洒落本コーパスの構造化　―仕様と事例の検討―

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

Structuring the Corpus of *Share-bon*

Taro Ichimura　(National Institute for Japanese Language and Linguistics)
Akihiro Kawase　(National Institute for Japanese Language and Linguistics)
Toshinobu Ogiso　(National Institute for Japanese Language and Linguistics)

１．はじめに

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』の XML 形式での電子化について、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検討し、モデルを示す。

市村・河瀬・小木曽(2012)では、洒落本に狂言を加え「近世口語テキスト」全体での基礎的な構造化仕様を検討した。本発表では、さらに実際の作業を経たうえでの再検討を加え、洒落本の文書構造や語・文字についてどのように処理し、その結果いかなるデータができるかを、いくつかの作品を例にとり、提示する。

２．洒落本のコーパス化の意義

洒落本は、登場人物の会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、近世後期の口語の実態を探る上での重要資料である。

大きく分けて江戸版と上方版があり、その口語体の会話部分はそれぞれの地域の言葉を反映する場合も多い。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いる。方言や中央語の形成を知る上でも、不可欠な資料である。

洒落本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」がある。上方の洒落本については「忍頂寺文庫洒落本データベース」に大阪大学忍頂寺文庫所蔵の洒落本類のデータがあり、これも貴重な資料である。いずれも、主に紙面にもとづく外形的な面でマークアップがなされており、また「忍頂寺文庫洒落本データベース」では、漢字を仮名に開いた「解釈データ」もあり、有用である。

しかしながら、現在のところ、品詞分解された索引や、形態論情報付きの大規模なコーパスはなく、また江戸・上方、宝暦期から文化・文政期まで広く見渡せるものはない。『洒落本大成』には幅広い作品が収められているが、利用に際しては、他の電子化データと重複する一部の作品を除き、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ことが期待できる。

３．本コーパスの設計方針

底本には『洒落本大成』を用いる。洒落本を対象とした大規模な叢書であり、多くの作品が収録されている。

コーパスの主な利用者としては、言語研究者を想定する。現在、上記のように、「大系本文データベース」のような、紙面にもとづいて外形的なマークアップがなされたテキストデータが存在するが、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。たとえば、単純なテキストデータで検索する場合、近世のような仮名遣いや

† tichimura@ninjal.ac.jp

漢字表記が多岐にわたる資料では、文字列検索でいちいちすべての表記を想定したうえで検索しなければならず、大きな足枷となる。その点「忍頂寺文庫洒落本データベース」は読みのデータを付記することで特に漢字表記の面での解決がはかられており、また、簡易なマークアップがされているため、ある程度の要素の抽出が可能である。

これらをふまえさらに、たとえば「ここからここまでがタイトルである」「ここからここまでが〇〇の発話である」などといった、文書構造情報を認定し、一定の構造で発話等がマークアップされることによって、得られた用例について、どのような要素の用例なのかを判断することができるのならば、さらに言語データとして有用なものになるだろう。用例が文書構造中のどの箇所で得られたものなのか（たとえば発話なのか地の文なのか）というのは、近世語研究にかぎらず、極めて重要である。さらにそこに形態論情報が付記され、どのような語のどの活用形か、などがあらかじめ特定されていれば、極めて質の高いデータが、電子データ検索によって瞬時に得られる。

本研究では、このような、いかなる要素の、いかなる性質を持つ語の表記体であるという情報が付された用例を一覧表として短時間で取り出すことが可能なコーパスを目指す。

そのため、記述には XML を用い、国語研が作成した『太陽コーパス』の仕様や BCCWJ の仕様、『明六雑誌コーパス』の仕様を継承しながら、TEI P5 を参考に必要なタグを選択・追加し、構造化する。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、形態論情報を付与する。

４．洒落本テキストの構造とタグセット

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)。

①序文・内題・署名等（＋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：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（時折小見出しを伴う複数セクション）
④後書き・尾題・出版情報等 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

作品によってばらつきはあるが、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、その間に地の文や割書が配置される。

