洒落本コーパス構築の試行

著者	市村 太郎, 村山 実和子
雑誌名	国立国語研究所論集
号	12
ページ	29-45
発行年	2017-01
URL	http://doi.org/10.15084/00000852

国立国語研究所論集 (NINJAL Research Papers) 12: 29–45 (2017)
ISSN: 2186-134X print/2186-1358 online

洒落本コーパス構築の試行

市村太郎[a]　村山実和子[b]

[a]常葉大学／国立国語研究所 共同研究員
[b]国立国語研究所 研究系 言語変化研究領域 非常勤研究員

要旨

筆者らは，現在，国立国語研究所で開発が進められている『日本語歴史コーパス』の一部として，近世洒落本を対象とするコーパスを開発しており，その試作版を『ひまわり版「洒落本コーパス」Ver. 0.5』（2015 年 10 月 28 日公開）として公開した。本コーパス構築にあたっては，他の『日本語歴史コーパス』所収のコーパス同様，文書構造に関する情報や形態論情報を付与するとともに，新たに所蔵版本への画像リンクや，詳細な話者情報を付与する試みを行った。これにより，近世資料の持つ地域差・位相差にも配慮した近世語コーパスのモデルを示すことができた *。

キーワード：日本語歴史コーパス，近世語，形態論情報，江戸語，上方語

1. はじめに

本稿で扱う，『ひまわり版「洒落本コーパス」Ver. 0.5』（以下「洒落本コーパス 0.5」[1]）は，現在国立国語研究所で『日本語歴史コーパス』[2] の一部として構築が計画されている，江戸時代・洒落本編の試作版である。

本コーパスの底本は，洒落本大成編集委員会編『洒落本大成』（1978–88 中央公論社）である。電子化に際して一部テキストを校訂し，そこに様々な情報を付加することで，XML データを構築した（市村・河瀬・小木曽 2012，市村 2014 参照）。今回公開した試作版はその XML データの一部を，構造化された言語資料を閲覧するのに適した全文検索システムである『ひまわり』（山口・田中 2005）上での利用を想定し，適合させたものである。

本稿では，以下第 2 節で本コーパスの概要について述べるとともに，第 3 節において特徴やデータの内容，課題について論じ，第 4 節で利用方法や活用例を提示する。

2.「洒落本コーパス 0.5」の概要

2.1「洒落本コーパス 0.5」の目指すところ

「洒落本コーパス 0.5」は，日本語史研究への貢献を主たる目的としたコーパスであり，『日本

* 本稿は 2015 年 11 月 28 日に国立国語研究所で行われた第 328 回日本近代語研究会にて発表したものを補筆・修正したものである。また，本稿は国立国語研究所機関拠点型基幹研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」（プロジェクトリーダー：小木曽智信）の研究成果である。また，本研究の一部は JSPS 科研費 15K16765 の助成を受けた。

[1] 国立国語研究所コーパス開発センター（市村太郎ほか）編（2015）は，2015 年 10 月 28 日に公開された。

[2] 国立国語研究所コーパス開発センター編（2015）参照。『日本語歴史コーパス』開発の経緯とこれからの拡張計画については，小木曽（2016），山崎編（2014）等を参照。

語歴史コーパス』の一部となる予定である。そのため『日本語歴史コーパス』所収作品（平安和文・虎明本狂言集）と同様に，本文校訂を経て XML データが構築され，互換性のある形態論情報（単語の情報）が付与されている。将来的には WEB 上のコーパス検索アプリケーション「中納言」上で，「通時コーパス」として，他の時代のコーパスと一括した利用を想定している。

「洒落本コーパス 0.5」の構築にあたっては，『洒落本大成』におけるテキストの状況を極力維持しつつ，ある語・ある語形の用例を確実に検索し，必要な情報を取得するとともに，コンピュータによる計量的研究にも活用可能な電子化資料の構築を目的としている。

対象としては『洒落本大成』から，江戸・京都・大坂それぞれを舞台とする作品を選定し，さらにその中から「画像が利用可能である」「構築上の難点が少ない」等を勘案して優先順位をつけ，順次構築に着手している。

2.2「洒落本コーパス 0.5」の対象作品

本コーパスは試作版として，表 1 の 3 作品を選定して先行公開を行った（以下『聖遊廓』『箱まくら』『花街鑑』）。

表 1 「洒落本コーパス 0.5」収録作品

作品名	著者	刊年	刊年（西暦）	舞台	参照用版本
『聖遊廓』	不明	宝暦七	1757	大坂	国語研蔵：小本 1 冊
『河東方言箱まくら』	大極堂有長	文政五	1822	京都	国語研蔵：中本 3 冊
『玉菊全伝花街鑑』	鼻山人	文政五	1822	江戸	国語研蔵：中本 3 冊

これらの作品はいずれも，現在国立国語研究所が版本を所蔵し，『日本語史研究資料』[3] として，資料画像が一般公開されているものである（図 1 参照）。試作版の本コーパスでは，『明六雑誌コーパス』[4] で実装された原本参照機能にならい，URL をコーパスデータ内に埋め込むことで，コーパスデータから『日本語史研究資料』の画像データを参照する機能を備えている。

[3] http://dglb01.ninjal.ac.jp/ninjaldl/（2015 年 11 月 26 日閲覧）
[4] 国立国語研究所コーパス開発センター，2012 年 10 月 31 日 Ver 1.0 公開。
http://pj.ninjal.ac.jp/corpus_center/cmj/meiroku/

