Vol.2012-CH-96 No.1
2012/10/12

近世口語テキストの構造化とその課題

市村太郎† 河瀬彰宏† 小木曽智信†

本稿では，国立国語研究所「通時コーパス」プロジェクトの一環として検討されている，『洒落本大成』『虎明本狂言』の電子化について，構造化仕様・文書型定義を示し，割書や発話表示等，資料特有の形式の扱いや，それに伴う課題等について論ずる．

Structuring Colloquial Early Modern Japanese Text and Its Issues of Definition

TARO ICHIMURA† AKIHIRO KAWASE†
TOSHINOBU OGISO†

This paper describes the specification and Document Type definition(DTD) for digitized documents of "Sharebon" and "Toraakira's Kyogen", as part of NINJAL Diachronic Corpus Project, and discusses its characteristic properties, styles and issues.

1. はじめに

国立国語研究所では基幹型共同研究プロジェクト「通時コーパスの設計」の一環として[1]，近世口語テキストを形態論情報付き XML 形式で電子化する計画である．本稿では，資料の電子化に際し，いかなる要素を認定し，どのように構造化するのが適切かについて，洒落本と狂言台本を対象に検討し，1 つのモデルを示す．

1.1 洒落本・狂言のコーパス化の意義

本研究で対象とする洒落本や狂言集は，その発話部分に当時の話し言葉が反映されているとされ，日本語史研究上，中・近世期の口語の実態を探る上での重要資料である[2]．

特に洒落本は，大きく分けて江戸版と上方版があり，その口語体の発話部分は，それぞれの地域の言葉を反映する場合もあり，また年代も 18C 後半から 19C 前半までと幅広く，近・現代語への過渡的状況を伺うのに適しており，方言や中央語の形成を知る上で，不可欠な資料である．しかしながら，今のところ主だった索引や上方を含めた全体を見渡すことが可能な大規模なコーパスはなく，利用に際しては，一部の作品を除き，個々の作品をその都度目視して用例を拾い集める他ない．もし一定の数量を持ち，アノテーションされた形態論情報付きコーパスが完成すれば，近世・近代語史研究に画期的な成果をもたらすであろう．

底本はそれぞれ『洒落本大成』[3]『大蔵虎明能狂言集翻刻注解』[4]を用いる．活字本としては現段階では最高水準のものでる．

1.2 本コーパス設計の方針

主な利用者として，第一に言語研究者を想定する．

近世口語テキストの電子化資料としては，先駆的なものとして国文学研究資料館の「大系本文データベース」[5][6]があり，主に紙面にもとづく外形的な面で詳細なマークアップがなされており，貴重な資料となっている．しかし，言語研究の観点からは，さらに言語構造面に重きを置いた構造化が求められる．そのため本研究では，形態論情報の付与を前提とし，言語上の区切りを重視しつつ，テキストの外形と折合をつけるという方針をとる．

具体的には，XML を用い，国語研が作成した『太陽コーパス』の仕様[7][8]や BCCWJ の仕様[9]を継承しながら，TEI P5[10]を参考に必要なタグを選択・追加し，構造化する．

また本稿では省略したが，構造化されたデータには，さらに形態素レベルでのタグを付し，品詞情報や活用形など，詳細な形態論情報を付与する予定である．

1 作品中に会話・地の文，セリフ・ト書き，序・後書き・注釈など，多様な要素を持つ洒落本・狂言を対象に 1 つのモデルを確立しておくことは，「通時コーパス」全体に汎用性をもつ仕様を作る上で，大きな足がかりとなるであろう．

2. 文書の構造と記述法

2.1 洒落本・狂言テキスト全体の構造

洒落本テキストは，会話部分を主とし，その他序文・前置き的な地の文・後書きで構成されることが多い（図 1）．

狂言テキストは，台本本文を中心とし，注釈が付されることがある（図 2）．

†国立国語研究所
National Institute for Japanese Language and Linguistics

①序文（＋目録・人物解説等）
構成要素：タイトル・本文・日付・署名（時折和歌・漢文等）

②状況描写など前置き的な地の文
構成要素：タイトル・本文・記号としての話者表示のない発話・引用

③会話部分（中心部）
構成要素：四角囲みなどの話者・発話・地の文・割書でのト書き

④後書き
構成要素：タイトル・本文・日付・署名・和歌等

図 1　洒落本テキストの構造概略

(著者の注釈)

台本本文
構成要素：話者・発話・ト書き・注釈

(著者の注釈)

図 2　狂言テキストの構造概略

狂言台本の場合，各々独立した演目ではあるが著者は変わらず，上演が前提とされており，序や後書がつかず，注釈が多くなる点で洒落本と異なる．しかし全体としては分量上，また構成上，会話部分が中心となり，またその会話部分も話者表示と発話が中心で，その間にト書きや割書が配置されるという点で，両者の文書構造は類似している．

