近世後期口語資料の形態素解析

—ルビ情報を利用した精度向上の試み—

村山 実和子（九州大学大学院生，国立国語研究所）
銭谷 真人（早稲田大学大学院生，国立国語研究所）
藤本 灯（国立国語研究所）
岡 照晃（国立国語研究所）

近世後期の口語資料では「心驚（びつくり）」「誘引（さそは|れ）」のような特殊かつ使用頻度の限られた漢字表記が多く，形態論情報付きコーパスを構築する際の自動形態素解析の精度向上の妨げとなっている．従来の手法では振り仮名は解析には一切使用されてこなかった．そのため特殊な漢字表記を解析しても，表記と振り仮名との乖離によって，振り仮名で示される語として解析結果が得られない問題があった．この問題に対し本発表では，主に以下の２つの方法を用いた対処について述べる．まず自動形態素解析の前処理として，特殊な漢字表記を平仮名で表記された振り仮名に置換する．そして自動形態素解析用辞書に仮名形のフィールドを基にした平仮名の書字形を追加する．この手法を用いた結果，辞書未登録の特殊な漢字表記だけでなく，「誘引（さそは|れ）」のような複数語にまたがる漢字表記も正しく解析できるようになった．

Morphological Analysis of Early Modern Japanese: An Approach to Improve the Analysis Precision Using the Ruby Information

Miwako MURAYAMA (Kyushu University, National Institute for Japanese Language and Linguistics)
Masato ZENIYA (Waseda University, National Institute for Japanese Language and Linguistics)
Akari FUJIMOTO (National Institute for Japanese Language and Linguistics)
Teruaki OKA (National Institute for Japanese Language and Linguistics)

We can see many special and rare usages of Kanji characters in the historical documents written in spoken language of Edo era, such as "心驚 (bikkuri) ", "誘引（sasowa-re）". Although we use an automatic morphological analyzer for creating a word-segmented and pos-tagged corpus, such notations prevent the analyzer from improving its performance. Ordinal automatic analysis does not use information of ruby. Therefore it does not treat the special notations; it does not understand the meaning of which ruby has from the Kanji notation. In this presentation, for dealing with this problem, we replace the special Kanji notations with its ruby characters and analyze by using morphological analysis dictionaries which has surface-forms created from kana-fields in the original dictionaries. As some results of these methods, we can correctly analysis not only special notations that are not resisted our dictionaries, but also notations that are across several word boundaries, such as "誘引（sasowa-re）".

1. はじめに

国立国語研究所では現在，**日本語歴史コーパス**（以下，CHJ）の構築に取り組んでいる．CHJでは古代語～近代語についての代表的な資料を集め，日本語の通時的研究に利用可能なコーパスとして順次公開を進めている．このうち「江戸時代編Ⅰ・Ⅱ」としてコーパス化の対象としている洒落本と人情本は，発話部分に当時の話し言葉が反映されており，近世語の実態を明らかにし，中世から近現代にかけての日本語の変化を探る上で極めて重要な資料である．これらは**洒落本コーパス**，**人情本コーパス**として，それぞれすでに試行版が公開されている．しかしながら，これらの試行版を構築する際に，近世特有の表記の多様性が，形態論情報付与の低コスト化のために導入された自動形態素解析を妨げるとして，以前より問題視されてきた[1]．特に顕著な事象として，以下のような現代では一般的ではない漢字表記と振り仮名との組み合わせがある（用例は洒落本『玉菊全伝花街鑑』より引用．括弧内は原本に付された振り仮名）．

