TEI P5 に基づく近世口語資料の構造化とその問題点

河瀬 彰宏　市村 太郎　小木曽 智信
国立国語研究所　国立国語研究所　国立国語研究所
コーパス開発センター　コーパス開発センター　言語資源研究系

国立国語研究所では，「日本語歴史コーパス設計」プロジェクトの一環として古典資料の形態素解析を実施している．形態素解析を行うためには，基礎資料となる古典テキストの電子化が必須である．これまでに様々な時代のテキストコーパスを電子化し，公開している．しかし，これらのテキストコーパスは，国立国語研究所が独自に考案したタグセットに基づく XML を用いてマークアップが行われているため，各コーパスを規定する要素は，基本的に統一されていない．そのため，複数のコーパス間の構造比較や計量分析を機械的に実施することが現状では難しいという問題を抱えている．したがって，複数のコーパスの構造を高次の視点から統一的に記述することが求められている．本稿では，この問題を解決するために，洒落本の一冊『傾城買二筋道』の版本を事例に，TEI P5 準拠の XML 形式による文書構造化を検討する．

Problems in TEI P5 Encoding on Colloquial Japanese Documents of the Early Modern Period

Akihiro Kawase
Ctr. Corpus Development
NINJAL

Ichimura Taro
Ctr. Corpus Development
NINJAL

Toshinobu Ogiso
Dept. Corpus Studies
NINJAL

The National Institute for Japanese Language and Linguistics (NINJAL) is conducting morphological analysis on Japanese classics. Digitization has been done thus far on the literature of several ages and various text corpora are published. However, each element (tag) of the text corpora is marked up under NINJAL's Document Type Definition, which is basically neither unified nor standardized. Under this circumstance causes problem with structural analysis and numerical analyses between several corpora. Thus it is necessary to design and mark up a unified definition from a higher level in order to conduct analyses concurrently. In this study, we examine the possibilities to convert documents of classical Japanese, an old block book from *Sharebon*'s "*Keisei-kai futasuji-no-michi*" (published in 1798) as a model case, with TEI-compliant XML and discuss its issues.

１．はじめに

国立国語研究所（以下，国語研）では，「日本語歴史コーパス設計」プロジェクトの一環として古典資料の形態素解析を実施している．形態素解析を行うためには，基礎資料となる古典テキストの電子化が必要となる．

これまでに平安時代を中心に和文[1]，漢文の要素が含まれる和漢混淆文[2]，近世口語テキスト[3-4]，などの電子化および形態素解析を進めている[5]．また，『太陽コーパス』[6]，『明六雑誌コーパス』[7]，BCCWJ[8-9]，などの様々なテキストコーパスを電子化し，公開している．

上記のテキストコーパスは，国語研が独自に考案したタグセットに基づく XML（Extensible Markup Language）を用いて文書構造のマークアップを行っている．しかし，各々のコーパスを規定する要素には，共通のタグが使用される場合が少なからずあるものの，基本的には共通のタグセットを使用していない．そのため，同一コーパス内での文書構造の比較や文字列の抽出は可能である一方で，複数のコーパス間の構造比較や計量分析を機械的に実施することが現状では難しいという問題を抱えている．したがって，複数のコーパスの構造を高次の視点から統一的に記述することが求められている．

本研究では，この問題を解決するために，TEI（Text Encoding Initiative）P5[10]準拠の XML 形式による文書構造化を検討する．具体的には，近世口語資料の洒落本に含まれる『傾城買二筋道』（1798 年刊行）の版本を事例に，タグセットを考案し，構造化を試みる．また，その過程で文書構造化の問題点を整理する．

２．洒落本の特徴と電子化の意義

洒落本は，おもに次の３点の特徴をもつことから，日本の近世における重要な言語資料であると考えられている．

（1）18 から 19 世紀前半までの幅広い年代に刊行されていること
（2）登場人物の発話（会話部分）に話し言葉が用いられていること
（3）江戸語・上方語の記述が豊富であること

したがって，洒落本を機械可読な形式に整備することは，近世後期の口語の実態を計量的観点から分析することを実現させ，日本語史や書誌学などの人文学研究を促進する意義がある．また，同時代には，洒落本と類似した構造をもつテキストが数多く存在しているため，これらの文献資料をアーカイヴ化するためのフォーマットを統一的な観点から新たに提供する意義がある．

