「人情本コーパス」の設計と構築

国立国語研究所論集 (NINJAL Research Papers) 12: 1–12 (2017)
ISSN: 2186-134X print/2186-1358 online

「人情本コーパス」の設計と構築

藤本　灯[a]　北﨑勇帆[b]　市村太郎[c]　岡部嘉幸[d]　小木曽智信[a]　高田智和[a]

[a]国立国語研究所 研究系 言語変化研究領域
[b]東京大学大学院人文社会系研究科 博士課程
[c]常葉大学
[d]千葉大学

要旨

現在，『日本語歴史コーパス』「江戸時代編」の一環として「人情本コーパス」を構築中である。2015 年 10 月には『比翼連理花廼志満台』を対象とした「人情本コーパス」の試行版（全文検索システム『ひまわり』版）を公開した。人情本のコーパス化は，(1) 原本表記に忠実な翻字テキストの作成，(2) (1) に最小限の校訂を加えた『ひまわり』版 XML テキストの作成の段階である。XML テキストの作成では，基本的に「洒落本コーパス」のタグセットに準拠し，合字や校訂にかかわるタグを追加した人情本用タグセットを用意した。また，『花廼志満台』初編上巻の形態素解析を行った結果，解析精度は約 87% であった。人情本に特徴的なイレギュラーな訓の多さが，精度の低さと関係している。今後，形態論情報付きコーパスを構築するにあたっての課題は，イレギュラーな訓を含む漢字に振られた「ルビ」を，どのように扱っていくかである *。

キーワード：人情本コーパス，日本語歴史コーパス，『比翼連理花廼志満台』

1. はじめに

現在，国立国語研究所では「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）を中心に『日本語歴史コーパス』[1] の構築が進められており，本年度までに「平安時代編」（『古今和歌集』，『源氏物語』など 16 作品），「鎌倉時代編 I 説話・随筆」（『今昔物語集』など 5 作品），「室町時代編 I 狂言」（『虎明本狂言集』），「江戸時代編 I」（洒落本），「江戸時代編 II」（人情本）が公開されてきた。また，近代語のコーパス構築も進められており，近代雑誌を対象とした『太陽コーパス』『近代女性雑誌コーパス』『明六雑誌コーパス』『国民之友コーパス』が公開されている [2]。このうち「江戸時代編」は，近世後期の口語資料とされる洒落本を対象とした「洒落本コーパス」（市村 2014）に次いで，同じく近世後期の長編恋愛小説である人情本を対象とした「人情本コーパス」の開発が計画され，いずれも試行版が 2015 年 10 月に公開されたものである。

* 本稿は，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」（プロジェクトリーダー：小木曽智信）および人間文化研究機構広領域連携型基幹研究「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化」（プロジェクトリーダー：高田智和）による成果の一部である。また日本語学会 2015 年度秋季大会で行ったブース発表（「人情本のコーパス化」）の内容に加筆修正を加えたものである。
[1] 小木曽（2016）参照。国立国語研究所 日本語歴史コーパス http://pj.ninjal.ac.jp/corpus_center/chj/
[2] 国立国語研究所 近代語のコーパス http://pj.ninjal.ac.jp/corpus_center/cmj/

幕末期の江戸語資料である人情本は，近世江戸語から近代東京語に至る口語史を知る上で欠かすことのできない資料であるが，従来日本語史研究において頻繁に利用されてきた『日本古典文学大系』（岩波書店）中の『春色梅児誉美』『春色辰巳園』や，『日本古典文学全集』『新編日本古典文学全集』（小学館）中の『春告鳥』，『梅暦』（岩波文庫）などは，いずれも為永春水の作品に限られており，言語資料としては偏りがあった。また，岡部嘉幸が作成した人情本刊行会編の活字テキストにもとづく 6 作品の電子テキストデータ，およびこれを全文検索システム『ひまわり』用に変換した『ひまわり』版『「人情本」パッケージ』[3] は，為永春水以外の人情本を利用できる点で貴重な資料であるが，その翻字テキストには原本からの大きな改変が加えられていることが知られており，扱いに注意が必要である（後述）。

