

OpenCHJ XML の設計と実践

— 日本語コーパス構築のための簡易 XML 文書定義 —

小木曾 智信^{1,2,a)} 呉 子凡^{2,1,b)} 吉安 良太^{1,c)}

概要：国立国語研究所で進める「オープン CHJ」プロジェクトの一環として、オープン CHJ のための簡易な XML 文書定義として、OCX mini と OCX Standard からなる二段階の OpenCHJ XML 仕様の策定を進めている。本発表では、その実践例として、プレーンテキストに対するマークアップ例（上代歌謡）、XHTML 文書に対するマークアップ例（青空文庫）、TEI 文書に対するマークアップ例（近代国定理科教科書）を示しつつ、OpenCHJ XML の仕様とその適用可能性について論じる。これにより、OpenCHJ XML が多様な文書形式に適用可能であり、日本語コーパス構築のための共通的な記述形式として有効であることを示す。

Design and Practice of OpenCHJ XML: A Simple XML Document Format for Building Japanese Corpora

1. はじめに：OpenCHJ の試み

国語研では、外部の研究者や日本語研究に関心を持つ人々が、日本語研究に有用な資料を共有・整備し、コーパスとして構築する「オープン CHJ」プロジェクトを進めている。本プロジェクトでは、オープンライセンスのテキストをもとに、国語研の既存コーパスと同様の形態論情報を付与し、コーパス検索アプリケーション「中納言」上で利用できる環境を整える。その一環として、オープン CHJ のための標準化された簡易な XML 文書定義である OpenCHJ XML の策定を進めている。

2. 日本語コーパスと XML

国立国語研究所が構築してきた大規模日本語コーパスでは、XML が本文と文書構造を保持する基礎形式として用いられてきた。『現代日本語書き言葉均衡コーパス』（BCCWJ）では、形態論情報を含まない「文書構造情報付き文字ベース XML」（C-XML）を基礎データとし、そこに

短単位・長単位の形態論情報を埋め込んだ「形態論情報付き統合形式 XML」（M-XML）を生成する独自形式が採用されている [1], [2]。『日本語歴史コーパス』（CHJ）でも、この構築基盤を継承し、資料の構造を独自の XML タグで記述するとともに、校訂・正規化した解析用本文と原文表記との対応をタグによって保持している [2], [3]。

これらのコーパスの構築・修正には、本文、XML タグ、短単位、長単位等を関係データベースの別個のテーブルに格納し、ファイル先頭からの文字位置をキーとして相互に関連付ける形態論情報データベースが用いられている。XML の取り込み時には、形態素解析の妨げとなるタグを除去して解析用文字列を生成し、各短単位には原文上の文字開始位置と文字終了位置を付与する [2]。この方式では、本文確定後の文字の追加・削除が、後続する形態論情報や文書構造情報との位置対応に影響するため、XML 文書から解析対象となる基底文字列を一意かつ安定して取り出せることが重要となる。

一方、TEI (Text Encoding Initiative) ガイドラインは、人文学資料の文書構造、書誌情報、翻刻、校訂、注記、言語情報等を共通の XML 語彙で記述するための包括的な枠組みであり、言語コーパスの記述、用途に応じた語彙のカスタマイズ、スタンドオフ・アノテーションの仕組みも備えている [4]。多様な資料の内容と構造を明示し、研究資源

¹ 国立国語研究所
NINJAL, Tachikawa, Tokyo 190-8561, Japan

² 総合研究大学院大学
SOKENDAI, Hayama, Kanagawa 240-0193, Japan

a) togiso@ninjal.ac.jp

b) gs2024r007@ninjal.ac.jp

c) r.yoshiyasu@ninjal.ac.jp

間の相互運用性を確保する点で、TEI は日本語コーパスの基礎データにも有用である。

ただし、TEI の一般的な記法を、文字位置を基盤とする上記の構築系にそのまま適用すると、形態素解析に inputs する本文の抽出規則が複雑になりやすい。たとえば、note は注記を要素内容として持ち、ruby は rb と rt によって基底文字列と読みを併記し、choice は orig/reg や sic/corr のような代替表現を同一箇所に保持する。XML 文書中の文字データを単純に連結すれば、注記、読み、代替表現が解析入力に混入する一方、これらを選択的に除外すれば、文書に含まれる文字データと解析用文字列との位置対応を別途管理しなければならない [4]。この問題は TEI のカスタマイズやスタンドオフ方式によって解決できるが、解析用本文への射影規則をプロジェクトごとに定める必要がある。そこで OpenCHJ XML では、既存の XML 資源を利用できる柔軟性を保ちながら、解析対象文字列を一意に生成し、文字位置による形態論情報との対応を安定して保つための最小限の規則を定める。

