近松浄瑠璃本のコーパス化—「語り」のテキストをどう扱うか

上野 左絵（国立国語研究所　言語変化研究領域）

日本の伝統的な演劇形式である浄瑠璃は、太夫によって語られる「語り物」である。そのためそれを文字化した浄瑠璃本は小説や戯曲とも異なる複合的な文体を持つ。本稿では日本語歴史コーパスの一環として構築中の「近松コーパス」における問題点を示す。特に浄瑠璃における「点」（句点）や文字譜、および掛詞の問題を取り上げ、これらの処理方法について検討を加える。

Construction of the Chikamatsu Joruri Corpus: Some Issues Arising from Its Narrative Styles

Sae Ueno (Language Change Division, NINJAL)

"Joruri", a traditional Japanese performing art, is a type of "Katarimono" (narrative) performed by a Tayu (narrator). The style of its text has complex factors that are different from those of novels, as well as plays. Therefore, some difficulties arise in constructing the Chikamatsu Joruri Corpus, as a part of The Corpus of Historical Japanese (CHJ). This article describes the problems caused by the discriminative style of Joruri texts, discussing, in particular, three issues: punctuation marks, Moji-fu (the musical notation), and Kakekotoba (a pivot word).

１．はじめに

国立国語研究所では「日本語歴史コーパス（CHJ）[1]」として、これまで平安、鎌倉、室町、近代の資料について形態論情報付きコーパスを構築・公開してきた。近世についても後期江戸語の言語実態を反映する資料として洒落本・人情本の試作版コーパスが公開されている。更に現在、近世前期上方語資料として近松門左衛門の世話浄瑠璃24作品に形態論情報を付した「近松コーパス」を構築中であり、これらの公開により日本語の研究資料が通史的に繋がることとなる。

浄瑠璃[2]のテキスト、特に近松門左衛門の世話物作品は、話しことばを反映した資料として日本語史、特に近世語・上方語研究における重要資料であるが、文体的には小説などの散文系テキストや演劇作品の台本と異なる部分が多い。そのためコーパス化にあたっては、テキストの構造化や形態論情報の付与といった過程において、浄瑠璃独自の様々な問題が生じるのである。

そこで本研究では、「近松コーパス」構築過程で明らかになってきた、浄瑠璃の文体の特殊性に起因する問題点を取り上げる。第２節でコーパスの概要および構築の手順を説明し、第３節で浄瑠璃テキストの文体的特徴を概説する。第４節・第５節では具体的な問題点として浄瑠璃の「点」とセンテンスの関係、および掛詞を取り上げ、形態素解析および形態論情報のデータベース化における処理について検討する。

２．「近松コーパス」の概要

「近松コーパス」は、近松門左衛門作の世話物浄瑠璃24作品を資料としたコーパスである。後述の通り浄瑠璃は演劇的・音楽的な側面を有するが、本コーパスの目的はあくまでも言語研究への利用である。底本には小学館の新編日本古典文学全集74巻『近松門左衛門集1』および75巻『近松門左衛門集2』を用いる。

本コーパスの構築において必要となる作業の中で大きなものは、「電子化」と「形態論情報の付与」の二つとなる[1]。

まず資料を電子テキスト化し、XMLにより文書の構造をマークアップする。「近松コーパス」の場合には、小学館より電子データの提供を受けることができ、紙テキストからのデータ入力という作業が省略された。また提供を受けたデータにはルビや頭注などの紙面情報、文書構造などがタグ付けされており、それらを利用しつつ、これまで国語研から公開されてきた「現代日本語書き言葉均衡コーパス」（BCCWJ）[2]やCHJ各サブコーパスの仕様を引き継ぐ形でタグの策定を行った。