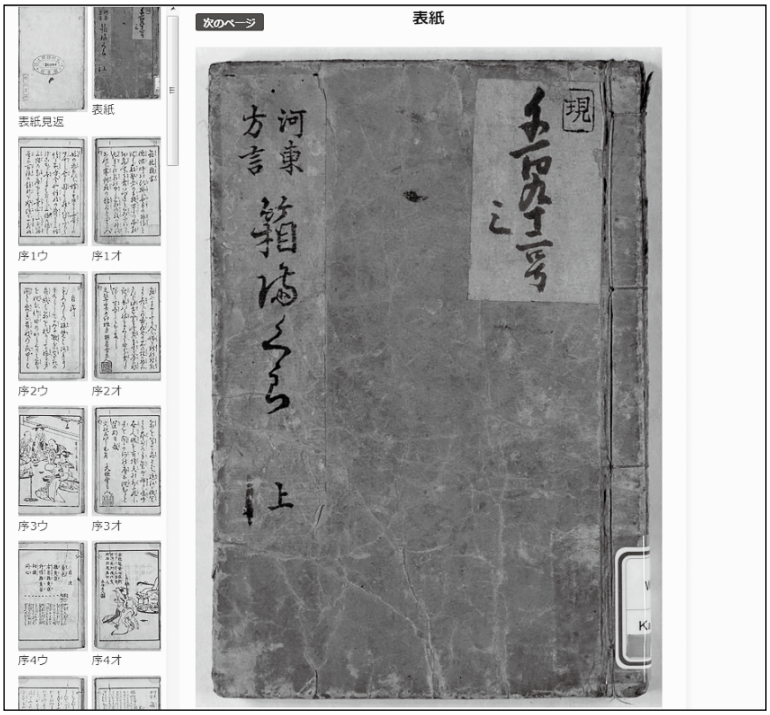

図 1　国立国語研究所所蔵『河東方言箱まくら』[5]

3.「洒落本コーパス 0.5」の特徴

本コーパスの大きな特徴は，形態論情報，XML タグによる文書構造情報，出典情報，話者情報が揃って付与されている点である。また，対応する版本画像がある場合は，検索結果から当該丁の公開画像リンク先にジャンプする機能を備えている。これにより，単なる文字列検索の用途にとどまらず，様々な情報による検索や，従来の索引に付されていたような，底本や画像の参照，高度な付帯情報の取得等を同時に行うことができる。

次頁の図 2（上）に挙げたものが実際の検索結果である。例えば語彙素「寝る」で検索した場合，その検索結果には，品詞や表記に関する情報，会話か割書きかなどの文書構造に関する情報，作品名・巻・ページ数などの出典情報，出現箇所が会話の場合，話者名や性別などの話者情報が同時に同行に表示される。また，検索結果の画像丁数をダブルクリックすると，当該丁の参照用画像にジャンプし，同行の他の箇所をダブルクリックすると，検索キーが含まれるテキストがブラウザで表示され（図 2（下）参照），当該箇所を含むテキストの全体を確認することができる。各情報の詳細や利用例は以下順次述べていく。

[5] http://dglb01.ninjal.ac.jp/ninjaldl/show.php?title=hakomakura&issue=001&num=1

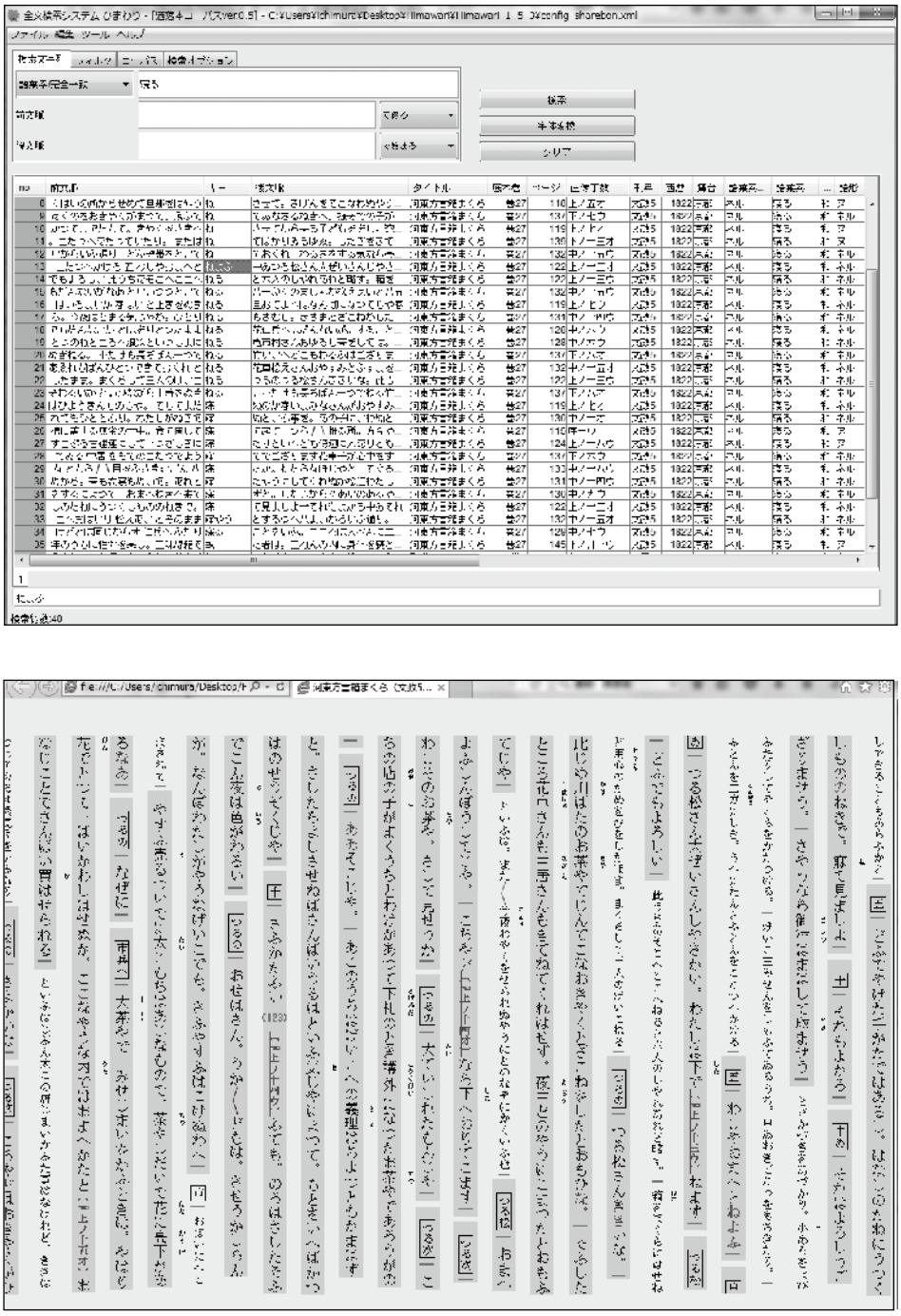

図 2　検索画面・検索結果と当該箇所のブラウザ上の表示

3.1 文書情報の付与

3.1.1 文字の入力

文字符号化方式は UTF-8，符号化文字集合は JISX0213 に準拠し，外字が出現した場合，代用可能な字がある場合は代用し，適切な代用字がない場合は「〓」で入力した。その際，底本の文字が外字となる箇所はタグを付与して表示している。