3. OpenCHJ XML：日本語コーパス向けの簡易な XML 文書定義

OpenCHJ XML (Open CHJ XML; 以下 OCX) は、日本語を中心とする人文系資料を、原資料の構造や表記情報を保持したまま言語解析に利用するための簡易な XML 文書定義である。仕様は、コーパス化に必要な最小限の処理規則を定める OCX mini と、文書構造・原文表記・解析方法の共通化を図る OCX Standard の二段階からなる*1。両者に共通する設計目標は、形態素解析等に inputs する本文文字列を一意に生成できること、解析結果を文字オフセットによって原文と対応付けられること、また、ルビ・注記・原表記などの情報を本文文字列から分離して保持できることである。OCX の名前空間 URI には <https://openchj.github.io/ns/ocx> を用いる。

なお、本稿でいう文字オフセットとは、XML ソース上の位置ではなく、OCX の規則によって生成された解析用本文文字列上の文字位置を指す。

3.1 OCX-mini

OCX mini は、文学作品・史料・資料テキストを少ない作業量でコーパス化するための最小限のマークアップ方式である*2。基本原則は、doc 要素内に本文として現れる文字を原則としてすべて解析対象とし、読み・注記・書誌情報など解析対象としない情報を属性値等へ分離することである。

OCX mini が定義するのは、XML 文書で利用できる要

素の全体ではなく、解析時に特別な処理を行う次の 4 種類の要素である。

- doc: コーパス化する本文の範囲を示す。textID 属性にテキスト識別子、corpusName 属性にコーパス名を記す。
- eos: 文末を示す空要素であり、解析用テキストの生成時には改行一文字に置き換えられる。OCX mini 準拠文書では、各文または文に相当する解析単位の終端をこの要素によって明示する。
- r: 本文文字列と、それに対応する解析対象外の文字列とを対応付ける。要素内容を解析対象とし、rt 属性の値は解析対象としない。通常のルビでは `<r rt="そ ばく">素朴</r>` のように読みを rt 属性に記すが、翻字資料では `<r rt="美">み</r>` のように、解析用の仮名表記を要素内容、対応する原字を rt 属性に記すこともできる。
- comment: 本文に含めない注記を表す空要素であり、注記内容を text 属性に記す。属性値は解析用テキストの文字数と文字オフセットに影響しない。

OCX mini は文書構造そのものを規定しない。このため、上記 4 種類以外の XML 要素を、資料の構造、レイアウト、校訂情報等の記述に併用できる。OCX mini 処理系は、これらの要素を解析上透明な構造要素として扱い、開始タグと終了タグ自体は解析用テキストに出力せず、その内部の文字データは通常どおり解析対象とする。たとえば、仏足石歌のデータで用いる lg, caesura, unclear 等は、歌のまとまり、句切れ、判読の不確実性を記述するための補助的な要素であり、OCX mini の 4 要素と併用できる。したがって、このようなデータも OCX Standard を必須とせず、OCX mini によって処理できる。ただし、unclear 等の要素内容も、別途解析制御を行わない限り本文文字列として解析対象になる。

OCX mini 文書から解析用テキストを生成する際には、doc 内の文字データを文書順に取り出し、XML の整形に用いられた半角空白・改行・タブを除いたうえで、eos を改行に変換する。r については要素内容だけを出力して rt 属性を除外し、comment は出力しない。OCX mini が定義しない要素についてはタグのみを除き、その内部の文字データを出力する。この単純な規則により、同一の XML 文書から一意の解析用テキストを生成し、形態論情報を文字オフセットで安定して管理できる。

OCX 単独文書では、

```
<doc xmlns="https://openchj.github.io/ns/ocx">
  ...
</doc>
```

のように OCX 名前空間をデフォルト名前空間として、接頭辞を省略できる。一方、XHTML や TEI 等の既存 XML 文書に埋め込む場合は、埋め込み先の名前空間と文書構造を維持し、ocx:doc, ocx:eos 等の OCX 要素を ocx: 接頭

*1 OpenCHJ XML: <https://openchj.github.io/ocx.html>

*2 OCX mini 文書定義: <https://openchj.github.io/ocx-mini.html>

辞付きで記述する。この二つの利用形態により、プレーンテキストに近い資料にも、すでに構造化された XML 文書にも、同じ解析規則を適用できる。

3.2 OCX-standard

OCX Standard (本稿執筆時点では v0.6 ドラフト) は、OCX mini を包含しつつ、mini では処理上の意味を定めない文書構造や原文表記について共通に利用できる語彙を定め、さらに資料内の範囲ごとに解析方法を指定できるようにした拡張仕様である^{*3}。すなわち、OCX mini が多様な XML 要素の併用を許すための共通最小層であるのに対し、OCX Standard は、その一部に共通の名称・意味・処理規則を与える層として位置付けられる。

