ひまわり版「洒落本コーパス」Ver. 0.5 利用案内

2015 年 10 月 23 日

市村太郎

１．はじめに

『ひまわり版「洒落本コーパス」Ver. 0.5』は、『日本語歴史コーパス』の一部として計画されている洒落本編の試作版である。

本コーパスは洒落本大成編集委員会編『洒落本大成』（1978-88 中央公論社）を底本としており、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（詳しくは市村・河瀬・小木曽 2013、市村 2014 参照）。

本稿では、本コーパスで利用できるテキストや、各種情報について、基礎的な利用を念頭に置き、概要を述べる。

なお、本コーパスは、試行版の段階であるため、**処理の揺れ等への対処や誤りのチェックが十分ではない**。そのため、目的の語を網羅的に検索する際は、複数の方法を用いて確認することを推奨する。また適宜「ページ番号」や画像情報を基に、底本『洒落本大成』の本文をご確認いただきたい。

◆検索対象（左上にある検索文字列タブのプルダウンメニューで選択⇒P. 4 図 1 画像参照。）

項目	説明	表示法・単位
①本文	本文文字列	
②ルビ	振り仮名	文字単位
③語彙素	単語の統合的単位	標準的漢字かな・終止形（現代語ベース）
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）
⑤語形	単語を音韻変化・活用等で区別するレベル	カタカナ・終止形
⑥品詞・活用型・活用形		UniDic・「中納言」に準拠 ※いわゆる「形容動詞」語幹は「形状詞」
⑦書字形	単語を表記で区別するレベル	終止形
⑧発音形・仮名形	発音や仮名表記の形	カタカナ・終止形
⑨語種	語の出自	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	一致する会話を表示

※『日本語歴史コーパス』を初めて利用される場合、「本文」の「文字列検索」でどのように情報付与されているかアタリをつけた上で利用されることを推奨します。

２．テキストの凡例

［１］外字の処理

本文テキストの文字入力は **JISX0213** に準拠している。

●外字となる箇所は、下記のように読みが同じで字形・用法の近い文字、または適切なものがないと判断した場合は**＝**に置き換えて入力した。

①**＝**妃（よきおんな・unicode 番号：36F9　『聖遊廓』）

②**＝**（けん・unicode なし　『聖遊廓』）さんずいに酉+華

③てつゝり（代用　『箱まくら』）「つ」に半濁点

④**＝**と（どつと・unicode 番号：35E2　『花街鑑』）

⑤玓**＝**（うるわしき・unicode 番号：74C5　『花街鑑』）

［２］テキストの校訂

●テキストの置き換え・補い

本文中一部を処理単位上の問題により、振り仮名（または傍記）の文字列をタグ付で本文と置き換える、または本文に補った。当該箇所は原則ブラウザ上で緑字で表示され、カーソルを合わせると「右ルビ置き換え」等が表示される（**単語単位を超えるカタカナ表示箇所には現在表示非対応**）。

ただし『聖遊廓』の下記の箇所については、例外的に、漢字本文の後に、振り仮名に基づき、文字列を補った。

①〈原文〉里仁為美択不処仁焉得知（さとはじんなるをよしとすゑらんてじんにおらすんはなんそちをゑん）

〈入力〉里仁為美択不処仁焉得知さとはじんなるをよしとすゑらんでじんにおらずんばなんぞちをゑん

②〈原文〉十目所視十手所指（ひとのみるてまへをおもふて）

〈入力〉目所視十手所指ひとのみるてまへをおもふて

③〈原文〉会者常離臨命終時不随者（あふはわかれしばしのたのしみじや）

〈入力〉会者常離臨命終時不随者あふはわかれしばしのたのしみじや

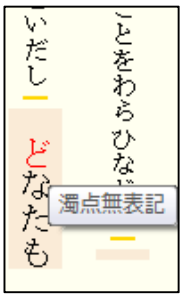
また、次の『聖遊郭』pp.333-334 の「かきをき」は、梵字にカタカナ振り仮名を付した表記であったが、振り仮名を平仮名表記した本文として入力した。

④かきをきうれしからぬみらいとてもながらへそいもならず。ままならぬ。あかぬしやばなれど。けふをかぎりのいのちぞと。をもひあきらめ。しぬるかくごのたびころも。こころせくままなまなかに。なれまじひとになれそめて。なばかりのこす。ふみづきのそらいつまでも。そひはふべきとをもひしに。をもひしかひも。しでのやまぢに