近年，鶴見人情本読書会（1998 〜 2000），浅川（2012）等が刊行されたことで，春水以外の作品を含めた人情本の利用が進みつつあるが，未だ言語研究において信頼するに足る人情本の活字テキストが十分にあるとは言えず，コーパス化は更に遅れている状況である。洒落本同様，人情本についても，信頼できるテキストによる電子化資料の構築が強く求められている。

「人情本コーパス」は，『日本語歴史コーパス』の中で初めて，「版本から翻字」したテキストを基に展開することとなるが，本稿では特に，形態論情報付与以前の段階となる，翻字テキストの作成および翻字テキストを基とした XML データの作成の過程に焦点を絞りつつ，「人情本コーパス」開発の背景と現状につき報告することとする。

2.『比翼連理花廼志満台』の翻字テキスト化・XML 化の過程

以上に述べた研究上の要請により，まずデータ化に着手した作品は，国立国語研究所が所蔵する人情本のうち，2015 年以降に各全編の画像を公開した次の 5 作品である [4]。

『小三金五郎仮名文章娘節用』（3 編 9 巻，曲山人，1831 〜 1834）
『春色梅児与美』（4 編 12 巻，為永春水，1832 〜 1833）
『梅暦余興春色辰巳園』（4 編 12 巻，為永春水，1833 〜 1835）
『比翼連理花廼志満台（はなのしまだい）』（4 編 12 巻，松亭金水，1836 〜 1838）
『おくみ惣次郎春色江戸紫』（3 編 9 巻，山々亭有人，1864 〜明治）

本節では，このうち 2015 年 10 月にテキスト版および『ひまわり』版を試験公開した『比翼連理花廼志満台』（以下『花廼志満台』）を対象としながら，データの作成方針および作成過程，また試験的に行った初巻の形態素解析の結果について述べることとする。

[3] 全文検索システム『ひまわり』用「人情本」パッケージ http://www2.ninjal.ac.jp/lrc/index.php?%C1%B4%CA%B8%B8%A1%BA%F7%A5%B7%A5%B9%A5%C6%A5%E0%A1%D8%A4%D2%A4%DE%A4%EF%A4%EA%A1%D9%2F%A5%C0%A5%A6%A5%F3%A5%ED%A1%BC%A5%C9%2F%A1%D6%BF%CD%BE%F0%CB%DC%A1%D7%A5%D1%A5%C3%A5%B1%A1%BC%A5%B8
[4] 国立国語研究所 日本語史研究資料 http://dglb01.ninjal.ac.jp/ninjaldl/

2.1『花廼志満台』の翻字テキスト化

『花廼志満台』は既に人情本刊行会による翻刻（1916 年出版）があるが，先述の通り，人情本刊行会による活字テキストには原本からの改変が加えられていることが知られている[5]。

図 1 に国立国語研究所蔵本『花廼志満台』の画像（初編中巻 7 ウ）を挙げ，図 1 をもとに筆者らが新たに翻字テキスト化したものを図 2 に，また比較対象として同範囲の人情本刊行会の活字テキストを図 3 に挙げる。