また，形態論情報の付与や，文字列検索へ対応するため，下記の校訂を行った。

A. 濁音・半濁音が期待される箇所に濁点が付されていない場合　⇒濁点付の仮名
B. 仮名 1 字分の踊り字　⇒対応する平仮名
C. 外来語を除くカタカナ表記箇所　⇒対応する平仮名

3.1.2 文書のマークアップ

上記の校訂箇所のような本文改変を行った箇所を含め，利用上必要な情報や研究上有用な情報について，次頁の表 2 に示すタグセットを設定し XML 形式のマークアップを行った。このタグセットは，市村・河瀬・小木曽（2012）で設定したものを，実際の構築状況や，「ひまわり」の特性等を勘案して改訂し，適用したものである。このタグ情報に基づいてテキスト内の情報を仕分けし，検索対象とするもの，表示させるものなどを指定している。

タグ構成の枠組みや，各タグの設定に関する基本的な考え方は，市村・河瀬・小木曽（2012）や市村（2014）で示したところであり，大きく分けると，文書構造を表すもの，校訂情報を表すもの，位置情報を表すものがある。

文書構造では，洒落本テキストの実態に合わせ，書籍という形状的単位ではなく，見出しに相当する単位を一記事と見た。また会話や割書きや引用を切り出し，その下位要素として文を配置した。「文」という単位は，「△は「○○。○○。」と言った。」のように，会話文の中の一文としての「文」を，引用を包括する大きな単位としての文が包む，というように本来重層的である。しかし本コーパスを近世口語資料としてみれば，会話文の利用が主体となることが想定され，会話文の中の単位としての「文」を切り出すことがより重要である。それに加え，データの複雑化を避ける意味もあり，引用をまたがない細かい単位としての文のみを設定することとした。

また，校訂情報として，テキスト構築時に濁点を付与した箇所，踊り字を開いた箇所等にタグを付した。これは，底本の原態を確認できるようにすることを意図している。

さらに，従来の書籍の索引類と同様の索引機能を備えることを目的とし，『洒落本大成』のページ数や参照用画像の丁数などの位置情報をタグに記録した。図 3 にはマークアップの実例を示す。

表 2 「洒落本コーパス 0.5」のタグセット

要素（タグ）名		説　明
<text>		1 作品全体
<front>		前付相当の箇所（序文等）
<body>		主本文相当の箇所
<back>		後付相当の箇所（跋文，刊記等）
<article>		1 記事の範囲（「回」相当）
<titleBlock>		記事とは認められない，<text> 直下レベルでの表題周り
<p>		段落を表す。タイトルや署名等を除く主本文
<block>		記事中のタイトルなど，主本文とは切り分けたい段落要素
<q>	@type=“会話”（<speech>）	ひとまとまりの会話文。ひまわり用に <speech> を <q> に統合。本要素に話者情報を付与。
	@type=“引用”（<quotation>）	文献等からの引用や手紙など。ひまわり用に <quotation> を <q> に統合。
	@type=“割書”（<warigaki>）	割書き箇所。ひまわり用に <warigaki> を <q> に統合。
<s>		文
<verse>		謡などの節付け箇所や和歌など韻文であることが明確な箇所
<delivery>		会話文の様式等を指定する記述
<speaker>		話者の表示
<corrSpan>		振り仮名等により文字列の置き換えを行った短単位以上の箇所
<hi>		小書き・傍線・囲みなどの文字列に対する装飾
<SUW>		語（短単位）
<lRuby>		本行の左側に振られた振り仮名等の文字列
<r>	（<ruby>）	本行の右側に振られた振り仮名文字列。ひまわり用に <ruby> を <r> に。
<add>		本文の補入箇所
<kanbun>		訓み下す際文字位置を置き換えた漢文等の箇所
<vMark>		底本原文が濁点無表記であった箇所
<odoriji>		底本原文が 1 字分の踊り字であった箇所
<corr>		誤字・脱字・衍字等の本文の修正
<g>		外字・絵文字等準拠する文字セットでは表示できない文字
<char>		1 字を表す単位，@script=“カタカナ”で，カタカナ表記箇所に使用
<info>		本文テキストに割って入れられなかった記号，丁付情報等
<pb><lb>		底本の改ページ位置・改行位置
<opb>		原本画像の丁や画像リンクとの対応