洒落本のコーパス化にあたっては、このような文書構造の各要素について、できるだけ均質に、過不足なくマークアップする枠組みを設定することが求められる。

４．１文書の階層構造に関する要素

表 1　文書の構造に関するタグ（太線は階層上の大きな切れ目）

タグ（要素）	説明	属性
<text>	作品（演目）全体、作品のシリーズ・タイトル等を開始タグ内に記述	@textID（必須） @series シリーズ名（必須） @title　作品タイトル（必須） @yomi 作品名の読み（任意） @year　　西暦成立年（必須） @year_w　和暦成立年（必須）
<front>	前付け	
<body>	主本文	
<back>	後付	
<article>	記事	@type（任意）
<titleBlock>	<article>レベルでのタイトル等の記述	
<p>	タイトルや注釈等を除く本文の塊	
<block>	<p>で記述された本文とは区別されるタイトル・注釈等のブロック要素	@type（必須）
<s>	文	
<SUW>	短単位（語）	（多岐にわたるため省略）

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つのテキスト全体を表す<text>と、それを構成する<front>①・<body>②③・<back>④の 3 要素から成る。作品に関する情報は、属性値として<text>内に記述する。

さらに、これらの内部は基本的には<article>に分割され、さらにその内部は<p>か<block>に、そしてその内部は<s>に分割される。さらにその文が、形態論情報を記述した短単位<SUW>に分割される。本コーパスでは、<p>は非常に大きな本文の塊に付されるだけであるので、テキストを分割する単位としては<s>と<SUW>が軸となる。

article 要素　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。また、前付や後付内には、自序とともに他人が記した文章や出版情報などが併存することがある。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・刊記の別等を記述する。

p 要素　<article>内の本文の塊全体で１つ付与する。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。本研究では「主たる本文かそれ以外か」に重点をおいている。

block 要素　視覚上または構成上、明らかに主本文の塊と区別される要素を表す。@type 属性で、タイトル・内題・尾題・小見出し・著者・日付・表・注釈等の別を記述する。

titleBlock 要素　テキストのタイトル（外題）のほか、序文等の後に再度作品のタイトル（内題）等や尾題等が示される場合がある。これらを厳密な階層構造の中に組み込むことは難しい。そのため、<article>と同階層でマークアップし、並列的に扱う。

s 要素　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。なお、<s>が<s>を含むような階層性は認めない。

SUW 要素　短単位（おおよそ語に相当）を表す。すべての文は短単位に分割される。本研究での基本的な単位である。語彙素・語形・書字形・活用型・活用形・発音形等語に関する多くの情報が、属性で記述される。開発中の「近世口語 UniDic」による解析結果を人手で修正して付与する。

キー	語彙素	発音形出現形	品詞	活用型	活用形
おゆき	オユキ	オユキ	名詞-固有名詞-人名--一般		
さん	さん	サン	接尾辞-名詞的-一般		
はやふ	早い	ハヨー	形容詞-一般	形容詞	連用形-ウ音便
お	御	オ	接頭辞		
いで	出でる	イデ	動詞-一般	下一段-ダ行	連用形-一般
わたし	私	ワタシ	代名詞		
も	も	モ	助詞-係助詞		
これ	此れ	コレ	代名詞		
から	から	カラ	助詞-格助詞		
かみゆひ	髪結い	カミユイ	名詞-普通名詞-一般		
さん	さん	サン	接尾辞-名詞的-一般		
に	に	ニ	助詞-格助詞		
かみ	髪	カミ	名詞-普通名詞-一般		
を	を	オ	助詞-格助詞		
ゆふ	結う	ユー	動詞-一般	五段-ワア行	連用形-ウ音便
て	て	テ	助詞-接続助詞		
もろ	貰う	モロ	動詞-非自立可能	五段-ワア行	連用形-ウ音便
て	て	テ	助詞-接続助詞		
こんや	今夜	コンヤ	名詞-普通名詞-副詞可能		
から	から	カラ	助詞-格助詞		
おしろい	白粉	オシロイ	名詞-普通名詞-一般		
も	も	モ	助詞-係助詞		
し	為る	シ	動詞-非自立可能	サ行変格	連用形-一般
て	て	テ	助詞-接続助詞		
べに	紅	ベニ	名詞-普通名詞-一般		
も	も	モ	助詞-係助詞		
つけ	付ける	ツケ	動詞-非自立可能	下一段-カ行	連用形-一般
て	て	テ	助詞-接続助詞		
おき	置く	オキ	動詞-非自立可能	五段-カ行	連用形-一般
ましよ	ます	マショ	助動詞	助動詞-マス	意志推量形
。	。		補助記号-句点		
ありや	彼れ	アリャ	代名詞		
いろ	色	イロ	名詞-普通名詞-一般		
め	奴	メ	接尾辞-名詞的-一般		
が	が	ガ	助詞-格助詞		
いに	往ぬ	イニ	動詞-一般	五段-ナ行	連用形-一般
おつ	居る	オッ	動詞-非自立可能	五段-ラ行	連用形-促音便
た	た	タ	助動詞	助動詞-タ	終止形-一般
これ	此れ	コレ	代名詞		
のふ	ノウ	ノー	感動詞-一般		
ををい	おい	オーイ	感動詞-一般		
/ \	/ \		補助記号-一般		