そのため，個々の要素名は異なるが，階層的に共通する仕様を作成することが可能であると考えられる．

2.2 タグセット

(1) 文書の構造に関する要素（表１）

タグ(要素)	説明	属性
<text>	作品全体	series, title, yomi, year, w_year
<front>	前付	
<body>	主本文	
<back>	後付	
<article>	記事, セクション	type
<p>	本文段落	
<block>	本文外の段落要素	type
<figureblock/>	図表 （入力不可）	type
<s>	文	

表 1　文書の構造に関するタグ

洒落本の図 1①～④のような，テキスト構造を表す大きな構成単位は，1 作品を表す**<text>**と，それを構成するものとして**<front>**:①・**<body>**:②③・**<back>**:④の 3 要素で表す．<text>は属性で作品に関する情報を記述する．

```
<text series="洒落本大成#12" title="阿闍陀鏡" yomi="おらんだかがみ" year="1798" year_w="寛政 10">
```

図 3　<text>形式化例（洒落本『阿闍陀鏡』）

<article>　前付内には自序とともに他人が記した序などが併存することがある．また，前付・後付を除いた中心的本文は，小見出し等を伴う複数の要素から成ることがある．このような階層の要素を表すものとして，<article>を用いる．type 属性で，序・跋・刊記等を記述する．

<p>　本文の塊全体で 1 つ付与する．当然ながら著作当時は改行１字下げで段落を表すという習慣はなく，後世の校注者が付さない限り，視覚上，また内容上いわゆる段落を認定するのは困難である．

<block>　視覚上また構成上明らかに主本文の塊と区別される要素．type 属性で，タイトル・小見出し・著者・日付・表・注釈等を記述する．

言語研究上，ある語について用例を比較する際，文体的に一定の条件のもとで比較することが求められるが，そのためには主本文で得られた用例であるか否かという点は非常に重要であり，明確に区別される必要がある．

なお狂言においては，発話のほかにやや小さい文字でト書きや注釈が付される．これらは視覚的に目立った区別はないが，内容としては所作を表す本文内的なものか，本文外のものかという大きな違いがある．そのため，注釈的な記述を<block>として切り出す（図 4）．

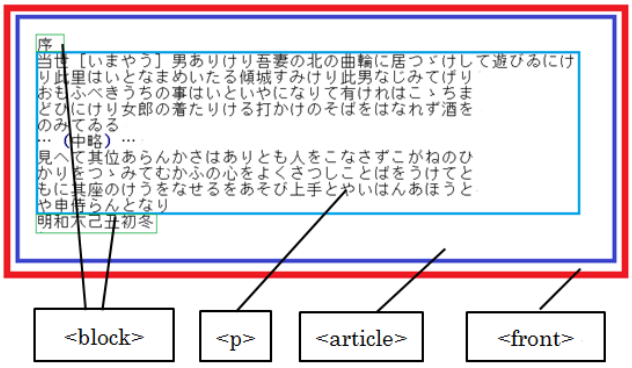
```
<speech><s><speaker>（大黒）</speaker></s><s>「其時大こくすゝみ出て、<lb/><ruby rubyText="一">いち</ruby><ruby rubyText="大">だい</ruby>三<ruby rubyText="千">ぜん</ruby>大千世界の<ruby rubyText="宝">たから</ruby>を是に、入おきたる、袋を汝に<ruby rubyText="取">と</ruby>らせつゝ、<pb n="6"/><lb/>猶もたからを打出す、<ruby rubyText="打">うち</ruby><ruby rubyText="出">で</ruby>の<ruby rubyText="小">こ</ruby><ruby rubyText="槌">づち</ruby>も汝にとらせ」</s></speech>
…（中略）…
<block type=”注釈”><s>「右<span type="傍線">大黒</span>のかたりに、古本にいはく、三面の<span type="傍線">大こく</span>を、<span type="傍線">むどうじ</span>にあんじしたると<lb />いへども、<span type="傍線">むどうじ</span>ハ<span type="傍線">伝教大師</span>よりのちにこんりうの所也、</s><s>其上いまに三面の<span type="傍線">大こ<lb/>く</span>ハ<span type="傍線">えいざん</span>の別所にありと云、是ふしんなるゆへに、<span type="
```

```
傍線">むどうじ</span>に<info originalPage="" />あんじしたる<lb
/>とハかゝず候<lb /> </s></block>
```

