タイトル	小学校国語教科書の電子化データの構成と利用に関する基礎的考察
著者	桃内，佳雄；立野，健人；MOMOUCHI，Yoshio；TATENO，Kento
引用	北海学園大学工学部研究報告(40)：129-138
発行日	2013-02-12

研究ノート

小学校国語教科書の電子化データの構成と利用に関する基礎的考察

桃内佳雄*・立野健人*

Fundamental Study of Constructions and Applications of Electronic Data of Japanese Textbooks for Elementary Schools

Yoshio MOMOUCHI* and Kento TATENO*

要　旨

小学校国語教科書には，様々な情報が含まれている．基本的な表現形は，文字列データ（文章）と画像データ（挿絵）である．文字列データは，音声源データとしても用いられうる．これらの種々のデータを電子化データとして構成し，利用するための基本的なしくみについて考察し，そのシステムとしての実現を試みる．

1．まえがき

小学校国語教科書には，様々な情報が含まれている．基本的な表現形は，文字列データ（文章）と画像データ（挿絵）である．文字列データは，内容に依存して，本文，てびき，言葉の練習，言葉の学習などの情報に分けられる．また，文字列データは，音声源データとしても用いられうる．これらのデータを電子化データとして構成し，利用するための基本的なしくみについて考察し，そのシステムとしての実現を試みる．最終的に構成される電子化データは，コード付きテキストベースとXML文書である．XML文書の利用システムは，JavaScriptコードを組み込んだ動的なHTMLプログラムとしての実現を試みた．また，コード付きテキストベースは，日本語文章の解析や解析アルゴリズムの設計のための貴重なデータとして利用した．

2．小学校国語教科書の構成

本考察で対象とする小学校国語教科書は，「教育出版」発行の「しょうがくこくご」／「小

* 北海学園大学工学部電子情報工学科
* Department of Electronics and Information Engineering, Faculty of Engineering, Hokkai-Gakuen University

学国語」[1]である．各学年上巻，下巻に分かれている．各巻の基本構成はほぼ同じであるが，低学年のほうがより簡単な構成である．基本的な構成要素が大部分出現していると思われる「6年生上巻」の概略構成を**図1**に示す．この具体例を一般化すると，小学校国語教科書の基本的な構成データは次のようになる．

・文字列データ
　① ＜タイトル＞（＜番号＞，＜表題＞，＜文章の種類＞，＜作者＞，＜訳者＞）
　② ＜本文章＞
　③ ＜てびき＞
　④ ＜補足＞（＜言葉の練習＞，＜言葉の学習＞，＜読書しよう＞，＜表現のために＞）
　⑤ ＜漢字＞（＜新しく出た漢字＞，＜前の学年までに出た漢字＞）
・画像データ
　⑥ ＜挿絵＞

文字列データの①と②が国語教科書の基本的な構成要素である．この二つを併せて本文章と呼ぶことにする．これらの構成データを漸進的に電子化する試みを進めている．

・目次
1　加代の四季　物語　杉　みき子
　・てびき
2　ふき子の父　物語　砂田　弘
　・言葉の練習
　・言葉の学習　言葉の調子
3　詩の世界　詩
　（1）詩
　（2）短歌と俳句
　・言葉の練習
4　生活を見つめて　作文
　・言葉の学習　日本の文字
5　科学者の目　説明文／記録文
　（1）せんこう花火　説明文　中谷宇吉郎
　・てびき
　（2）貝の村の人口調査　記録文　阿部　襄
　・てびき
　・表現のために　とじこみ文集
6　道産子　物語　吉田　元
　・てびき
　・言葉の練習
　・読書しよう　目的を持って本を選ぼう
7　調査したことをまとめて　作文
8　考えを深める　論説文／作文
　（1）美を求める心　論説文　小林秀雄
　（2）自分の花，自分の木　作文
　・てびき
　・言葉の練習
　・言葉の学習　語句の組み立て
9　川とノリオ　物語　いぬい　とみ
　・てびき
　・表現のために　好きな表現を書きぬこう
新しく出た漢字
前の学年までに出た漢字