文書構造の記述には、front、body、back、div、head、p、s、lg、l、caesura、pb、lb、quote、sp、speaker、unclear 等、TEI P5 の要素名と基本的な意味の一部を借用した語彙を用いる。ただし、OCX Standard は TEI 文書そのものを作成することを目的とせず、これらの語彙を利用する際にも TEI 名前空間への厳密な適合を必須としない。OCX 単独文書では、OCX 名前空間をデフォルト名前空間として、借用語彙を含む各要素を接頭辞なしで簡潔に記述できる。既存の TEI 文書を利用する場合は、TEI 名前空間に属する要素をそのまま保持し、OCX 固有の処理要素だけを ocx: 接頭辞付きで追加できる。OCX 処理系は、仕様で対応付けた借用語彙に限り、両者を同等の構造概念として扱う。

OCX Standard で OCX mini に追加される主要な機能は、次のとおりである。

- 解析方法の指定: proc 要素によって、範囲ごとに形態素解析で使用する辞書を dic 属性で指定し、必要に応じて norm 属性で解析前の正規化を指定する。
- 通常の形態素解析を行わない範囲の処理: skip 要素によって、漢文、外国語、URI、プログラムコード等を通常の形態素解析から除外する。tokenize="single"では範囲全体を一短単位とし、tokenize="space"では空白によって擬似的に単位分割する。分類は pos 属性に記す。
- 原文表記の保持: odoriji 要素では、解析に用いる展開後の文字を要素内容に置き、原表記の踊り字を orig 属性に保持する。割書は warigaki 要素、割書内部の改行は wbr 要素で表す。
- 文書構造の共通化: 文、段落、歌、引用、発話、改ページ、改行、判読の不確実性等を、TEI から借用した語彙によって共通に記述する。注記と本文文字列の対応には、mini から継承した comment と r を用いる。文境界は、s 要素の終端または eos 要素によって示す。

OCX mini との互換性を保つため、既存データのように s 要素内の文末に eos を置くことも許容し、この場合は eos と直後の s 要素終端を一つの文境界として扱う。解析用テキストの生成時には、mini の規則を継承したうえで、proc、skip、odoriji、warigaki 等に定められた処理を適用し、speaker など仕様上解析対象外とした要素の文字列は除外する。

以上のように、OCX mini は既存テキストを最小限の変更でコーパス化するための仕様であり、任意の XML 文書構造と併用できる。OCX Standard は、構造要素を使用可能にするための仕様ではなく、複数の資料・作成者・処理系の間で文書構造や解析制御の意味を共有する必要がある場合に用いる拡張仕様である。二段階の仕様を設けることで、プレーンテキスト、XHTML、TEI を含む多様な資料に対し、必要な範囲だけ段階的に OCX を適用できる。

4. OpenCHJ XML の活用例

4.1 テキストデータの OpenCHJ 化

プレーンに近いテキストデータを OCX mini にしたがってコーパスにした例として、上代歌謡 (仏足石歌・歌経標式例歌) を取り上げる。

「仏足石歌」は薬師寺に伝わる仏足石の隣にある歌碑の碑文で、770 年頃の成立とされる。「歌経標式」は藤原浜成が著した歌論書で、序文によれば 772 年の成立である。34 首の和歌を例として引用しており『万葉集』に記載のない歌が含まれている。いずれも短いものだが、資料が極めて限定的な時代の貴重な日本語資料である。

これらを OpenCHJ で利用可能にするために、テキストデータを作成し、OCX 形式で最低限のアノテーションを施した。

「仏足石歌」のテキストは Wikipedia「仏足跡歌碑」^{*4}の記載をもとに Oxford NINJAL Corpus of Old Japanese (ONCOJ)[5] ^{*5} を参照して整備した。「歌経標式」のテキストは ONCOJ のローマ字のテキストをひらがなに変換して作成した。

XML 文書のルート要素は <doc> を用い、デフォルト名前空間として OCX mini の名前空間 (<https://openchj.github.io/ns/ocx>) を指定している。各和歌は一首を単位として <lg type="waka"> 要素で構造化し、歌番号は id 属性に付与した。