図 4　注釈の形式化例（狂言「ゑびす大黒」）

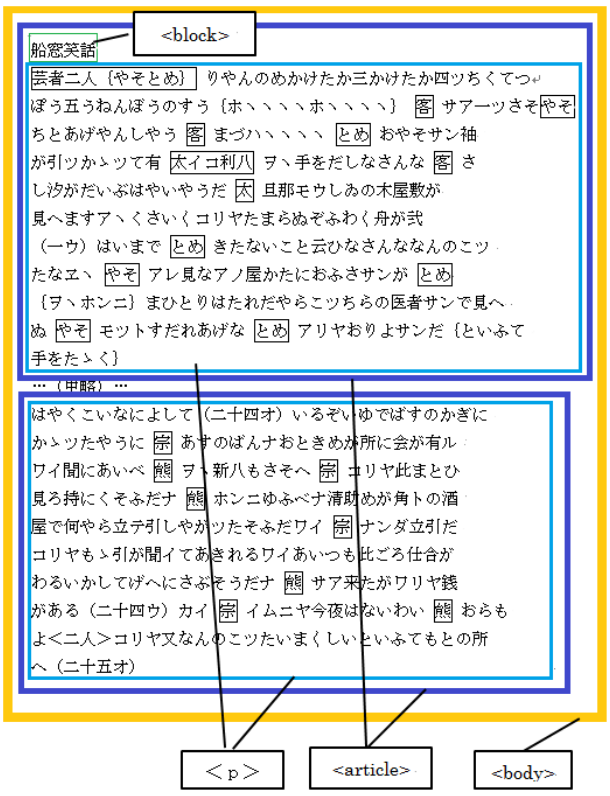
<s>　本コーパスで想定される最も基本的な単位で，形態素解析上も極めて重要であり，すべてのテキストは文に分割される．ただしいわゆる「文」とは完全に同一ではなく，発話や割書の区切りでも切る．これは，たとえば発話の連続を引用の「と」等で受ける場合，「と」がどこまでをマークするのか不明確なことが多く，場合によっては巨大な文が出来上がってしまうためである。

また人物等の表示によっても区切る．これは，後続するその人物の発話が複数文ある場合や，そもそも発話でない場合に，先頭の文のみに人物表示が含まれるのは，論理的に合わないことによる（各形式化例参照）．




```
<front>
<article type="序">
<block type="section"><s><pb n="297"/><cb n="1"/><lb/>序
<lb/></s></block>
<p><s><ruby rubyText="いまやう">当世</ruby>男ありけり
</s><s>吾妻の北の曲輪に居つゞけして遊びゐにけ<lb/>り
</s><s>此里はいとなまめいたる傾城すみけり</s><s>此男な
じみてげり<lb/></s><s>おもふべきうちの事はいといやにな
りて有けれはこゝちま<lb/>どひにけり</s><s>女郎の着たり
ける打かけのそばをはなれず酒を
…（中略）…
<lb/>見へて其位あらんかさはありとも人をこなさずこがね
のひ<lb/>かりをつゝみてむかふの心をよくさつしことばを
うけてと<lb/>もに其座のけうをなせるをあそび上手とやい
はん</s><s>あほうと<lb/>や申侍らん</s></speech><s>とな
り<lb/></s></p>
<block type ="date"><s>明和六己丑初冬<info originalPage="二
ウ"/><lb/><cb n="1"/><lb/></s></block>
</article>
</front>
```

図 5　<front>の形式化例（洒落本『郭中奇譚』）




```
<body>
<article>
<block type="section"><s>船窓笑語<lb/></s></block>
<p><speech><s><speaker><span type="囲み">芸者二人やそとめ
</span></speaker></s><s>りやんのめかけたか三かけたか四ツちく
てつ<lb/>ぽう五うねんぼうのすう</s><s>ホヽヽヽヽホヽヽヽヽ
</s></speech><speech><s><speaker><span type="囲み">客</span>
</speaker></s><s>サアーツさそ</s></speech>
<speech><s><speaker><span type="囲み">やそ</span></speaker>
<lb/></s><s>ちとあげやんしやう</s></speech><speech>
<s><speaker><span type="囲み">客</span></speaker></s><s>まづ
</s><s>ハヽヽヽヽ</s></speech><speech><s><speaker><span type="
囲み">とめ</span></speaker></s><s>おやそサン袖<lb/><cb n="2"/>
<lb/><pb n="298"/>が引ツかゝツて有</s>
</speech><speech><s><speaker><span type="囲み">太イコ利八
</span></speaker></s><s>ヲヽ手をだしなさんな</s>
</speech><speech><s><speaker><span type="囲み">客
</span></speaker></s><s>さ<lb/>し汐がだいぶはやいやうだ
</s></speech><speech><s><speaker><span type="囲み">太
</span></speaker></s><s>旦那モウしゐの木屋敷が<lb/>見へます
</s><s>アヽくさい／＼</s><s>コリヤたまらぬぞ</s><s>ふわ／＼
舟が弐<lb/><info originalPage="一ウ"/>はいまで
</s></speech><speech><s><speaker><span type="囲み">とめ
</span></speaker></s><s>きたないこと云ひなさんな</s><s>なん
のこツ<lb/>たなヱヽ</s></speech>
<speech><s><speaker><span type="囲み">やそ</span></speaker>
</s><s>アレ見な</s><s>アノ屋かたにおふさサンが
```

```
</s></speech><speech><s><speaker><span type="囲み">とめ
</span></speaker><lb/></s><s>ヲヽホンニまひとりはたれだやら
こツちらの医者サンで見へ<lb/>ぬ</s>
</speech><speech><s><speaker><span type="囲み">やそ
</span></speaker></s><s>モツトすだれあげな</s></speech>
<speech><s><speaker><span type="囲み">とめ
</span></speaker></s><s>アリヤおりよサンだ</s></speech>
<warigaki><s>といふて<lb/>手をたゝく</s></warigaki>
…（以下略）…</p></article></body>
```


