科教科書で時期の早い『物理階梯』（1872）、『小学化学書』（1875）、『初学人身窮理』（1876）である。

まず、記号などを除く語彙項目を、語の出自を表す語種情報に基づいて和語・漢語・外来語・混種語・固有名に分け、各資料における構成比を比較した（図 4）。これにより、同じ理科教科書における時期差をみる通時的比較と、同時期の国語教科書と理科教科書の違いをみる共時的比較の両面から、本資料の特徴を検討することができる。

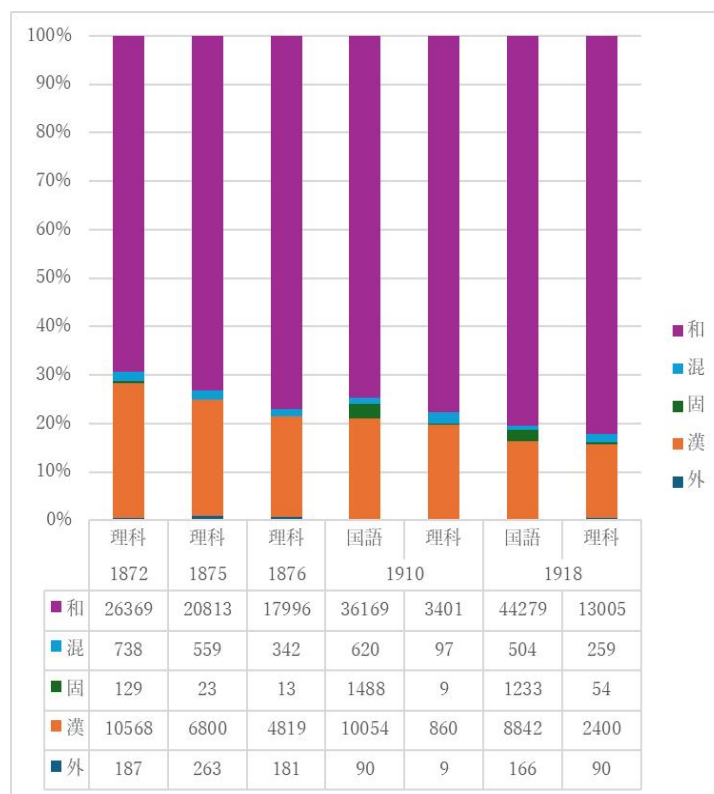


図 4 比較対象資料における語種別構成比

通時的にみると、本稿の対象資料の範囲では、明治初期の理科教科書ほど漢語の割合が高い。たとえば、1872 年の『物理階梯』では漢語が全体の約 28% を占め、1875 年の『小学化学書』では約 24%、1876 年の『初学人身窮理』では約 21% である。これに対し、1910 年・1918 年の国定理科教科書では、それぞれ約 20%、約 15% であった。一方、和語は明治

⁷ <https://github.com/togiso/OpenCHJ-KokuteiRikaTextbooks>

初期の3資料で約70～77%、1910年・1918年の国定理科教科書で約78%、約82%である。したがって、明治初期の資料から明治末・大正期の国定理科教科書にかけて、漢語の割合が低下し、和語の割合が上昇する傾向が認められる。

共時的にみると、1910年・1918年とも、国語教科書と理科教科書の語種構成に大きな差はない。ただし、固有名（人名・地名など）の割合は、国語教科書が1910年約3%、1918年約2%であるのに対し、理科教科書はいずれも1%未満である。この差には、国語教科書の読本に人名・地名を含む文章が多いことが関係する可能性がある。なお、資料数・語数が異なるため、結果の一般化には慎重を要する。

表1 各比較対象資料のMSTTR（1000語単位）

成立年	教科	作品名	segment_count	tokens_used	MSTTR_1000
1872	理科	物理階梯	42	42000	0.2912
1874	理科	小学化学書	29	29000	0.2831
1876	理科	初学人身窮理	23	23000	0.3169
1910	国語	小学校5年	23	23000	0.3663
1910	国語	小学校6年	32	32000	0.3659
1918	国語	小学校5年	29	29000	0.3593
1918	国語	小学校6年	33	33000	0.3652
1910	理科	第5学年	5	5000	0.2796
1918	理科	第5学年	8	8000	0.2619
1918	理科	第6学年	9	9000	0.2682