本文の記述においては、形態素解析等の自然言語処理における利便性を考慮し、解析対象となるひらがな表記のテキストを <r> 要素の要素内容として配置した。一方、一字一音の万葉仮名による原文情報は rt 属性へと退避させる設計をとっている。これにより、後述するコーパス検索アプリケーション上で原文を表示できるようにしている。

^{*3} OCX Standard : <https://openchj.github.io/ocx-standard.html>

^{*4} <https://ja.wikipedia.org/wiki/仏足石歌碑>

^{*5} <https://oncoj.ninjal.ac.jp/>

OCX mini は、独自に定義する 4 種類の要素以外の文書構造を規定しない。このため、本データでは、資料の構造や翻刻上の情報を表す補助的な XML 要素を OCX mini の要素と併用した。これらの補助的要素は解析上透明に扱われ、その内部の文字列は通常の本文文字列として解析対象となる。具体的には以下の通りである。

- 歌の構造: 一首の範囲を `<lg type="waka">` で囲み、歌番号を `id` 属性に記した。
- 句切れの表示: 和歌内部における韻律的・構造的な区切りを示す空要素として `<caesura/>` を使用した。空要素であるため、解析用テキストには影響しない。
- 欠損・判読不能箇所の翻刻: 判読不能文字、本文の欠損、あるいは空白箇所については `<unclear>` 要素を用い、元データの記述状態を保持した。この要素の内容は、OCX mini の本文原則に従って解析対象となる。
- 解析単位の明示: 各 `<lg type="waka">` の末尾に文末を示す `<eos/>` 要素を配置した。これにより、OpenCHJ を用いた形態素解析や検索処理において、一首をひとつの文 (sentence) 相当の処理単位として定義している。

サンプルとして仏足石歌の冒頭を示す。

```
<doc xmlns="https://openchj.github.io/ns/ocx"
textID="仏足石歌" corpusName="上代歌謡">
<lg id="1" type="waka"><r rt="美">
み</r><r rt="阿">あ</r><r rt="止">と</r><r rt="都">つ</r><r rt="久">く</r><r rt="留">
る</r><caesura/><r rt="伊">い</r><r rt="志">
し</r><r rt="乃">の</r><r rt="比">ひ</r><r rt="鼻">
び</r><r rt="伎">き</r><r rt="波">
は</r><caesura/><r rt="阿">あ</r><r rt="米">
め</r><r rt="尔">あ</r><r rt="伊">い</r><r rt="多">
た</r><r rt="利">り</r><caesura/><r rt="都">
つ</r><r rt="知">ち</r><r rt="佐">
さ</r><r rt="閑">へ</r><r rt="由">ゆ</r><r rt="須">
す</r><r rt="礼">れ</r><caesura/><r rt="知">
ち</r><r rt="々">ち</r><r rt="波">
は</r><r rt="々">は</r><r rt="賀">が</r><r rt="多">
た</r><r rt="米">め</r><r rt="尔">
に</r><caesura/><r rt="毛">も</r><r rt="呂">
ろ</r><r rt="比">ひ</r><r rt="止">と</r><r rt="乃">
乃</r><r rt="多">た</r><r rt="米">
め</r><r rt="尔">に</r><eos/></lg>
```

美阿止都久留 伊志乃比鼻伎波 阿米尔伊多利 都知佐閑
由須礼 知々波々賀多米尔 毛呂比止乃多米尔
(御跡作る 石の響きは 天に到り 地さへ揺すれ 父母がため
に 諸人の為に)

この XML ファイルを対象として、MeCab と上代語 UniDic[6]^{*6} を用いて形態素解析を行ったのち、誤りを修正したものをコーパスの形態論情報データとした。「歌経標式」の形態論情報の整備にあたっては [7] を参考にした。XML ファイルと形態論情報は GitHub 上で公開している。^{*7}

4.2 XHTML 文書の OpenCHJ 化

すでに構造化された XHTML 文書に OCX mini を埋め込んだ例として、青空文庫^{*8}所収の近代文学作品を取り上げる。今回対象としたのは、芥川龍之介『蜘蛛の糸』、宮沢賢治『やまなし』、梶井基次郎『檸檬』、新見南吉『手袋を買いに』、山川方夫『夏の葬列』の 5 作品である。

これらを OpenCHJ で利用可能にするため、青空文庫から各作品の XHTML ファイルを取得し、コーパス化の対象となる本文部分に対してのみ OCX mini によるマークアップを施した。

具体的には、各 XHTML 文書の `<body>` 要素のうち形態素解析の対象とする本文部分を `<ocx:doc>` で囲んだ上で、その範囲内に Python スクリプトを用い以下のような最小限の改変を加えた。