図 6　<body>の形式化例（洒落本『郭中奇譚』）

(2) 文・語の機能に関する要素（表２）

	タグ（要素）	説明	属性
文を含む	<speech>	発話	type
	<quotation>	引用(非発話)	
	<warigaki>	割書	
	<stage>	狂言等のト書き	
文内部	<speaker>	発話者	
	<delivery>	発話のスタイル	
	<verse>	韻文	

表 2　文・語の機能に関するタグ

文や文連続の機能を表すものに 4 つの要素を認める．

<speech>　1 発話者の 1 回の発話連続を表す．洒落本・狂言共通であり，両テキストとも，話者表示と一体となって現れることが多い．そのため<speaker>は発話と一体として扱う（図 6）．なお時折，他の発話者と同じ外形で示されている人物表示とその人物の発話との間に割書等が入るケースがあるが，後続の発話と対応する場合は，そのような人物表示も<speaker>と認定する（図 7）．


```
<s><speaker><span type="囲み">店</span></speaker></s><s>髪は嶋田に繻子の帯後にしやんとむすび下いまやうのこ<lb/>ぴちや染素ぬいに少し金でいあしらいはでならぬふうにて<lb/>すゝみ出いふやうは</s><speech><s>顔で風切位もなくてつねに<info originalPage="三オ"/>首と<lb/>腰と<vMark>で</vMark>拍子とり下駄にはさへなきを引ずりてあちこちと歩<lb/>行事を自由にするゆへ我／＼をば<ruby rubyText="みせ">店</ruby><ruby rubyText="つき">付</ruby>／＼と腰ぬけのやう<lb/>に下さげにいわしやんすれど此方は生れ付のむくなを表に<lb/>集りゐて見せて思ひ付を取てお出るお客を大切にする故に<lb/><cb n="2"/><lb/><pb n="333"/>見せつきとは云なり</s>…（中略）…<s>かまわぬ事ながら勤する身は同じやうに客さん方の思ひ入<lb/>がはづかしい</s></speech><s>といひ出せば<info originalPage="四オ"/><lb/></s>
```


図 7　<speech><speaker>の構造化例（洒落本『夢中生楽』）

<quotation>　手紙等発話以外の引用要素を表す。

<warigaki>　洒落本における割書は，多くは本文中に細字二行で，会話部分における地の文または注釈として，発話間に現れる．一概に地の文とも，注釈ともいえず，割書にならない地の文と共存するケースもあり，多種多様である．そのため，<warigaki>を認定する（ただし笑い声や間投詞の類が小書きで 2 行にわたっているようなものは割書とは認めない）（図 8，図 9）．