図1　小学校国語教科書6年生上巻：概略構成

3．小学校国語教科書の本文章テキストベースの構成と利用

小学校国語教科書の基本的な構成要素である本文章をテキストデータとして電子化し，本文章テキストベースを構成し，本文章を構成する各要素に適切なコードを付与して，本文章テキストベースからの情報抽出，情報検索ツールの作成を行った．なお，本文章の個数は，１年上（12)・下（８)，２年上（９)・下（８)，３年上（９)・下（８)，４年上（９)・下（８)，５年上（７)・下（８)，６年上（９)・下（８）で，全部で103個となっている．

3.1　コードを付与した本文章テキストベースの構成

本文章の各要素にコードを付与する．**図２**に例を示す．この本文章は，１年生上巻の５番目の本文章である．

```
1105001000038：／Sおむすび　ころりん
1105001001038：　おじいさんが、おむすびを　おとして　しまいました。
1105001002038：　おむすびは、ころころ　ころがって、すっとんと、あなに　おちました。
1100501003038：　あなから、うたが　きこえて　きました。
1100501004038：「おむすび　ころりん　すっとんとん。」
1100501005038：　おじいさんは、うれしく　なって、おむすびを、みんな　あなに　おとしました。
1100501006040：　しまいに、おじいさんも、すっとんとんと、あなの　なかに　おちました。
1105001007040：「おじいさん、おもちを　ついて　あげましょう。」
1105001008040：　ねずみたちは、また、すっとんとんと　うたいながら、おもちを　ついて　くれました。
1105001009043：　ねずみが、おみやげを　くれました。
1105001010043：「よっこらしょ。」
1105001011043：　うちへ　かえって　みると、なかには、こばんが　いっぱい　はいって　いました。
```

図２　「おむすびころりんすっとんとん」(１年生上巻)

コードの詳細構成は次のようである．

・１桁目：学年　：１～６
・２：上・下巻　：上巻　１，下巻　２
・３－６：文章番号　：最初の２桁　文章番号；
　　３桁目　部分文章番号；４桁目　埋め込み文章番号
・７：文章の種類　：１　童話，昔話，物語；　２　説明文；　３　作文
　　４　詩　　；　５　記録文；　６　論説文
　　７　脚本　　；　８　伝記　；　９　随筆
　　０　その他
・８－10：文番号　：文章中の番号
　　文章のタイトルなど特殊な情報を表現する文については，文番号を０とする．

文番号が［000］である文はタイトル行である．
タイトル行は次のような記号を用いて表現する．
／N　番号，／S　表題，／A　著者，／T　翻訳者
［1201001000004：／N一／Sおじさんのかさ／A佐野洋子］

・11－13：ページ番号　：文が含まれているページ

文のコードは，国語教科書文章における文の位置づけを表すための基本的な情報である．文が教科書の中のどこにあって，どんな種類の文章の中にあるのかという基本的な情報がコード化されている．タイトルのコードは，学年，上・下巻，文章番号，文章の種類を含み，またタイトル部分には，タグが付与されている．これらの情報を利用して，次のような情報表示，情報抽出ツールを構成することができる．

① 教科書目次の作成ツール
② 文章種類別タイトルリストの作成ツール
③ 文章の骨組みの抽出（主文章，部分文章，埋め込み文章の構成の抽出）ツール

また，次節で述べる検索ツールKWICの利用とともに，本文章テキストベースを，例えば，次のような日本語文章解析に関する研究において，基礎的な解析データとして利用した．

① 日本語文章における複数の指示対象を持つ名詞句の解析[2)]
② 日本語文章における「の格」要求名詞の解析[3)]
③ 日本語文章におけるゼロ代名詞の解析[4)]

3.2 本文章テキストベース解析ツールKWICの作成

本文章テキストベースを文字列データとして，文字列のパターンマッチング処理によって，テキストベースに含まれるキーワード検索を行うツールKWIC（Key Word In Context）をJavaによるGUIプログラムとして作成した．ある語（Key Word）を検索し，その前後の文脈（Context）と共に出力する．たとえば，“山”という語をKey Wordとして，ファイルtext21n.txt（２年上巻テキストベース）をKWIC検索すると**図 3** のような結果が得られる．保存ボタンを押して，結果を保存することもできる．

上のKWICを基本として，いくつかの発展形を作成している．例えば，①複数キーワード指定KWIC，②複数キーワード順序指定KWIC，③否定隣接文字列指定KWIC，④否定キーワード指定KWICなどである．**図 4** は否定キーワード指定KWICの実行例である．文字列として，［山＃田］が入力されている．記号＃の後のキーワードは含んではいけない（否定）キーワードである．