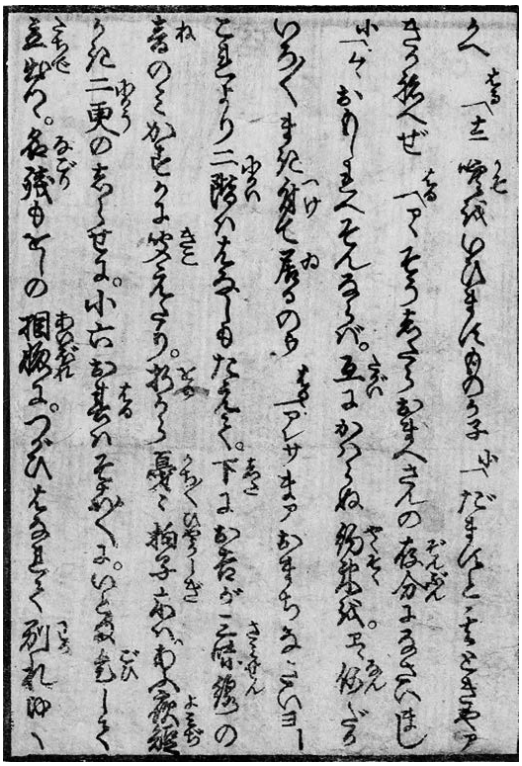

図 1　国立国語研究所蔵本『花廼志満台』（初編中巻 7 ウ）

[5] 浅川 (2014) によれば, 版本の改変は「版本の本文の表記の改変」「版本の本文の削除」「版本にない本文の追加」の三種に分類される。

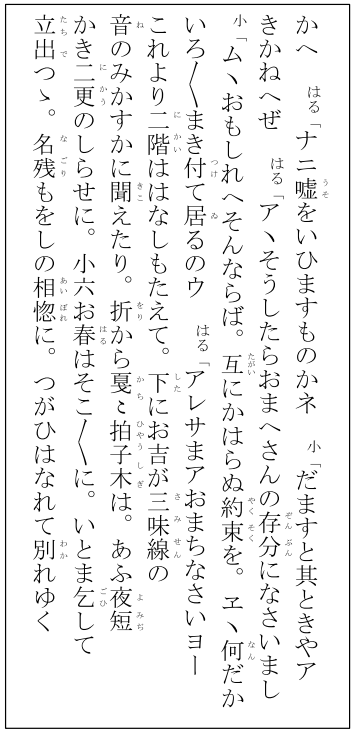
かへ　春「ナニ嘘をいひますものかネ　小「だますと其ときやア
きかねへぜ　春「アヽそうしたらおまへさんの存分になさいまし
小「ムヽおもしれへそんならば。互にかはらぬ約束を。ヱヽ何だか
いろ〳〵まき付て居るのウ　春「アレサまアおまちなさいヨー
これより二階ははなしもたえて。下にお吉が三味線の
音のみかすかに聞えたり。折から戞〻拍子木は。あふ夜短
かき二更のしらせに。小六お春はそこ〳〵に。いとま乞して
立出つゝ。名残もをしの相惚に。つがひはなれて別れゆく

図 2　筆者らによる翻字テキスト

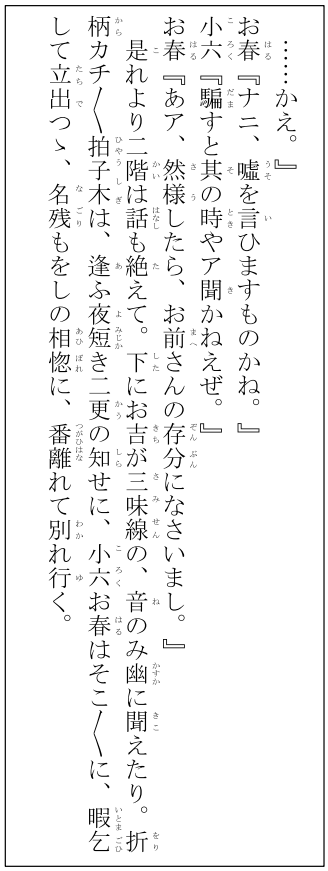
……かえ。』
お春『ナニ、嘘を言ひますものかね。』
小六『騙すと其の時やア聞かねえぜ。』
お春『あア、然様したら、お前さんの存分になさいまし。』
是れより二階は話も絶えて。下にお吉が三味線の、音のみ幽に聞えたり。折
柄カチ〳〵拍子木は、逢ふ夜短き二更の知せに、小六お春はそこ〳〵に、暇乞
して立出つゝ、名残もをしの相惚に、番離れて別れ行く。