次に、テキスト長の影響を抑えて語彙の多様性を比較するため、MSTTR（mean segmental type-token ratio）を算出した。本稿では、各資料を1000語ずつの区間に分け、各区間の異なり語数を延べ語数で割ったTTRを求め、その平均をMSTTRとした。1000語に満たない末尾の端数は集計から除いた（表1）。

その結果、1910年・1918年の国語教科書ではMSTTRがおおむね0.36前後であるのに対し、今回対象とした国定理科教科書では0.26～0.28程度であった。また、明治初期の理科教科書では0.28～0.32程度であり、今回の国定理科教科書と比較するとやや高い値となっている。同時期の比較では、国語教科書のほうが理科教科書よりも語彙の多様性が高い傾向にある。理科教科書では、説明対象となる事物や現象に関わる語が繰り返し用いられるため、相対的に語彙の反復性が高くなる可能性がある。また、理科教科書内部の通時的比較では、明治初期の理科教科書より明治末・大正期の国定理科教科書のほうが語彙の多様性がやや低くなる傾向も見られる。

以上のように、語種構成と語彙多様性を指標とすることで、同じ理科教科書における通時的変化と、同時期の教科間差の双方を検討することができる。今回整備した国定理科教科書データは、既存のCHJ資料を拡充するとともに、明治・大正期の教科書における語彙・文体の比較分析に資することが期待される。

5. おわりに

本稿では、国立教育政策研究所（教育図書館・文部科学省図書館）で公開されている明

治・大正期の国定理科教科書 3 冊を対象に、画像取得、Gemini3 による OCR・TEI 準拠 XML 生成、人手修正、OCX mini への変換、形態素解析・修正、公開準備までの工程を報告した。また、既存の CHJ 資料との比較により、本データの活用例を提示した。

本事例を通して、研究者が自ら整備した資料データを OpenCHJ 上で公開・利用するための一連の手順は、一定の見通しをもって実施できることを示した。とくに、生成 AI を活用した本文入力・構造化作業と、OCX mini を用いた OpenCHJ 向けのデータ整備とを接続することで、資料データ公開に至るまでの作業負担を軽減できると考えられる。

今後、形態論情報の最終確認を行い、2026 年秋の OpenCHJ 公開を目指す。本事例が、歴史日本語資料をはじめとする未整備資料のデータ化・公開と、OpenCHJ における利用可能な資料の拡充に向けた一つの参考となることを期待したい。

謝 辞

本研究は国立国語研究所のプロジェクト「開かれた共同構築環境による通時コーパスの拡張」による成果である。

文 献

金賢眞・小木曾智信 (2026) 「LLM を用いた近代教科書画像データの XML 文書化 —Gemini 3 による TEI 自動アノテーションの評価と分析—」研究報告人文科学とコンピュータ (CH), Vol. 2026-CH-140, No. 18, pp. 1-6.

小木曾智信 (2025). OCX mini 文書定義, <https://openchj.github.io/ocx-mini.html> (2026 年 7 月 9 日確認).

堤智昭・小木曾智信 (2023). 「複数の UniDic 辞書による形態素解析支援ツール『Web 茶まめ』の実装と運用」情報処理学会論文誌, 64:3, pp. 749-757.

小木曾智信・小町守・松本裕治 (2013) 「歴史的日本語資料を対象とした形態素解析」, 自然言語処理, 20:5, pp. 727-748.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. (2004) “Applying conditional random fields to Japanese morphological analysis”, In EMNLP, Vol. 4, pp. 230-237.

小木曾智信・中村壮範 (2014). 「『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用」自然言語処理, vol. 21, No. 2, pp. 301-332.

国立国語研究所コーパス開発センター(近藤明日子) 編 (2013) 『近代文語 UniDic 短単位規程集 Ver. 1.1』.

小椋秀樹・小磯花絵・富士池優美・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 (上) (下).