●濁点の校訂

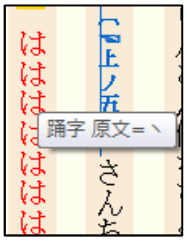
濁音が期待される箇所に濁点が付されていない場合は、諸資料を参考に検討の上、必要な箇所は濁点を補った。

当該箇所はブラウザ上では赤字で表示され、カーソルを合わせると「濁点無表記」と表示される（**右図**）。



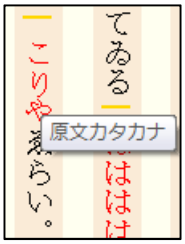
●踊り字の校訂

仮名１字分の踊り字は、想定される仮名に置き換えた。当該箇所はブラウザ上では赤字で表示され、カーソルを合わせると「踊り字　原文＝ゝ」などと表示される（**右図**）。



●カタカナ表記箇所

底本本文中カタカナで表記された箇所は、外来語を除いては平仮名に置き換えた。当該箇所はブラウザ上赤字で表示され、「原文カタカナ」と表示される（**右図・<u>単語単位を超えるカタカナ表示箇所には現在表示非対応</u>**）。



●漢文等

返り点などが付されており訓読が明確に可能な漢文等については、訓読した形を本文とした。その際、原文位置と番号で対応を取り、ブラウザ上では「※」にカーソルを合わせることで、移動後の文字の元あった位置を表示した（**右図**）。

返り点等のない漢文箇所は「未知語」として扱い、品詞「漢文」とした。

●捨て仮名は本文から除いた。

３．Himawari 上の表示項目と内容

『ひまわり版「洒落本コーパス」Ver. 0.5』の本文には様々なタグ（本稿末参考表）や単語情報（後述）が付されており、その情報は、コーパス検索ツール「Himawari」上に、検索結果として表示される（**図１**）。

検索結果中**「画像丁数」以外の項目をダブルクリック**すると、ブラウザ上で作品全文と共に確認することができる（**図２**）。また、インターネットに接続した状況で「画像丁数」をダブルクリックするとブラウザが開き、対応する版本画像掲載ページへジャンプする。