図 2　短単位解析済みデータの例（一部項目省略・8 巻『風流裸人形』p.277 上段 2 行～）

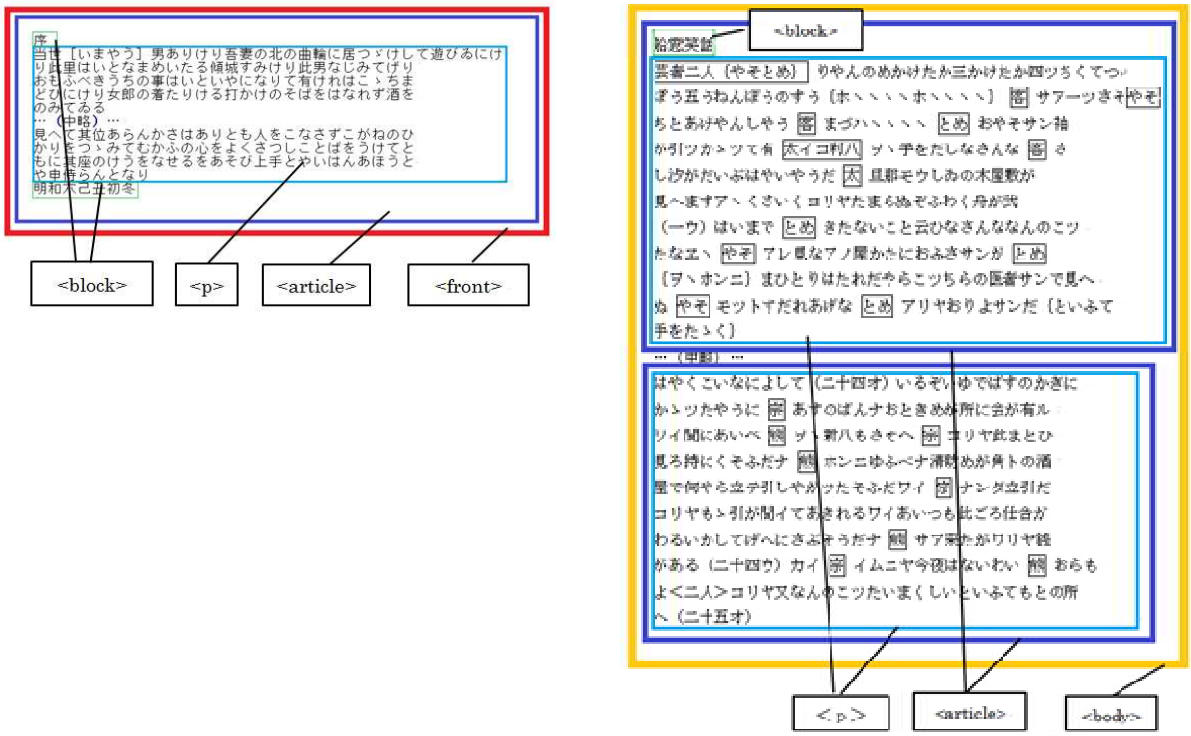

図 3　作品冒頭～本文のマークアップ（4 巻『郭中奇譚』pp.297-308）

４．２文・語の機能に関する要素

表 2　文・語の機能に関するタグ

タグ（要素）	説明	属性	
<speech>	会話	@source（任意） @type（任意）	
<quotation>	単純な発話以外の引用要素	@type（任意） @source（任意）	
<warigaki>	割書き		↑文以上
<speaker>	話者		↓文未満
<delivery>	発話等のスタイルの表示		
<verse>	韻文		