3．洒落本の構造

国文学研究資料館は，洒落本の紙面に外形的なマークアップを施し，「大系本文データベース」[11]を構築している．しかし，コーパス言語学の観点から資料を分析するためには，外形的な情報だけでなく，文書構造および言語構造についても精緻にマークアップを施すことが求められる．

図 1（左）は，『傾城買二筋道』の版本から抜粋した画像であり，図 1（右）は，その文字を読みやすくワープロ印字し直したものである．

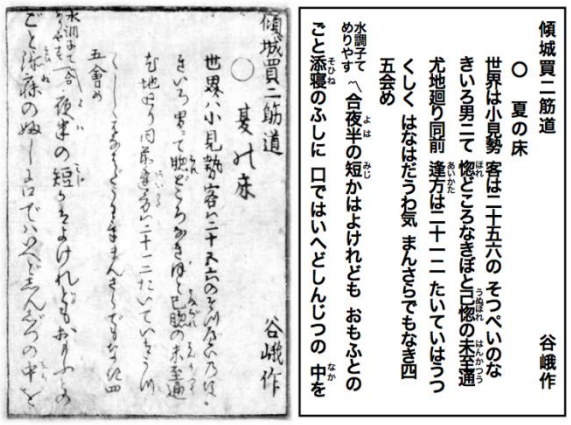

傾城買二筋道
○ 夏の床
谷峨作
世界は小見勢 客は二十五六の そつぺいのな
きいろ男にて 惚どころなきほど口惚の未全通
尤地廻り同前 達方は二十一二 たいていはうつ
くしく はなはだうわ気 まんさらでもなき四
五会め
水調子て へ合夜半の短かはよけれども おもふとの
めりやすく ごと添寝のふしに 口ではいへどしんじつの 中を

図 1 『傾城買二筋道』（国立国語研究所所蔵）の抜粋（左）版本の画像と（右）そのワープロ印字
Figure 1 (left side) Excerpt image of "*Keisei-kai futasuji-no-michi*" from the original block book (owned by NINJAL), and (right side) its characters typed with a word processor.

一般に洒落本は，（a）前付け部分，（b）会話と地の文を混ぜた物語本文，（c）後付け部分の順に構成される．例えば，『傾城買二筋道』の場合，（b）物語本文は「○夏の床」および「○冬の床」の二章に分かれ，前章に，初めは熱くやがて冷めて離れる男女の関係を，後章に，初めは冷たくやがて熱く結ばれる男女の関係を，対比させるように描いている（図 2）．

作品によって多少の違いはあるものの，洒落本の基本構成は，登場人物同士の会話―とりわけ話者表示とその発話―が中心であり，それらの行間に地の文が配置される．

以下では，このような構造をもつ洒落本のテキストを精緻にマークアップしていく．

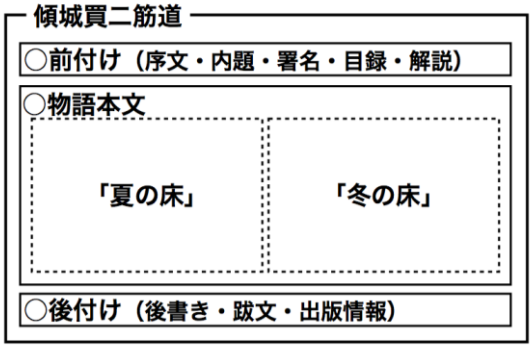


図 2 『傾城買二筋道』の基本構成
Figure 2 The basic structure of "*Keisei-kai futasuji-no-michi*"

4．文書全体の構成に関わる構造化

上述のように洒落本は，テキストの構成として（a）前付け部分，（b）物語本文，（c）後付け部分をもつ．

この構成は，一般的な欧文の写本と一致するため，TEI P5 準拠の要素を用いるならば，テキスト全体は<text>，（a）前付け部分は<front>，（b）物語本文は<body>，（c）後付け部分は<back>をそれぞれ対応させることができる．そして，<front>，<body>，<back>の内部に置かれる序文，内題，導入文などの各記事の構成は， （text division）によってそれぞれ規定することができる．さらに 以下の記事は，基本的にタイトル部分と本文の塊であるパラグラフによって構成されるので，それぞれ<head>と<p>（paragraph）を対応させることができる．

以上 7 つの要素を階層構造に留意してまとめた一覧を表 1 に示す．

表 1 <text>から<p>までの要素の一覧
Table 1 List of elements from <text> to <p> level.

