『高等小学読本』の形態論情報付きコーパス

近藤 明日子† 小木曽 智信† 加藤 文明子††
†人間文化研究機構 国立国語研究所 ††成蹊大学大学院 文学研究科 院生

明治期から昭和期にかけて設置された高等小学校で用いられた，国定の国語科教科書『高等小学読本』の全文コーパス「高等小学読本コーパス」の概要について報告する。本コーパスは一般に公開する予定である。
本コーパスはXMLを用いて，本文テキストにその言語的階層構造や表記等に関する情報を併せて記述する。特に，形態素解析辞書「近代文語UniDic」による形態素解析結果に基づく形態論情報の付与に特長がある。
また，本コーパスを簡便に利用してもらうため，全文検索システム「ひまわり」に搭載した形式も用意する。「ひまわり」では本文テキストに対する文字列検索だけでなく，形態論情報に対する検索も可能である。
最後に，本コーパスの形態論情報を用いた語彙分析の例を紹介する。

The Morphologically Annotated Corpus of "*Koto Shogaku Tokuhon*"

KONDO Asuko† OGISO Toshinobu† KATO Fumiko††
†National Institute for Japanese Language and Linguistics
††Facalty of Humanities, Seikei University

In this paper, we describe the corpus of "*Koto Shogaku Tokuhon* (高等小学読本)": Japanese textbooks used in higher elementary schools from the Meiji era to the Showa era. This corpus is characterized by morphological annotations based on *Kindai Bungo UniDic*: an electronic dictionary for morphological analysis which aims to modern Japanese language written in classical style. Moreover, this corpus is included in the full text searching system *Himawari*, in order to make it easy for users to look up characters and words in it. In addition, we show an analysis of the vocabulary of this corpus using annotations of it.

1. はじめに

発表者は『高等小学読本』の形態論情報付きコーパス（以下，「高等小学読本コーパス」）を作成しており，整備完了後，一般に公開する予定である。「高等小学読本コーパス」は，尋常小学校に続く教育課程として，明治期から昭和期に設置された高等小学校で用いられた国定の国語科教科書『高等小学読本』をコーパス化したものである。コーパスの規模は，延べ語数で約84,000語（記号類は除く）である。

このコーパスはXMLを用いて，本文テキストにその言語的階層構造や表記等に関する情報を併せて記述するが，最大の特長は，語単位で見出し語形・品詞・活用型・活用形・語種などの形態論情報を付与している点である。

本発表では，「高等小学読本コーパス」の概要について報告し，さらに，本コーパスを用いた語彙分析の例を紹介する。

2. 『高等小学読本』とは

『高等小学読本』の用いられた高等小学校とは，1886（明治19）年から1941（昭和16）年まで設置された，尋常小学校に続く教育課程である[1]。そこで用いられた国語科教科書（読本）は，当初は文部省の検定を経た複数種類が刊行されたが，1903（明治36）年に小学校教科書の国定制度が確立したことを受けて，1904（明治37）年以降は，国定の読本である『高等小学読本』1種に統一された。この国定『高等小学読本』は，その使用時期により4期に分けることができるが（[4], p.8），本コーパスは，1904（明治37）年度から1911（明治44）年度の改訂まで使用された第1期（全8冊）のものに基づいている。

第1期『高等小学読本』の使用が開始された当初，義務教育課程の尋常小学校（4年制）に続く高等小学校は，2年制・3年制・4年制の3種があったが，1908（明治41）年以降は，尋常小学校の6年制化に伴い，高等小学校は2年制または3年制の2種となった。新旧の制度での高等小学校の学年と各学年で使用された『高等小学読本』の冊番号を表1に示す[2]。