電子化されたデータは形態素解析辞書UniDicを用いて解析し、語単位で形態論情報を付す。語の単位には国立国語研究所で規定した短単位

1 http://pj.ninjal.ac.jp/corpus_center/chj/

2 本論文中で「浄瑠璃」という場合、近松以降のいわゆる新浄瑠璃を指す。

（SUW）を用いて CHJ 全体としての一貫性を保つ。

「近松コーパス」ではここまでの作業を以下のように進めてきた。

1） 試行版としてまず 2 作品につき、時代的・文体的に比較的近いと考えられる狂言 UniDic を用いて解析
2） 解析データに人手修正を施し、2 作品分（約 16,500 語）の試行版データを作成
3） 試行版データと人手修正済みの洒落本データを学習用データとして用い、洒落本用 UniDic による追加 9 作品の再解析

このあと、

4） 人手による解析結果の修正
5） 再学習解析

を繰り返して解析精度の向上を図ることになる。

なお、再解析に際して洒落本用 UniDic を用いたのは、最初の狂言 UniDic による解析で期待したほどの精度が出なかったためであるが、CHJ 洒落本の形態素解析と同様、地の文と会話文を<speech>要素により区別して解析する手法[3]により、再解析時には精度が各段に向上した。

効率的な解析手法の開発に加え、校訂本文を用いるという点も形態素解析に有利に働いたと考えられる。たとえば CHJ 洒落本に関して市村・小木曽[3]では仮名遣いや漢字表記の多様性による解析精度への影響が指摘されていたが、新編日本古典文学全集では歴史的仮名遣いへの統一、異体字や宛て字の包摂が行われており、解析に有利に働くと考えられる。また人手によるデータ修正においても、校訂者による読みや解釈を利用することができるため、効率よく修正作業を行うことができる。「近松コーパス」は、条件的には比較的低コストでのコーパス構築が可能であるといえるだろう。

3．浄瑠璃テキストの特徴

浄瑠璃はそもそも「語り物」であり、散文や韻文、更には戯曲などの要素を兼ねた複合的な文体を持つ。そのため前項で述べてきた構築過程で、独自の問題が生じることとなった。具体的な問題点を挙げる前に、まずは浄瑠璃本テキストの特徴を明らかにしておきたい。以下に浄瑠璃本の持つ文体的性格を列挙する。

① 戯曲・台本としての性格
浄瑠璃は読み物とは違い人形劇を「演じる」ための台本である。そのため多数のせりふが挿入されたテキストになっている。

② 語り物的性格
歌舞伎台本などの一般的な戯曲・台本は話者表示とせりふ、ト書きなどから構成されることが多い（図 1）。浄瑠璃も①に挙げたように基本的性質は演じるための台本なのだが、「せりふ」「ト書き」という表示形式を取らない。これは基本的に太夫が一人で語る[3]という浄瑠璃の「語り物」としての性格によるものであろう。そのため「せりふ」「地の文」という区別が曖昧な部分も多く、構造化にあたっては一定の方針を定める必要がある。
戯曲以外にも、せりふが多用される作品においては話者表示を伴う場合があるが（図 2）、このような表示も浄瑠璃本では行われない。地の文、会話、ト書きがすべて語りの中で、見た目上の分かちのないまま文字化されている（図 3）。