図 3　KWIC

図 4　否定キーワード指定KWIC

4．小学校国語教科書の本文章XML文書の構成と利用

小学校国語教科書文章について，より柔軟な構造的，意味的処理を可能とするために，国語教科書文章をXML文書として構成する．まず，教科書中の本文章に関する基本的な構成は次のようである．この構成単位を＜教材＞と呼ぶことにする．

［　＜本文章番号＞　＜題名＞　＜作者＞　＜本文章＞　：　＜分野＞　＜ページ数＞　］

例えば，２年生の国語教科書文章上・下を，**図5. 1**のようなタグの構成によるXML文書としてまとめることができる．本文章本体は，別に本文章XML文書として作成する．**図5. 2**で，教材タグの属性urlの値として与えられている“f004. xml”や“f008. xml”がその例である．

このXML文書は，本文章本体の情報は含んでいなくて，目次情報と考えることができるので，目次XMLと呼ぶことにする．正確には，目次XML文書と表記すべきところ，文脈より明らかなので目次XMLと略して表記する．

本文章XML文書は，本文章ごとに**図 6** のような構成で作成する．本文章XMLと略記する．

基本的な目次XMLと本文章XMLを構成し，併せて，XSLTプログラムやHTMLプログラムを開発し，利用することで，本文章XMLに含まれる様々な情報の表示と利用が可能になる．

5．小学校国語教科書のXML文書の構成と利用

5. 1　小学校国語教科書XML文書の構成

前章で構成した目次XML，本文章XMLを基本的なデータとして，小学校国語教科書に含まれている他の情報，③手引き（＜てびき＞），④補足情報（＜言葉の練習＞＜言葉の学習＞＜読書しよう＞＜表現のために＞），⑤漢字情報（＜新しく出た漢字＞＜前の学年までに出た

```
＜小学国語教科書２年＞
　＜上＞
　　＜教材　url="本文章XML">
　　　＜本文章番号＞・・・＜／本文章番号＞
　　　＜副文章番号＞・・・＜／副文章番号＞
　　　＜題名＞・・・＜／題名＞
　　　＜作者＞・・・＜／作者＞
　　　＜分野＞・・・＜／分野＞
　　　＜ページ数＞・・・＜／ページ数＞
　　＜／教材＞
　　＜教材＞
　　・・・
　　＜／教材＞
　　・・・
　＜／上＞
　＜下＞
　　・・・
　　・・・
　　・・・
　＜／下＞
　＜付属　url="Top.html" >
　　＜戻る＞戻る＜／戻る＞
　＜／付属＞
　＜／小学国語教科書２年＞
```

図5.1　目次XML文書の構成

```
〈?xml version="1.0" encoding="Shift_JIS" ?〉
〈?xml−stylesheet type="text/xsl" href="mokuji01.xsl"?〉
〈小学国語教科書２年〉
　〈上〉
　　〈教材 url="f004.xml"〉
　　　〈本文章番号〉１〈/本文章番号〉
　　　〈副文章番号/〉
　　　〈題名〉はるの　くまたち〈/題名〉
　　　〈作者〉神沢　利子〈/作者〉
　　　〈分野〉どうわ〈/分野〉
　　　〈ページ数〉４〈/ページ数〉
　　〈／教材〉
　　〈教材 url="f008.xml"〉
　　　〈本文章番号〉２〈/本文章番号〉
　　　〈副文章番号/〉
　　　〈題名〉ひっこして　きた　みさ〈/題名〉
　　　〈作者〉安藤　美紀夫〈/作者〉
　　　〈分野〉どうわ〈/分野〉
　　　〈ページ数〉８〈/ページ数〉
　　〈/教材〉
　・・・
　・・・
　〈付属 url="Top.html"〉
　　〈戻る〉戻る〈/戻る〉
　〈/付属〉
〈/小学国語教科書２年〉
```

図5.2　目次XML文書の具体例

漢字＞），⑥画像データ（＜挿絵＞），⑦読み上げ音声データなどを追加していくことによって，統合的な小学校国語教科書XML文書が構成される．ただし，現時点で，小学国語教科書に含まれている上記のすべての情報を含むには至っていない．