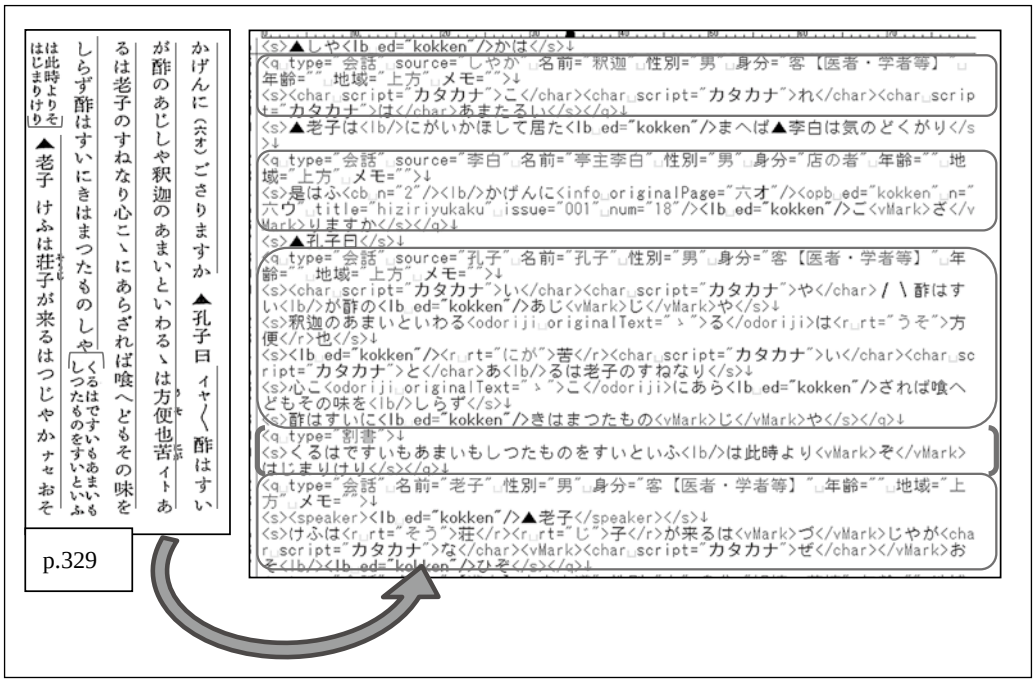

図 3 『聖遊廓』本文と XML データ

市村・河瀬・小木曽（2012）では，会話・引用箇所・割書きについて，<speech>，<quotation>，<warigaki> と，要素を分けて設定していたが，本仕様では，本文中の性質の異なるテキスト要素を <q> でマークし，会話文を表す「会話」と，文献引用等を表す「引用」，地の文や注記に相当する「割書」の 3 種を属性として書き分けた。なお, 無表示の物は「割書」ではない本文である。割書き内の会話文は認定していない。

これは，コーパス検索ツール「ひまわり」の仕様に合わせ，これらを検索結果中の 1 つの列にまとめて表示させるためである。

このように，タグ設計の発想は，市村･河瀬･小木曽（2012）を踏襲し，『日本語歴史コーパス』内での汎用性を維持しつつ，ツールやコーパス本文に合わせて応用し，タグを運用している。

3.2 話者情報の付与

本コーパスにおいては, これまでの『日本語歴史コーパス』や『太陽コーパス』等に先駆けて，本文情報と同時に，試験的に話者情報の詳細を付与した（次頁表 3)。また，検索結果では，形態論情報の右側に，次の図 4 のように表示される。本文種別として，「割書」「会話」「(文献等) 引用」の別が表示され（空白は，その他の地の文等），その右側に表 3 の「項目」にあるような情報が表示される。

表 3 「江戸時代編 I 洒落本」の話者情報

項目	説明	表示内容
①話者 （表示を統一）	一人の話者に対する統一した話者名。判明するものに限り，【遊女，芸子】の別を示している。	花車，旦那，【芸子】つるの，八五郎，白楽天，【遊女】太夫大空…等
②性別	話者の性別。「その他」は人間以外の登場人物等を想定している。	男，女，その他
③身分	A 店の者，B 客，C その他の 3 別を基に，【】の記述でさらに細分化している。	A　娼妓・芸妓，店の者，禿，太鼓持
		B　客，客【上層・むすこ・通人】，客【武士】，客【医者・学者等】，客【町人】
		C　その他，その他【町人】，その他【上層町人】，その他【武士】，その他【医者・学者等】，その他【神・仙人・僧侶等】，その他【侠者】，使用人
④年齢	原則老人・子供の別のみ。数としての年齢は「メモ」に記載し，現在保留中。	
⑤地域	会話から推定される話者の使用言語（出身地）を 3 種で認定。上方の洒落本における江戸話者等も想定している。	江戸，上方，田舎

本文種別	話者	性別	身分	年齢	地域	メモ
割書						
引用-手紙						
会話	【遊女】玉菊	女	娼妓・芸妓		江戸	玉菊 遊女時代
会話	滝三郎	男	客【上層・むすこ・通人】		江戸	滝三郎
会話	ばば	女	その他【町人】	老人	江戸	
引用-典拠						
会話	屋根介	男	その他【町人】		江戸	
会話	【遊女】玉菊	女	娼妓・芸妓		江戸	玉菊 遊女時代
会話	玉蔵夫婦	男	その他【町人】		江戸	
引用-典拠						
会話	番頭堅兵衛	男	客【上層・むすこ・通人】		江戸	番頭・親仁株
会話	【遊女】玉菊	女	娼妓・芸妓		江戸	玉菊 遊女時代
割書						
会話	ばば	女	その他【町人】	老人	江戸	
会話	延寿	男	客【上層・むすこ・通人】		江戸	延寿
引用-典拠						
会話	延寿	男	客【上層・むすこ・通人】		江戸	延寿
割書						
会話	滝三郎	男	客【上層・むすこ・通人】		江戸	滝三郎
会話	延寿	男	客【上層・むすこ・通人】		江戸	延寿

図 4 「洒落本コーパス 0.5」における本文情報と話者情報

これにより，検索によって得た用例を，性別や話者，地域ごとに集計することなどが可能になる。また，作品ごとに統一された話者表示がなされることにより，個人の言語使用量が格段に計数しやすくなった。例えば遊女「玉菊」が本文上で「たま」「玉」「玉菊」などと異なる表示をされていたとしても，同一作品内の同一人物は統一され，全例について「【遊女】玉菊」と表示される。

ただ，上記の情報はあくまで洒落本の範囲に限っての規則化を目指しているものである。このような記述を「通時コーパス」へと拡張・一般化することを考えたときに，普遍的な情報としていかなる項目を設定し，いかに記述するのかという点は今度の課題である。