(1) 恋に憔悴（やつるゝ）面影を
(2) 夕べおやしきの衆に．誘引（さそはれ）て

自動形態素解析においては，本行の漢字表記を解析対象とするため，専用の解析用辞書に登録済である「憔悴（ショウスイ）」「誘引（ユウイン）」

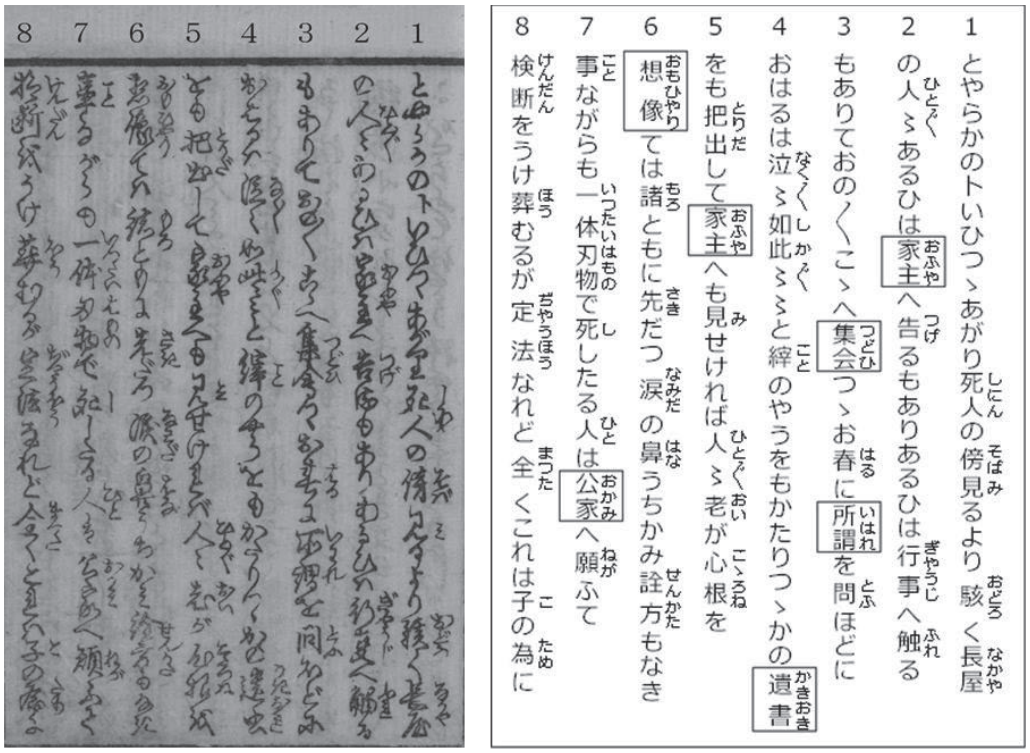

図１　版本画像（左）とその翻刻テキスト（右）
底本：国立国語研究所所蔵『比翼連理花廼志満台』二編上17丁表

という語に解析されることになるが，これは原本に付された振り仮名とは対応せず，誤解析となる．さらに(2)の場合，本文の読みとしては動詞「誘う」＋助動詞「れる」の二単位に解析されるのが望ましいのに対し，解析結果は一単位相当になるという，単位数の齟齬も生じる．

本発表では近世後期口語資料に特徴的，かつ誤解析の原因となりやすい特殊な漢字表記，いわゆる**当て字**の実態を明らかにし，自動形態素解析に際して生じる課題と，その解決手法を提案する．

2. CHJ 江戸時代編の概要

2.1 CHJ 江戸時代編の現状

CHJ のうち，江戸時代編では，近世後期の有用な口語資料である洒落本と人情本のコーパス化を進めている．このうち洒落本コーパスは，『洒落本大成』を底本とした [2]．文書構造や話者情報を XML 形式で付与し，さらにテキストデータを国立国語研究所の規定した言語単位である**短単位**に分割し，各短単位に**形態論情報**（品詞，活用形，読みなど）付与を行っている（短単位の詳細は後述する）．現在，『聖遊郭』,『河東方言箱まくら』,『玉菊全伝花街鑑』の三作品について，「ひまわり版『洒落本コーパス』Ver.0.5」[3] を試作公開中である．人情本コーパスについては，江戸期の版本を底本として新たに翻刻を行い（図 1），振り仮名情報付きの XML データを構築している（図 2）．こちらも「ひまわり版「人情本

<pb n="一七オ" num="43"/>とやらかの</s></speech><s><char scriptt="カタカナ">と</char>いひつ<odoriji originalText="ゝ">つ</odoriji>あがり<r rt="し">死</r><r rt="にん">人</r>の<r rt="そば">傍</r><r rt="み">見</r>るより<r rt="おどろ">駭</r>く<r rt="なか">長</r><r rt="や">屋</r><lb/>の<r rt="ひと">人</r><r rt="〳〵">〻</r>あるひは<r rt="おふや">家主</r>へ<r rt="つげ">告</r>るもありあるひは<r rt="ぎやう">行</r><r rt="じ">事</r>へ<r rt="ふれ">触</r>る<lb/>もありて