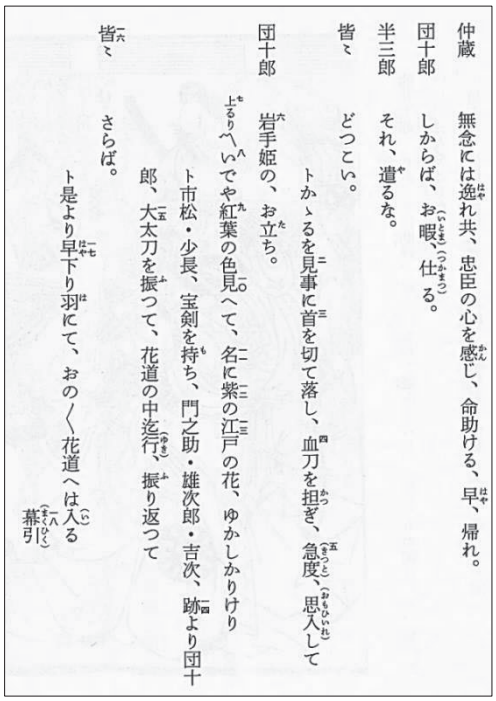

仲蔵　無念には逸れ共、忠臣の心を感じ、命助ける、早、帰れ。
団十郎　しからば、お暇、仕る。
半三郎　それ、遣るな。
皆〳〵　どつこい。
トかゝるを見事に首を切て落し、血刀を担ぎ、急度、思入して
団十郎　岩手姫の、お立ち。
上るりへいでや紅葉の色見へて、名に紫の江戸の花、ゆかしかりけり
ト市松・少長、宝剣を持ち、門之助・雄次郎・吉次、跡より団十郎、大太刀を振つて、花道の中迄行、振り返つて
皆〳〵　さらば。
ト是より早下り羽にて、おの〳〵花道へは入る
幕引

図 1　岩波『江戸歌舞伎集』「御摂勧進帳」一部（p.136）

③ 音曲としての性格
浄瑠璃は音楽劇であり、その本文には節付けがされている。そのため語りの本文の他に、曲節が記号化して示されている（文字譜・節博士）。その他にも語る際の息継ぎ、場面転換のための間合いなどといった情報が様々な符号で記されており、上演や伝承に際して重要な意味を持つ。

④ 韻文としての性格

[3] 場面によっては複数の太夫により役を担当して交互に詞を語ることもあるが、その場合も会話部分に続く地の文は同じ太夫に続けて語られる。

特に「景事」「道行」などと呼ばれる音楽的章段は多く七五調であり、語の選択にもリズムを重視した部分が見られ韻文的である。また掛詞や物尽くしといった修辞、謡や小唄の引用といった特徴が多く見られる。

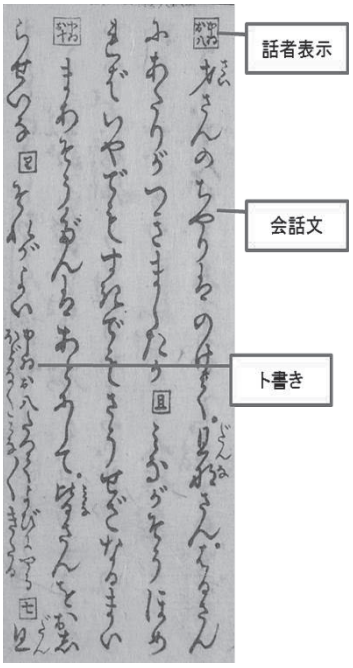
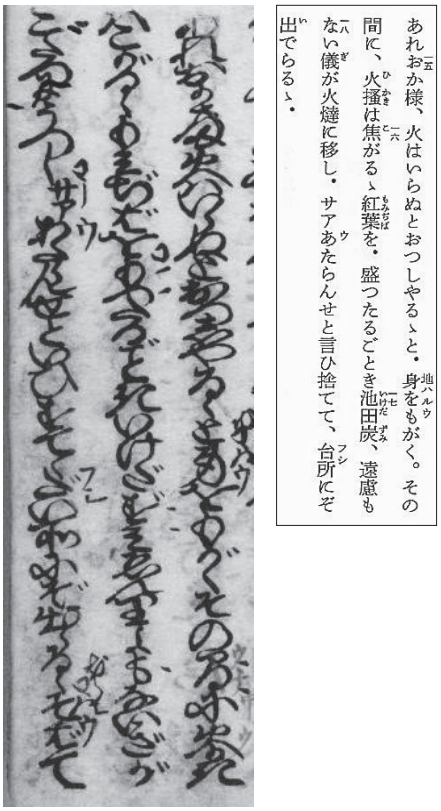