3.3 形態論情報の付与

3.3.1「洒落本コーパス 0.5」における形態素解析

本コーパスでは，文書構造や話者についての情報を XML 形式で付与するとともに，テキストを言語単位に分割し，形態論情報（品詞や活用形，読みなどの情報）を付与している。形態素解析器「MeCab[6]」と，形態素解析用辞書「近世口語 UniDic」で形態素解析した結果を人手で修正し，短単位データを構築した。近世口語 UniDic は，現代語向けに開発された形態素解析辞書 UniDic（伝ほか 2007）をもとに，洒落本の本文データに応じて修正･拡張したものである（小木曽･市村･鴻野 2013）。

近世後期口語資料は表記のバリエーションが豊富であり，とりわけ洒落本は作品ごとに作者が異なるため，傾向を見出すことが難しい。また口語体と文語体が混在しており，会話文には江戸･上方といった地域差が反映されることもある。そのような文体の特徴が解析精度の向上を阻んでいたが，洒落本専用の解析辞書を作成したことに加え，会話と地の文それぞれに適した形態素解析を施す，という手法を取り入れることで，現在およそ 90% の解析精度を実現している（市村・小木曽 2016）。今後，学習用データが蓄積されるにしたがい，より精度が向上するものと期待される。

3.3.2 形態論情報の特徴と意義

文字列検索による用例収集では，先にも挙げたように近世資料ならではの表記の複雑さが大きな課題となるが，形態論情報が付与されていることで，その問題がほぼ解消される。例えば「大概（タイガイ）」という形式を検索する場合，文字列検索であれば，「大概」「たゐがい」「大がい」「てへげへ」など様々な語形，表記を想定しなければ，すべての用例にあたることはできない。このような場合に，UniDic の特徴の 1 つである，見出し語の階層構造（語彙素，語形，書字形，発音形）が効果的にはたらく。例として挙げた図 5 のとおり，異語形・異表記が，代表の見出し語（語彙素）に紐づけられているため，階層ごとの検索に限らず，あらゆるパターンを網羅的に検索することが可能になるのである（なお，活用語の場合，下二段活用，下一段活用などの活用型は「語

[6] 工藤拓（2006–）MeCab: Yet Another Part-of-Speech and Morphological Analyzer （http://taku910.github.io/mecab/）

形」のレベルに追加されるため，文語と口語の用例をまとめ上げることもできる)。近世期の資料にとっては特に重要な機能であるといえよう。

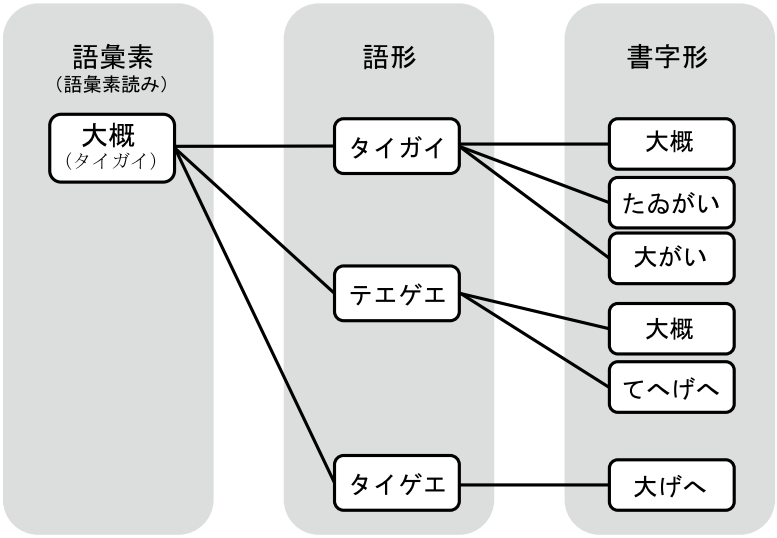

図 5　UniDic における階層構造の例（発音形は省略）

また，特定の語を検索するだけでなく，例えば「お（接頭辞)」＋形容詞連用形＋「御座る」のように語と品詞を組み合わせる検索や，品詞が接尾辞に分類されるものだけを抜き出すといった操作も，形態論情報が付与されることで初めて可能となる。

3.3.3 洒落本の言語状況に対応した形態論情報の付与と課題

公開した 3 作品に加え，今後公開予定の他作品について作業を進める中で，近世語資料全体に共通する問題や，今後さらに複雑化すると予想される課題が浮かび上がってきた。ここでは，「活用」に関わる事例を報告する。

「日本語歴史コーパス」において，活用型は大きく「文語活用」「口語活用」に分けられている。『平安時代編』『室町時代編Ⅰ　狂言』では主に文語活用，『現代日本語書き言葉均衡コーパス』では口語活用が採用されており，その過渡期に位置する「洒落本コーパス 0.5」では，会話は「口語」ベース，それ以外は「文語」ベースとしている（市村 2015)。

ただし，以下に述べる指定の助動詞は，時代差・地域差を特に反映するものであり，個別の検討を必要とした。ここで取り上げる形式は，「なり」「じゃ」「だ」の 3 形式だが，すでに公開されている歴史コーパスの場合，『室町時代編Ⅰ　狂言』では「なり」「じゃ」を，『明治・大正編』では「だ」「じゃ」を，それぞれ併用している。洒落本でも，助動詞「だ」を用いることが想定されるが，本コーパスが対象とするのは宝暦年間（1751–1764）から文政年間（1818–1830）にかけての，江戸・京都・大坂を舞台とする作品群であり，一般に江戸語としての性格が強いといわれる助動詞「だ」を，例えば宝暦期の上方版に当てはめることは不適当であろう。しかし，上方

を舞台とする作品に，どの時期から「だ」を認めるかという線引きもまた困難である。

そこで本コーパスでは，助動詞「だ」について，基本的に江戸を舞台とする作品の会話部分のみに認めることとし，それ以外は狂言の方針に準拠して，助動詞「なり」「じゃ」を適宜用いることとした。これは先に述べた「文語」ベース・「口語」ベースの方針にも対応したものである。この方針に基づくと，例えば見た目だけでは判別困難な（1）～（4）は，次の ⇒ のように区別することになり，それをまとめたものが（5）である。