図2　振り仮名の構造化タグの例
<r rt="振り仮名">本行の文字列</r>

コーパス」Ver.0.1」[4]として，『比翼連理花廼志満台』の試作版を公開中である．現時点では，本文および振り仮名を対象とした文字列検索のみが可能であり，今後，形態論情報を付与したコーパスを公開する予定である．いずれのコーパスも全文検索システム「ひまわり」[5]上で利用可能である（図 3）が，将来的には，その他の CHJ 作品と同様，Web 上のコーパス検索アプリケーション**中納言**[1]における利用を想定している．中納言上での公開時には洒落本・人情本ともに対象作品を拡充し，また人情本は形態論情報付きのコーパスとして公開すべく，現在も開発を進めている．

[1] https://chunagon.ninjal.ac.jp/

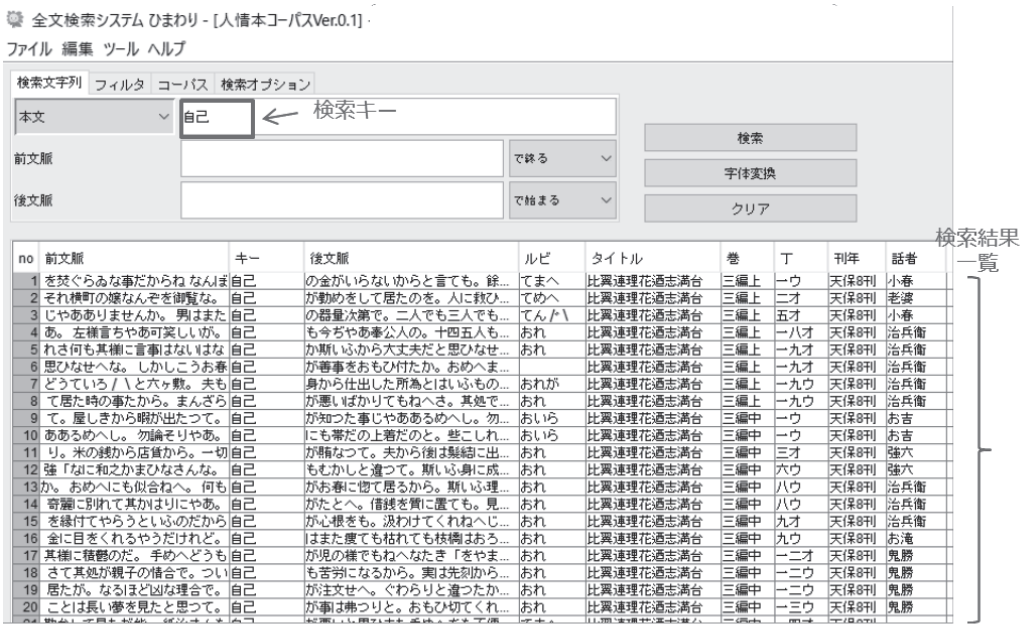

図3 ひまわり版「人情本」コーパスVer.0.1 検索画面

2.2 形態論情報の付与

生テキストに対し一から人手で形態論情報を付与することは人員と時間の非常にかかる作業である．そのため CHJ では，まず自動形態素解析器 MeCab[6]および自動形態素解析器用の辞書 UniDic を用いた自動解析を行い，その解析結果を人手修正していくという作業方針を採用している[7]．江戸時代編でも，辞書に近世口語 UniDic を使用し，人手による修正作業を実施している．UniDic はもともと現代語のコーパスへの形態論情報付与のために開発された辞書である[8]．すべての見出し語について**短単位**という斉一な言語単位が設定されており，各見出し語は，語彙素，語形，書字形，発音形という階層構造（図 4）を持っている．この階層構造により，コーパス検索の際，表記のゆれや語形の変異にかかわりなく，網羅的に用例を収集することが可能となっている．現在，古文解析用の辞書においても，これらの特性はそのままに，必要な見出し語を補充，語の単位を修正するなどして，各時代のテキストに対応させている．