機能としては会話や本文のつなぎ，挿入句のような役割を果たしているため，<speech>と同階層で認定する．

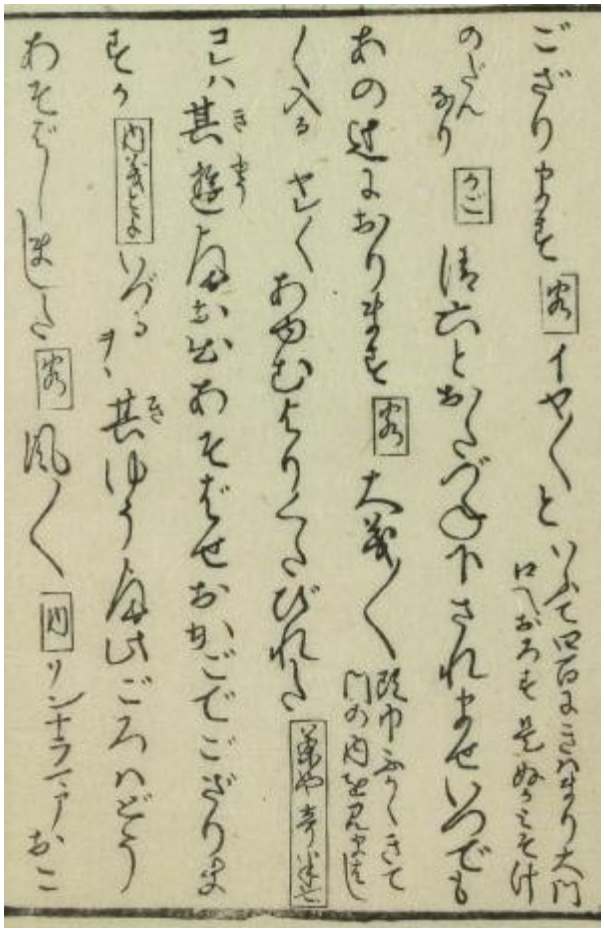


図 8　参考：版本紙面における割書・話者表示等
（早稲田大学図書館蔵『郭中奇譚』[11]）


```
<speech><s><speaker><span type="囲み">客</span></speaker></s>
<s>イヤ／＼</s></speech><warigaki><s>と<lb/>いふて四百にきはまり大門口へおろす</s><s>是ぬかみそ汁のだんなり</s>
</warigaki><speech><s><speaker><span type="囲み">かご</span>
</speaker></s><s>清六とおたづね下されませ<lb/></s><s>いつでもあの辻におります</s></speech><speech><s><speaker><span type="囲み">客</span></speaker></s><s>大義／＼</s></speech>
<warigaki><s>頭巾ふかくきて門の内を見まはし／＼入る
<lb/></s></warigaki><speech><s>ヤレ／＼あゆむよりくたびれた
</s></speech><speech><s><speaker><span type="囲み">茶や亭半七
</span></speaker></s><s>コレハ<ruby rubyText="き">其
```

```
</ruby><ruby rubyText="ゆう">遊</ruby>様お出<lb/>あそばせ
</s><s>おかごでござりますか</s></speech><s><span type="囲み">
内義とよ</span></s><warigaki><s>いづる</s></warigaki>
<speech><s>ヲヽ<ruby rubyText="き">其</ruby>ゆう<lb/>様此ごろ
はどうあそばしました</s>
```

図 9　『洒落本大成』による図 8 対応箇所の形式化例

会話文から続く場合，割書内で文が終止し，次の文が割書内で始まるケースがある（図 9）．

なお，やや例外的な事例として，割書内に，ごく簡単な発話が出現することもある．

ただ，割書に入れられている事時点で補足的な扱いであり，言語研究上も主本文中の発話と同等に扱うことは難しいであろう．また分量的に僅かであり，本仕様では割書中の発話は<speech>としては認定しない．

<stage>　狂言等のト書きを表す．洒落本の地の文や同内容の割書に相当する，本文内的な要素である（図 10）→<block>の項参照．

```
<speech><s><speaker>（男）</speaker></s><s>「是<lb/>ハ<ruby
rubyText="つの">津</ruby><ruby rubyText="くに">国</ruby><span
type="傍線"><ruby rubyText="芦">あし</ruby></span><ruby
rubyText="屋">や</ruby>の里の者にて候、</s>
…(中略）…
<s>是へくわんじやう申、<ruby rubyText="御">み</ruby><ruby
rubyText="注連">しめ</ruby>を<ruby rubyText="引">ひ</ruby>か
ばやとぞん<lb/>ずる</s></speech><stage><s>「しめをひくまねを
して、大こうちの所にいる、</s><s>さがりはにて、<span type="
傍線">ゑびす</span>ハさき、<lb/><span type="傍線">大こく
</span>ハあと、はしがゝりにて<span type="傍線">ゑびす
</span><span type="傍線">大こく</span>うたふなり</s></stage>
```

図 10　<stage>の形式化例（狂言「ゑびす大黒」）

単文や語連続・語の機能を表す要素は以下を認める。

<speaker>　発話の前に付属する，話者の表示である．主に囲みや小書きで表される．

<delivery>　発話の冒頭には，話者だけでなく，発話のスタイルを小書き等で記してある場合がある．

<verse>　韻文は，歌・俳句等について文以下の単位で付す．

```
<speech><s><speaker><span type="小書き">久
</span></speaker></s><s>「そらしりまへんぜ</s><s>何ぞおう
<lb/><cb n="1"/><lb/>たい</s></speech><warigaki><s>トいゝなが
ら</s></warigaki><speech><s><delivery><span type="小書き">歌
</span></delivery></s><s><verse>「水は下ゑとながるゝけれどみづ
にこと<lb/>づけなるものか</verse></s></speech><warigaki><s>お
松里かとかを見合して一寸わらう。</s><s>よふ子の有こと也
</s></warigaki>
```

図 11　<delivery><verse>の形式化例（洒落本『興斗月』）

(3) 語・文字単位で外形等を表す要素（表３）

タグ（要素）	説明	属性
	囲み	type
<lRuby>	左ルビ	rubyText
<ruby>	ルビ	rubyText
<vMark>	濁点付きに変換	
<gaiji>	外字	memo,sub,unicode