- ルビ形式の変換: 青空文庫の XHTML 文書においてルビは `<ruby><rb>漢字</rb><rp>(</rp><rt>ルビ</rt><rp>)</rp></ruby>` の形式で示されるが、これを OCX 形式に変換した。
- 外字の Unicode 変換: 青空文庫においては JIS X 0208, JIS X 0201 に含まれない文字が外字として扱われている^{*9}が、XHTML 文書内では各外字について JIS X 0213 の面区点番号が示されているため、対応する Unicode 文字に自動変換した。
- 文境界の明示: 形態素解析における一文の範囲を示すため、文末に置かれうる約物 (句点・感嘆符・疑問符・閉鉤括弧) の直後 (句点と閉鉤括弧が連続するときは最後にのみ) に `<ocx:eos/>` を挿入した。例外的に読点が文末に置かれている箇所などは手作業で挿入した。

サンプルとして、OCX mini 埋め込み後の『やまなし』の XHTML 文書を示す。`<head>` 要素を始め、もともと本文書内に記載されていたメタデータの類は一切編集していない。なお、紙幅の都合上、本文部分は冒頭と末尾を残し省略した。

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN"
"http://www.w3.org/TR/xhtml11/DTD/xhtml11.dtd">
<html xmlns="http://www.w3.org/1999/xhtml"
xml:lang="ja" >
<head>
<meta http-equiv="content-style-type"
content="text/css" />
<link rel="stylesheet" type="text/css"
href="../../aozora.css" />
<title>宮沢賢治 やまなし</title>
<script type="text/javascript"
src="../../jquery-1.4.2.min.js"></script>
<link rel="Schema.DC"
href="http://purl.org/dc/elements/1.1/" />
<meta name="DC.Title" content="やまなし" />
<meta name="DC.Creator" content="宮沢賢治" />
```

^{*6} https://clrd.ninjal.ac.jp/unidic/download_all.html#unidic_jodai

^{*7} <https://github.com/togiso/OpenCHJ-JodaiKayo>

^{*8} <https://www.aozora.gr.jp/>

^{*9} https://www.aozora.gr.jp/annotation/external_character.html

```
<meta name="DC.Publisher"
  content="青空文庫" />
</head>
<body>
<ocx:doc textID="宮沢賢治 やまなし"
  corpusName="青空文庫"
  xmlns:ocx="https://openchj.github.io/ns/ocx"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-
    instance"
  xsi:schemaLocation="https://openchj.github.io/
    ns/ocx
    https://openchj.github.io/schema/ocx/mini/
    1.0/ocx-mini.xsd">
<div class="metadata">
<h1 class="title">やまなし</h1>
<h2 class="author">宮沢賢治</h2>
<br />
<br />
</div>
<div id="contents" style="display:none"></div>
<div class="main_text"><br />
  小さな谷川の底を写した二枚の青い<ocx:r rt="げん
  とう">幻燈</ocx:r>です。<ocx:eos/><br />
<br />
<div class="jisaige_3" style="margin-left: 3em">
<h4 class="naka-midashi">
<a class="midashi_anchor" id="midashi10">一、五
  月</a>
</h4>
</div>
<br />
  二<ocx:r rt="ひき">疋</ocx:r>の<ocx:r rt="かに
  ">蟹</ocx:r>の子供らが青じろい水の底で話していま
  した。<ocx:eos/><br />
  『クラムボンはわらったよ。』<ocx:eos/><br />
  『クラムボンはかぶかぶわらったよ。』<ocx:eos/><br />
  ; (中略)
  親子の蟹は三疋自分<ocx:r rt="ら">等</ocx:r>の穴
  に帰って行きます。<ocx:eos/><br />
  波はいよいよ青じろい焰をゆらゆらとあげま
  した、<ocx:eos/>それは又<ocx:r rt="こんごうせ
  き">金剛石</ocx:r>の粉をはいているようでした。
  <ocx:eos/><br />
  <br />
  *<br />
<br />
  私の幻燈はこれでおしまいであります。
  <ocx:eos/><br />
<br />
<br />
<br />
</div>
</ocx:doc>
<div class="bibliographical_information">
<hr />
<br />
  底本:「新編風の又三郎」新潮文庫、新潮社<br />
    1989 (平成元) 年 2 月 25 日発行<br />
    1989 (平成元) 年 6 月 10 日 2 刷<br />
  初出:「岩手毎日新聞」岩手毎日新聞社<br />
    1923 年 (大正 12 年) 4 月 8 日<br />
  入力: 蔭龍<br />
  校正: noriko saito<br />
  2008 年 4 月 15 日作成<br />
  2013 年 7 月 8 日修正<br />
  青空文庫作成ファイル: <br />
  このファイルは、インターネットの図書館、
  <a href="http://www.aozora.gr.jp/">青空文庫
  (http://www.aozora.gr.jp/) </a>で作られました。入
  力、校正、制作にあたっては、ボランティアの皆さんで
```

す。

; (中略)