要素	説明
<text>	作品（演目）全体
<front>	前付け
<body>	物語本文
<back>	後付け
	序文，内題，章など
<head>	タイトル
<p>	パラグラフ

５．パラグラフ以下の要素の構造化

ここでは，洒落本の<p>（paragraph）以下の構造について述べる．

（ａ）前付け部分<front>および（ｃ）後付け部分<back>に含まれる<p>以下の内容は，目次，自序，他者による導入文などであった．これらのうち文の体裁をとるものは，欧文の電子化と同様に，TEI の<s>（sentence unit）を用いる．目次は箇条書きの体裁をとる場合が多いため，リスト<list>と項目<item>を組み合せて規定する．

次に，洒落本の本体にあたる（ｂ）物語本文<body>に含まれる<p>以下は，おもに（ｂ$_1$）地の文，（ｂ$_2$）会話文，（ｂ$_3$）割書きによって構成される．

（ｂ$_1$）地の文は，物語の会話以外の情景描写と説明部分である．これも<s>によって規定することができる．

（ｂ$_2$）会話文は，話者が語る部分である．基本的に話者は五郎のように囲み文字によって示されるが，前後の文脈から自明の場合は示されないこともある．また，話者の会話文の範囲は，基本的に次の囲み文字が出現するまで続くという規則を徹底している．したがって，ここでは会話文の範囲に話者情報をもたせながら規定し，囲み文字が存在する場合—すなわち，話者が明示される場合—は，その範囲内に囲み文字を置く方針をとる．会話文の範囲は，会話・思考内容の表現のために準備されている<q>（quoted）に，属性@who=”話者”をもたせて規定する．そして囲み文字は名称を参照する際に準備されている<rs>（referencing string）を用いて規定し，会話そのものは<s>を活用する．図 3 に二人の会話部分を示す．これを XML 形式で表現すると図 4 のようになる．

（ｂ$_3$）割書きは，人物の登場・動作・仕草などの説明を表す部分であり，行内に多段組の文として出現する．割書きは，洒落本の同時代の狂言にあるト書きの影響を受けており，機能もほぼ変わらないため，TEI が脚本のト書き用に準備した<stage>を用いて規定できる．ここでは，割書きの内容そのものは，会話部分の会話文<s>そのものと区別するために<l>（verse line）を用いて規定しておく．また，多段組という外形的な情報は，<cb>（column break）を用いて表現する．ただし，作品によっては，割書きの中に話者を表す囲み文字が出現することもあるので，<stage>の範囲内では，（ｂ$_2$）会話文の<q>以下で用いた囲み文字<rs>の出現も許す．図 5 は，会話文にはさまれた割書きの例である．これを XML 形式で表現すると図 6 のようになる．