表 3　語・文字単位で外形等を表す要素

****　○や□で囲まれる，傍線が付される，小書きなど外形的特徴を持った文字列を表す（図 9 など）．必ずしも話者等と対応するわけではなく，機能は一定ではない．

狂言で，マーカ「○」等で対応する本文がある場合，本文をその位置に入れ，type="挿入本文"とする．

<lRuby>　文字列に沿って小書きされる文字は，右側の振り仮名だけでなく，左側に付されることがある．例えば本文の方言形に対応する語を左側に記すなど，概して注釈的性質があり，語単位で付されることが多く，多くの場合文字単位で付される振り仮名と比較すると，大きな単位である．rubyText 属性内にルビ文字列が記述される．

```
茂佐左ヱ門は猶<lb/>もはらたて。<ruby rubyText="はな">鼻
</ruby><ruby rubyText="ごゑ">声</ruby>にてなまりかけ
</s><speech><s><speaker><span type="囲み">茂佐左ヱ門
</span></speaker></s><s><ruby rubyText="さい">最</ruby><ruby
rubyText="ぜん">前</ruby>から<lRuby rubyText="いつかう">い
<lb/>ちやい</lRuby>／＼<ruby rubyText="わか">別</ruby>り申さ
ない。</s><s>幸次さんうまいかとは<lRuby rubyText="なに">あん
</lRuby>だ<lb /><cb n="1"/><lb/>ア。</s><s><ruby rubyText="くら
ひ">喰</ruby>物だべいか。</s><s><lRuby rubyText="あゝ">うつし
ゆ</lRuby>それよ。</s><s>うらアが<lRuby rubyText="大夫">すべ
た</lRuby><lb/>が</s></speech>
```

図 12　<lRuby>形式化例（洒落本『阿闍陀鏡』）

<ruby>　多くは文字列の右側に付され，文字・文字列の読みを表す振り仮名等を指す．rubyText 属性内にルビ文字列が記述される．狂言の右側漢字傍記も含む．

<vMark>　原拠テキストにはなく，電子化に際して新たに濁点を付与した箇所（図 7　5 行目等）に付す．

近世期以前の資料では，発音上濁点付きの仮名の音で読まれることが期待される仮名に，必ずしも濁点が付与されているとは限らない．本コーパスは形態素解析辞書 UniDic を基に形態素解析されるのを前提としており，表

記の負担を軽減するため，データベースに取り込む前の段階で，濁点つきの仮名に変換し，本タグを付与する．

<gaiji> 本コーパスの文字集合は JISX0213 であるが，稀にこれに含まれない文字が使用されることがあり，文字単位で付す．絵文字等も含む．memo 属性で外形等を記述．unicode 番号があるばあいは unicode 属性で番号を記述．また，他の入力可能文字で代用可能かを判断し， 属性で sub="1"（可），sub="0"（不可）と記述する．

(4) 位置情報（表４）

タグ(要素)	説明	属性
<pb/>	ページ開始	n
<cb/>	段開始	n
<lb/>	行開始	

表 4 底本テキストの位置情報を表すタグ

電子化の直接の対象となる底本での，位置情報を指す．

<pb/> ページの開始位置に挿入される，空要素．n 属性によってページ番号が記述される．

<cb/> 段組みの格段開始位置に挿入される，空要素．n 属性によって底本のページ内の何段目かが記述される．

<lb/> 紙面上の各行の開始位置に挿入される，空要素．

(5) その他の要素（表５）

タグ(要素)	説明	属性
<info/>	本文外情報	originalPage, text

表 5 その他の要素を表すタグ

(1)～(4)ではカバーしきれない，本文外の情報を空要素 **<info/>**で表す．例えば『洒落本大成』においては，翻刻対象とした原拠本の丁付等の位置情報が「(ニオ)」(＝二丁オモテ）と本文内に表示されており，このような原典の視覚的位置情報を originalPage 属性で記述する．