高等小学校への進学率は1911（明治44）年に

表1 新旧高等小学校の学年と『高等小学読本』冊番号との対応関係

旧制高等小学校（1907年度以前）	新制高等小学校（1908年度以降）	『高等小学読本』冊番号
第一学年		一・二
第二学年		三・四
第三学年	第一学年	五・六
第四学年	第二学年	七・八
	第三学年	（新制第三学年用）

```
<sentence>
  <ruby rubyText="みこと">
    <SUW orthToken="命" lForm="ミコト" lemma="尊" pos="名詞-普通名詞-一般" Form="ミコト" pronToken="ミコト"
      wType="和" start="1140" end="1150" morphID="800" BOS="True"/>
    命
  </ruby>
  <SUW orthToken="は" lForm="ハ" lemma="は" pos="助詞-係助詞" Form="ハ" pronToken="ワ" wType="和"
    start="1150" end="1160" morphID="810"/>
  は
  <SUW orthToken="重い" lForm="オモイ" lemma="重い" pos="形容詞-一般" Form="オモシ" cType="文語形容詞-ク"
    cForm="連体形-イ音便" pronToken="オモイ" wType="和" start="1160" end="1180" morphID="820"/>
  重い
  <ruby rubyText="ふくろ">
    <SUW orthToken="袋" lForm="フクロ" lemma="袋" pos="名詞-普通名詞-一般" Form="フクロ" pronToken="フクロ"
      wType="和" start="1180" end="1190" morphID="830"/>
    袋
  </ruby>
  <SUW orthToken="を" lForm="ヲ" lemma="を" pos="助詞-格助詞" Form="ヲ" pronToken="オ" wType="和"
    start="1190" end="1200" morphID="840"/>
  を
  <SUW orthToken="せおっ" lForm="セオウ" lemma="背負う" pos="動詞-一般" Form="セオウ" cType="文語四段-ハ行"
    cForm="連用形-促音便" pronToken="セオッ" wType="和" start="1200" end="1230" morphID="850"/>
  せおっ
  <SUW orthToken="て" lForm="テ" lemma="て" pos="助詞-接続助詞" Form="テ" pronToken="テ" wType="和"
    start="1230" end="1240" morphID="860"/>
  て
  <SUW orthToken="をら" lForm="オル" lemma="居る" pos="動詞-非自立可能" Form="オル" cType="文語四段-ラ行"
    cForm="未然形-一般" pronToken="オラ" wType="和" start="1240" end="1260" morphID="870"/>
  をら
  <SUW orthToken="れ" lForm="レル" lemma="れる" pos="助動詞" Form="ル" cType="文語下二段-ラ行" cForm="連用
    形-一般" pronToken="レ" wType="和" start="1260" end="1270" morphID="880"/>
  れ
  <SUW orthToken="ます" lForm="マス" lemma="ます" pos="助動詞" Form="マス" cType="助動詞-マス" cForm="終止
    形-一般" pronToken="マス" wType="和" start="1270" end="1290" morphID="890"/>
  ます
  <SUW orthToken="の" lForm="ノ" lemma="の" pos="助詞-準体助詞" Form="ノ" pronToken="ノ" wType="和"
    start="1290" end="1300" morphID="900"/>
  の
  <SUW orthToken="で" lForm="ダ" lemma="だ" pos="助動詞" Form="ダ" cType="助動詞-ダ" cForm="連用形-一般"
    pronToken="デ" wType="和" start="1300" end="1310" morphID="910"/>
  で
  <SUW orthToken="、" lForm="" lemma="、" pos="補助記号-読点" Form="" pronToken="" wType="記号"
    start="1310" end="1320" morphID="920"/>
```

図 1　「高等小学読本コーパス」の XML 形式の例

は 48%[3] に達しており，『高等小学読本』は当時の日本人の国語能力の形成に一定の影響力を持っていたと考えられ，その日本語資料としての価値は高いものであると言える。

3．コーパスの形式

「高等小学読本コーパス」は，マークアップ言語である XML を用い，本文テキストに言語的階層構造や表記等に関する情報を併せて記述する（図 1）。

本文テキストについては，文字集合 JISX0208:1997（JIS 第一・第二水準）を用いて電子化する。また，検索や形態素解析での便宜のため，原文表記に次のような校訂を行う。校訂前の原文表記の情報は，XML タグによって記述する。