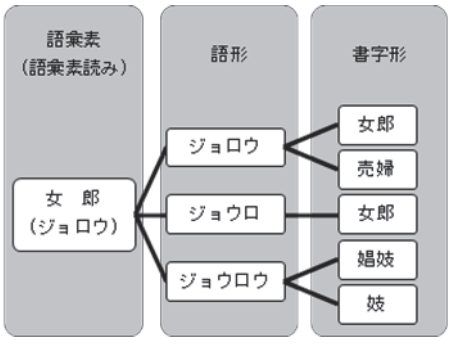

図 4 UniDic の階層構造（発音形は省略）

3. 近世後期口語資料の特徴と問題

3.1 自動形態素解析時の問題とその要因

CHJ のうち既に公開されている平安時代編，室町時代編 I の狂言では，コーパスの拡充に比例して各時代用 UniDic の解析精度も向上していった．洒落本でも，2015 年 3 月時点での解析精度（F-1 値：発音形まで推定）は約 86 であったが，専用の解析辞書の作成，および文体に合わせた辞書の使い分けにより，現在およそ 90 の解析精度を実現している[9]．しかし公開済の他の時代のコーパス（解析精度約 96）[10]に比べると，この値は低い．人情本も同様であり，試験的な自動形態素解析の結果，精度は約 87 であった[11]．そこでエラー分析を実施したところ，誤解析の要因として特に目立ったのが，(1)(2)のような本行の漢字表記に付された振り仮名が，国語辞書に登録されているような一般的な読みと一致しない例であった．

3.2 振り仮名と漢字表記

近世後期口語資料の表記上の特徴として，振り仮名付きの漢字が多用されることが挙げられる[12]．中でも人情本の振り仮名付与率は突出しており，調査対象中の漢字の 86%に振り仮名が付されていたという報告もある[13]．洒落本や人情本などの通俗小説類は大衆向けの出版物であり，難しい漢語が読めない読者でも，振り仮名と本行の平仮名の部分だけを目で追っていけば，内容を理解できるように工夫されていた．ただし，現代において漢字表記とその読み方が固定的であるのに対し，近世における漢字表記と振り仮名の組み合わせは，より自由度の高いものであった．図 1

に挙げた版面を見ても，「家主（おふや）」「遺書（かきおき）」「想像（おもひやり）て」といった現代語では一般的でない組み合わせが存在する（翻刻テキスト内，枠で囲んだもの）．「江戸の戯作の場合，誤読難読の恐れのある語に振り仮名を付けるだけでなく，作者の表現手法として漢字表記と並立するような振り仮名が付けられる」[14]ため，近世の版本における振り仮名は，本行文字列に対して従属的あるいは補足的なものではなく，本行文字列と併せて本文の一部，もしくは本文そのものとして捉える必要がある．

しかしながら形態論情報を付与する際は，本行文字列を XML の本文文字列として使用している．そして自動形態素解析の際には本文文字列のみが解析対象となり，振り仮名の情報は一切参照しない．その結果，表記が多様かつ，漢字とその読みが一般的な対応をしない，いわゆる**当て字**の場合，その表記が UniDic に未登録であるために誤解析が生じやすくなる．

3.3 当て字の実態

近世版本においては，判読のために漢語にわかりやすい振り仮名を付したと考えられる訓もあれば，その場面において最も相応しい漢字を当てたと思われる振り仮名もある．それが近世において通用であったか，イレギュラーなものであったかを明らかにするには，近世日本語に対する網羅的な調査が必要になる．しかし解析精度を向上させようとする以上，イレギュラーであると考えられるものを一旦規定する必要がある．そこで本研究においては，以下の特徴を有する漢字表記を当て字として取り上げる．

A) 振り仮名と漢字が文字単位で対応しない
B) UniDic に書字形が未登録である
C) 国語辞書等でも通用と認められていない

なお，通常の音訓からはずれるものとしては，「何処（どこ）」「明日（あした）」といった熟字訓も存するが，このように広く一般に通用しており，すでに解析辞書に登録が見られるものは当て字の例からは除外する．当て字には，次に挙げるように，語単位のもの（表 1），短単位をこえるもの（表 2）など，様々なパターンが見られる．（『玉菊全伝花街鑑』『比翼連理花廼志満台』から例を挙げた）