図 3　人情本刊行会による翻字テキスト

従来のテキストでは，図 3 に挙げた部分のみを見ても，仮名遣いや仮名／漢字表記，話者名の改変，ルビの増減，句読点や鉤括弧の追加，更には原本からの省略（原本 3 ～ 4 行目）も行われていることが分かる[6]。このことは旧テキストの在り方が一概に劣っていることを示すものではないものの，特に，表記を含めた近世日本語の研究をしようとする者にとって，非常に使用し辛いものであったことは事実である。そこで筆者らはまず，語学研究に堪え得る翻字テキストの作成，すなわち原本の表記を可能な限り復元することを試み，そこに多少の手を加え[7]，図 4 のようなテキストを新たに用意し，公開した。本稿ではこれを「原本翻字テキスト」と呼ぶ。

[6] 本書の他の箇所では，原本の「綯繰（しなだれ）」を「撓垂」とするような漢字表記の改変もまま行われる。
[7] 会話時の鉤括弧や句点の付与，原本にある話者名を【　】で，合字を〔　〕で括る等の改変を施した。ただし行取り，誤字･脱字･衍字や濁点の有無などは原本の通りとし，更に口絵の頁にある詞書等の翻字も行い，原本画像との対照が可能となるように努めた。なお，これらのテキストおよび凡例は，国立国語研究所「日本語史研究資料」→「比翼連理花廼志満台」にて試験公開中である。

（7 ウ）
かへ」【はる】「ナニ嘘｛うそ｝をいひますものかネ」【小】「だますと其ときやア
きかねへぜ」【はる】「アヽそうしたらおまへさんの存分｛ぞんぶん｝になさいまし」
【小】「ムヽおもしれへそんならば。互｛たがい｝にかはらぬ約束｛やくそく｝を。エヽ何｛なん｝だか
いろ／＼まき付｛つけ｝て居｛ゐ｝るのウ」【はる】「アレサまアおまちなさいヨー」
これより二階｛にかい｝ははなしもたえて。下｛した｝にお吉が三味線｛さみせん｝の
音｛ね｝のみかすかに聞｛きこ｝えたり。折｛をり｝から戞ゝ｛かち／＼｝拍子木｛ひやうしぎ｝は。あふ
夜｛よ｝短｛みぢ｝
かき二更｛にかう｝のしらせに。小六お春｛はる｝はそこ／＼に。いとま乞｛ごひ｝して
立出｛たちで｝つゝ。名残｛なごり｝もをしの相惚｛あいぼれ｝に。つがひはなれて別｛わか｝れゆく。

図 4 『花廼志満台』原本翻字テキスト

しかし，このように原本の表記をなるべく忠実に再現するということは，未校訂であることと同義である。未校訂のデータは，近世日本の表記・音韻・出版等に関する研究には適した面がある一方で，例えば，漢字表記が特定できない語を隈なく調べることなどは困難であるし，仮名遣いの揺れや濁点の有無，誤字・脱字等を一々考慮，想定して複数のパターンで検索することも手間である。この問題を克服するために，将来的には形態論情報の付与されたデータの公開を計画しているが，今回はその中間段階として，最小限の校訂（2.2 参照）を施した本文を，全文検索システム『ひまわり』のパッケージとして試験公開することとした。その仕様を次節に述べる。

2.2 XML による構造化

上（図 4）に挙げた原本翻字テキスト版を『ひまわり』上で円滑に利用するため，誤字・脱字・衍字の修正，濁点の追加，カタカナの平仮名への統合などの校訂を行い，校訂後のテキストに対し，XML によって文書構造，文・語の機能，文字の外形などの情報を付与した（これらの修正については，修正前の情報が XML 内に保持されている）。

タグは基本的に，同じ近世の戯作を扱った「洒落本コーパス」の構造（市村ほか 2013）に準拠しているが，「人情本コーパス」では前述の通り，底本を「原本翻字テキスト」とした経緯があるため，整合性の保持・原本表記の再現のために細部を改めた。具体的には，合字の情報を残すために <goji> を追加し（図 5），本文修正箇所を示す <corr>（図 6）の属性として校訂を行ったことを示す @revision を追加した。

```
・どのやうな事するものか<r rt="よふ">様</r><r rt="す">子</r>はしれ<lb/><pb n="七ウ" num="24"/>ぬ<goji>こと</goji>ながら<r rt="なん">何</r>でも<r rt="つぢ">辻</r>に<r rt="たつ">立</r>て居て
・<s>はる</s><speech source="お春"><s>「<char script="カタカナ">を</char><char script="カタカナ">や</char>お<r rt="とつ">爺</r>さんとした事が。とんだ<goji>こと</goji>をおつ<lb/>しやるねへ。</s>
```