①漢字カタカナ交じり文はカタカナをひらがなに校訂する。
②踊り字の一部（例：「流しゝかば」「ことゝて」）を通常の仮名に校訂する。
③棒引き仮名遣いは現代の仮名遣いに校訂する。棒引き仮名遣いとは，漢字音や感動詞を仮名書きする場合に長音記号「ー」を用いる表記法であり，例えば「ちょうちょ（蝶々）」は「ちょーちょ」，「ああ」は「あー」となる。1900（明治 33）年の小学校令施行規則により学校教育で用いられたが，1908（明治 41）年の文部省令で廃止された。本コーパスの対象とする第 1 期『高等小学読本』もこの棒引き仮名遣いが用いられている。

言語的階層構造や表記等に関する情報を記述するための XML タグセットは『現代日本語書き

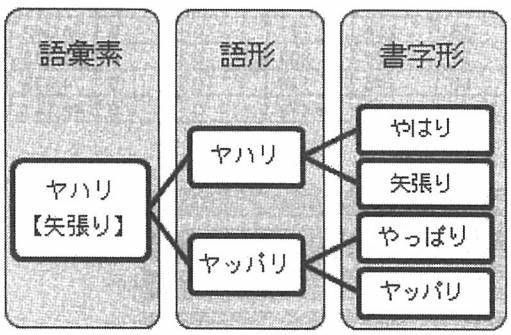
言葉均衡コーパス』[5]に準じ，一部は『高等小学読本』特有の情報を記述するため独自の仕様とする。

本コーパスの最大の特長として，語単位で見出し語形・品詞・活用型・活用形・語種などの形態論情報が付与されている点が上げられる。これまで，日本の近代以前の言語資料のコーパスで，形態論情報を付与して公開されたものはほとんどない。

形態論情報は，近代文語資料を対象とした形態素解析辞書「近代文語 UniDic」[6]を用いて本文を形態素解析した後，人手修正を加えたものを付与する。その特色として次のような点があげられる。

(1) 語の単位に「短単位」を採用している。「短単位」は「近代文語 UniDic」や現代語用の形態素解析辞書「UniDic」[7]で採用されている解析単位で，単位や品詞等の認定に関して厳密な規程[8]を持つ，ゆれの少ない斉一な単位である。よって，本コーパスと「近代文語 UniDic」や「UniDic」で解析した他の資料の解析結果との相互比較が可能である。

(2) 表記の揺れや語形の変異にかかわらない見出し語が付与されている。これも各 UniDic に共通の辞書構造に由来するものであるが，見出しが語彙素・語形・書字形・発音形の 4 つのレベルの階層構造を持つ。例えば，語彙素（見出し語に相当）「ヤハリ【矢張り】」の下のレベルには，語形「ヤハリ」「ヤッパリ」が布置され，さらにその下のレベルには，各語形に対する書字形「やはり」「矢張り」と「やっぱり」「ヤッパリ」が布置される（図 2）。



※発音形レベルは省略

図 2　形態論情報の階層構造の例

よって，研究目的に応じ，同語異語判別のレベルを自由に選択することができる。

(3) 語種情報が付与されている。和語・漢語・外来語・混種語といった語種は，日本語の語彙分析において基本的な観点としてとりあげられてきた。各 UniDic は解析結果にこの語種情報の付与が可能であり，本コーパスもその情報を取り入れている。

以上のように，本コーパスの形態論情報は日本語研究に適したものとなっている。

4．XML タグセットの概要

本コーパスで使われる XML タグセットの概要は以下のとおりである。

① sample
読本 1 冊を表す。以下の属性を持つ。
sampleID：読本名と冊番号を表す。

② cluster
読本 1 冊は 20〜22 の課から成っており，その 1 課分を表す。以下の属性を持つ。
ka　：課番号を表す。
title　：課の題名を表す。
buntai　：文体の種類（文語／口語）を表す。
kana　：表記に使用する仮名の種類（ひらがな／カタカナ）を表す。

③ title
冊や課の題名部分を表す。

④ sentence
文に相当するまとまりを表す。原則として，句点などの表記上の手がかりに基づいて自動認定する。以下の属性を持つ。
type　：値が quasi の場合，文区切り文字以外の基準により自動付与されたものを表す。