また，狂言テキストにおいてはママ注や注などの傍記が本文脇に付されることがある．このような本文外の傍記等については text 属性で記述する．

2.3 文書型定義（DTD）

2.2 で提示したタグセットの文書型定義を示す．

```
<!-- A -->
<!ELEMENT text (front|body|back)*>
<!ATTLIST text series CDATA #REQUIRED>
<!ATTLIST text title CDATA #REQUIRED>
<!ATTLIST text yomi CDATA #IMPLIED>
<!ATTLIST text year CDATA #IMPLIED>
<!ATTLIST text year_w CDATA #IMPLIED>
<!ELEMENT front (article)*>
<!ELEMENT body (article)+>
<!ELEMENT back (article)*>
<!ELEMENT article (p|block|figureBlock)*>
<!ATTLIST article type CDATA #IMPLIED>
<!-- P -->
<!ELEMENT p (s|speech|quotation|warigaki|stage)*>
<!ELEMENT block (s|speech|quotation|warigaki|stage)*> <!ATTLIST block type CDATA #IMPLIED><!ELEMENT figureBlock EMPTY>
<!ATTLIST figureBlock type CDATA #IMPLIED>
<!-- Q -->
<!ELEMENT speech (s)*>
<!ATTLIST speech type CDATA #IMPLIED>
<!ELEMENT quotation (s)*>
<!ELEMENT warigaki (s)*>
 <!ELEMENT stage (s)*>
<!-- S -->
<!--<!ENTITY % characterElements "gaiji|vMark|pb|cb|lb|info"> -->
<!--ELEMENT s (%inlineElements;)*-->
<!ELEMENT s
(#PCDATA|verse|speaker|delivery|span|lRuby|ruby|gaiji|vMark|pb|cb|lb|info)*>
<!ELEMENT verse (span|ruby|gaiji|vMark|pb|cb|lb|info)*>
<!ELEMENT speaker (span|ruby|gaiji|vMark|pb|cb|lb|info)*>
<!ELEMENT delivery (span|ruby|gaiji|vMark|pb|cb|lb|info)*>
<!-- SPAN -->
<!ELEMENT span (#PCDATA|ruby|gaiji|vMark|pb|cb|lb|info)*>
<!ATTLIST span type CDATA #IMPLIED>
<!ELEMENT lRuby (#PCDATA|ruby|gaiji|vMark|pb|cb|lb|info)*>
<!ATTLIST lRuby rubyText CDATA #IMPLIED>
<!-- LUW,SUW -->
<!-- morph
<!ELEMENT LUW (#PCDATA|%inlineElements;)*>
<!ELEMENT SUW (#PCDATA|%inlineElements;)*>
 -->
<!ELEMENT ruby (#PCDATA|gaiji|vMark|pb|cb|lb|info)*>
<!ATTLIST ruby rubyText CDATA #REQUIRED>
<!-- CHAR -->
<!ELEMENT gaiji (#PCDATA)>
<!ATTLIST gaiji memo CDATA #IMPLIED>
<!ATTLIST gaiji sub CDATA #IMPLIED>
<!ATTLIST gaiji unicode CDATA #IMPLIED>
<!ELEMENT vMark (#PCDATA)>
<!-- EMPTY -->
<!ELEMENT pb EMPTY>
<!ATTLIST pb n CDATA #REQUIRED>
```

```
<!ELEMENT cb EMPTY>
<!ATTLIST cb n CDATA #REQUIRED>
<!ELEMENT lb EMPTY>
<!ELEMENT info EMPTY>
<!ATTLIST info originalPage CDATA #IMPLIED>
<!ATTLIST info text CDATA #IMPLIED>
```