```
</div>
</body>
</html>
```

以上のように、OCX mini のタグを埋め込むことによって、青空文庫の XHTML 文書に最小限の編集で形態素解析を施してコーパス化することができる。修正した XHTML 文書と形態論情報は GitHub 上で公開しており、*¹⁰、OpenCHJ 中納言においても利用可能である。

なお、OCX 要素の追加後も XML としての整形式性は維持されるが、元の XHTML DTD または TEI スキーマに対する妥当性を保証するには、OCX を組み込んだ複合スキーマまたはカスタマイズが必要となる。

4.3 TEI 文書の OpenCHJ 化

すでに構造化された TEI 文書を OCX mini でアノテーションしてコーパス化した例として、明治・大正期に刊行された国定理科教科書を取り上げる。

対象としたのは、文部省著『尋常小學理科書：兒童用 第五學年』(1910 年刊)、『尋常小學理科書：兒童用 第五學年』(1918 年刊)、および『尋常小學理科書：兒童用 第六學年』(1918 年刊) の 3 資料である。これらは教育図書館・文部科学省図書館 OPAC で画像が公開されている*¹¹。本事例では、公開画像から本文テキストを作成し、構造化と形態素解析を行ったうえで、OpenCHJ で利用可能なコーパスデータとして整備した。

まず、教科書画像を PDF 形式で取得し、作成したプロンプトに基づいて、生成 AI を用いた OCR と TEI 形式 XML データの生成を半自動的に行った。生成されたデータについては、原資料との照合を行い、本文の誤認識を修正するとともに、ファイル間で異なっていた要素の配置や記述形式を手で統一した。

生成された TEI 文書では、書誌情報を<teiHeader>に、書籍の内容を<text>に記述し、さらに<text>の内部を、表紙・目録に相当する<front>、本文に相当する<body>、奥付に相当する<back>に分けている。このほか、本文では、各課を区切る<div>、課の題目を示す<head>、文を記述する<s>、ページの変わり目を示す<pb/>などを用いている。これらの TEI のタグは、OpenCHJ 化に際して特に変更する必要がなく、そのまま利用できる。このように、資料の構造があらかじめ明示された TEI 文書は、既存の文書構造を保持したまま OCX mini に変換しやすい。

OpenCHJ 化にあたっては、TEI 文書全体の構造を維持しつつ、形態素解析の対象となる本文部分に OCX mini の

*¹⁰ <https://github.com/togiso/OpenCHJ-Aozora>

*¹¹ https://nierlib.nier.go.jp/opac/opac_search/?lang=japanese

要素を追加した。具体的な変換方針は以下の通りである。

- 解析対象範囲の明示: 形態素解析の対象を教科書本文に限定するため、<body>内に<ocx:doc>を配置した。<teiHeader>、<front>、<back>に記述された書誌情報、目録、奥付などは保持するが、形態素解析の対象とはしない。
- 文境界の明示: 形態素解析における一文の範囲を示すため、各文の末尾に<ocx:eos/>を付与した。
- ルビ情報の保持: 原資料に付されたルビは<ocx:r>要素でマークアップし、読みを rt 属性に記述した。

この設計により、TEI によって記述された書籍全体の構造と、OpenCHJ での形態素解析・検索に必要な情報とを、一つの XML 文書内に共存させることができる。

変換後の本文部分の例を以下に示す。

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>尋常小學校理科書 第五學年兒童用</title>
        <author>文部省</author>
      </titleStmt>
      <publicationStmt>
        <publisher>文部省</publisher>
        <date when="1910">明治 43 年</date>
        <pubPlace>東京</pubPlace>
      </publicationStmt>
      <sourceDesc>
        <bibl>
          <title>尋常小學校理科書 第五學年兒童用</title>
          <date>明治 43 年 11 月 24 日發行</date>
        </bibl>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <front>
      <div type="title_page">
        <s xml:id="s000_1">第五學年兒童用</s>
        <s xml:id="s000_2">尋常小學校理科書</s>
        <s xml:id="s000_3">文部省</s>
        : (中略)
      </div>
    </front>
    <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">
        : (中略)
      <div type="chapter" n="6">
        <head>六 竹</head>
        <p>
          <s xml:id="s042">竹の幹は多くの<ocx:r rt="ふし">節</ocx:r>を有し、節と節との間は空なり。
          <ocx:eos/></s>
          <s xml:id="s043">葉は枝の節に着き、その本は枝を包む。
          <ocx:eos/></s>
          <s xml:id="s044">葉には並行せるすぢあり。
          <ocx:eos/></s>
          <s xml:id="s045">幹の下端に連りて地中に地下莖といふ莖あり。
          <ocx:eos/></s>
          <s xml:id="s046">これにも多くの節あり。
          <ocx:eos/></s>
          <pb n="11"/>
        </p>
      </div>
    </body>
  </text>
</TEI>
```

```
<s xml:id="s047">幹の下端及び地下莖より多くの細き根出づ。
<ocx:eos/></s>
<s xml:id="s048">筍は地下莖の節より出づる若き幹にして、多くの皮にて包まる。
<ocx:eos/></s>
<s xml:id="s049">此の皮を「竹の皮」といふ。
<ocx:eos/></s>
<s xml:id="s050">竹は地下莖を伸し、筍を生じて次第に繁茂するものにして、花を生ずること稀なり。
<ocx:eos/></s>
<s xml:id="s051">竹の幹は諸種の器具を造るに用ふる等其の用途甚だ多し。
<ocx:eos/></s>
</p>
```

