2017 年 11 月 1 日

ひまわり版「人情本コーパス」

ver.0.2

（『日本語歴史コーパス 江戸時代編』）

仕様書

文責　藤本灯　北﨑勇帆　福山雅深

1. はじめに

『ひまわり版「人情本コーパス」ver. 0. 2（『日本語歴史コーパス 江戸時代編』）』は、web 上で原本画像が公開されている、国立国語研究所蔵本を始めとする江戸期の版本を底本とする。

・現在公開中の作品
［底本が国立国語研究所蔵本であるもの］
・比翼連理花廼志満台　4 編 12 巻　松亭金水作　1836～1838 序
http://dglb01.ninjal.ac.jp/ninjaldl/bunken.php?title=hananosimadai
＊上記 URL 内にて、JPEG 画像、PDF 画像、全文テキストを公開。
＊四編下のみ東京大学国語研究室蔵本を使用。
http://kokugo.l.u-tokyo.ac.jp/data/bunken.php?title=hananoshimadai

本コーパスでは、各作品の XML ファイルおよび『ひまわり』用データを提供する。『ひまわり』の利用により、本文・ルビの文字列検索の一覧や該当箇所の翻字テキスト・原本画像の参照が可能となる。

2.文字の処理

・本文テキストのうち、Unicode4.0 で表現できない文字は**=** で示した。
・不明字は■で示した。
・漢字は現行の字体によることを原則としたが、次のものについては原表記に近似の字体を用い、区別した。「云／言」「开／其」「貟／貌」「匕／匙」「吊／弔」「哴／囁」「哥／歌」「売／殻」「帋／袋」「无／無」「栂／梅」「皈／帰」「艸／草」「計／斗」「弐／二」「餘／余」

3.データの種類と形式

本コーパスの公開形式は以下の 2 種類である。
・テキストファイル
・『ひまわり』用データ

3.1『ひまわり』へのインストール方法

既に『ひまわり』を使用している場合、以下のフォルダ・データを『ひまわり』フォルダに移動することによって、使用中の『ひまわり』内に本コーパスを組み込むことができる。
・Corpora フォルダ
・config_ninjobon.xml

3.2『ひまわり』を使ったコーパスの検索方法

基本的な検索方法は「全文検索システム『ひまわり』 利用者マニュアル」（ver.1.5）を参照されたい。

本コーパスにおいては、「丁」列のセルをダブルクリックすることにより該当頁の原本画像を（図 1）、「行」列のセルをダブルクリックすることにより該当箇所が強調された原本画像を参照できる（IIIF ビューア Mirador を利用）。また、その他の任意の列のセルをダブルクリックすることにより、テキスト画面を参照することができる（図 2）。

図 1：原本画像の参照

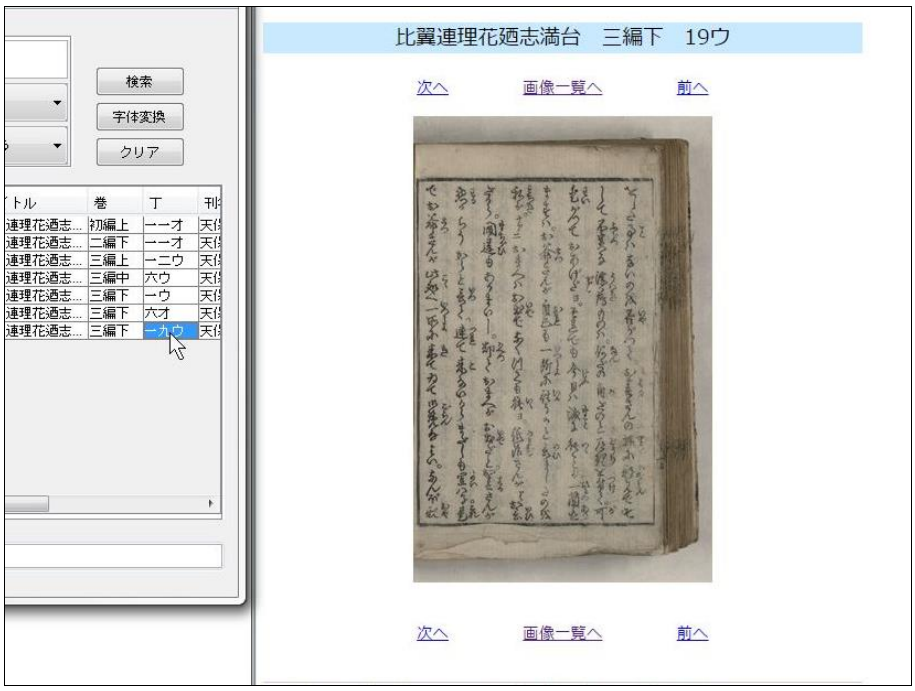
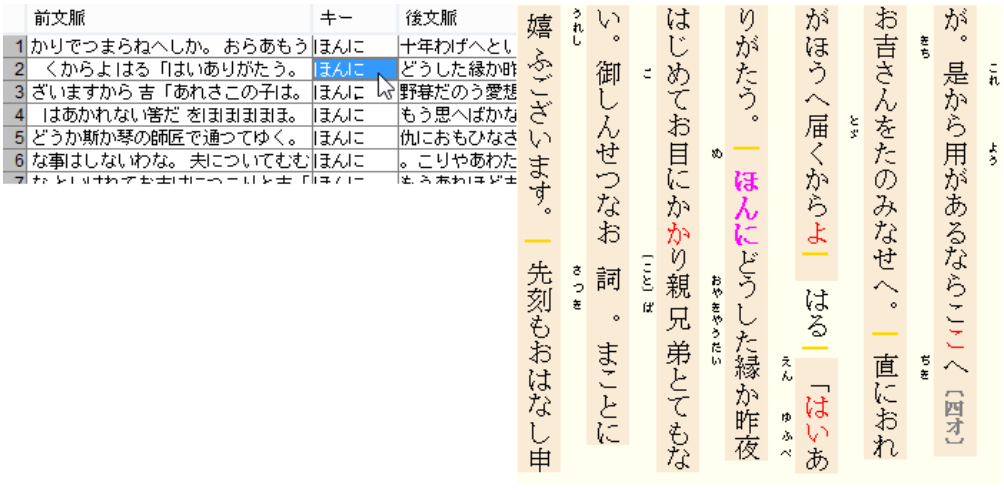