speech 要素　1 回的な会話文の連続を表す。<speaker>を内部に認定し、一体として扱う。会話文内に話者が示されていない場合には@source 属性で話者を可能な限り記述する。なお、割書内にもごく簡単な会話文が出現することもあるが、割書中では認定しない。

quotation 要素　和歌・手紙等、単純な会話文以外の引用要素を表す。@type でどのような種の引用かを、@source で出典を記述する。

warigaki 要素　多くは細字二行で、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。ただし笑い声や間投詞の類が小書きで 2 行に渡るものは割書とは認めない。

speaker 要素　会話文に付属する、話者の表示である。主に囲みや小書きで表される。

delivery 要素　会話文の冒頭には、話者だけでなく、「歌」などと、スタイルを小書き等で記してある場合がある。その場合に本要素を付与する。

verse 要素　和歌・俳句・歌等明らかな韻文ついて、文未満の単位で（主に文毎に）付す。

４．３主に語・文字単位で外形・機能等を表す要素

表 3　語・文字単位で外形等を表す要素

タグ(要素)	説明	属性	
<hi>	文字列（語）に対する装飾	@rend（必須）	↑短単位以上
<lRuby>	左ルビ	@rubyText（必須） @rubyBase（任意）	↓短単位未満
<ruby >	ルビ	@rubyText（必須） @rubyBase（任意）	
<odoriji>	踊り字を開いた文字	@originalText（必須）	
<gap/>	抹消・破損等で判読できない文字の存在（空要素）		
<corr> <corr/>	本文修正	@type（必須） @originalText（任意）	
<unclear>	推読された文字	@originalText（任意） @type（任意）	
<vMark>	濁点・半濁点付仮名に変換した箇所		
<g>	外字（JISX0213 外）絵文字等	@type（必須） @ref（任意）	
<kana>	片仮名を平仮名に変換した箇所		
<kanbun> <kanbun/>	漢文（返読）	@type（任意）返読前 返読後等 @originalText （ 任 意 ） @id（任意）	

hi 要素　○や□で囲まれるなど外形的特徴を持った語以上の文字列を表す。囲みの人物表示は必ずしも話者になるわけではなく、機能は一定ではない。このようなものを外形的にマークアップし、@rend 属性で様態を記述する。「マアマア」等間投詞的なものは除く。

ruby 要素　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を記述し、複数短単位に対して付されている場合は、先頭の短単位のみに認定し、@rubyBase に実際の対象文字列を付す。いわゆる宛漢字等も含む。

lRuby 要素　文字列の左側に付される小書き。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る。右側ルビと共存する場合は、右側ルビよりも比較的対象範囲が大きい。語単位で付されることが多い。

vMark 要素　電子化に際して新たに濁点を付与した箇所に付す（踊り字の箇所は除く）。

４．４位置情報と本文外情報

表４　底本テキストの位置情報や本文外の情報を表すタグ

タグ(要素)	説明	属性
<pb/>	ページ開始（空要素）	@n（必須）
<cb/>	段開始（空要素）	@n（必須）
<lb/>	行開始（空要素）	
<info/>	本文外情報（空要素）	@originalPage（任意）@ text（任意）

```
<text textID="洒落本大成_024_京都_興斗月" series="洒落本大成#24" title="興斗月" yomi="きょとつき" year="1836" year_w="天保 7"><front><article type="序"><p><s><pb n="131"/><cb n="1"/><lb/>年<vMark>ご</vMark>ろ我勝れて河東を好め<vMark>ど</vMark>も価高きゆへうとまれて行こ<lb/>と稀也</s><s>只老留誌の類を見て鬱を散る而已なりし<vMark>が</vMark>或夜東方<lb/>に見馴ぬ光あり</s><s>これなむ興斗つきといふ<info originalPage="一オ"/></s><s>唖の川に輝<lb/>舟あり</s><s>是に乗て蚓ののたくり一冊としはりに至ま<vMark>で</vMark>自作せ<lb/>りと慢して馬鹿の底をた<odoriji originalText="ゝ">た</odoriji>く</s><s>是を名号て興斗つゐて何処ま<lb/><vMark>で</vMark>乗て行<info originalPage="一ウ"/>と云<lb/></s></p><block type="date"><s>天保七年<lb/>申<kana>の</kana>孟夏<lb/></s></block><block type="author"><s>前代未聞<lb/>武木右衛門<lb/>自序<lb/><info originalPage="二オ"/></s></block></article><titleBlock><block type="内題"><s><pb n="132"/><cb n="1"/><lb/>興斗月<lb/></s></block><block type="author"><s>武木右衛門戯作</s></block></titleBlock></front>（以下略）
```