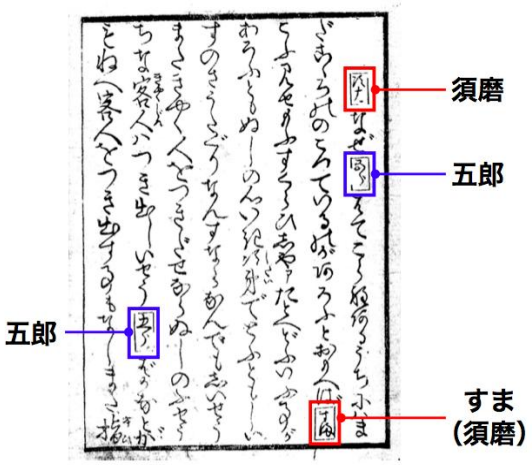


図 3　二人の会話部分の例
Figure 3　Example of a conversation.

```
<q who="#須磨"><rs>須磨</rs>
    <s>なぜ</s>
</q>
<q who="#五郎"><rs>五郎</rs>
    <s>はて こら程あるうちには まだ心ののこつているのがあろふとおもへば</s>
</q>
<q who="#須磨"><rs>すま</rs>
    <s>こふ見せもふすぐらひじやア たとへどふいふ事があろふとも
        ぬしの心いき次第で どふともしいすのさ</s>
    <s>うたぐりなんすなら なんでもしいせう</s>
    <s>また客人をつきだせなら ぬしのふせうちな客人はつき出しいせう</s>
</q>
<q who="#五郎"><rs>五郎</rs>
    <s>ばかな</s>
    <s>科もねへ客人をつき出す事もなし</s>
    <s>また指切…</s>
</q>
```

図 4　二人の会話部分（図 3）の XML 表現
Figure 4　Description of a conversation between two characters (Figure 3) using TEI-based XML.

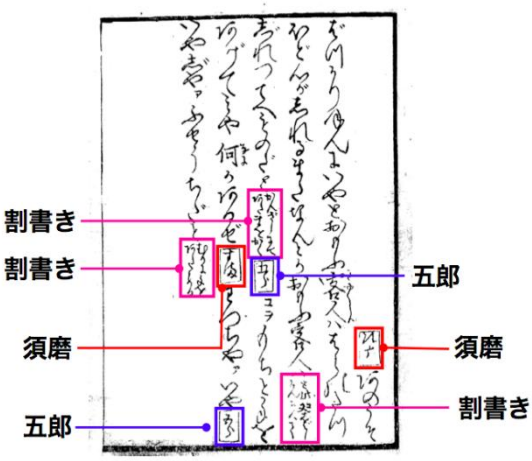


図 5　割書きの例
Figure 5　Example of an interlinear note *warigaki*.

ここまでの文書構造をタグ間の階層関係として整理すると図 7 のようになる．

```
<q who="#須磨"><rs>須磨</rs>
    <s>あの うそばつかり</s>
    <s>ほんに いやとおもふ客人は はらのたつほど心がしれる</s>
    <s>また なんとかおもふ客人は</s>
</q>
<stage><l>と此処しばし<cb />かんがへて</l></stage>
<q who="#須磨">
    <s>じれつてへものだ</s>
    <s>と</s>
</q>
<stage><l>かんざしにて<cb />あたまをかく</l></stage>
<q who="#五郎"><rs>五郎</rs>
    <s>コヲもちと手をあげてみや</s>
    <s>何かあるぜ</s>
</q>
<q who="#須磨"><rs>すま</rs>
    <s>わつちやアいや</s>
</q>
<q who="#五郎"><rs>五郎</rs>
    <s>いやじやア不承知だ と</s>
</q>
<stage><l>むりに手を<cb />あらためる</l></stage>
```

図 6　割書き（図 5）の XML 表現
Figure 6　Description of *warigaki* (Figure 5) encoded with TEI-based XML.

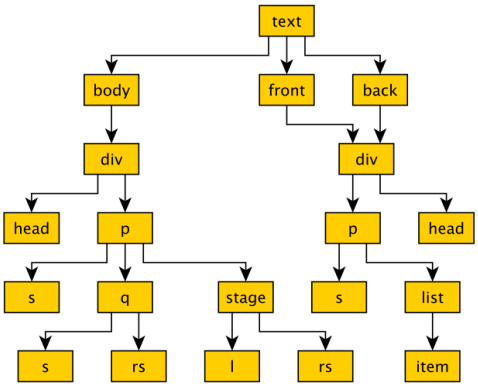


図 7　<text>から<p>以下までの要素の階層構造
Figure 7　Hierarchy within a group of elements from <text> to <p> level.

以上，パラグラフ以下の階層で用いた要素を階層構造に留意してまとめた一覧を表 2 に示す．

表 2　<p>以下の要素の一覧
Table 2　List of elements below <p> level.