（1） をもてのこたつでよう寐てでござります（『箱まくら』，京都，1822）⇒「なり」
（2） ああふ便な事でござります（『花街鑑』，江戸，1822）⇒「だ」
（3） あのやうに無常な事ばつかりいふてでござんすわいな（『聖遊廓』，大坂，1757）⇒「なり」
（4） かならず気鬱せぬやうに。心強くおもふてゐよ。（『花街鑑』，江戸，1822）⇒「だ」

（5） 洒落本における，指定の助動詞「だ」「なり」の区別

会話以外→			助動詞「なり」
会話→	舞台が上方	→	助動詞「なり」
→	舞台が江戸	→	助動詞「だ」

この方針は話者の位相や，刊年に拠るものではないため，実際には混在していると言ってよく，次のように例外的に処理したものもある。

（6） 病人の死だ跡へ見廻たやうなものだ。（『阿蘭陀鏡』，京都，1798）⇒「だ」
（7） 手ばかりじやない足もなにも達者だ。（『昇平楽』，京都，1800）⇒「だ」

（6）の話者は，文中で「江戸化言（なまり）」の医者と紹介されていることから，その人物の会話のみ，江戸を舞台とする作品と同様に処理した。（7）は特に注記は見えず，同一人物の会話内に上方語的特徴（助動詞「じゃ」の使用，連用形ウ音便など）も見られたため，終止形「だ」の例のみを助動詞「だ」，それ以外は「なり」で対応した。今のところ，（7）の類例は他に見られないが，今後，他作品の解析が進めば，同様の「例外」が現れるものと推測される。しかし現状では，個々の例を正確に解釈することは困難であり，不要な誤りを防ぐためにも（5）のような機械的な処理にとどめた。「洒落本コーパス 0.5」を利用するにあたっては，このような背景に留意する必要があるが，本コーパスには前節までで示したとおり，本文情報と話者情報が実装されている。それら複数の情報を併せて検討することで，より詳細に実態を記述することができるだろう。

以上の事例のように，「洒落本コーパス 0.5」では，その会話部分に地域差・位相差が色濃く反映されることが，形態論情報を付与する上での課題となっている。また通時的に見たときに，中世・近現代のコーパスとどのように整合性をとるかが，1 つの指針となる一方で難しさも生んでいる。それは洒落本に限らず，近世の口語資料を扱う上で共通の課題といえよう。

3.4 テキストの二重性に関わる課題

近世末の資料では，矢野（1987）で指摘されるように，本行の漢字文字列に対して，左右傍記で意味的に関連のある語を当てていると思われる表記が見られる。下記の A・B は，『花街鑑』において，試みに調査し，列挙したものである。

A.『花街鑑』（1822）中，右振り仮名を本文と置換した文字列（複数短単位を含んでしまう為）
①恭しからざる与（とは）②随身神（じんやだい）門を抜る。③よう〳〵の事で全快（よくなつ）て。④さぞ獣子（ばかげた）。もんだらうね

B.『花街鑑』における非置換当て字風の語（漢字 1 字の場合は『日本国語大辞典（第二版）』の「表記」にないもの）
未至（やぼ） 僻説（いきすぎ） 娼妓（けいせい） 滑稽（しやれ）て 借心（むしん） 紅葉（もみぢ） 蒼天（うらゝか） 家業（なりわい） 朝夕（あけくれ） 温泉（とうじ） 彷彿（ぼうぜん） 何所（どこ） 彼方（あつち） 只管（ひたすら） とり携（つい）て 不測（ふしぎ） 幼稚子（おさなご） 携（すが）り 容貌（すがた） 嬌態（けだかく） 欣然（よろこぶ） 識人（ちかづき） 一貼（いつぷく） 時行（はやり） 同家店（あいながや） 可愛想（かあいそう） 媒妁（なかうど） 愕然却（あきれかへつ）て 数多（たんと） 心驚（びつくり） 所為（ていたらく） 不造化（ふしあはせ） 懐中（ふところ） 容貌（みめかたち） 令郎（むすこ） 株（かぶ） 顔色（かんばせ） 彼所（かしこ） 弾妓（げいしや） 恍惚（ほれられ）る 饌（たべ）ませなんだ 長寿（ながいき） 佳味（うめへ） 食（まんま） 一寸（ちよつ）くら 睨（よこめ） 婦人（をんな） 中旬（なかば） 発（おこ）して 方便（てだて） 倒（どう）と 速（たち）まち 調市（でつち） 光景（ありさま） 風評（うはさ） 一面（べた） 愁懐（すご〳〵） 離離（ぐれはま） 幼稚（いとけなき） 誘引（さそはれ）て 同道（あいび）なせへ 今晩（こんや） 一寸（ちよい）と 浮気（そゝり）ぶし 両個（ふたり） 雛妓（しんぞう） 賄（せは） 閨閉（ぴつしやり） 如斯（かう）話説（はなし） 働（かな）しい 幼稚（ちいさい） 両親（ふたり） 与所（よそ） 催促（せめはた）られ 工（こしらへごと） 花街（さと） 誑（だま）される 自恍（うぬぼれ）て 温補（あつため）て 嫌疑（うたがひ） 斗（はからひ） 鐺（あしなへ） 本性（ぢがね） 顕然（あらはるゝ） 出来難（にく）い 強顔（つれなく） 偽惑（うたぐり） 快然（こゝろよく） 邂逅（たまさか）に 憔悴（やつるゝ） 安堵（やすまら）ね 連（あつぱ）れ 居（すは）らず 辞（いな）み 最（いと） 媚貌（いろけ） 粧容（みえ） 散乱（ちら）がつて 腮褶（えら） 美（よく） 歎（かな）しい 正直（ほんとう） 何卒（とふぞ） 外悋気（おかやきもち） 好望体（ほしきてい） 横雲（しのゝめ） お悪（いや） 紀念（かたみ） 佳美（はなやか） 終事（しまい） 蚤（はや）く 父（おや）

「紅葉＝もみぢ」のような読みとしてある程度一般的と推測されるものから,「温補＝あつため」のように，語として当てていると推測されるものまで様々である。対処として，前者は傍記された読みの語の語彙素の書字形として UniDic に登録することが考えられる。また後者は，ひとま