図 5　<goji> の例
（「こと」が合字であることを示す。）

```
<s>はる</s><speech source="お春"><s>「<char script="カタカナ">ゑ</char><odoriji originalText="ヽ">ゑ</odoriji><r rt="こ">小</r><r rt="ろく">六</r><lb/>さんが<corr type="revision" source="東大本">心</corr>がはりとはへ。</s>
```

図 6　<corr> の例
（「心」の箇所が国立国語研究所蔵本で判読不能であるため，東京大学文学部国語研究室蔵本により校訂（補読）したことを示す。）

また，テキストの一単位が洒落本と比べて大部であるため，便宜的に <text> を一巻分とし，全体のコーパスを示すために上位タグ <corpus> を用いることとした。すなわち，本コーパスは図 7 のような構造を取ることになる。

```
<corpus>
  <text title="比翼連理花廼志満台" volume="初編上" ……>
                              ⋮
  </text>
  <text title="比翼連理花廼志満台" volume="初編中" ……>
                              ⋮
  </text>
  <text title="比翼連理花廼志満台" volume="初編下" ……>
                              ⋮
  </text>
                              ⋮
  <text title="比翼連理花廼志満台" volume="四編下" ……>
                              ⋮
  </text>
</corpus>
```

図 7　<corpus>・<text> の例

「人情本コーパス」のタグセットを表 1 に示す。このタグを基に XML 化を行った『花廼志満台』を『ひまわり』上で検索した結果を図 8 に，XML を変換して文脈を表示した文脈閲覧画面を図 9 に示す。また，<text> の属性として国立国語研究所データベース「日本語史研究資料」の URL と巻次を，<pb> の属性として丁数の通し番号を与えることにより（図 10），『ひまわり』上から原本画像へとアクセスすることができるようになっている（図 11）。これらの仕様の詳細については，藤本・北崎（2015）を参照されたい。

表 1 「人情本コーパス」のタグセット

	階層	タグ	説明
文以上	1	corpus	コーパス全体
	2	text	テキスト一冊のまとまり
	3	front	序文
		body	本文
		back	跋文
	4	article	記事
		titleBlock	全体のタイトルの記述
	5	p	本文のひとかたまり
		block	内題などのブロック要素
	6	speech	会話文
		warigaki	割書き
		quotation	字下げ，手紙など
文	7	s	一文
文未満	8	speaker	話者
		hi	囲み，傍線
		r	ルビ
		lr	左ルビ
		odoriji	踊り字
		vMark	濁点無表記箇所
		goji	合字
		corr	本文修正箇所
		unclear	原本の不鮮明箇所
		gap	判読不明箇所
位置情報		pb	頁開始位置
		lb	行開始位置

図 8　『ひまわり』の検索画面

図 9　文脈閲覧画面

```
<text title="比翼連理花廼志満台" volume="初編上" year="1836" year_w="天保 7 刊"
url="http://dglb01.ninjal.ac.jp/ninjaldl/show.php?title=hananosimadai" vol="001">
                                    ⋮
                                    ⋮
<pb n="四オ" num="17"/>いふものだからいくぢや<char script="カタカナ">あ</char>ねへはな
</s>
<s><r rt="やす">安</r><char script="カタカナ">い</char>ものは<r rt="こめ">米</r>ばかり<r
rt="しよ">諸</r><lb/><r rt="しき">色</r>の<r rt="ね">直</r>はだん／＼あがるしいけねへ
<char script="カタカナ">の</char><char script="カタカナ">う</char>。</s>
```

図 10　原本画像へのリンクの例
（url="..." により底本の親 URL を，vol="001" により初編上巻を，num="17" により表紙から数えて 17 頁目（n="四オ" により 4 丁表）であることを示す[8]。）

