『日本語歴史コーパス 江戸時代編Ⅱ人情本』
テキストの凡例と『中納言』表示項目について

2019 年 3 月 29 日 村山実和子

1. はじめに

人情本は，幕末期の江戸語が描写された会話文を含む文学作品であり，近世江戸語から近代東京語への過渡的状況を研究する上で欠かせない資料の一つである。人情本のコーパス化に際しては，まずは言語資料として信頼性の高いテキストを作成することに主眼が置かれ，原本に忠実な翻刻と，その電子化が進められた[1]。

『日本語歴史コーパス 江戸時代編Ⅱ人情本』は，その翻刻テキストにもとづいて構築している。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（詳しくは藤本ほか 2017 参照）。その多様な情報が反映されたものが中納言版の本コーパスである。本稿では中納言上に表示される各種情報について，テキストの校訂状況や表示情報に関する概要を述べる。なお，本コーパスでは，研究上必要と思われる情報をできるだけ底本の状況に即して記述するよう努めたが，完全に反映できているわけではなく，また誤りが残存している可能性もある。そのため，適宜「ページ番号」を基に底本の本文を確認されることを推奨する。

2. コア・非コアデータの認定

コーパス構築に際しては，作品全体をコアデータ（機械による形態論情報付与後，すべてのデータに人手による修正を行ったもの）として扱うことが望ましいが，人情本 8 作品を対象とした場合，全体で約 40 万語と大部のデータであったため，その一部をコアデータとして集中的に精度を高め，残りを非コアデータとして公開することとした。

(1) コアデータ

コアデータには各作品の初編（前編）を設定した。これは約 10 万短単位のデータからなり，全体のおよそ 25%に相当する。

(2) 非コアデータ

初編をのぞく二編（後編）以降が対象である。これは約 30 万短単位のデータからなる。なお，非コアデータについては，江戸板の洒落本および人情本のコアデータを学習用コーパスとして作成した解析用辞書で再解析を行い，その一部に人手による修正を加えた。コアデータより精度は劣るため，必要に応じて文字列検索などを組み合わせて検索することを推奨する。

[1]原本に忠実に翻刻したテキストデータは，国語研の「日本語史研究用テキストデータ集」（https://textdb01.ninjal.ac.jp/dataset/）で公開されている。

3. テキストの凡例

3.1 外字等の処理

3.1.1 外字

本文テキストの文字入力は JISX0213 に準拠している。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（表 1 参照）。

表 1 『江戸時代編 II 人情本』における外字一覧（コアデータのみ）

	キー	振り仮名	作品名	Unicode 番号	備考
1	〓り	たど	花迺志満台	U+36F9	「漂」＋りっとう
2	〓	たとり	花迺志満台	-	「漂」＋りっとう
3	〓伸	しんしん	仮名文章娘節用	U+4C0E	にんべん＋「晋」
4	〓	さかな	恋の花染	-	「肴（偏）」＋「攵」
5	〓	まち	恋の花染	U+21742	ぎょうにんべん＋「矣」
6	〓る	ふけ	連理の梅	-	もんがまえ＋「更」
7	〓	となり	明烏後の正夢	-	にんべん＋糞

3.1.2 絵文字・記号

絵文字や記号など，電子化にあたって置き換えが不可能なものについても，同様に「〓」で表示している。未知語扱いとして，「絵文字・記号等」の品詞を与えた淵。原文文字列については，「外部リンク」の底本画像によって直接参照することが可能である。

3.2 テキストの校訂

3.2.1 濁点

〔1〕濁点の付与

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。ただし，清濁両形あり判断に迷う場合には極力濁点を付与しない方針をとった（「室町時代編 I 狂言」「江戸時代編 I 洒落本」に同じ）。濁点を補う前の文字列は，「原文文字列」に表示される。