図 4『興斗月』冒頭の形式化例（大成 29 巻 pp.131-132）

```
<body><article><p><s><ruby rubyText="おゝ">大</ruby><ruby rubyText="き">木</ruby><ruby rubyText="ど">戸</ruby>の<ruby rubyText="ちり">塵</ruby>は<ruby rubyText="みづ">水</ruby><ruby rubyText="うり">売</ruby>の<ruby rubyText="しづく">雫</ruby>にしめり<ruby rubyText="てん">天</ruby><ruby rubyText="りう">竜</ruby><ruby rubyText="じ">寺</ruby>の<ruby rubyText="かね">鐘</ruby>は<ruby rubyText="ひぐらし">蜩</ruby>の<ruby rubyText="こへ">声</ruby>にひ<odoriji originalText="ゞ">び</odoriji><lb/>く</s><s><kana><hi rend="囲み">くつわのをと</hi></kana></s><s>ちやんらん／＼</s><speech><s><speaker><hi rend="囲み">馬士二人歌</hi></speaker></s><s><verse>お<odoriji originalText="ゝ">お</odoriji>れへと<info text="上"/>な<kana>あ</kana>引い<lb/>かぬ<kana>あ</kana>う</verse></s><s><verse><kana>そ</kana><kana>れ</kana>そうだにな<kana>あ</kana>引</verse></s></speech><speech><s><speaker><hi rend="囲み"><kana>あ</kana><kana>と</kana><kana>の</kana>馬士</hi></speaker></s><s>かみ<ruby rubyText="むら">村</ruby>の
```

<kana>う</kana><ruby rubyText="ゑ">江</ruby><ruby rubyText="ご">五</ruby>右<lb/><ruby rubyText="ゑ">衛
</ruby><ruby rubyText="む">門</ruby>が<kana>あ</kana>よめ<ruby rubyText="じよう">女</ruby><kana>なあ
</kana><ruby rubyText="うみ">産</ruby><ruby rubyText="づき">月</ruby>だ<kana>あ</kana>といつけがどふだ<kana>あ
</kana></s><s>まだひり<lb/><ruby rubyText="だ">出</ruby>さねへかな<kana>あ</kana></s></speech>（中略）
<speech><s><speaker><hi rend="囲み">金</hi></speaker></s><s><kana>あ</kana><kana>い</kana>さあ。おさらば／＼</s>
</speech><s>○<ruby rubyText="なつ">夏</ruby>の<ruby rubyText="よ">夜</ruby><lb/>は。まだ<ruby rubyText="よひ">宵
</ruby>ながら。<ruby rubyText="あけ">明</ruby>ぬるを。<ruby rubyText="し">知</ruby>らせよふとて。<ruby rubyText="
からす">烏</ruby>がか<lb/>あ／＼。<ruby rubyText="かね">鐘</ruby>がごん／＼。<ruby rubyText="つき">春</ruby><ruby
rubyText="ごめ">米</ruby>屋ががつたり／＼<info originalPage="丁付なしオ"/></s> </article></body>
<back><article type="跋"><block type="section"><s><lb/>跋</s></block> <s><lb/><cb n="1"/><pb n="311"/><ruby
rubyText="すい">粋</ruby>とは<ruby rubyText="うめ">梅</ruby><ruby rubyText="ぼし">干</ruby><ruby rubyText="や">野
</ruby><ruby rubyText="ぼ">父</ruby>とは<ruby rubyText="にはとり">鶏</ruby>の名かときくやうな<ruby rubyText="し
ん">新</ruby><ruby rubyText="じゆく">宿</ruby>田舎にあや<lb/>め咲とはしほらしとぞめきの<ruby rubyText="こえ">声
</ruby><ruby rubyText="う">有</ruby><ruby rubyText="てう">頂</ruby><ruby rubyText="てん">天</ruby>にひ<odoriji
originalText="ゞ">び</odoriji>き（中略）
<s>鳴<lb/><ruby rubyText="あゝ">呼</ruby><ruby rubyText="わが">吾</ruby><ruby rubyText="とう">党</ruby>いきちよ
んの君子をしてこれにあそはしめば<ruby rubyText="すなはち">則</ruby>其<lb/><ruby rubyText="しり">尻</ruby>つま
らざるにちか<odoriji originalText="ゝ">か</odoriji>らん<ruby rubyText="ずい">随</ruby><ruby rubyText="いき">行
</ruby><ruby rubyText="さん">散</ruby><ruby rubyText="じん">人</ruby><ruby rubyText="ずい">随</ruby><ruby
rubyText="がへり">帰</ruby>の<ruby rubyText="まくら">枕</ruby><ruby rubyText="もと">上</ruby>に<ruby rubyText="ば
つ">跋</ruby>す<lb/></s> <block type="date"><s>安永乙未秋<lb/></s></block><block type="publisher"><s>新甲館蔵書
<lb/><info originalPage="丁付なしオ"/></s></block></article></back>