図　１　検索画面と検索結果

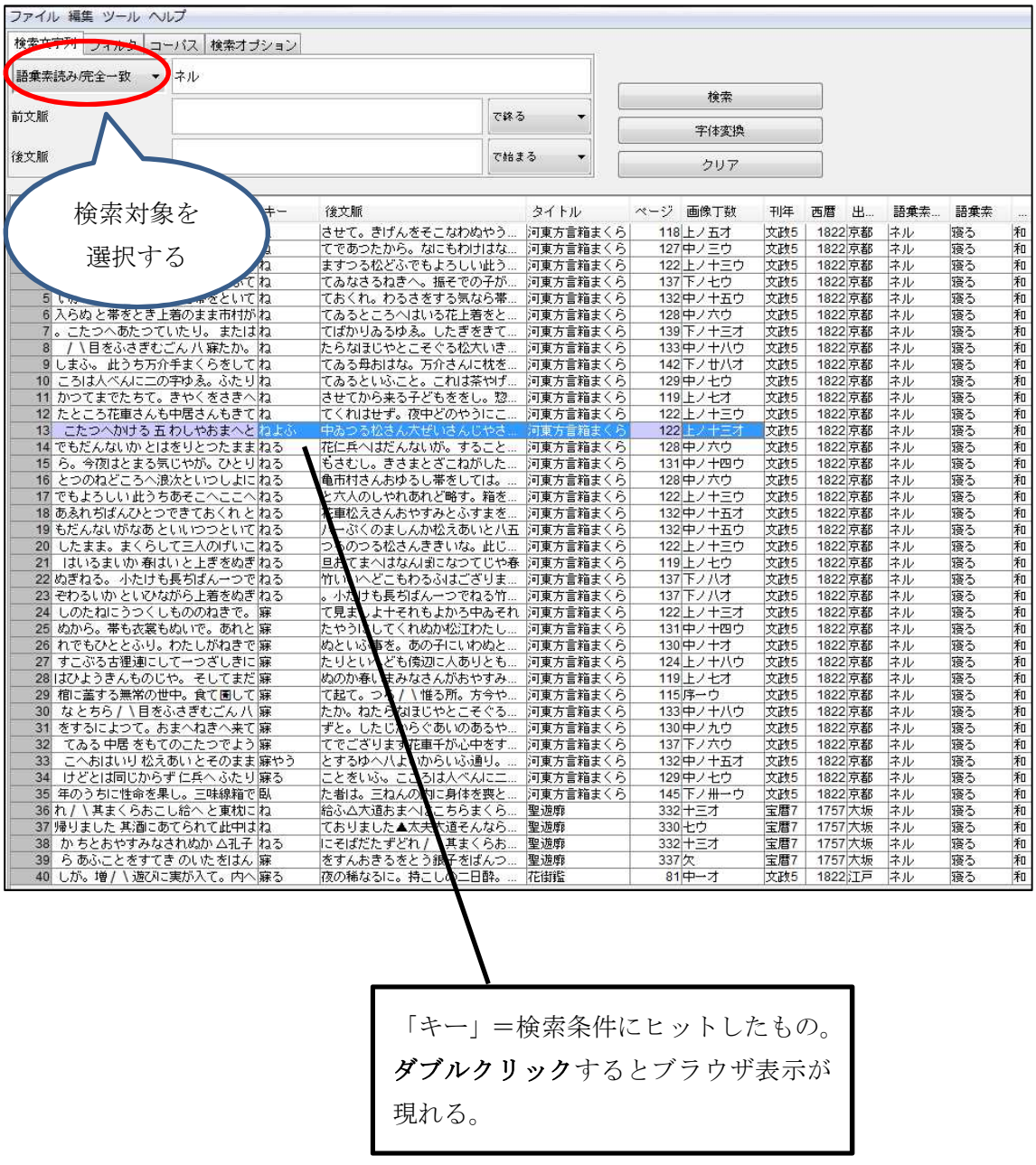
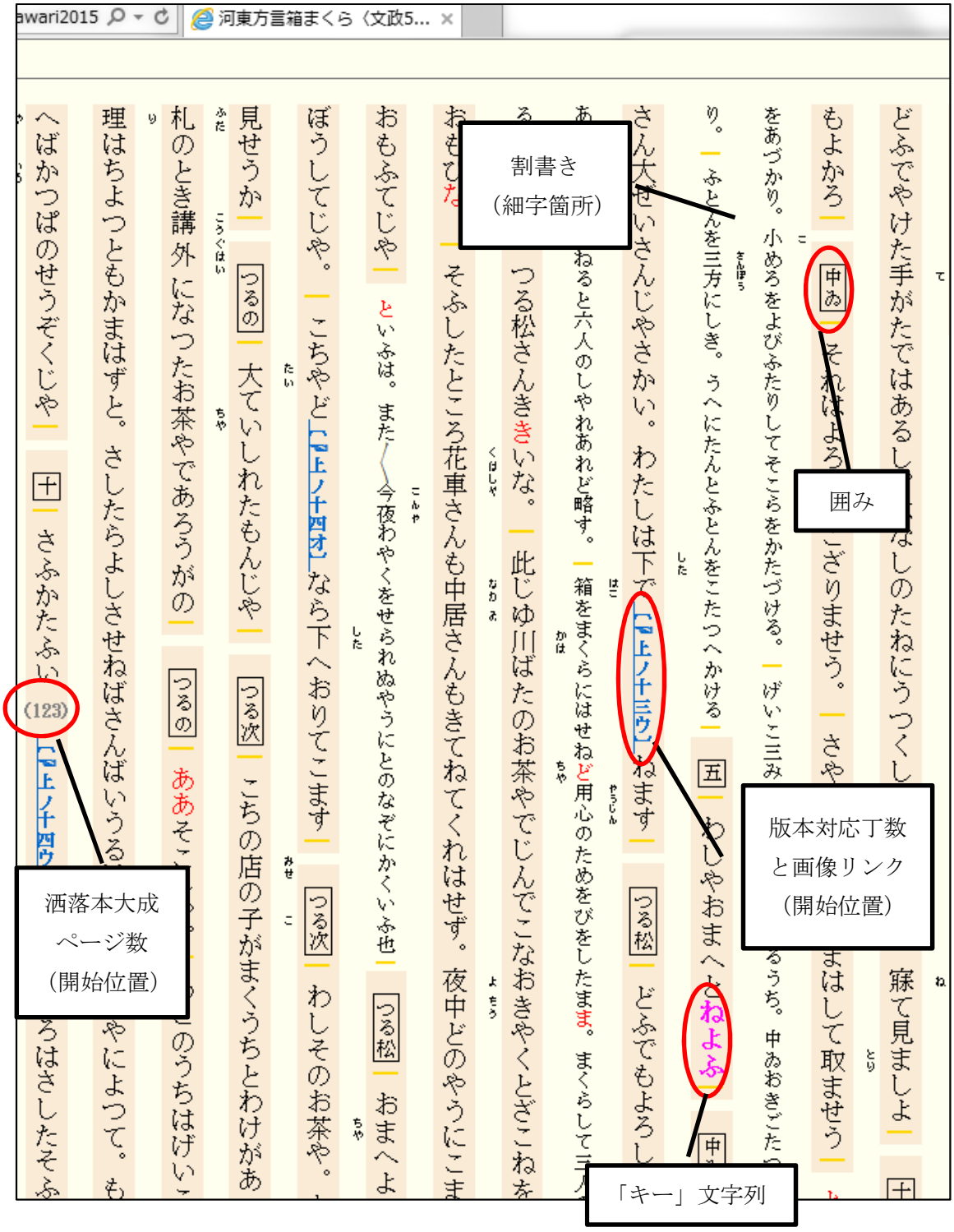