【例】

(1) 〈原文〉そのやうに泣ものしやない（花迺志満台）

〈入力〉そのやうに泣ものじやない

〔2〕濁点の排除

清音が期待される箇所に濁点が付与されている場合，それにより形態論情報の付与が困難な形式は，タグ付きで濁点を排除した。変換前の文字列は「原文文字列」および「原文

KWIC」に表示される。

【例】

(2) 〈原文〉モウほんに十年わげへといゝ商売がある（花迺志満台）
〈入力〉もうほんに十年わ**け**へといい商売がある

(3) 〈原文〉アノ児の親父に。ふづつかつて（花迺志満台）
〈入力〉あの児の親父に。**ぶつ**つかつて

(4) 〈原文〉米八さ゛ん今の通りだがら。(春色梅児与美)
〈入力〉米八さ゛ん今の通りだ**か**ら。

(5) 〈原文〉篝火といふのを見じやせう（春色江戸紫）
〈入力〉篝火といふのを見**じ**やせう

(6) 〈原文〉いつそのことにさつばりと切て（春色辰巳園）
〈入力〉いつそのことにさつ**ぱ**りと切て

3.2.2. 踊り字

仮名 1 字分の踊り字（ゝ，ゞ，ヽ，ヾ）は，想定される仮名に置き換えた。変換前の文字列は「原文文字列」および「原文 KWIC」に表示される。対して，2 字分以上に相当するくの字点，また漢字の連続は，置き換えの対象としない。また，振り仮名内の踊り字も置換しない。

【例】

(7) 〈原文〉あなたをのけて余の人に。添ますこゝろはござりません（仮名文章娘節用）
〈入力〉あなたをのけて余の人に。添ます**こころ**はござりません

(8) 〈原文・入力同じ〉いやありやあ御免だ**／＼**（恋の花染）

3.2.3 カタカナ

底本本文においてカタカナで表記された箇所は，平仮名で表示した。変換前の文字列は「原文文字列」に表示される。ただし，振り仮名内のカタカナは置換していない。

【例】

(9) 〈原文〉アレサこの子は。ホンニ野暮だのう（花迺志満台）
〈入力〉**あれさ**この子は**ほんに**野暮だのう

3.2.4　誤字・脱字・衍字等

コーパス本文には，前述のとおり，底本に忠実に翻刻したテキストを用いている。なるべく原態を維持するよう努めたが，検討のうえ必要と判断された箇所については，タグ付きでテキストの訂正を行った。いずれも,「原文文字列」には変換前の文字列が表示される。

【例：脱字】

(10)〈原文〉おめへ方親子二人で。どうか斯かくらていく位のことは。ほねを折て見やう。
（花迺志満台）

〈入力〉おめへ方親子二人で。どうか斯かくら**し**ていく位のことは。ほねを折て見やう。

(11)〈原文〉善兵衛事の由縁を。妻のお貞に**譚**（ものがた）しに。（春色江戸紫）

〈入力〉善兵衛事の由縁を。妻のお貞に**譚り**（ものがた）しに。

【例：衍字】

(12)〈原文〉思ひやらるゝ男の心。惚**れれ**た女の心には。（春色梅児与美）

〈入力〉思ひやらるる男の心。惚**れ**た女の心には。

(13)〈原文〉源之助ので**ござりござり**ますね。（仮名文章娘節用）

〈入力〉源之助ので**ござり**ますね。

【例：誤字】

(14)〈原文〉叔母がわし故に後**脂**さゝれると思へば（明烏後の正夢）

(15)〈入力〉叔母がわし故に後**指**さされると思へば

(16)〈原文〉身上の為にも**宣**には違ないとは思ふものゝ（春色連理の梅）

〈入力〉身上の為にも**宣**には違ないとは思ふものの

3.2.5 テキストの置き換え・補い

〔1〕振り仮名にもとづく補読

本文中，期待される送り仮名・活用語尾・助詞等が振り仮名に含まれる場合，タグ付きで本文に補った。その際，振り仮名自体は，元のまま保持するようにした。置換前の文字列は，「原文文字列」に表示される。

【例】

(17)〈原文〉何角（なにかと）力になつて進ぜたうち　→　〈入力〉｜何角（なにか）｜**と**｜力｜（春色江戸紫）

(18)〈原文〉フンうまく言（いや）あがらあ。　→　〈入力〉｜言（いや）｜**やあがらあ**｜
（春色連理の梅）

(19)〈原文〉お父様が没（おかくれ）あそばして　→　〈入力〉｜**お**｜没（おかくれ）｜　（同上）

〔2〕臨時的な補い

コーパス本文において，テキストを補う際には原則的に振り仮名にもとづいているが，以下の「来（きた）り」については処理単位上の問題から，例外的にタグ付きで仮名に置き換え／文字を補っている。「原文文字列」には変換前の文字列が表示される。

【例】

(20)〈原文〉おかめの部屋へそつと**来(きた)り**　→　〈入力〉｜**来**｜**たり**｜（仮名文章娘節用）

3.2.7. 漢文

漢文（風）の箇所について，返り点や振り仮名などで訓読が可能な漢文等については，訓読した形を本文とした。返り点のない漢文箇所や，長大な部分（跋文等）は，「未知語」として扱い，品詞を「漢文」とした。また置き字についても，品詞「漢文」としている。