5.2　小学校国語教科書XML文書の利用

小学校国語教科書XML文書の利用の基本的な機能として，次のような機能の実現を試みた．本文章XMLへの挿絵の挿入，また文章の読み上げ機能の付加を行った．この二つの工夫によって，挿絵を見ながら本文章を読んだり，聞いたりすることができる．

① 目次の表示
② 分野別文章検索
③ キーワード検索
④ 漢字の検索と表示
⑤ 本文章と挿絵の表示

```
＜章題 章題＝"　1　はるの　くまたち　"＞
 ＜教材＞
  ＜本文章番号＞1　＜/本文章番号＞
  ＜題名＞はるのくまたち　＜/題名＞
  ＜作者＞神沢　利子＜/作者＞
 ＜/教材＞
 ＜本文＞
  ＜文＞　山に、こぶしの　花が　さきました。＜/文＞
  ＜文＞　くまの　かあさんは、ふゆごもりの　あなから　出たばかり。＜/文＞
  ＜文＞　二ひきの　こぐまは　生まれたばかり。＜/文＞
  ＜文＞　木の　め、くさの　め、かあさんは　おいしい　ものを　さがします。＜/文＞
  ＜文＞　空が　あかるく　まぶしくて、足の　うらが　くすぐったくて、子ぐまたちは　くふくふ　わらいます。＜/文＞
  ＜文＞　かあさんは　木に　のぼります。＜/文＞
  ＜文＞　なんて　じょうずな　こと。＜/文＞
  ＜文＞　ほら、もう、あんなに　たかい　ところ。＜/文＞
  ＜文＞　青空に　ゆれる　こぶしの　花を　たべて　います。＜/文＞
  ＜文＞　子ぐまたちにも、「さあ、おたべ。」と　おとして　やります。＜/文＞
  ＜文＞　子ぐまたちは、もしゃもしゃ、ぺちゃぺちゃ、花を　たべます。＜/文＞
  ＜文＞　子ぐまたちの　上に、花は　ゆきのように　おちて　きます。＜/文＞
  ＜文＞　たべあきた　子ぐまたちは、こんどは、ころころ　おすもうごっこ。＜/文＞
  ＜文＞　かあさんは、ゆっくり　木から　おりて　きます。＜/文＞
  ＜文＞　子ぐまたちは、もう　すぐ　大きくなるでしょう。＜/文＞
  ＜文＞　木のぼりを　して、ひとりで　たべものを　さがすでしょう。＜/文＞
  ＜文＞　おや、空に　一つ、たべのこしの　花。＜/文＞
  ＜文＞　それは、白い　ひるの　お月さまです。＜/文＞
 ＜/本文＞
 ＜A url="mokuji01.xml"＞
  ＜戻る＞戻る＜/戻る＞
 ＜/A＞
＜/章題＞
```

図 6　本文章XML文書の具体例

⑥　文章の読み上げ

このような機能を組み込んだ小学校国語教科書XML文書の利用システムについて，以下に実行画面とともに説明する．システムは，1年生から6年生までのこれまでに構築されている情報を統合的に利用するものとして構成され，JavaScriptコードを組み込んだ動的なHTMLプログラムとして実現されている．

（1）トップページ

ここから，1年生から6年生までの各学年のトップページへリンクする．また，文章検索のページを別に設けてリンクするようにした．例えば，6年生のトップページへのリンクをクリックすると，6年生のトップページが表示される[5)]．