図 5『甲駅新話』本文・後付の例（大成 6 巻 pp.295-311）

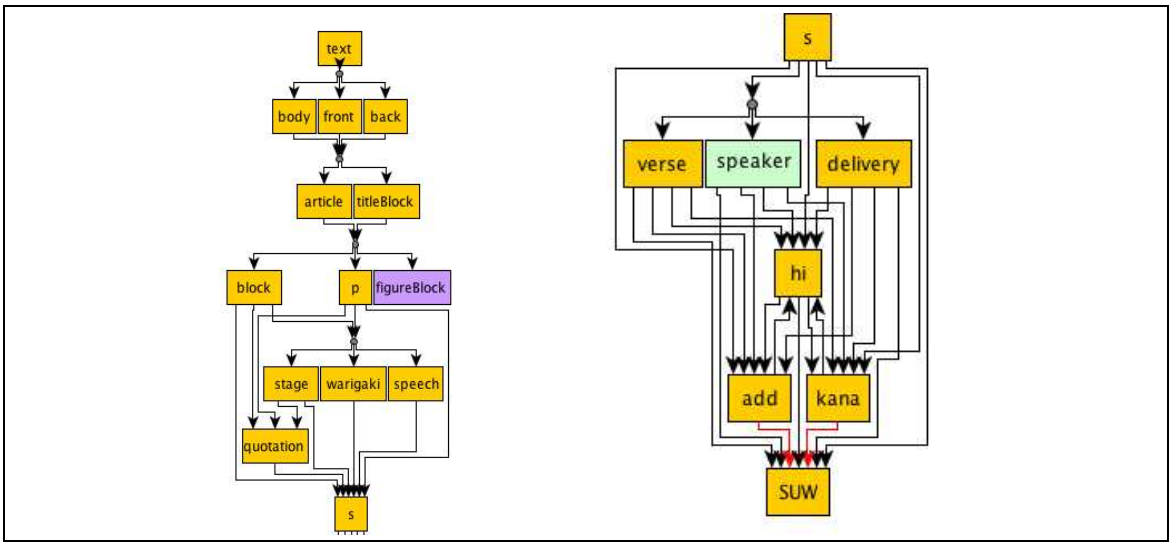

図 6「近世口語コーパス」の文書型定義図（SUW 要素まで）

５．本研究における課題

５．１　割書きの扱いをめぐって

割書や引用の前後で文を区切ると、「という」など半端な文が生じるという問題がある。一方で、テキストの構造からみれば〔会話―割書―会話―割書…〕という直感的に定式化した流れがある。このような複数の構造を厳密にカバーするのは困難である。

しかし、言語学的な利用を考慮すると、「割書・引用で文を区切る」ことは、多くの場合「発話と地の文を区別する」ことと合致する。また割書を解体するにしても、分類や単位の切れ目の認定は困難であり、洒落本においては、最大公約数的に「主本文に対する付属的な何か」を表す「割書」であるほうが利用の便を考慮しても現実的ではないかと考える。