⑤ SUW
語（短単位）を表す。以下の属性を持つ。
orthToken　：書字形を表す。
lForm　：語彙素読みを表す。
lemma　：語彙素（見出し語に相当）を表す。
pos　：品詞を表す。
Form　：語形を表す。
cType　：活用型を表す。
cForm　：活用形を表す。
pronToken　：発音形を表す。
wType　：語種を表す。
start　：語の始まる文字位置を表す。
end　：語の終わる文字位置を表す。
morphID　：語の通し番号を表す。
BOS　：値が true の場合，文頭に現れる語であることを表す。

⑥ ruby
振り仮名を表す。以下の属性を持つ。
rubyText：振り仮名の文字列を表す。

⑦ correction
踊り字・棒引き仮名遣いを現代の表記に校訂したことを表す。以下の属性を持つ。

type　　：値が odoriji の場合，踊り字を校訂したことを，値が boubiki の場合，棒引き仮名遣いを校訂したことを表す。
originalText：原文を表す。

⑧ warigaki
割書された文字列を表す。

⑨ kogaki
小書きされた文字列を表す。

⑩ position
原本での位置を表す。以下の属性を持つ。
page：ページ番号を表す。

⑪ br
論理的改行を表す。

5．「ひまわり」でのコーパス利用

本コーパスは XML 形式によるテキストファイルでの公開だけでなく，簡便に利用してもらうため，全文検索システム「ひまわり」[9]に搭載した形でも公開する予定である。「ひまわり」では，まず，本文テキストや振り仮名に対する文字列検索が可能である。検索結果は KWIC により表示され，冊番号，課番号，課の題名等の情報も併記される（図 4，稿末）。

また，語彙素（見出し語）・語彙素読み・品詞等の形態論情報に対する検索も可能である（図 5，稿末）。ただし，検索結果ではマッチした語は後文脈に接続した形で表示され，完全な KWIC 形式とはならない。これは，形態論情報の付与に用いられる SUW 要素が空要素であるために起こる現象である。

さらに，マッチした文字列・語について Web ブラウザでより広い文脈で閲覧することもできる。その画面では前後文脈中の語の形態論情報も確認することができる（図 6，稿末）。

6．語彙分析の例

ここで，「高等小学読本コーパス」を用いた語彙分析の例を紹介する。コーパスに付与されている，課ごとの文語・口語の文体情報と，語の語種情報を利用し，文体別の語種比率を見てみよう。比較資料として，①「近代文語 UniDic」開発のために作成した旧民法と福沢諭吉・山路愛山・北村透谷各人の論説文のデータ（短単位による形態論情報付与済み），②「近代文語 UniDic」Ver.1.1 で形態素解析した『太陽コーパス』[10]の文語記事（NDC 第 1 次区分が「9」（文学）」の記事を除く）のデータをとりあげる。

各資料の延べ語数（記号類・未知語を除く）と自立語の語種別異なり語数を表 2 に示す。また，表 2 から和語・漢語・外来語・混種語の異なり語数について比率を求め，グラフに示したものが図 3 である。

表 2　近代文語資料の延べ語数・異なり語数

		延べ語数	異なり語数							
			和語	漢語	外来語	混種語	固有名	記号	不明	合計
高等小学読本	文語	49,216	2,494	2,781	35	161	410	0	0	5,881
	口語	34,360	2,027	1,459	21	95	214	0	0	3,816
民法		47,723	337	1,062	2	59	1	0	0	1,461
福沢諭吉		82,560	1,601	4,298	36	236	171	0	0	6,342
山路愛山		21,795	1,059	2,006	32	131	264	0	0	3,492
北村透谷		39,039	1,370	2,951	66	145	146	0	0	4,678
太陽		3,016,558	9,122	26,879	1,495	1,173	9,478	0	0	48,147