ず傍記はさておき，まず本文テキストの語を認定し，傍記は掛詞のような「二重テキスト」として，別途記録することが考えられる（小木曽 2016: 79）。ただ，処理のありかたを決定したとしても，両者とも何を基準に，どこで線を引くのかという課題が残る。

また，現在，単位の問題があるため，上の A タイプは，右傍記を読みと見て，本行と右傍記を置換することで対処しているが，元の本文テキストの置き換えられた文字列は検索対象として漏れてよいのかという問題がある。

例えば後年の『安愚楽鍋』にみられる「開化文明（よのひらける）」などの場合，現在の方法では，「開化文明」という文字列に対して，傍記の「よ／の／ひらけ／る」という 4 短単位が対応するため，右傍記と置換することとなる。本コーパスは，すべてのテキストが「短単位」に分割され線条的に配置される。そしてその短単位による検索が前提となるため，傍記と単位数のずれ，あるいは単位が対応する箇所の前後関係のずれが生じた場合は，最終的にどちらかの文字列を本文テキストとし，その単位のみを認定せざるを得ないのである。

その結果，「開化文明」という元の本文文字列は，ブラウザ上での表示対象とはなるものの，検索対象とはならない。「開化」も「文明」も，近代初頭における重要漢語であるが，検索対象にならないただの「置き換えられた文字列」となってしまうのは，言語資料としての損失と言ってよい。

このような文字列への対処としては，先ほど挙げた「二重テキスト」と同様の対処が考えられる。ひとまず本文テキストとしては「開化」「文明」を認定しつつ，別途「よ」「の」「ひらけ」「る」をデータとして持たせ，ID 等で本文テキストと関連を持たせておく，といった方法が想定される。

今後，人情本や，近代初期資料を扱うことも視野に入れ，このような表記上の問題について，「二重テキスト」を処理するシステムを整備し，日本語史資料としての価値をできるだけ損なわないレベルでの着地点を探る必要がある。

4.「洒落本コーパス 0.5」の利用法と利用例

4.1「洒落本コーパス 0.5」の検索対象

本コーパスでは，次頁にまとめた表 4 の要素によって，用例検索を行うことができる。③から⑧については 3.3 で述べたところである。ある短単位の用例を収集する際に，③語彙素の表記が推測しがたい場合は，④語彙素読みで検索することにより補完することができる。

表 4　検索対象（市村 2015 を加筆・修正）

項目	説明・検索の対象	検索対象データの例
①　本文	本文の文字列。	
②　ルビ	振り仮名。本文の文字単位または本文の短単位。文字種や仮名遣いは原則原文どおりで入力されている。	文字単位：き（器）りやう（量） 語単位：けいせい（娼妓） ※（　）はルビが傍記された本文テキストの文字列。
③　語彙素	単語の統合的単位。現代語ベースで標準的漢字かな表記，終止形で入力されている。	寝る・為る・成る・有る・見せる 悪しい・有り難い・可笑しい・寂しい 仮令・是非・唯・一寸 事・物・者・人
④　語彙素読み	「語彙素」の読み。カタカナ・終止形（原則現代語ベース）。	ネル・スル・ナル・アル・ミセル アシイ・アリガタイ・オカシイ・サビシイ タトエ・ゼヒ・タダ・チョット コト・モノ・ヒト
⑤　語形	単語を音韻変化・活用型などで区別した単位。	チョット―チト―チョイト ミセル―ミス
⑥　品詞・活用型・活用形	出現した例（キー）の，品詞や活用型，活用形に関する情報。UniDic に準拠。	（キー：あつ） 品詞：動詞－非自立可能 活用型：五段－ラ行 活用形：連用形－促音便
⑦　書字形	キーに出現した表記法で当該短単位の終止形を表示。	（キー：おかしかつ　語彙素：可笑しい） 書字形：おかしい
⑧　発音形・仮名形	発音や仮名遣いの形。カタカナ・終止形で表示。	（キー：おめへ　語彙素：御前） 仮名形：オメヘ 発音形：オメー
⑨　語種	語の出自。漢字一字で表示される。	和（和語）・漢（漢語）・外（外来語）・混（混種語）・固（固有名詞）
⑩　話者名	会話の話者。話者情報に入力されている統一的な話者名。	【遊女】玉章，【芸子】春野，八五郎，髪結の弟子，あまたの子供

検索に際しては任意の情報によるフィルタリングも可能である。また検索結果を表計算ソフトで処理することによって，他の情報と複合した集計が可能である。

例えば，「ちょいと」という語を検索したい場合，語彙素「一寸」で検索すると，「チョット」や「チト」などの異語形も検索結果に混入することになるが，語形「チョイト」で検索すれば，「チョイト」のみを検索結果に表示させることができる。

4.2「洒落本コーパス 0.5」の利用例―ワ行五段（ハ行四段）動詞連用形の音便形―

最後に，コーパスの利用例を示しておく。

以下，本コーパスを利用して「ワ行五段（ハ行四段）動詞連用形の音便形」を調査した。検索条件は，［検索文字列（活用形／部分一致）＝［ウ促］音便］とし，活用型を限定するため，フィルタに［活用形＝［五四］段 -［ワハ］を含む］とした。また，会話文に限定するため，［本文種別＝会話で始まる］も設定した。その検索結果を表計算ソフトで集計したのが表 5 である。

表 5　ワ行五段（ハ行四段）動詞連用形の音便形

	連用形－ウ音便	連用形－促音便	総計
江戸話者	10	23	33
『花街鑑』	10	23	33
会う		1	1
願う	1		1
見繕う		1	1
言う	2	9	11
思う	4	5	9
揃う		1	1
奪い散らがう		1	1
弔う	1		1
適う	1		1
買う		2	2
付き合う		2	2
慕う	1		1
貰う		1	1
上方話者	83	2	85
『箱まくら』	72	2	74
違う	1		1
会う	2		2
言う	32		32
仕舞う	2		2
思う	24		24
笑う	1		1
食う		1	1
酔う	5		5
叩き合う	1		1
買う	2	1	3
貰う	2		2
『聖遊廓』	11		11
違う	1		1
言う	6		6
思う	4		4
田舎話者	4		4
『箱まくら』	4		4
言う	1		1
思う	2		2
酔う	1		1
総計	97	25	122