【例】

(21)〈原文〉何事も手につき不申　→　〈入力〉手につき**申不**　（春色江戸紫）

(22)〈原文〉不及ながら斯して苦労して　→　〈入力〉**及不**ながら斯して苦労して　（春色辰巳園）

(23)〈原文〉稗史小説之為著述也　→　〈入力〉ママ　（明烏後の正夢）

4. 中納言における表示項目と内容

『日本語歴史コーパス 江戸時代編Ⅱ人情本』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，WEB 上のコーパス検索ツール「中納言」上に検索画面・結果画面として表示される（図１・図２）。 以下では，中納言上の主な表示項目とその内容に関して概説する。

図 1 検索画面

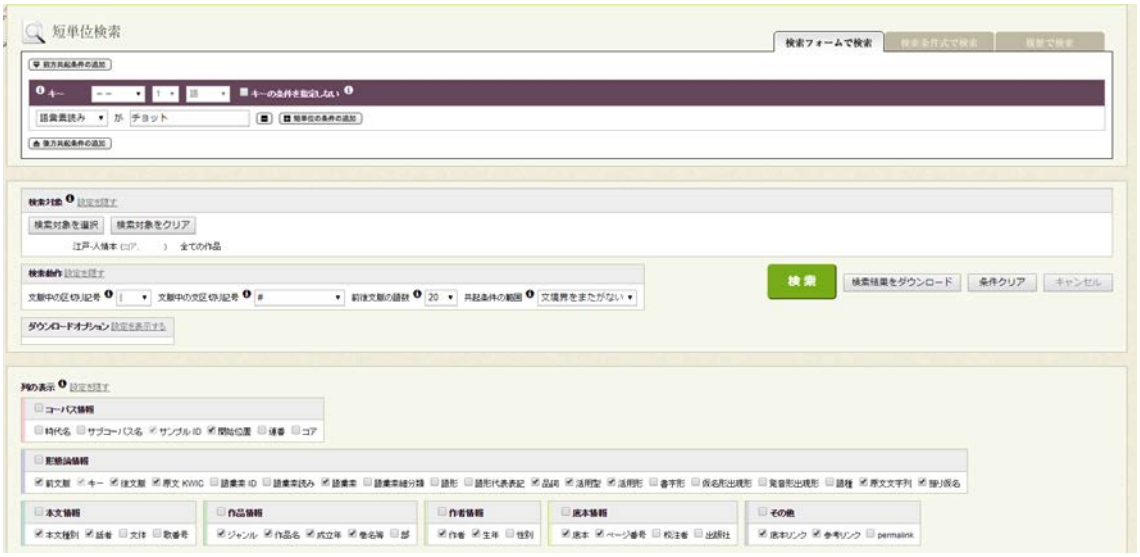
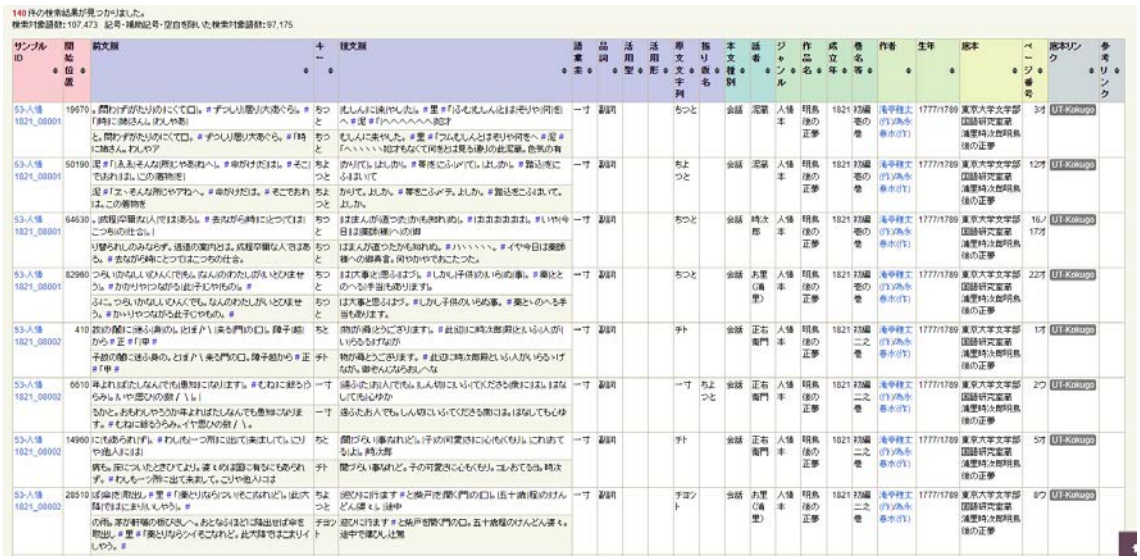