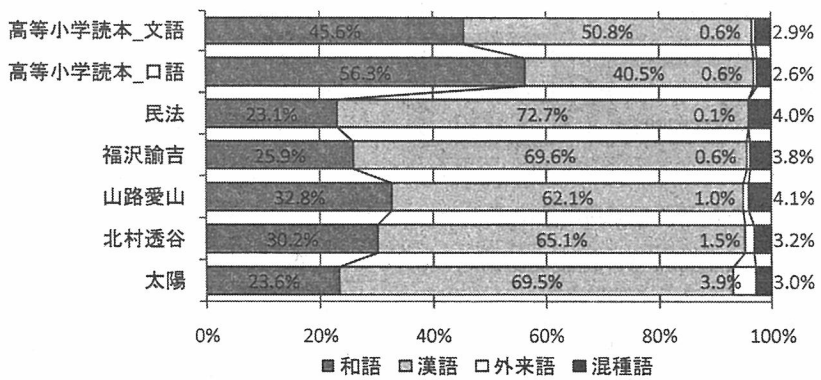


図 3　近代文語資料の語種比率（異なり語数ベース）

これを見ると，『高等小学読本』内部では，文語のほうが口語より漢語の比率が高く，逆に和語の比率が低い。また，『高等小学読本』と他の近代文語資料を比較すると，『高等小学読本』の文語よりも他の近代文語資料のほうがさらに漢語の比率が高く，和語の比率が低くなっていることがわかる。

このように，本コーパスの XML タグ情報を活用することで，語彙等の計量的分析が可能であり，また，本コーパスと同様の短単位による形態論情報を付与した他の資料との比較分析を容易に行うことができる。

7．おわりに

以上，『高等小学読本』の形態論情報付きコーパスの概要の報告と，それを用いた語彙分析例の紹介を行った。将来，『高等小学読本』以外の資料でも形態論情報付きコーパスが公開されることで，現代日本語同様，近代以前の日本語でもコーパス言語学的手法による研究が盛んになることが期待される。

注

1 以下，当時の学制・教科書制度に関しては[1][2][3]等を参照した。

2 新制第三学年用の『高等小学読本』は制度改正に併せて 1908 年度から用いられた。ただし，本コーパスには収録していない。

3 1910 年度の尋常小学校卒業者 749,994 人に対する 1911 年度の高等小学校入学者 358,089 人の比率。人数は各年度の『日本帝国文部省年報』に拠る。

参考文献

[1] 文部省（編・監修）：学制百年史，1981, http://www.mext.go.jp/b_menu/hakusho/html/hpbz198101/より閲覧可

[2] 国立教育研究所（編）：日本近代教育百年史 第四巻 学校教育 2，文晶堂，1974

[3] 日本国語教育学会（編）：国語教育辞典，朝倉書店，2001

[4] 中村紀久二（解説）：復刻 国定高等小学読本 解説，大空社，1991.

[5] http://www.ninjal.ac.jp/kotonoha/

[6] http://www2.ninjal.ac.jp/lrc/index.php?UniDic

[7] http://download.unidic.org/

[8] 小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・原裕：『現代日本語書き言葉均衡コーパス』形態論情報規程集 第 3 版，国立国語研究所，2010