: (中略)

```
</div>
</ocx:doc>
</body>
<back>
  <div type="colophon">
    <s xml:id="s_colophon1">明治四十三年十一月廿一日印刷</s>
    : (中略)
    <s xml:id="s_colophon10">印刷所 東京書籍株式会社工場</s>
    <s xml:id="s_colophon11">發賣所 國定教科書共同販賣所</s>
  </div>
</back>
</text>
</TEI>
```

この XML 文書を対象として、形態素解析器 MeCab と近代文語 UniDic[6]^{*12}を用いて形態素解析を行った。解析結果は形態論情報修正用のデータベースに取り込み、原資料および形態論情報規程集を参照しながら、人手による修正を行った。

以上のように、TEI 文書の構造を維持しながら、本文の解析対象範囲や文境界などを OCX mini によって明示することで、既存の TEI データを OpenCHJ の検索・分析環境で利用できるコーパスデータへと変換できる。作成した国定教科書の XML データは GitHub 上で公開しており^{*13}、形態論情報を付与したデータも公開する予定である。

5. OpenCHJ 中納言

「中納言」は、国立国語研究所が開発・運用する Web ベースのコーパス検索アプリケーションである [8], [9]。特別なソフトウェアを導入せずにブラウザ上で利用でき、検索対象となるコーパスに応じて、文字列検索に加え、UniDic に基づく短単位・長単位の形態論情報を用いた検索を行うことができる。語彙素、語形、書字形、品詞、活用型・活用形などを条件として指定し、前後の語との共起条件を組み合わせて用例を抽出できるほか、検索結果には前後文脈、形態論情報、著者情報、出典情報等が表示される。得られた結果はタブ区切りテキストとしてダウンロードでき、集計や統計分析に利用することができる。中納言には、現代

^{*12} https://clrd.ninjal.ac.jp/unidic/download_all.html#unidic_kindai

^{*13} <https://github.com/togiso/OpenCHJ-KokuteiRikaTextbooks>

図 1: OpenCHJ 中納言の短単位検索画面 (中納言 2.7.2、データバージョン 2026.05)

語・歴史語、書き言葉・話し言葉、方言、学習者言語などを対象とする多様なコーパスが搭載され、共通のインターフェースから検索できる。利用には無償のユーザ登録を要するが、日本語研究・日本語教育におけるコーパス利用の基盤として広く用いられている [10]。

OpenCHJ 中納言^{*14}は、この中納言上で OpenCHJ のデータを検索・利用するための公開環境である。2025 年 3 月に、源氏物語 (渋谷栄一版)、青空文庫所収作品、速記叢書講演説集を収録して最初の版を公開した。続いて、同年 11 月に短単位版「関西弁コーパス」を追加し、2026 年 5 月には虎寛本狂言集、上代歌謡 (仏足石歌・歌経標式例歌)、青空文庫所収小説等 13 作品を追加した [11]。これにより、上代から現代までの異なる時代・ジャンルの資料を、共通の短単位体系と検索インターフェースによって横断的に利用できるようになっている。