図 2　洒落本『河東方言箱枕』



あれおか様、火はいらぬとおつしやるゝと・身をもがく。その間に、火搔は焦がるゝ紅葉を・盛つたるごとき池田炭、遠慮もない儀が火燵に移し・サアあたらんせと言ひ捨てて、台所にぞ出でらるゝ・

図 3　『心中重井筒』原本と校訂本文の例

このように様々な文章の性格が混在しているところから、浄瑠璃の構造化や形態素解析に際しては様々な問題点が生じてくる。以下ではその中から、上記③から派生する句点や文字譜の問題と、④に関連する掛詞の問題について詳述する。

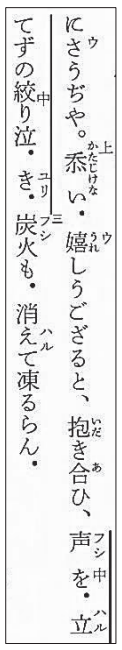
4．構造化に際しての問題

浄瑠璃のテキストに形態素解析を施す際に生じる問題の第一は、「点」の問題である。近代以降文章においては文の区切りを表すのに句点を用いるが、浄瑠璃本においても本文中に句点に近い記号が用いられている。ただし記号としては「。」だけではなく、「.」や「丶」（なみだ点）なども用いられ、その用法も現代のいわゆる「句点」とは異なる部分がある。そのためここでは浄瑠璃本における句点相当の記号を単に「点」と呼ぶ。

浄瑠璃の「点」については、一般に以下のように説明される。

> “一、浄瑠璃の句点の「。」は、音曲的な句切りを示すもので、あながち文意にそって付されたものではない。”[4]
>
> “一、浄瑠璃の句点の「。」は息継ぎの意味があり、散文の場合とは異なる（下略）”[5]

つまり浄瑠璃における「点」は必ずしも文の切れ目を意味しない。近代的文章においては、句点をマークしておく（<s>：TEI5 に準拠）ことにより、UniDic を用いた形態素解析時、これを境界認定に利用することができるのだが、浄瑠璃本においては単に句点記号のある箇所を文境界とみなすことはできない。そればかりか、時には一形態素を分かつ場合も見られ（図 4）、形態素解析の支障となる場合もある。



にさうぢや。糸い・嬉しうござると、抱き合ひ、声を・立てずの絞り泣・き・炭火も・消えて凍るらん・

図 4　語を分断する点の例

なお浄瑠璃では形態素を分かつことがある記号類として、他に場面転換の間などを示すいおり点（〽）がある。たとえば一般的な戯曲であれば、場面の転換はト書きで表され、それが語の内部に入り込んで分断することはない。浄瑠璃のこのような例は、「語り」による表現・演出の特殊性を示すものといえるだろう。

このような「点」への対処として、付されてい

る「点」について人手で(a):<s>相当のもの　(b):<s>に合致しない、音曲的な句切りや息継ぎを意味するものに分類した。

前者については<s>タグを付すことにより文境界としての機能を持たせたが、後者(b)については別のタグの設定が必要となった。しかし、上記のような例を想定した要素はTEI P5[6]では見当たらない。強いていえばspoken Textsに対する<pause/>が近いが、<pause/>は談話における実際の「間」を想定するもので、音楽的指示を示すものではない。また<caesura/>は韻文における区切りを示すものであるが、韻律のlineを区切るものであり、意味単位の分断は想定されていない。浄瑠璃の「点」(b)に関していえば、出現環境としては<pause/>が近いが、タグの表す意味としては<caesura/>が近いということになる。

なおCHJにおける「近松コーパス」は、先述の通りあくまで言語資料としてのコーパスの提供を目的としており、演劇・音曲研究に直接資することを意図していないため、その設計には文字譜・曲節情報の公開は含まれていない。しかし形態素の途中であらわれるこのような「点」や記号は、浄瑠璃の演出をあらわすものであり、また息遣い（ブレス）を注記するということは芸能の伝承という側面にも大きく関わってくるものである。言語処理上はともかく、データの価値としては単純に無視することはできない。