図 2 検索結果の表示イメージ



4.1 形態論情報

基本的に BCCWJ や『江戸時代編 I 洒落本』と同様であり，小椋他(2011)などを参照されたい。中納言において表示される形態論情報（短単位）は，Unidic の見出しに対応している。以下には利用に際して注意すべき点を幾つか挙げる。

〔1〕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**したものである。語彙素で検索することで，同語彙素内の各種語形・活用形・書字形等の異なるもの一括して取得することができる(図 2 は語彙素読み「チョット（副詞「一寸」)」による検索結果である)。

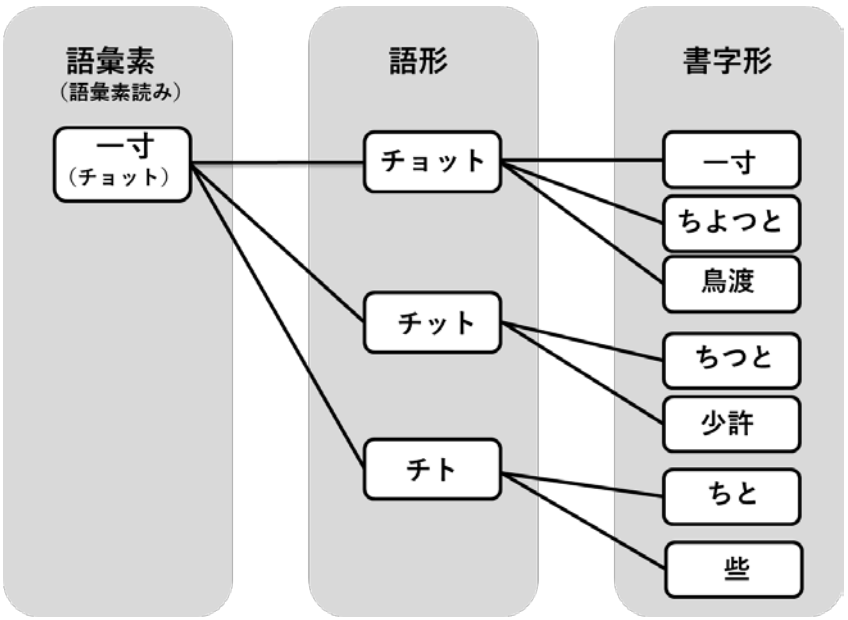
〔2〕語形

「**語形**」は，異語形を区別するレベルであり，例えば「チット」「チョト」など音形の異なるものは，語彙素「一寸」の語形として認定される。(図 3 参照)

〔3〕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 参照)

図 3 語彙素「一寸」の語彙素・語形・書字形



〔4〕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。

〔5〕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」も同様）。

〔6〕活用形

「活用形―小分類」の「融合」は，前部の活用語に後続する助詞等が取り込まれるなどし，短単位として分割しがたいものである。例えば形容詞「若い」の仮定形「若けれ」に助詞「ば」の接した「若ければ」の変化した「わかけりや」は，下記のように処理されている。

【例】

(24)｜おばさん｜おめへ｜まあ｜**わかけりや**｜何｜を｜する｜の｜だ｜（花迺志満台）

→形容詞「若い」・形容詞-一般・**仮定形-融合**

〔7〕原文文字列

踊り字・濁点・片仮名等を校訂する前の本文は「原文文字列」および「原文 KWIC」に表示される（図 4）。

図 4 本文テキスト（出現書字形）と原文文字列

サンプルID	開始位置	前文脈	キー	後文脈	語彙素	品詞	活用型	活用形	原文文字列	振り仮名
53-人情1833_04002	20120	だ|へ＃其様|にふさい|でから|に|両方|がだんまり|か|へ＃と＃いひながら|丹次郎|にむかひ＃「| もし	ゑ|初|に|お|目|に|かかつ|てはだ|おなじみ|もねへ|わたい|がふしつけ|らしい|わけだ|けれど	もし	感動詞-一般			モシ	
		「ヲヤどうしたんだへ＃其様|にふさい|でから|に両方がだんまりかへ＃ト＃いひなから丹次郎にむかひ＃「	モシ	エ初|にお目|にかゝつてまだおなじみもねへわたいがふしつけらしい						