図 2　ブラウザ上の表示（Internet Explorer の場合）



以下では、Himawari・ブラウザ表示上の、主な表示項目とその内容に関して概説する。

［1］作品・底本の情報（タイトル～出版地）

●タイトル

書名タイトルは、原則『洒落本大成』の表示に従っている。

●底本巻

『洒落本大成』における対応巻

●ページ

『洒落本大成』における対応巻のページ数

●画像丁数

国立国語研究所蔵本等、対応する版本画像がある場合、「画像丁数」が半丁ごとに表示される。

インターネット接続環境下にある場合、この箇所を**ダブルクリック**すると、ブラウザが開き、対応する丁の画像にジャンプする。

［2］形態論情報（語彙素～発音形）

表示される形態論情報（短単位）は、形態素解析辞書 UniDic の見出しに対応している。

単位認定基準は、原則「中納言」上の『BCCWJ』・『平安時代編』・『室町時代編 I 狂言』や、ひまわり版『明六雑誌コーパス』等と同様である。小椋他(2011)などを参照されたい。

活用の型に関しては、『平安時代編』や『室町時代編 I 狂言』では主として「文語」活用、「BCCWJ」では無表示の口語活用が採用されていた。

これらの中間に位置する洒落本では、本文の状況により、**「会話」は「口語」ベース、「会話」以外は「文語」ベース**としている。

これは、会話文における文語二段動詞の一段化や、文語四段動詞の五段化が進んでいる一方、地の文・割書きにおいては未だ文語的表現が主流であったことによる。

例：語彙素「寝る」の場合

①会話文の場合

八 よいからいふ通り。どふぞ帯をといて**ね**ておくれ。

（河東方言箱まくら・動詞-一般・**下一段-ナ行**）…口語で処理

②割書きの場合

よふね入らぬと帯をとき上着のまま市村が**ね**てゐるところへはいる

（河東方言箱まくら・動詞-一般・**文語下二段-ナ行**）…文語で処理

以下、形態論情報に関して、注意すべき点を幾つか挙げる（**図３参照**）。

●語彙素・語彙素読み

「語彙素」は単語の各種語形・活用形・書字形（表記）を統合した辞書の見出しレベルの階層であり、一般的な漢字・仮名で表記される。「語彙素読み」はその読みを**カタカナ表記**した物である。語彙素で検索することで、同語彙素内の異語形・異活用形・異表記形等を、一括して取得することができる（**図１**は語彙素読み「ネル」で検索）。

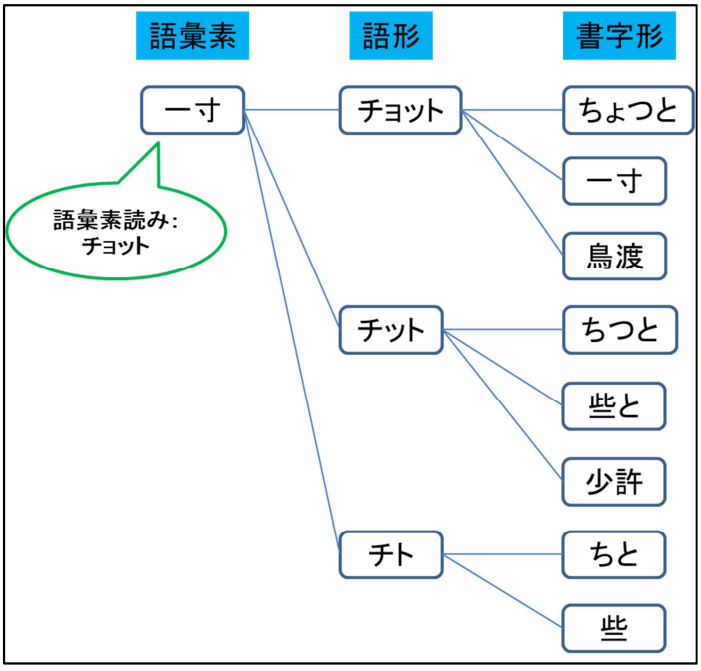
●語形

「語形」は、異語形を区別するレベルである。ただし、2015 年 10 月時点では、文語四段活用と文語上下二段活用は、別語形ではなく**別語彙素**として認定しているため注意。