５．２文認定と解釈の問題

根本的な問題であるが、『洒落本大成』は注釈や句読点等が付された校訂本文ではなく、形態論情報はもちろん、文区切りを与える際には、高度な文解釈が求められる。活用語の終止形と連体形が統一される時期でもあり、文末の認定にはしばしば困難が伴う。話者が長々と話すもの、道行などのように語調がかかわるものは特に難解である。文脈・活用形・ソ系指示詞などが手がかりだが、それでも不明確な箇所については強いて文区切りを付すことはしない方針である。

例を挙げる。下のように、名詞が連続する場合、並列的に述べられているのか、文が切れるのか、厳密に判断するのは難しい。

（例）店もなく揚屋もなく商ひ場といふてはうゑもなき雲天のざしきいつこさためぬ枕の数（大成 7 巻『無論里問答』p.50 下段）

また、引用周りの扱いも問題となる。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、また一口に文と言っても、場合によっては幾重にも階層ができ、巨大な文が出来上がってしまい、著しく均質性を損なうことがある。そのため、原則直接的な引用や割書き、話者表示の前後では文を区切る。

発話内の引用については、間接引用なのか直接引用なのかが不明確な場合が多い。ただし、手紙を読み上げる箇所等があり、このような明確に直接引用とわかるものについては、文区切りを付す方針である。

このように可能な限り客観的な根拠を探るのが原則だが、細部に決定的な規則を設けるのは困難である。個別に判断し、場合によっては保留せざるを得ないのが現状である。

５．３　修辞・言葉遊びへの対応

掛詞や洒落のような言葉遊びや、文や語と区切りとは関係ない七五調などが見られる。これらは近世期に限らず歴史的作品では重要な修辞技法の一種であるが、階層構造から逸脱したパラレルなものである。これに対しては、索引等では二重に採取する方針がとられているものもあるが、コーパスの開発に利用している現状のシステムでは形態論情報の二重付与に対応していない。今後、システム拡張を含め、対応を検討していく必要がある。

（A）我搗栗といわゐそめ老もわかいも<u>よろ昆布</u>（大成 7 巻『三幅対』p.352 下段 16 行）

（B）＜知暁＞ごさまのかぎ迄預けしは＜青蜉＞<u>づけし</u>は物を思はざりけり＜几石＞<u>ざりけり</u>のちじよくをすゝぐ（大成 2 巻『穿当珍話』p.207 上段 3～5 行）

６．おわりに

文や引用の認定・解釈は、歴史的な資料をコーパス化する際の大きな課題である。また、５．３のような修辞・言葉遊びの類は、今後和歌集や歌舞伎・浄瑠璃を積極的に扱うことを考慮すると、大きな課題である。

歴史的資料を対象にコーパスを構築するにあたっては、外形と機能、言語の線条性と版面がもつ構造のバランスをとり、適切にラベルを与えていくことが重要である。その上で「何を拾いたいか、どこまで期待されているか」という利用者のニーズに沿う必要がある。

1 作品中に会話・地の文・割書き・序・後書き・手紙など、比較的多様な要素を持つ洒落本を対象に 1 つの記述モデルを確立しておくことは、「日本語歴史コーパス」全体に汎用性をもつ仕様を作る上での一つの足掛かりになると考える。

文　献

市村 太郎、河瀬 彰宏、小木曽 智信(2012)『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会報告(CH96)　pp.1-8

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 国立国語研究所

近藤泰弘(2012)「日本語通時コーパスの設計について」『国語研プロジェクトレビュー』3 pp.84-92 国立国語研究所

田中牧郎(2005)「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」『国立国語研究所報 122　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』pp.1-48　博文館新社

田中牧郎、小木曽智信(2000)「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』8　pp.141-152　国立国語研究所

安永尚志(1998)『国文学研究とコンピュータ』勉誠社

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(2011)『特定領域研究「日本語コーパス」平成 22 年度研究成果報告『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』　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

洒落本大成編集委員会(1978-88)『洒落本大成』中央公論社

関連 URL

「大系本文（日本古典文学・噺本）データベース」http://base3.nijl.ac.jp/

「忍頂寺文庫洒落本データベース」http://www.let.osaka-u.ac.jp/~iikura/Ninjoji_Ono/syarebon.html

「Text Encoding Initiative」（ガイドライン P5 日本語版）
http://docsci.infon.org/stack/P5JA/index-toc.html