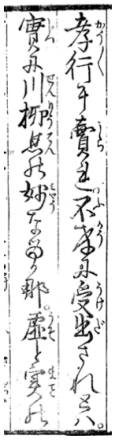
要素	説明
<list>	箇条書き
<q>	会話文の範囲
<stage>	割書き
<figure>	画像，挿絵，捺印
<item>	項目
<rs>	話者
<l>	割書き中の文
<s>	本文

6．センテンス以下の要素の構造化

ここでは，洒落本の<s>（sentence unit）以下の構造について述べる．

（a）前付け部分<front>および（c）後付け部分<back>に含まれる<s>以下の内容は，本文以外に日付と署名が頻出する．これらは厳密に<date>と<name>を用いて規定できる．

次に，（b）物語本文<body>に含まれる<s>以下では，ルビ付き文字が頻出する．ルビ（ruby annotation）とは，任意の文字に対する小さな文字による読み・説明の表記である．ここでは，ルビ付き文字に対して，単語の文法情報を付与するために TEI が準備した<w>（word）に，属性@ana=”ルビ文字”をもたせて規定する．具体的には，図 8 のように表現する．



```
<s>
    <w ana="かう/ゝ">孝行</w>に
    <w ana="うら">売</w>れ
    <w ana="ふかう">不孝</w>に
    <w ana="うけだ">受出</w>されとは<lb />
    <w ana="じつ">実</w>に
    <w ana="せんりうでん">川柳点</w>の
    <w ana="みやう">妙</w>なるかな
</s>
<s>
    <w ana="うそ">虚</w>と
    <w ana="まこと">実</w>の<lb />
    …
</s>
```

図 8　ルビ付き文字の XML 表現
Figure 8　Example of a ruby annotation encoded with TEI-based XML.

また，JISX0213 外の文字や絵文字については，外字<g>（gaiji）を用いて規定する．

以上，センテンス以下の階層で用いた要素を階層構造に留意してまとめた一覧を表 3 に示す．

表 3　<s>以下の要素の一覧
Table 3　List of elements below <s> level.

要素	説明
<date>	日付
<name>	署名
<w>	ルビ付き文字
<g>	外字

7．位置情報と本文以外の情報

上述のタグに加えて，文の改行位置とページの開始位置には，それぞれ<lb>（line break）と<pb>（page break）を割り当てる．ただし，洒落本の版本のページは欧文と異なり，紙面の表裏によって示される．ここでは，<pb>に属性@n=”紙面の表裏”と@xml:id=”ページの通し番号”をもたせる．

洒落本の版本には，ページをまたぐ挿絵が印刷されたり，署名やタイトルに捺印を伴ったりすることがある．ここでは画像全般について，国語研

のサーバ上から該当する画像ファイルを参照・表示できるようにリンクを設定する．具体的には，TEI が準備した<figure>と<graphic>を用いて表現する．例えば，『傾城買二筋道』では，挿絵は（ａ）前付け部分<front>の中の序文と目次の間に置かれている．一枚の用紙の表裏に渡って印刷されているため，図 9 のように表現する．

```
<front>
    <div type="preface">
        <!-- 序文1 -->
    </div>
    <div>
        <!-- 序文2 -->
    </div>
    <div type="illustration">
        <pb n="序二ウ"/>
        <figure n="04"><graphic url="http://~.jpg" /></figure>
        <pb n="序三オ"/>
        <figure n="05"><graphic url="http://~.jpg" /></figure>
    </div>
    <div type="index">
        <!-- "目次部分" -->
    </div>
</front>
```

図 9　挿絵の XML 表現
Figure 9　Example of encoding an image encoded with <figure> and <graphic>.

また，捺印は，外字<g>に分類することもできるが，ここでは挿絵と同様に画像として扱う．例えば，『傾城買二筋道』では，捺印は，タイトル<head>やパラグラフ<p>の内部に出現するため，<figure>と<graphic>を使用できる範囲を 直下の階層だけでなく，さらにその下の<head>や<p>においても許す．

以上，位置情報と本文以外の情報について用いた要素を階層構造に留意してまとめた一覧を表 4 に示す．

表 4　位置情報と本文以外の情報に関する要素の一覧
Table 4　Elements referring locations, et cetra.