図 13　近世口語コーパスの DTD

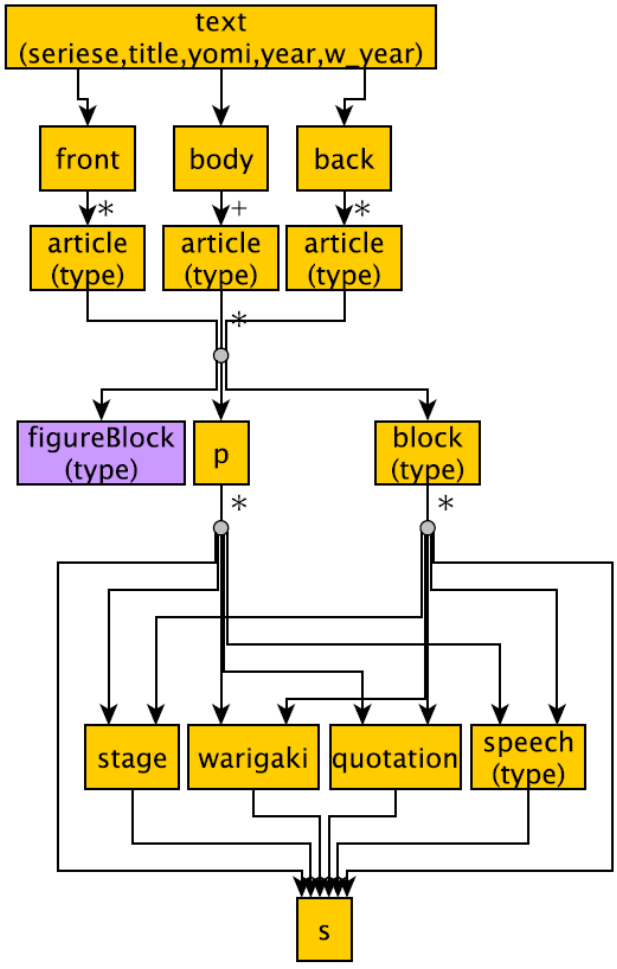


図 14　文以上の階層構造

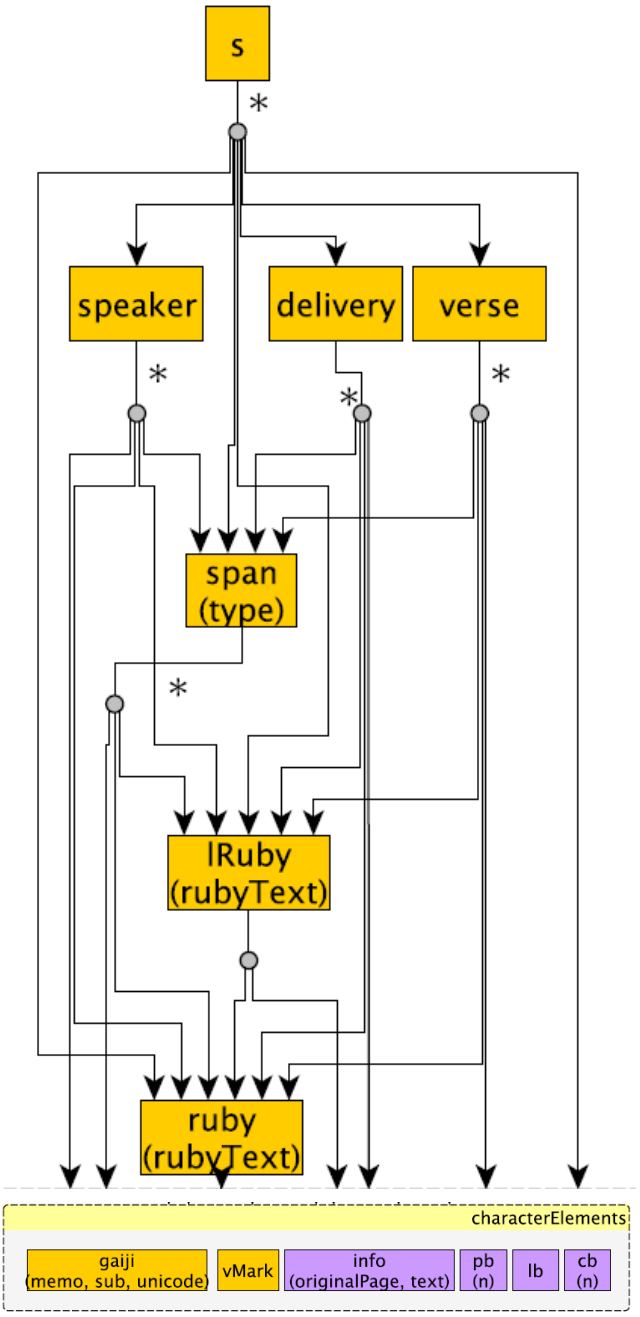


図 15　文以下の階層構造

3. 問題点と今後の課題

<p>の付与範囲

本稿では，<p>の付与方針として，<block>と同レベルのものとして，本文の塊全体で１つ付与することとした．つまり，タイトルや注釈を除いた主本文の大きなまとまりを示すことになる．

しかしながら，現代の改行一字下げによる「段落」は直感的にはより狭い範囲であるし，TEI[10]でももう少し小さな単位が想定されているようである．

例えば，洒落本や狂言において連続する会話と，「と」による引用マーカなど，それに付随する割書・ト書きの集合を段落と認定することも可能であろう．

外形と言語的機能の齟齬

本稿では割書の前後で文を区切ることとしたが，そうなると当然「という」などという半端な文が出来上がってしまうことにもなる．

しかしながら，テキストの構造からみれば〔会話―割書―会話―割書…〕，〔会話―ト書き―会話―ト書き…〕というのは直感的に定式化した流れである．現状ではテキストの階層性と線条性を二重にカバーするのは難しい．

外形と機能に齟齬が生じる場合は少なくなく，その都度妥協点を探ることになる．

解釈の問題

根本的な問題であるが，『洒落本大成』は注釈や句読点等を付された校訂本文ではなく，特に文区切りを与える際には，高度な文解釈が求められる．また近世期は活用語の終止形と連体形が統一される時期であるため，しばしば困難が伴う．

おわりに

以上のように残された課題は多いが，本稿では近世口語テキスト構造化の１つのモデルを示した．

歴史的言語資料としてコーパスを構築するにあたっては，外形と機能とのバランスをとることが極めて重要である．さらにその上で「何を拾いたいか，どこまで期待されているか」に沿う必要もある．

外形と機能を，「余計なことをしない」レベルで研究上のニーズに沿い，バランスよく構造化することが求められる．

参考文献

1) 近藤泰弘:日本語通時コーパスの設計,NINJAL「通時コーパス」プロジェクト・Oxford VSARPJ プロジェクト合同シンポジウム 通時コーパスと日本語史研究予稿集,pp.1-10(2012).
2) 飛田良文（編）,日本語学研究事典,明治書院(2007).
3) 洒落本大成編集委員会,洒落本大成,中央公論社,(1978-88).
4) 大塚光信,大蔵虎明能狂言 翻刻注解 上下,清文堂出版(2006).
5) 大系本文（日本古典文学・噺本）データベース, http://base3.nijl.ac.jp/
6) 安永尚志:国文学研究とコンピュータ,勉誠社(1998).
7) 田中牧郎:言語資料としての雑誌『太陽』の考察と『太陽コーパス』の設計,雑誌『太陽』による確立期現代語の研究 『太陽コーパス』研究論文集,国立国語研究所報告 122,pp.1-48(2005)
8) 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,pp.141-152(2000).
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 ver.2.2,特定領域研究「日本語コーパス」平成 22 年度研究成果報告,文部科学省 科学研究費 特定領域研究 「日本語コーパス」データ班(2011).
10) Text Encoding Initiative／TEI ガイドライン P5 日本語版, http://docsci.infon.org/stack/P5JA/index-toc.html
11) 早稲田大学図書館古典籍総合データベース, http://www.wul.waseda.ac.jp/kotenseki/html/he13/he13_01963_0006/index.html