表 1 のような例に関しては，UniDic に書字形を追加することで処理は可能になる．しかし，作品によって，作者や時代背景，舞台となる地域が異なる洒落本・人情本はテキストの均質性が低く，表記のバリエーションは莫大なものとなることが予想される．そのように特殊かつ使用頻度の少ない語を新たに登録し続けることは UniDic の構造上も負担が大きく，そもそも通常の音訓・表記と同レベルのものとして扱うべきかどうか，検討する必要があろう．また人情本コーパスでは，一作品のテキスト量が膨大であることから，大部分を非コアデータ（人手修正を行わず，自動解析のみを行ったデータ）として公開する予定であり，自動解析の精度の低さは大きな問題となる．表 2

表 1　語単位の当て字の例

漢字表記	振り仮名
花街	さと
一図	ひとすじ
一面	べた
横雲	しのゝめ
温泉	とうぢ
佳美	はなやか
佳味	うめへ
家業	なりわい
快然	こゝろよく
懐中	ふところ
滑稽	しやれ
偽惑	うたぐり
仇心	うきな
強顔	つれなく
嫌疑	うたがひ
心驚	びつくり
蒼天	うららか
両親	ふたり
得心	きゝいれ
風評	うはさ
完爾	につこり
沢山	たんと
女児	むすめ
息災	たつしや
白眼	にらみ

表 2　短単位をこえる当て字の例

漢字表記	振り仮名
与	と｜は
全快て	よく｜なつ｜て
獣子	ばかげ｜た
看官	みる｜ひと
密通	わるい｜こと
誘引	さそは｜れ

のように短単位をこえるパターンも，現状の漢字表記のままでは適切な形態論情報の付与ができない．このように当て字の頻出する近世後期口語資料を用いたコーパスの構築にあたっては，振り仮名の情報を利用した新たな解析方法の開発が求められる．

4. XML 構造化タグを利用した形態素解析

当て字を自動形態素解析するため，本研究では当て字の本行文字列をXMLのルビタグの属性として格納されている振り仮名に置換して（以下，この操作を「ルビを開く」と表現）解析を実施する．置換対象となるルビタグはあらかじめ前述の基準で選定し，type属性に「当て字」という値を設定した．この作業は今回すべて人手で実施したが，[15]で提案されている当て字の自動検出を用いた自動化を検討中である．

[9]では，洒落本の形態素解析を文体別に分け，それぞれ専用の辞書を使った解析を実施していた．本研究でも同様に口語，文語それぞれ専用の辞書を用意する．ただしルビを開いた平仮名表記にも対応するため，辞書の仮名形出現形のフィールドから仮名表記（カタカナ）を取り出し，平仮名に置換した後，辞書のキーとなる表層形，および書字形出現形のフィールドと置換し，辞書の新たなキーとして追加を行った．これにより辞書の登録キー数はおよそ2倍のサイズとなった（約300万）．また各短単位に対し，新たに1個フィールド（列）を追加し，そこに置換前の漢字表記での書字形を残した．置き換えを行っていない場合は，キーと同じ文字列がここに格納されている．

上記の辞書をそれぞれ下記の短単位情報アノテーション済みコーパス上で文語・口語の文体別に学習を行った．

甲駅新話，阿闍陀鏡，北華通情，興斗月，新月花余情，陽台遺編・妣閣秘言，風流裸人形，異本郭中奇譚，京都_箱まくら，粋の曙，花街鑑，花街寿々女，跖婦人伝，遊子方言，傾城買四十八手，繁千話，傾城買二筋道，郭中奇譚，侠者方言，聖遊廓，月花余情

またコーパスは通常の本行表記だけでなく，本行表記を平仮名化した仮名形出現形に置き換えたものも併用した．使用したコーパスを訓練9：評価1に分割して精度を評価した.分割の結果，文語の訓練用コーパスは12,447文，107321短単位，評価用コーパスは1383文，12,351短単位となった．また口語の訓練用コーパスは13176文，165534短単位，評価用コーパスは1464文，17274短単位となった．評価結果を表3に示す[2].