（2）6年生のトップページ

左のフレームで，全体の一覧と分野別一覧へのリンク，キーワード検索へのリンクなどが可

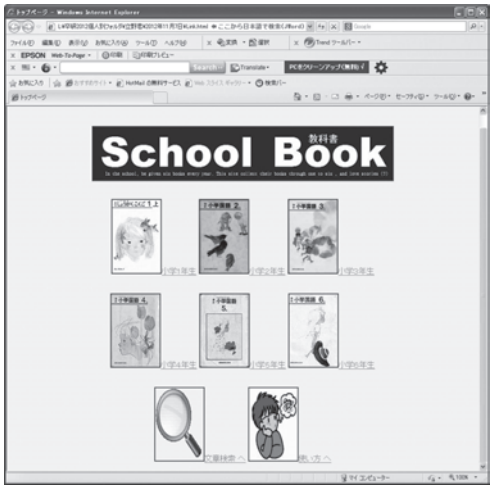

図 7　（1）トップページ

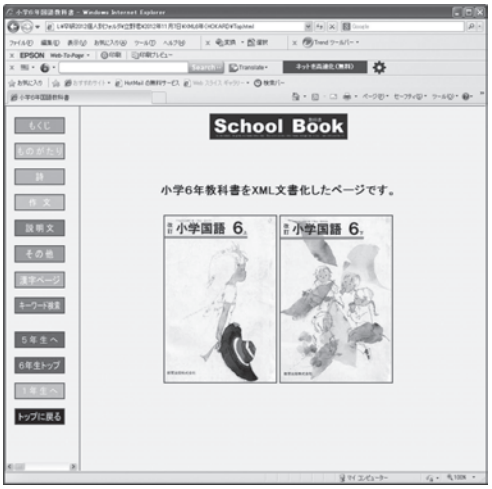

図 8　（2）6年生のトップページ

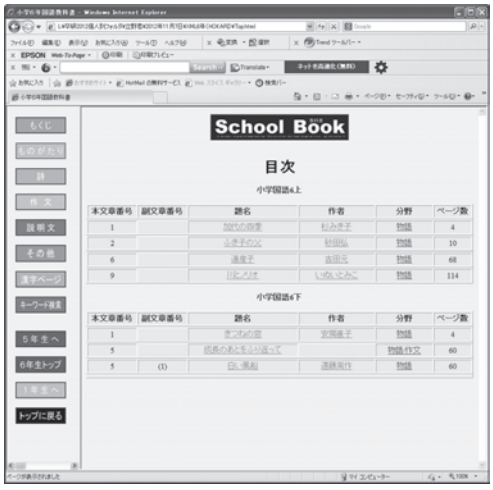

図 9　（3）[ものがたり] 目次

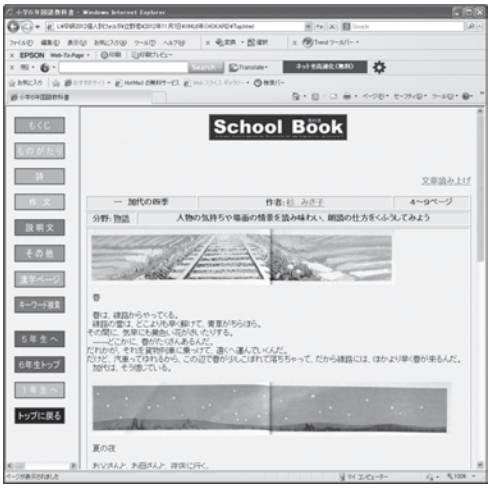

図10　（4）[加代の四季] 本文章

能である．

（3）分野別目次のページ

分野別一覧の中から，[ものがたり] をクリックすると（3）のような6年生の教科書に含まれている [ものがたり] の一覧表が表示される．[加代の四季] をクリックすると，[加代の四季] の本文章が挿絵付きで表示される．

（4）本文章のページ

[加代の四季] の本文章と挿絵のページが表示されている．右上の [文章読み上げ] をクリックすると，文章の読み上げが行われる．[作者] をクリックすると，作者に関する情報が表示される．本文章のページには，テキストデータ，挿絵（画像）データ，読み上げ音声デー