●書字形

「書字形」は異表記を区別するレベルである。同語形でありながら、活用語尾を除いた箇所に別の文字符号が与えられる場合、それぞれ別の書字形となる。

図　３　語彙素「一寸（チョット）」の語彙素・語形・書字形



●品詞

学校文法における形容動詞は、語幹は「**形状詞**」、語尾は「助動詞」に分割されている。

●活用型

文語活用として処理されているものには「文語下一段」のように「文語」が表示されるが、口語活用には「下一段」のように「口語」は表示されない。（「文語四段」「五段」なども同様）

［３］本文情報と話者情報

本コーパスにおいては、本文情報と同時に詳しい話者情報が付与されている。検索結果では、次の**図４**のように表示される。

図 ４「洒落本コーパス」における本文情報と話者情報

本文種別	話者	性別	身分	年齢	地域	メモ
会話	白楽天	男	客【医者・...		上方	
割書						
割書						
会話	旦那	男	客【上層・...		田舎	四十余
割書						
会話	中居	女	店の者		上方	
会話	【芸子】つ...	女	娼妓・芸妓		上方	
割書						
割書						
会話	五兵へ	男	客【町人】		上方	手代風市兵衛仲間
会話	五兵へ	男	客【町人】		上方	手代風市兵衛仲間
引用-野暮輔評						
会話	【芸子】み...	女	娼妓・芸妓		上方	
割書						
割書						
会話	中居	女	店の者		上方	
割書						

●本文種別

性質の異なる本文を<q>でマークし、本項目で表示している。

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」が表示される。無表示の物は「割書」ではない本文である。

なお、割書き内の会話文風のものは、認定の難しさや性質の違いから、「会話」と認定していない。

●話者

「洒落本コーパス」における各会話文について話者表示の統一をはかり、「話者」覧に表示した。

例：　原文：花　→　「話者」列表示：置屋花車くま　（同一作品内で統一）

なお、「身分」が「娼妓・芸妓」の場合、判明する限りで、遊女の場合は【遊女】、芸子の場合は【芸子】と、それぞれ話者欄の先頭に付与した。

●性別

話者の性別を記述。

★記述内容：[男, 女, その他]

「その他」は、動物以外等による性別不明の生命体（例えば『鳩翁道話』に見られる「松茸」）や、複数人の会話文などを想定している。

●身分

話者の身分・職業等を記述。娼妓／芸妓、店員、客（【】にて詳細)、その他（【】にて詳細）等を大別。現状は洒落本に特化した記述であり、汎用性は薄い。今後も妥当性等を検討し、改良する予定である。

★記述内容：[娼妓・芸妓,店の者, 禿, 太鼓持, 使用人,
客, 客【上層・むすこ・通人】, 客【武士】, 客【医者・学者等】,客【町人】,
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】]

●年齢

子供、老人など、判明する限りで、比較的有標な事項のみを記述。今後も妥当性等を検討し、改良する予定である。

★記述内容：[子供, 老人 ,年少【こめろ・若者・でっち】 ,年長【母親など】]

●地域

話者の使用言語（出身地）について、本文記述や会話の言葉遣い等から、判明する限り記述。「上方板洒落本における江戸話者」などの抽出を想定している。

★記述内容：[江戸, 上方, 田舎]

●メモ

話者情報付与過程における作業上のメモ（年齢など）を表示している。

図５　参考　本コーパスのタグセット

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。本タグに話者情報を付与。
	@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

※ なお本稿および本コーパスに関するご意見・ご質問等がありましたら、筆者までお寄せください。

参考文献

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』 2012(1) 1-8

市村太郎・河瀬彰宏・小木曽智信(2013)「洒落本コーパスの構造化―仕様と事例の検討―」第 3 回コーパス日本語学ワークショップ予稿集 pp.249-258

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　明治書院

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011）「『現代日本語書き言葉均衡コーパス』形態論情報規定集第 4 版（下）」特定領域研究「日本語コーパス」平成 22 年度研究成果報告書　国立国語研究所

小椋秀樹・須永哲矢(2012)「中古和文 UniDic 短単位規程集」基盤研究(C)「和文系資料を対象とした形態素解析辞書の開発」研究成果報告書 2　国立国語研究所