文字譜のような音曲・実演に関わる情報は本コーパスでは要素として採用しないが、現時点では元データで使用されたいた<fushi>というタグをそのまま採用し、文字譜の種類をtext属性値に収めて出現位置を示している（図5）。

```
<fushi text="フシ"/>声<fushi text="中"/>を
<caesura text="."/><fushi text="ハル"/>立てず
の<fushi type="下つき" text="中"/>絞り泣
<caesura text="."/><fushi text="ユリ"/>き
<caesura text="."/>
```

図 5　図4傍線部分XML

これらの文字譜は、西洋音楽における楽譜のような、音階やリズムをあらわすものとは全く異なるものである。文字譜の種類には、たとえば「ウ」「ハル」といった音の高低に関するもの、「詞」「地」などのせりふ廻しに関するもの、「謡」「順礼」といった、音曲に関するものなどがある。どちらかといえば音符・楽譜というよりは、演出指示と言った方が近いかもしれない。

図6は「近松コーパス」における文字譜の出現度数トップテンを示したものだが、文字譜総数10,151のうち、約半数は音の高低を表すものであることがわかる。これらの譜は、基本的には本文における位置（どの音が高く発音されるか）が示されていればよく、現状の仕様で情報としては十分である。しかし「詞」「色」といった語り方や音曲に関する譜では、「その語り方／音曲はどこまで有効か」という範囲が必要となる。

文字譜	出現度数
ウ	4029
ハル	1661
詞	946
色	877
フシ	765
中	734
地ハル	371
地色中	283
地色ハル	265
地色ウ	220
総計	10151

図 6　文字譜上位10種

今回は言語研究のためのコーパス作成を目的としているためすべての<fushi>がempty elementになっている。しかし例えばある曲節の終わりを示すための「ナヲス」という譜が存在する、「詞」と「地」は交互にあらわれる場合が多いなど、語り方や音曲の範囲を示す手掛かりは文書上に示されている。少し手を加えれば、これらの要素をある程度自動的にマークアップすることは可能であろう。

邦楽においては歴史的に、テキストに直接曲節や譜を書き入れ、そのテキストを出版することが行われてきた。浄瑠璃をはじめ謡曲などの構造化にも、こういった音曲的要素を示すための要素の検討が必要となるのではないか。

こういった文字譜の種類や対応についての研究とマークアップへの応用は、今後の課題としたい。

5．掛詞の問題（データベースの拡張）

浄瑠璃テキストにおけるもう一つの大きな問題は、掛詞・物尽くしといった修辞技巧の多用である。中でも掛詞は、一つの語単位が多重の形態論情報を持つものであり、複数の情報をどのようにコーパスに反映させるかが問題となってくる。

「掛詞」は和歌において発達してきた修辞法であり、CHJ「万葉集コーパス」では和歌における掛詞の情報付与について検討されている[7]が、浄瑠璃における掛詞は、伝統的な作法に基づく和歌に比べて自由度が高く、言葉と言葉の「掛け方」

のパターンも、また掛詞に用いられる語の種類も多い。

表 1 坂本（1987）の掛詞分類

分類	音	拍	用例
A	○	○	はねと／＼をあはせのその（合せ／袷） （曾）
B	×	○	今は又よそにあらしの身にぞしむ（有らじ／嵐）（絵）
C	○	×	内包型：たがひにこひ茶のはつむかし（恋ひ／濃茶） （潤） 鎖型 ：心はうちやう天王寺（有頂天／天王寺）（紅）
D	×	×	たよりなぎさに立めなみ（無き／渚） （薩）
E	「物尽し」の類における兼用語		互の心ふとぬのゝなごりは一たんたちきつて（一旦／一反）（薩）