図 11　原本画像の表示

[8] なお，国立国語研究所蔵本は四編下第 11 丁を欠くため，以下のように URL を指定することにより，四編下全体を東京大学文学部国語研究室蔵本にリンクした。
<text title="比翼連理花廼志満台" volume="四編下" series="" textID="" year="1838" year_w="天保 9 序" url="http://kokugo.l.u-tokyo.ac.jp/data/show.php?title=hananoshimadai" vol="012">

2.3 形態素解析の試行

初編上巻を対象とし，形態素解析[9]とその結果の人手修正を試験的に行った結果について，次に述べる。まず，テキストの性質が大きく異なる序の部分を含めた場合と含めない場合について，四つのレベルで評価した解析精度（F 値）を表 2 に示す。

表 2　形態素解析の試行結果

	序なし	序あり
Lv.1：単位境界の認定	0.8726	0.8740
Lv.2：Lv.1 ＋品詞・活用形の認定	0.8360	0.8346
Lv.3：Lv.2 ＋語彙素の認定	0.8214	0.8189
Lv.4：Lv.3 ＋発音形の認定	0.8168	0.8139

語を区切る範囲での大きな修正から，発音単位の小さな修正までを含む修正箇所は 4331 短単位のうち 541 であり，主な修正箇所は，名詞・動詞の書字形単位での未登録箇所[10]であった。未登録箇所の多いことは，総ルビであることを利用して敢えてイレギュラーな訓を多く宛てることにより文学的な効果を狙った人情本の特質[11]によるものと考えられる（参考までに，原本翻字テキストより抽出した「通常の音訓でないルビを持つ熟語」を含むバリエーション全例を稿末の付録に掲げた）。なお，このような表記や音訓のバリエーションの存在は，単に形態素解析の精度を下げる点でのみ問題となるのではない。これらの音訓を，その漢字語の「通常の音訓」と同列のものとして，形態素解析用辞書に新たに登録すべきか，するとすれば，「小春（あいつ）」のような一回的なものは除くとしても，「自己（あれ）」「同道（いつしよ）」「悪女（あま）」のようなものの要不要の線引きをどこで行うかといったことは課題として残る。またそれに関連して，本行，ルビのいずれを検索用本文として設定するかといったことも，ルビの語形を前提として読ませる人情本のスタイルにおいては大きな問題となる。本行を本文とすれば，特殊な表記形の語（「同道（いつしよ）」など）の解読が難しくなり，ルビを本文とすれば，同音訓異義語の認定が難しくなるためである。無論，新たに両者を併用するスタイルの採用を検討する余地もある。今回は，『ひまわり』版の試行までについて述べたが，次に形態論情報を付与する段階においては，ルビの扱いに関する議論が必須となろう。

なお，これまでの『日本語歴史コーパス』構築の過程において，語彙素単位で辞書に登録されていなかった語は，「親人」「願籠」「気扱い」「苦艱」「後悔い」「城下（しろした）」「溺惑」「編次」「旧り行く」「娘気」などであるが，「溺惑」「倶利伽羅」（倶利伽羅紋紋の意）などは『日本国語大辞

[9] 形態素解析器として「MeCab」を，形態素解析用辞書に「洒落本コーパス」制作用の「近世口語 Unidic」（試行版，Ver 0.9）を用いた。Unidic を用いた形態素解析については小木曽（2013）を参照。

[10] 例えば，本書に現れた「看取（みとり）」について，これまでの『日本語歴史コーパス』の構築の過程では，語彙素「見取り」に書字形「見取り」「見取」および発音形「ミトリ」が形態素解析用辞書に登録されていたが，書字形「看取」については未登録であった。

[11] 人情本とそのルビについてのまとまった研究としては矢野（1987）があり，（人情本の）「漢字の使用に際しては，振り仮名の効用を最大限に活用していると思われる」との言及がある。