キーに表示された「もし」が，原文文字列では片仮名表記であることが示される。また，前後文脈の原文 KWIC 表示を見ると，「いひながら」が「いひなから」，「かかつて」が「かゝつて」であったことがわかる。ただし，これらはあくまで補助的な表示であるため，底本の詳しい状況を確認する場合は，外部リンク先の底本を直接ご参照いただきたい。

〔8〕振り仮名

本行の文字列の右側に小書きされた文字については「振り仮名」列に表示されるが，本コーパスでは，左に付された注釈的な語句は対象外とした。ただし，元の XML データには，<lRuby>（左ルビ）というタグで情報が保持されている。

4.2 本文情報

次に，本文の会話・ト書きの別，話者等の表示方法について述べる（表 2）。

〔1〕本文種別

性質の異なる本文を区別し，「本文種別」として表示している。具体的な分類は下記のとおりである。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を考慮し，「会話」と認定していない。

表 2　本文種別の対応

本文種別		説明
会話		本文中の会話部分
割書き		割書きされた本文
地の文（明示なし）		地の文相当のものから割書きをのぞいた本文
その他	**内題**	各作品の内題
	尾題	各作品の尾題
	見出し	序・跋等作品内の section を表す
	話者	著者による話者表示

〔2〕話者

本コーパスにおける会話文について話者表示の統一をはかり，「話者」欄に表示した。

【例】

〈原文〉里　→　〈「話者」列の表示〉**お里（浦里）**　　※同一作品内で統一

なお，先に公開した『江戸時代編 I 洒落本』では，ほかのサブコーパスに先駆け，話者の属性について詳細な情報を付与していた。しかしながら，『江戸時代編 II 人情本 (Ver.0.8)』では，話者名のみが表示されるという差異があることに注意されたい。

4.3 作品情報

〔1〕ジャンル

「ジャンル」には一律に「人情本」が表示される。

〔2〕作品名

「作品名」には，個別の作品名が表示される（角書は省略）。

〔3〕成立年

「成立年」には個別の作品の成立年が表示される。

〔4〕巻名等

「巻名等」には，初編上巻／中巻／下巻などの別が表示される。

〔2〕～〔4〕については，「作品リスト」〈http://pj.ninjal.ac.jp/corpus_center/chj/edo.html〉として対応表を作成した。適宜ご参照いただきたい。

4.4 底本情報

〔1〕底本

「底本」には，国語研または東京大学国語研究室所蔵の底本の書名が表示される。

〔2〕ページ番号

「ページ番号」には，対応する底本の丁数が表示される。

4.5 外部リンク

全作品について，その底本である版本の画像が公開されている。この項目には，検索結果に応じて，そのリンク先が表示される。リンク先の表示内容は以下のとおり。

表 3 外部リンク先一覧

表示名	参照先
Ninjal	国立国語研究所 日本語史研究資料 http://dglb01.ninjal.ac.jp/ninjaldl/
UT-Kokugo	東京大学文学部国語研究室 資料画像 http://kokugo.l.u-tokyo.ac.jp/

なお，参照の際には以下の点に注意されたい。

① 花迺志満台：基本的には国語研所蔵本を底本とするが，四編下第 11 丁を欠いている。翻刻本文および画像リンクは東京大学国語研究室蔵本による。

② 明烏後の正夢：三編下（巻之九）23 丁・24 丁を欠いている。翻刻本文は早稲田大学図書館蔵本（へ 13 02909 0004）による。

本コーパスのタグセット

	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
<speech >	ひとまとまりの会話文。本タグに話者情報を付与。
<quotation>	文献等からの引用や手紙など。
<warigaki>	割書き箇所。
<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
<ruby>	本行の右側に振られた振り仮名文字列。
<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

参考文献

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

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（2018 年 3 月 28 日確認）

市村太郎（2016）「「江戸時代編」の構築と課題」『日本語学会 2016 年度春季大会予稿集』日本語学会

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』言語処理学会

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

市村太郎・村山実和子（2017）「洒落本コーパス構築の試行」『国立国語研究所論集』12

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』93-5

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』国立国語研究所

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』形態論情報規程集 第 4 版（上）（下)』国立国語研究所

藤本灯・北﨑勇帆（2015）「ひまわり版「人情本コーパス」ver.0.1（『日本語歴史コーパス 江戸時代編』）仕様書」http://pj.ninjal.ac.jp/corpus_center/chj/doc/ninjobon0.1-doc.pdf

藤本灯・北﨑勇帆・市村太郎・岡部嘉幸・小木曽智信・高田智和（2017）「「人情本コーパス」の設計と構築」 『国立国語研究所論集』 12，pp.1-12.