近松世話物浄瑠璃の掛詞については、先行研究において掛ける語と掛けられる語の音・拍数の移動により分類されている[8]。表 1 は[8]の記述をまとめたものである。

A～Eの分類のうち、Eの「物尽し」というのは、単にある語の裏にテーマに基づいた別の意味を持たせるという趣向である。例では副詞の「一旦」に、裁縫に関する語として布の長さの単位「一反」が隠されているのだが、「一反」の意味が表に出なくても文脈上は問題がない。本コーパスではこのような例は掛詞として扱わず、複数の意味がそれぞれ文脈に関わってくる場合に限り多重の情報を付すこととした。

「近松コーパス」の大きな特徴は、掛詞のような多重の構造を持つ文章に対応できるよう、データベースの多層的拡張を行う点にある。主となる形態論情報を収めたテーブルに加え、掛詞により加えられた二つ目の情報を収めるテーブルを設けることにより、多重の形態論情報を同時に扱い、また検索の際にもそれぞれの語が検索できるような設計を行う（図 7）。

なお、コーパス構築のコストの面から考えれば、掛詞として用いられている語のどちらかだけを採用するというやりかたもあるだろう。しかしCの例、特に鎖型のように二つの語が一部で繋がる形である場合、どちらかの語を採ると意味をなさない語の断片が残ってしまう。

わかりやすい例として図 7 では『五十年忌歌念仏』の「とはかは口にぞ着きにける」という部分を挙げている。この場合、下に続く「川口」を語として切り出すと、上から続いている「とはかは」の一部である「とは」が意味をなさない断片として残り、副詞の「とはかは」が語彙データから抜け落ちてしまうことになる。このような事態を防ぐためにも、多重の情報付与とその管理が必要となるのである。

作業の準備としてはまず、掛詞にあたる語の認定を行うが、これは各種校訂や底本の頭注などに拠っている。次に認定されたそれぞれの掛詞のペアについて主（メイン）と副（サブ）を設定する必要があるのだが、本コーパスでは、下に続いていく語を主、上から続いていく語を副と定めた。