表3 自動形態素解析精度（F1-値）

		境界認定	品詞認定	語彙素認定	発音認定
文語	ALL	96.36	92.39	91.20	90.71
	書字形レベルの未知語作成	91.48	85.59	83.81	83.14
	語彙素レベルの未知語作成	95.52	91.39	90.14	89.71
口語	ALL	97.06	93.53	92.73	92.43
	書字形レベルの未知語作成	94.15	89.36	88.39	88.01
	語彙素レベルの未知語作成	96.81	93.25	92.43	92.14

上記で作成した辞書を使い，当て字属性の付与されたルビタグを開いたXMLの本行文字列を解析した．その結果，辞書未登録の「口誑（こうぢやう）」のような特殊な漢字表記を「口上，名詞-普通名詞-一般」のように解析できるようになっただけでなく，「誘引（さそは|れ）」のような複数語にまたがる漢字表記も正しく解析できるようになった．

5. おわりに

本研究では，誤解析の原因となりやすい特殊な漢字表記（当て字）が多数出現するテキストに対して，(1) XML構造化タグを利用して，解析対象を漢字表記から振り仮名に置き換え，(2) 形態素解析用辞書に平仮名の書字形を追加することで，解析精度の向上が可能であることを明らかにした．特に，斉一な言語単位を設定している歴史コーパスにおいて，「誘引（さそは|れ）」のように，本行文字列としては一短単位（誘引），付された読み方としては二短単位（誘は｜れ）となる場合の処理方法が喫緊の課題となっていたが，本手法の導入により解決されることが分かった．

洒落本・人情本のコーパスは，現在試行版を公開中であり，今後，公開作品の拡充を計画している．当て字が頻出するテキストとしては，同じく近世後期の滑稽本や，明治期の小説類も挙げられ，今後それらの資料をコーパス化する際にも，この手法が応用可能である．

なお今回，研究対象とした当て字は全て人手で抽出，タグ付けを行ったが，[15]で提案されるような手法を用いれば，将来的にはより精確な自動判別が可能になる．引き続き，解析手法の検討や解析用辞書の開発を進める必要がある．

付記

本研究は，国立国語研究所共同研究「通時コーパスの構築と日本語史研究の新展開」（リーダー：小木曽智信）ならびに，人間文化研究機構広領域連携基幹研究プロジェクト「異分野融合による総合書物学の構築」のユニット「表記情報と書誌形態情報を加えた日本語歴史コーパスの精緻

[2] 評価時にルビを開く処理はしていないことに注意

化」（リーダー：高田智和）の研究成果を報告したものである．

参考文献

1) 市村太郎：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, Vol. 33, No. 14, pp.96-109, 明治書院（2014）．

2) 洒落本大成編集委員会編：洒落本大成，中央公論社（1978-88）．

3) 国立国語研究所コーパス開発センター（市村太郎ほか）編：『ひまわり版「洒落本コーパス」Ver.0.5』，〈 http://pj.ninjal.ac.jp/corpus_center/chj/edo.html#share〉（参照 2016-11-01）．

4) 国立国語研究所コーパス開発センター（藤本灯・高田智和ほか）編：『ひまわり版「人情本コーパス」Ver.0.5』，〈 http://pj.ninjal.ac.jp/corpus_center/chj/edo.html#ninjou〉（参照 2016-11-01）．

5) 山口昌也：構造化テキストに対応した全文検索システム『ひまわり』，国立国語研究所報告 122，pp.49-82，博文館新社（2002）．

6) Taku Kudo, Kaoru Yamamoto(Titech), Yuji Matsumoto：Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)，pp.230-237（2004）．

7) 伝康晴，小木曽智信，小椋秀樹ほか：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，Vol. 22，国立国語研究所（2007）．

8) 前川喜久雄：KOTONOHA『現代日本語書き言葉均衡コーパス』の開発，日本語の研究, Vol. 4, No. 1，pp.82-95，日本語学会（2008）．

9）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.4-5，言語処理学会（2016）．

10) 小木曽智信・中村壮範：『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用（コーパスアノテーション:新しい可能性と共有化にむけての試み），自然言語処理 21-2，pp.301-332（2014）．

11）藤本灯，北﨑勇帆，市村太郎ほか：「人情本コーパス」の設計と構築, 国立国語研究所論集, Vol. 12（2017 未公刊）．

12) 小松寿雄：江戸時代の国語，東京堂出版（1985）．

13) 矢野準：人情本の漢字，漢字講座 7 近世の漢字とことば，pp.199-218，明治書院（1987）．

14) 土屋信一：式亭三馬の漢字使用―『浮世風呂』を資料として―，日本語学，Vol.5，No.5，pp34-40（1986）．

15）岡照晃：文字単位の多対多自動アライメントを用いた日本語歴史コーパスのルビアノテーションの自動修正，じんもんこん 2016（2016 未公刊）．