典〔第二版〕』（小学館）の初出より遡る語であり[12]，「通時コーパス」の一環としての「人情本コーパス」も日本語研究に資するものである点，言うを俟たない。

3. おわりに

以上，『比翼連理花廼志満台』を例に，「人情本コーパス」構築の背景および現状を述べた。

今後は，本稿 2 節冒頭で挙げた人情本 5 作品をはじめとして，翻字テキスト作成から形態論情報の付与までを行っていく予定である。その構築の過程においては，完成形である形態論情報付きコーパスに至る以前の翻字テキストや，XML による構造タグ付きテキストといった中間段階のデータの提供も可能となる。コーパス化の完成を目指すとともに，コーパス化に際して人情本の「ルビ」をいかに扱っていくかは，今後の検討課題としたい。

付録

参考までに，原本翻字テキストより抽出した「通常の音訓でないルビを持つ熟語」を含むバリエーション全例 *（『比翼連理花廼志満台』全編を対象とする）を，「代名詞」「類似の表現が複数の仮名／漢字表記を持つもの」「同じ語に対して複数の漢字表記があるもの」「同じ漢字語に対して複数のルビがあるもの」に場合分けし（重複するものは上位を優先），原本表記のまま以下に挙げる。

* よって「らうにん（退糧・浪客）」の項目に同義の「浪人」の語が挙がっていないなどの点には留意されたい。またここでは表記のバリエーションを示すことを目的とするため，仮名遣いや活用形の差異しか持たない語群や，一種の音訓（熟字訓）しか持たない語群は除いてある。

■代名詞

□あれ・おいら・おら・おらあ・おれが・じぶん・てまへ・てめへ（自己），てんべ（自我・自己），てめへ・そつち（其方）　□あいつ（小春・彼女・彼奴・彼様奴），あなた（貴君・貴嬢・貴僧・此方・彼方），あれ（彼女），きやつ（彼女・彼奴）　□あち・あつち・かなた（彼方），あすこ・かしこ（彼処），こち・こなた（此方），こつち（此辺・此方・彼方），そこ（此処・其処・其所）　□かう（箇様・這般），かやう（這般）　□いつく・いづく（何方），とこ（何処・何所），どこ（何処・何所・何方）

■類似の表現が複数の仮名／漢字表記を持つもの

□あね・あねへ・おむす・をんな（処女），がき（女児），むすめ（処女・女児・少女・娘女），をとめ（処女・少女）　□いつか（先頃・先日），いつぞや（先頃・先日・先外）　□いひわけ（分解），わか・わかつ・わかり・わかる（分解），わけ（情合・分解・訳合・有理）　□うたて（薄情），うはき（多性・薄情・浮薄），うわき（多性・薄浮）　□うまれつき（性質・生質），かたぎ（性質）　□おつと・ていし・ていしゆ・をとこ（良人），ていし（主女）　□おば（老婆），ば・はア・はゞ・ばゝ・ばゝあ（老婆），ばゝ（媼さ），ばあ・ばゝ（老母）□おやぢ（親父・爺父），ちやん（親父），とゝ（爺父）　□かたち・くはたち・なり（形容），みなり（形容・身形）　□きやうたい・きやうだい（姉弟・姉妹・姉婦），きやうでへ（姉妹）　□きりやう（標致・標緻・容緻），きれう（標致）　□しごと（活業・針線），しやうばい・せうばい・てわざ・なりわい（活業）　□じやうだん・じようだん（雑談・戯談），じやうだん（串戯）　□たばかつ・たばかり（詐偽），たばかる（変詐・変訛）　□とろばう・どろぼう（盗人・盗賊），ぬすびと（盗賊）　□なく・よにない（死亡），なくなり（死去・辞世）　□にこ・につこ・につこり（完爾・莞爾），にこり（微笑）　□ふうふ（夫婦），めうと（女夫・夫婦）　□ほんたう（実正），ほんとう（実情・実正・信実・真実・本体），ほんとふ（真実）