要素	説明
<pb/>	ページ開始位置
<cb/>	割書きの分割位置
<lb/>	改行位置
<graphic/>	画像ファイルへのリンク
CDATA	文字（character data）

８．まとめと今後の課題

本研究では，歴史的な日本語資料のコーパス化を網羅的かつ汎用的に進めていくことを究極的な目標として，TEI P5 準拠のタグセットを考案し，資料の外形と機能の構造化を試みた．ここでは，近世口語資料に多くみられる形式に従う洒落本の版本を事例にマークアップを行い，仕様を確立するための指針を示した．本研究で使用したタグセットの階層関係を，紙面の都合上<s>の上下に分割し，図 10 にまとめた．図 10（左）は<text>から<s>，図 10（右）は<s>から文字 CDATA（character data）までの要素である．

しかし，網羅性の観点では，近世口語資料の電子化だけでは不充分であり，より多くの時代・ジャンルのテキストを精査・検証し，内容を反映させていく必要がある．その上で，今後解決すべき課題について以下に２点記す．

形態論情報の付与

本文<s>や割書き文<l>の内容について，形態論情報（品詞・活用形・読みなど）を付与することにより，言語資源として質の高いコーパスの設計を目指す．

洒落本には，（イ）本来は濁点を付けるべき文字，（ロ）カタカナが混在しているために辞書が形態素として認識できない語，（ハ）あらゆる種類の踊り字（々・ヽ・ゝ・ゞ・ヾ・ㇾ・く・ぐ）が出現するため，形態素解析を正しく行うためにこれらをどのように整形していくのか検討する必要がある．

現在，（イ）（ロ）（ハ）に対して，国語研が独自に考案した<vMark>，<kana>，<odoriji>という要素を用いた本文整形を施しているが，本研究の目的に示したように，今後は，テキストの構造を維持したまま，形態素解析が正しく実施できるように TEI 準拠の要素を用いた表現方法を考案していく．

ルビの拡張

通常縦書きの文書では，ルビは文字の右側に置かれる．本研究では，これを<w>（word）で規定した．しかし，近世口語資料では，文字の左側にも同様にルビを付与することがある．右側ルビには文字の読みや宛て語を，左側ルビには右側ルビ以外の読みや語の意味を記す傾向がある．

とくにルビは，日本の古典から現代の漫画に至るまで幅広く利用されており，日本語資料にとって必要不可欠な表記法である．しかしながら，現状の TEI 準拠のタグセットでは，ルビを適切に構造化することが困難である．現状 CSS，XHTML，HTML5 の技術を用いて次のように表現する[12]：

```
<ruby>
    漢<rp>(</rp><rt>かん</rt><rp>)</rp>
    字<rp>(</rp><rt>じ</rt><rp>) </rp>
</ruby>
```

ここで<ruby>，<rt>，<rp>の要素は，それぞれルビを振るテキストの範囲，ルビの文字列，ルビ表示に対応していない場合に表示する文字列を表している．しかし，この方針では，田舎（いなか），十字街（よつつじ）といった熟字訓や宛て字を XML で表現する際に問題が生じる．今後は，ルビについて汎用性を追究した新たな構造化を考案し，TEI に提言していく必要がある．