[9] http://www2.ninjal.ac.jp/lrc/

[10] 国立国語研究所（編）：太陽コーパス 雑誌『太陽』データベース，博文館新社，2005

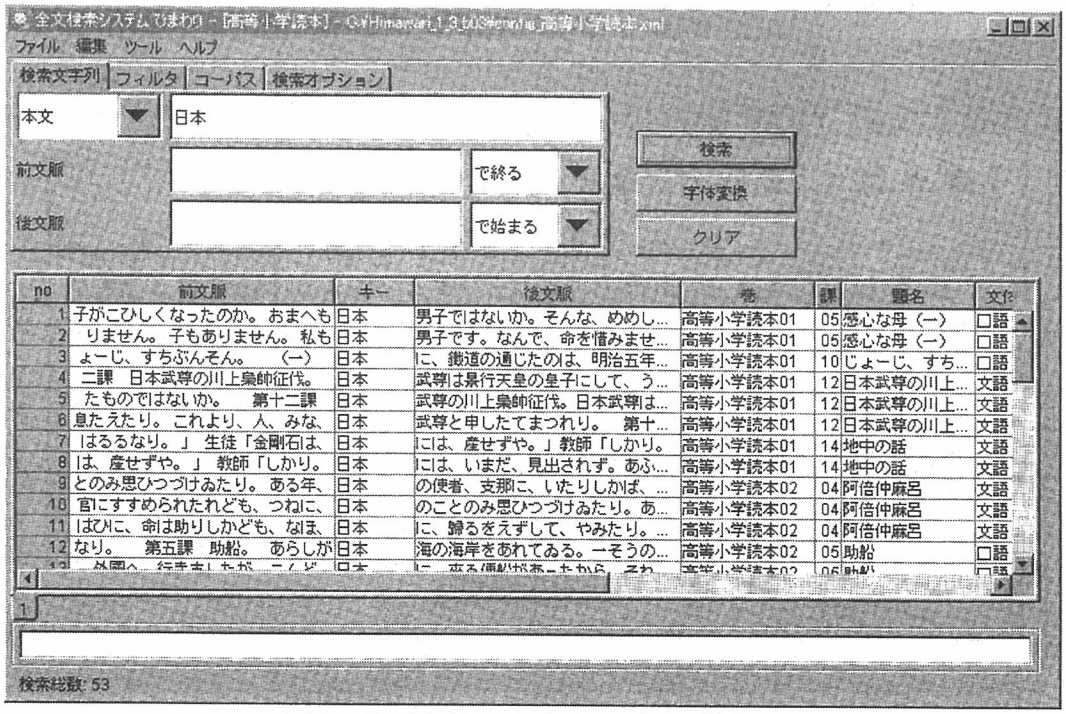


図 4 「ひまわり」の検索画面（文字列検索）

図５　「ひまわり」の検索画面（語検索）

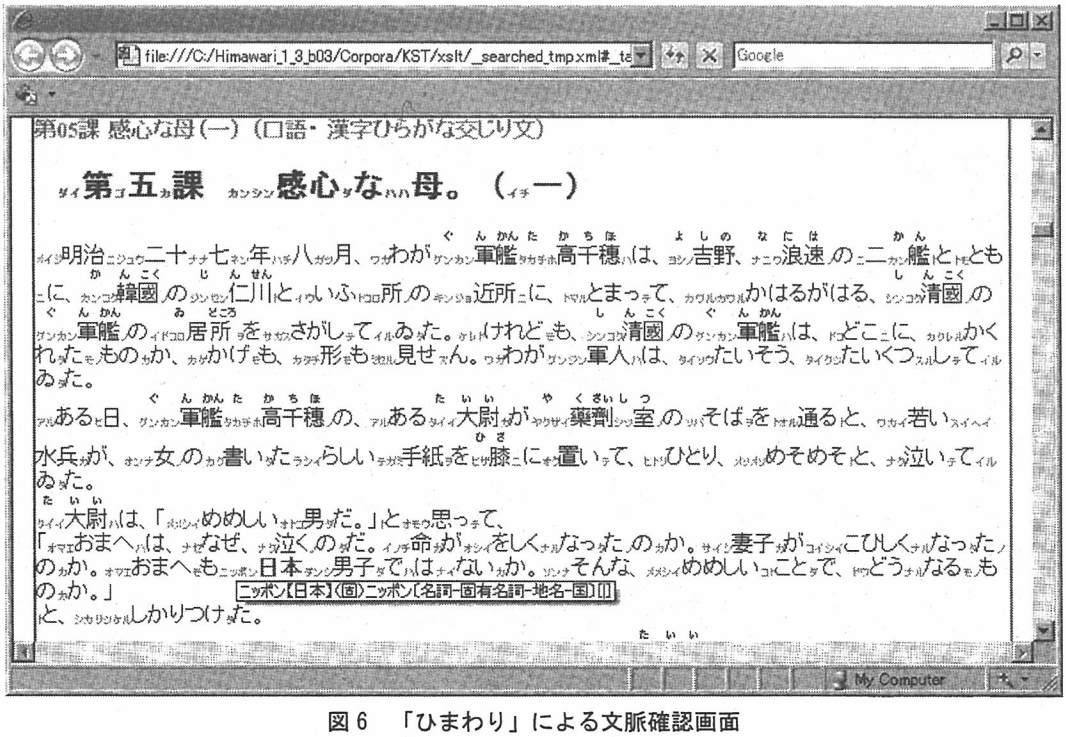


図６　「ひまわり」による文脈確認画面