UniDic による形態素解析後、人手で短単位修正を行う際に、主になる語の情報を短単位（主）

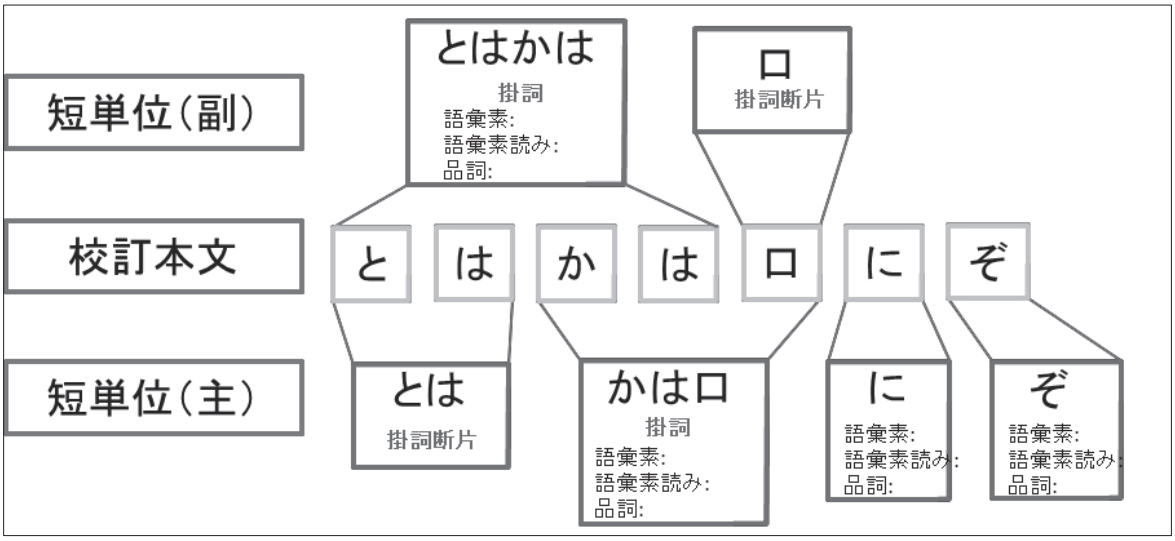


図 7 掛詞拡張イメージ

テーブルに、副の意味を短単位（副）テーブルに入れる。

C・鎖型の掛詞の場合には、まず掛詞の主となる語を切り出して情報を付与し、残りの部分には「掛詞断片」という情報を付けて解析から外す。掛詞の副となる部分に対しても、拡張テーブル（図 7 の短単位（副）テーブル）で同様の処理を行う。こうしてC・鎖型の掛詞にも、多重の形態論情報の付与が可能となるのである。

パターンの自由度が高い掛詞の処理方法を策定し、多重性を持つ語の扱いが可能になることにより、近世後期資料、特に CHJ 洒落本コーパスでも懸案であった（[9][10]）、地口や洒落といった修辞の処理への応用も期待できる。

６．まとめ

本稿では、浄瑠璃という文体の複合的な特徴について検討し、コーパス化に際してそのテキストをどのように扱うべきか、「点」と文字譜、掛詞という三つの大きな問題を取り上げつつ述べた。

その中で「点」や「文字譜」の問題は音曲に絡み、テキストのマークアップを目的とする TEI の基準に合致しない部分もあるが、こういった要素を策定することにより謡曲などの歴史的文献や音楽関連のテキストへの応用も期待される。

また掛詞を目的とした、データベースの拡張は本コーパスの大きな特徴である。まだ設計段階ではあるが、掛詞にとどまらず広く応用可能なシステムとなるであろう。

「語り」の文体を持つ資料は、浄瑠璃にとどまらず、謡曲や平曲、落語など多様である。特に落語に関しては近代以降多くの速記本が残されており、口語資料としての価値も高い。「近松コーパスの」設計は、これらのテキストにも援用できよう。それぞれの文体性や時代による違いを精査しつつ、他資料への応用の可能性を検討していきたい。

参考文献

[1]田中牧郎：『日本語歴史コーパス』の構築，『日本語学』，33-14，pp.56-67（2014）

[2]山口昌也，高田 智和，北村 雅則，間淵 洋子，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. 2.0，（2008）

[3]市村太郎，小木曽智信：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，pp.107-110（2016）

[4]信田純一校注：凡例，新潮日本古典集成『近松門左衛門集』，p.3（1986）

[5]土田衞校注：凡例，新潮日本古典集成『浄瑠璃集』，p.3（1985）

[6]Text Encording Initiative: TEI Guidelines, http://www.tei-c.org/Guidelines/

[7]鴻野知暁：万葉集コーパスの設計，国立国語研究所「通時コーパス」国際シンポジウム発表ポスター（2015）

[8]坂本清恵：義太夫節の掛詞―近松世話物浄瑠璃譜本を資料にして―，国文学研究，Vol.93，pp.42-56（1987）

[9]市村太郎，河瀬彰宏，小木曽智信：洒落本コーパスの構造化 ―仕様と事例の検討―，国立国語研究所 第 3 回コーパス日本語学ワークショップ予稿集，pp.249-258（2013）

[10]河瀬彰宏，市村太郎，小木曽智信: TEI：P5 に基づく近世口語資料の構造化とその問題点，じんもんこん 2013 論文集，Vol.2013，No.4，pp.7-12（2013）

以上の文中に挙げた用例は小学館『新日本古典文学全集』を底本としている。各用例に作品名を略称で記載した。以下の通りである。

用例出典一覧

（曾）『曾根崎心中』　（絵）『心中二枚絵草紙』
（潤）『卯月潤色』　（紅）『卯月紅葉』
（薩）『薩摩歌』

本研究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）の研究成果を報告したものである。