■同じ語に対して複数の漢字表記があるもの

□あたり（近所・四辺）　□あま（悪女・女子）　□いくら（何程・幾干・幾許）　□いつ（何時・幾日）　□いつしよ（一室・同室・同床・同道・同伴）　□うかべ（瓢蕩・放心）　□うそ（偽言・虚言）　□おかみ（内義・内室）　□かくて（却説・再説）　□かし（河岸・川岸）　□さかや（酒坊・酒楼）　□たつしや（健息・息災）　□ちやうづ（浄水・手水・小便）　□ぢようろ（妓女・女郎・娼妓）　□てがみ（手簡・手翰・書翰）　□てだて（手術・手便・方便）　□となり（合壁・隣家）　□のろけ（痴情・恋情）　□ひとり（一個・一人）　□ふてへ（大胆・不届）　□ますらを（壮士・壮男）　□めうと（女夫・夫婦）　□もくろむ（計較・計掠）　□もとより（元来・固来）　□やまと（大和・日本）　□らうにん（退糧・浪客）　□わるいこと（密事・密通）

[12]「溺惑」の初出は『広益熟字典』（1874），「倶利迦羅紋紋」の意としての「倶利伽羅」は『柳多留』一五二（1838–40）が初出として挙げられており，いずれも本作（1836–38）の例が古い。

□わるもの（兇児・凶児）　□をとこ（漢士・侠者・雄士・良人）
■同じ漢字語に対して複数のルビがあるもの
□あからさま･すつぱり（明さ地）　□あゆび･あるき･あるく･あるひ（歩行）　□いぶか･いぶかり･おつ（不審）　□おちぶれ･おちめ（零落）　□おつくり･みじまひ（化粧）　□このかた･こんど（以来）　□さいわい･さいわひ･しあはせ（僥倖）　□せへ･わざ（所為）　□ぢんすけ･やきもち（嫉妬）　□のこらず･みんな（不残）　□みて・みるひと（看官）　□わかいもの・わかうど（弱官）

参照文献

浅川哲也（2012）『春色恋廼染分解 翻刻と総索引』東京：おうふう.
浅川哲也（2014）「江戸時代末期人情本の活字化資料にみられる諸問題―「あるのです」は「あるです」―」『日本語研究（首都大学東京）』34: 1–14.
藤本灯･北崎勇帆（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
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　臨時増刊号・特集「日本語史研究と歴史コーパス」』33(14): 96–109.
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』249–258.
小木曽智信（2013）「中古仮名文学作品の形態素解析」『日本語の研究』9(4): 49–62.
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): 72–85.
鶴見人情本読書会（1998 ～ 2000）「〈翻刻〉『仮名文章娘節用』前編（･後編･第三編）」『鶴見日本文学』2（～ 4）.
矢野準（1987）「人情本の漢字」佐藤喜代治（編）『漢字講座 7　近世の漢字とことば』199–218.　東京：明治書院.

Design and Construction of the *Ninjobon* Corpus

FUJIMOTO Akari[a]　KITAZAKI Yuho[b]　ICHIMURA Taro[c]
OKABE Yoshiyuki[d]　OGISO Toshinobu[a]　TAKADA Tomokazu[a]

[a]Language Change Division, Research Department, NINJAL
[b]Graduate Student, Humanities and Sociology, The University of Tokyo
[c]Tokoha University
[d]Chiba University

Abstract
The *Ninjobon* Corpus is currently under construction as a part of the Edo Period Collection of the Corpus of Historical Japanese. In October 2015, a trial version of the *Ninjobon* Corpus (full text search system in the *Himawari* edition) focusing on the *Hiyokurenri Hana no Shimadai* was publicly released. The *Ninjobon* Corpus creation is at the stage of (1) faithful transcription of the original printed book into text, and (2) creation of the "Himawari" XML texts with minimal revisions to (1). In the creation of the XML texts, the tag set is fundamentally based on the *Sharebon* Corpus, though a tag set with tags related to ligatures and revisions was prepared for the *Ninjobon*. Further, the results of a morphological analysis of the first volume of *Hana no Shimadai* showed an analytical precision of approximately 87%. The low precision is caused by the large number of characteristically irregular readings in the *Ninjobon*. One challenge in a corpus construction with annotated morphological information is on how to address the "rubies" attached to *kanji* characters with irregular native Japanese readings.

Key words: *Ninjobon* Corpus, Corpus of Historical Japanese, *Hana-no-Shimadai*