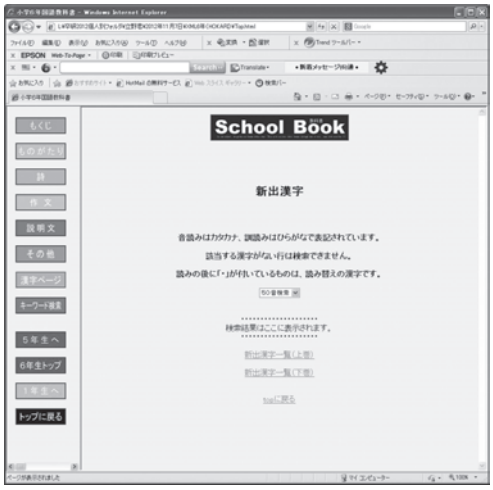

図11　（5）漢字ページ

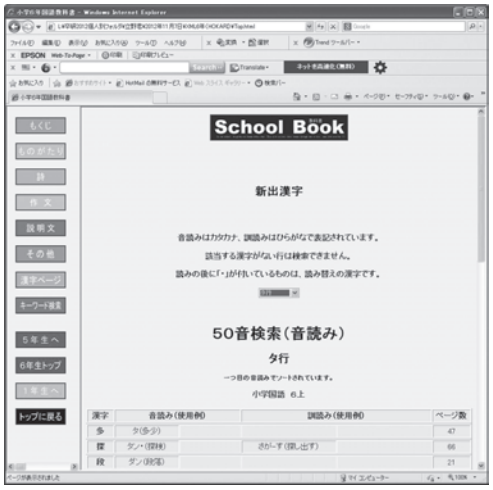

図12　（6）漢字50音検索「タ行」

タへのリンク情報が含まれている．

（5）漢字ページ

6年生のトップページで，［漢字ページ］をクリックすると，新出漢字に関するページへリンクする．漢字50音検索と新出漢字一覧表を表示することができる．

（6）漢字50音検索「タ行」の表示

（7）キーワード検索

6年生のトップページで，［キーワード検索］をクリックすると，キーワード検索のページが開く．**図13**の画面は，キーワード「電気」を入力したところである．キーワード「電気」を含む文章が表示されている．（1）のトップページからも全学年のデータを参照する形での

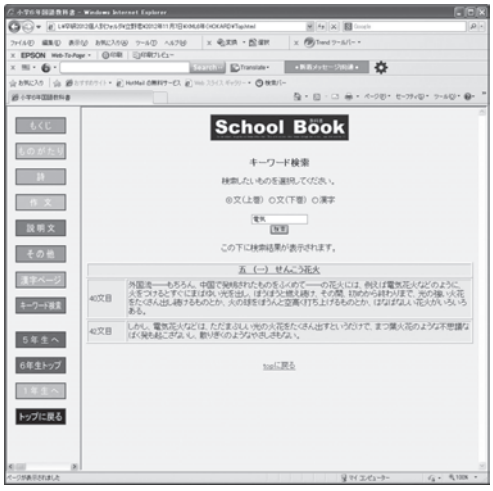

図13　（7）キーワード検索

図14　（8）トップページからのキーワード検索

キーワード検索が可能である（8）.

7．あとがき

小学校国語教科書を出発点となる原データとして，まず，コード付き本文章テキストベースを構成した．次に，目次XML文書，本文章XML文書を構成して，そこに含まれる情報を漸進的に増やしながら，同時に，利用のしくみも構築していくという試みについて報告した．その試みの過程で，小学校国語教科書に含まれているすべての情報を電子化することはできていない．例えば，補足情報や挿絵の一部，また，本文章の欄外の注釈的情報や本文章ごとの漢字情報などが電子化されていない．これらのデータの補充，電子化のための工夫，改良が残されているが，小学校国語教科書に対応する電子化データと利用システムの一つのひな形を構成することができた．

8．謝辞

本考察の基礎となるデータとシステムは，学部4年生による卒業研究の中で，段階的に作成が進められてきました．作成に携わってくれたすべての卒業研究生に感謝の意を表します．

参考文献

1）木下順二，松村明，柴田武監修：改訂しょうがくこくご1上下，1983；改訂小学国語2上下，1983；改訂小学国語3上下，1983；改訂小学国語4上下，1983，新訂小学国語5上下，1987；改訂小学国語6上下，1983，教育出版.
2）桃内佳雄：日本語物語文章における名詞句の指示対象の数の同定のための基礎的考察，電子情報通信学会／ソフトウェア科学会・言語とその環境シンポジウム論文集，pp.107-114，1990.
3）桃内佳雄：日本語文章における「の」格要求名詞について，北海学園大学工学部研究報告，21，pp.107-119，1994.
4）桃内佳雄・柴田更紗：CENTERモデルによる日本語ゼロ代名詞解析に関する基礎的考察，北海学園大学大学院工学研究科紀要工学研究，No.5，pp.85-92，2005.
5）保苅祐太：2011年度卒業研究報告書「小学六年国語教科書のXML文書の構成と利用」，2012.
6）高羽実：XML & JavaScriptシステム開発，秀和システム，2001.
7）山田祥寛：10日で覚えるXML入門教室第2版，翔泳社，2004.