近世期の上方語では，ワ行五段（ハ行四段）動詞は，「た」や「て」に接続する場合，連用形がウ音便化することが知られている（湯澤 1963 等）。一方，促音便は，「言う」や漢籍等の文語調を除けば「京阪地方では普通ではな」く（湯澤 1963: 93），江戸語に特徴的なものとされている。

表 5 を見ると，おおむねそのとおりの結果となっているのだが，近世末の上方洒落本『箱まくら』では，2 例ほど促音便形の例が見られる。

(8) 千　そしてかうがいでなんぼほどかつてある。〔買う〕
(9) 梅　はいどこなと　くつて見なされ。〔食う〕

上方洒落本でも江戸語話者が登場することが時折あるため，話者情報を参照すると，例(8)は「話者：千太郎，性別：男，身分：客【上層・むすこ・通人】，地域：上方，メモ（年齢）：二十六・七」であることがわかる。また，例（9）は，「話者：芸子梅野，性別：女，身分：娼妓・芸妓，地域：上方，メモ（年齢）：二十七・八」であり，20 代男女の上方語話者（という設定）であることが確認される。前後文脈を確認する限りでは漢文調・文語調でもない。

このことから，幕末期の洒落本，『箱まくら』では，上方語話者による，ワ行五段（ハ行四段）動詞連用形の促音便形の例が見られることが確認される。

もちろん，表 5 にみるように，上方洒落本はウ音便形の方が圧倒的多数であるため，本例で通説が否定されるわけではないが，コーパスを利用することにより，このようなわずかな例をも網羅し，話者情報を参照しつつ，直ちに分析の俎上に載せることができる。この意義は極めて大きいであろう。また，作品全体が一定の単位に区切られていることにより，ただ用例を得るだけではなく，資料体全体でのその用例・数値の位置づけを確認することができる。

5. おわりに

洒落本を対象としたコーパスの構築は，着手当初困難が予想されていたが，本コーパスの公開によって，見通しを立てられた。また並行して開発が進められている「人情本コーパス」等，近世後期口語資料を対象とするコーパスのモデルを示すことができた。

今後も，構築中のコーパスの精度を向上させるとともに，コーパス外情報との接続の充実や，必要な情報の追加や修正も視野に入れ，『洒落本コーパス』の開発を進め，「通時コーパス」の構築に寄与したい。

参照文献

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵（2007）「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』22: 101–123.
市村太郎（2014）「近世口語資料のコーパス化―狂言・洒落本のコーパス化の過程と課題―」『日本語学 11 月臨時増刊号　日本語史研究と歴史コーパス』33(14): 96–109. 東京：明治書院.
市村太郎（2015）「ひまわり版「洒落本コーパス 0.5」利用案内」http://pj.ninjal.ac.jp/corpus_center/chj/doc/sharebon0.5-doc.pdf
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』2012(1): 1–8.

市村太郎・小木曽智信（2016）「文書構造を利用した近世期洒落本の形態素解析」『言語処理学会第 22 回年次大会発表論文集』107–110. 言語処理学会.
国立国語研究所コーパス開発センター（編）（2015）『日本語歴史コーパス』（バージョン 2015.3, 中納言バージョン 2.0.1）https://chunagon.ninjal.ac.jp/（2015 年 11 月 28 日確認）
国立国語研究所コーパス開発センター（市村太郎ほか）（編）（2015）『ひまわり版「洒落本コーパス」（日本語歴史コーパス江戸時代編）』http://pj.ninjal.ac.jp/corpus_center/chj/edo.html#share（Ver. 0.5）（2015 年 11 月 28 日確認）
小木曽智信（2016）「『日本語歴史コーパス』の現状と展望」『国語と国文学』1110: 72–85.
小木曽智信・市村太郎・鴻野知暁（2013）「近世口語資料の形態素解析の試み」『第 4 回コーパス日本語学ワークショップ予稿集』146–150. 東京：国立国語研究所.
山口昌也・田中牧郎（2005）「構造化された言語資料に対する全文検索システムの設計と実現」『自然言語処理』12(4): 55–77.
山崎誠（編）（2014）『講座日本語コーパス 2 書き言葉コーパス―設計と構築―』東京：朝倉書店.
矢野準（1987）「人情本の漢字」佐藤喜代治（編）『漢字講座第 7 巻 近世の漢字とことば』199–218. 東京：明治書院.
湯澤幸吉郎（1963）『徳川時代言語の研究』東京：風間書房.

例文出典

『阿蘭陀鏡』［1798］：洒落本大成編集委員会（編）（1982）『洒落本大成 第十七巻』東京：中央公論社.
『昇平楽』［1800］：洒落本大成編集委員会（編）（1983）『洒落本大成 第十九巻』東京：中央公論社.

A Trial Construction of the *Sharebon* Corpus

ICHIMURA Taro[a] MURAYAMA Miwako[b]

[a]Tokoha University / Project Collaborator, NINJAL
[b]Adjunct Researcher, Language Change Division, Research Department, NINJAL

Abstract

This paper presents an overview, the features, and utility of the *Sharebon* Corpus. We attempted to construct a corpus of Early Modern Japanese text, which is a part of the Corpus of Historical Japanese (CHJ) built by The National Institute for Japanese Language and Linguistics. We released a trial version of the *Sharebon* Corpus on October 28, 2015.

This corpus has not only annotated morphemes and document information, just as the other corpora of the CHJ, but also realized the following new functions. First, we implemented the reference function that displays images of original books printed from woodblocks. Second, we made detailed annotations of information about speakers. Early Modern Japanese texts are written in various styles because of the differences, such as region, social class, and generations among others. In this article, we will illustrate that this corpus, which provides voluminous information, will be effective for such texts.

Key words: Corpus of Historical Japanese, Early Modern Japanese, morphological information, Edo dialect, *Kamigata* (*Kinki*) dialect