OpenCHJ では、オープンライセンスの原資料をもとに

整備した本文と形態論情報を、公開可能な範囲でリポジトリからも提供している [12]。この原データへのリンクは、作成者の情報とともに中納言の検索結果に表示される。したがって、利用者は中納言によって用例を検索できるだけでなく、基礎データを取得して独自の分析や再利用を行うこともできる。外部の研究者等が整備した資料を、国語研の既存コーパスと共通する形式で検索・公開するという OpenCHJ の構想において、中納言はデータ構築と研究利用とを結ぶ公開基盤として位置付けられる。

図 1 に、OpenCHJ 中納言の短単位検索画面を示す。検索キーとして語彙素、語形、書字形、品詞等を指定できるほか、前後の短単位との共起条件を組み合わせることができる。また、検索結果に表示する情報として、形態論情報だけでなく、作品・作者・底本等の情報を選択できる。

6. おわりに

本稿では、日本語コーパスの基礎データとして XML を

^{*14} <https://chunagon.ninjal.ac.jp/open-chj/>

用いる意義と、形態素解析および文字位置に基づく形態論情報管理の観点から生じる課題を整理し、そのための簡易な文書定義として OpenCHJ XML を提案した。OCX mini は、既存の文書構造をできる限り維持しながら、解析対象範囲、文境界、ルビ、注記を最小限の規則によって処理できるようにするものである。OCX Standard はこれを拡張し、文書構造や原文表記、解析方法について複数の資料・作成者・処理系で共有可能な記述方法を与える。上代歌謡、青空文庫 XHTML、近代国定理科教科書 TEI への適用例から、資料の形態や既存の XML 語彙が異なる場合にも、共通の処理規則によってコーパス化できることを示した。

今後は、OCX v0.6 の仕様とスキーマを確定するとともに、文書の検証、既存 XML からの変換、解析用テキストの生成、形態論情報の取り込みを支援するツールを整備する必要がある。また、具体的な適用例と作成手順を蓄積し、外部の研究者が資料を提供・修正しやすい共同構築の仕組みを充実させることが重要である。多様なオープンライセンス資料が OCX によって継続的にコーパス化され、原資料・XML・形態論情報が再利用可能な形で公開されるとともに、OpenCHJ 中納言から統一的に検索できるようになることで、OpenCHJ が日本語研究のための持続的で開かれた共通基盤へ発展することが期待される。

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」および JSPS 科研費 26K21717, 25H01243 の成果の一部である。

参考文献

- [1] 国立国語研究所コーパス開発センター: 『現代日本語書き言葉均衡コーパス』利用の手引 第 1.1 版, 国立国語研究所 (2015). https://repository.ninjal.ac.jp/record/3244/files/BCCWJ_Manual_v1_1.pdf.
- [2] 小木曾智信, 中村壮範: 『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 (2014). DOI: 10.5715/jnlp.21.301.
- [3] 小木曾智信, 岡 照晃, 中村壮範, 八木 豊: 『日本語歴史コーパス』における原文 KWIC 表示機能の実装, 言語資源活用ワークショップ発表論文集, Vol. 2, pp. 252–257 (2017). DOI: 10.15084/00001526.
- [4] TEI Consortium: TEI P5: Guidelines for Electronic Text Encoding and Interchange (2026). Version 4.11.0, <https://tei-c.org/release/doc/tei-p5-doc/en/html/>.
- [5] Frellesvig, B., Horn, S. W. et al.: Oxford-NINJAL Corpus of Old Japanese (2025).
- [6] 小木曾智信, 小町 守, 松本裕治: 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, Vol. 20, No. 5, pp. 727–748 (2013).
- [7] 沖森卓也, 佐藤 信, 矢嶋 泉: 歌経標式——注釈と研究, 桜楓社 (1993).
- [8] 中村壮範, 小木曾智信: Web 版コーパス検索アプリケーション「中納言」の公開, 言語処理学会第 17 回年次大会発表論文集, pp. 344–347 (2011).
- [9] 国立国語研究所: 中納言オンラインマニュアル: 中納

言の特徴, <https://clrd.ninjal.ac.jp/chu-00.html>. 2026 年 7 月 27 日参照.

- [10] 国立国語研究所: 国立国語研究所で公開されているコーパスの利用状況, <https://clrd.ninjal.ac.jp/data.html>. 2026 年 7 月 27 日参照.
- [11] OpenCHJ Project: OpenCHJ 中納言 更新履歴, <https://openchj.github.io/history.html>. 2026 年 7 月 27 日参照.
- [12] OpenCHJ Project: OpenCHJ Project, <https://openchj.github.io/>. 2026 年 7 月 27 日参照.