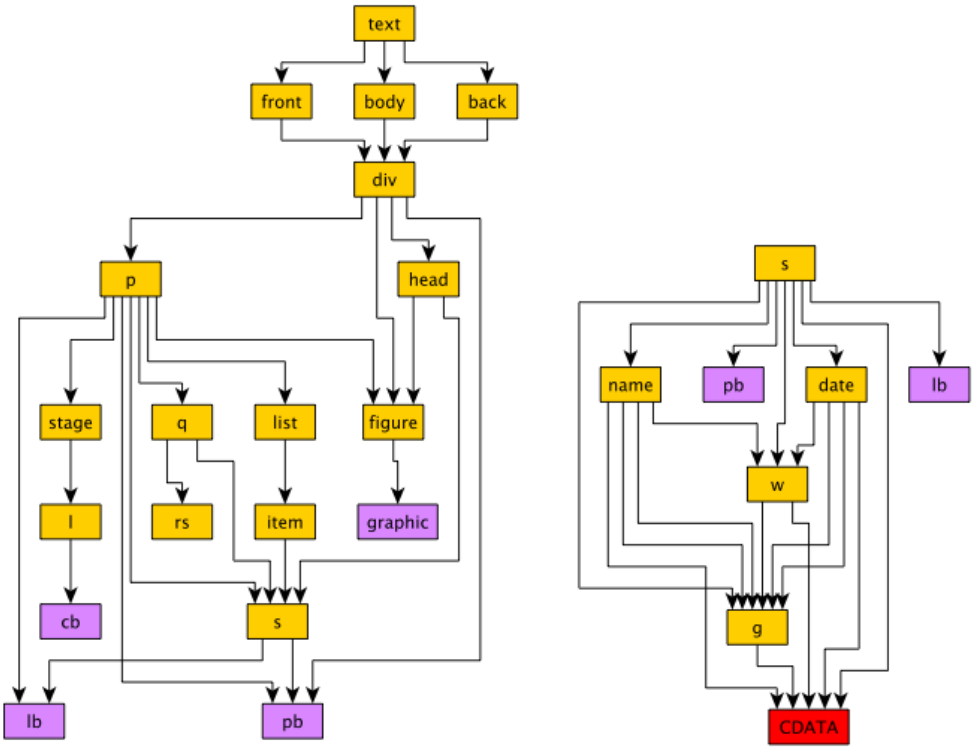


図 10　（左）<text>から<s>までと（右）<s>から CDATA までの要素の階層構造
Figure 10　Hierarchy within a group of elements
(left side) from <text> to <s> level, and (right side) from <s> to CDATA (character data) level.

謝辞

本稿は，国立国語研究所の共同研究プロジェクト「通時コーパスの設計」および「統計と機械学習による日本語史研究」に基づいている．

This paper is based on the collaborative research project “Design of a Diachronic Corpus” and “Study of the History of the Japanese Language Using Statistics and Machine-Learning” carried at the National Institute for Japanese Language and Linguistics.

参考文献

1)　近藤泰弘：日本語通時コーパスの設計について，国立国語研究所プロジェクトレビュー，Vol. 3，No. 2，pp.84-92（2012）．
2)　田中牧郎：説話のパラレルコーパスの設計，3 回日本語学コーパスワークショップ予稿集，pp.259-268.（2013）.
3)　市村太郎，河瀬彰宏，小木曽智信：近世口語テキストの構造化とその課題，情報処理学会研究報告人文科学とコンピュータ研究報告，CH96，pp.1-8.（2012）．
4)　市村太郎，河瀬彰宏，小木曽智信：洒落本コーパスの構造化，第 3 回日本語学コーパスワークショップ予稿集，pp.249-258.（2013）．
5)　Toshinobu Ogiso Mamoru Komachi Yasuharu Den and Yuji Matsumoto : UniDic for Early Middle Japanese, In Proceedings of the 8th International Conference on Language Resource and Evaluation(LREC), pp.911-915(2012).
6)　田中牧郎，小木曽智信：総合雑誌『太陽』の本文の様態と電子化テキスト，日本語科学，Vol. 8，pp141-152（2000）．
7)　近藤明日子，田中牧郎:『明六雑誌コーパス』の仕様，国立国語研究所共同研究報告 12-03 近代語コーパス設計のための文献言語研究成果報告書，pp.118-143（2012）．
8)　前川喜久雄：KOTONOHA『現代日本語書き 言葉均衡コーパス』の開発，日本語の研究，Vol. 4，No.1，pp.82-95（2008）．
9)　山口昌也，高田智和，北村雅則，間淵洋子，大島一，小林正行，西部みちる：『現代日本語書き言葉均衡コーパス』における電子化フォーマット，Ver2.2，LR-CCG-10-04〈http://www.ninjal.ac.jp/corpus_center/bccwj/doc.html#02〉(参照 2013-08-01).
10)　Text Encoding Initiative TEI P5 Guidelines〈http://www.tei-c.org/Guidelines/P5/〉（参照 2013-08-01）．
11)　国文学研究資料館：「大系本文(日本古典文学・ 噺本)データベース」〈http://base3.nijl.ac.jp/〉（参照 2013-08-01）.
12)　Benoit, G：Expanding, Facilitating, and Applying Ruby to Explore User Engagement with Encoded Texts〈http://web.simmons.edu/~benoit/rc/ruby/Ruby-Poster.pdf〉（参照 2013-08-01）．