図 2：テキスト画面の参照



3.2.1 本文検索

本文にある文字列を漢字または平仮名により検索することができる。

※平仮名／片仮名の踊り字、片仮名はそれぞれ平仮名に正規化してあるため、本文の検索に用いることはできない。例えば「ハヽ」（原本）は「はは」として検索可能である。ただし、いわゆる「くの字点」は「/ \」として検索可能である。

※濁点の有無は原本表記に従ったものであり、正規化を経ていない。

※「2.文字の処理」で示した字群を含む語については、各字で検索する必要がある。

3.2.2 ルビ検索

原本に存在するルビ（振り仮名）を検索することができる。

※ルビの踊り字は原本表記に従ったものであり、正規化を経ていない。平仮名の踊り字は「ゝ」、いわゆる「くの字点」は「/ \」として検索可能である。

3.2.3 話者検索

以下の統一話者名（または以下の統一話者名の一部）により、当該人物による会話文等の冒頭部を検索することができる。

・『比翼連理花廼志満台』の話者名一覧（登場順）

お春・老婆・内儀・兵衛・弱冠・小六・お吉・和之一・強六・人々・皆々・糸助・悪者・娘・お滝・治兵衛・おさん・女中・船頭・小春・女・琴歌・鬼勝・浪人・お浜・男・若者・下女・和尚

3.3『ひまわり』を使ったコーパスの検索結果の見方

・前文脈／後文脈

検索した語の前後 20 字を表示する。

※「検索オプション」のタブより表示文脈の長さを変更可能。

・キー

検索した語を表示する。

・ルビ

原本に存在する振り仮名を表示する。

※表示されるのは原則として検索キー（が漢字文字列の場合）の最初の一字に対するルビである。「其方」に「そつち」のルビが付される場合、「其」「方」「其方」のいずれで検索してもルビ欄には「そつち」と表示される。

・**タイトル**

資料名を表示する。

・**巻／丁**

検索結果の所在を表示する。

※「丁」列のセルをダブルクリックすると該当箇所の原本画像が参照できる。

・**行**

検索結果の所在を表示する。

※「行」列のセルをダブルクリックすると該当箇所の前後を含む数行を青枠で囲った原本画像が参照できる（IIIF ビューア Mirador を利用)。

・**刊年／西暦**

資料の刊年を和暦／西暦で表示する。

・**話者**

該当箇所が会話文である場合、統一話者名を表示する。

・**vol ／p**

通し巻数／巻ごとのページ数を表示する。

※他の列と同様に、「vol」「p」をクリックすることにより並び替えが可能。

4 タグセット

本コーパスでは本文に文書構造・文字・表記などに関する情報を XML によって付与している。使用したタグは次表の通りである。

表 1：タグセット

タグ	説明	属性
corpus	コーパス全体	
text	テキスト一冊のまとまり	title, volume, year, year_w url, vol
front	序文	
body	本文	
back	跋文	
article	記事	type
titleBlock	全体のタイトルの記述	
p	本文のひとかたまり	
block	内題などのブロック要素	
speech	会話文	source
warigaki	割書き	
quotation	字下げ、手紙など	
s	一文	
speaker	話者	
hi	囲み、傍線	rend
r	ルビ	rt
lr	左ルビ	rt
odoriji	踊り字	originalText
vMark	濁点無表記箇所	
goji	合字	
g	外字などの特殊文字	type
char	カタカナなど	script
corr	本文修正箇所	type, subType, originalText
unclear	原本の不鮮明箇所	
gap	判読不明箇所	
pb	頁開始位置	n,num
lb	行開始位置	

参考文献

藤本灯・北﨑勇帆・市村太郎・岡部嘉幸・高田智和、「人情本のコーパス化」、日本語学会2015年度秋季大会予稿集、pp169-174