『日本語歴史コーパス』における原文KWIC表示機能の実装

著者	小木曽 智信, 岡 照晃, 中村 壮範, 八木 豊
雑誌名	言語資源活用ワークショップ発表論文集
巻	2
ページ	252-257
発行年	2017
URL	http://doi.org/10.15084/00001526

『日本語歴史コーパス』における原文 KWIC 表示機能の実装

小木曽 智信（国立国語研究所言語変化研究領域）・岡照晃（国立国語研究所コーパス開発センター）・中村壮範（マンパワーグループ株式会社）・八木豊（株式会社ピコラボ）

Implementation of “Original Text KWIC” Display Function in the Corpus of Historical Japanese

Toshinobu Ogiso (NINJAL), Teruaki Oka (NINJAL), Takenori Nakamura (Manpower Japan Co., Ltd.), Yutaka Yagi (Picolab Co., Ltd.)

要旨　日本語史研究の基礎資料は、残された文献に見られる用例であるが、その原文は今日一般に用いられる表記とは大幅に異なる形である場合が少なくない。例えば、『万葉集』は万葉仮名で、キリシタン資料は当時のポルトガル語式のローマ字で表記されている。こうした資料をコーパスとして形態論情報を付与し、現代人に読みやすいものとするためには、原文を校訂して漢字ひらがな交じりにした読み下し本文を用意する必要がある。一方で、読み下し本文では失われてしまう情報も少なくないため、用例には原文を併せて表示することが求められる。『日本語歴史コーパス』では従来、原文情報を保持しつつ必要な修正を行った上で形態論情報を付与して公開してきたが、原文情報の提供方法は限定的だった。今回新たに、コーパス検索アプリケーション「中納言」上で、原文の前後文脈付きで検索結果を表示できる機能を実装した。本発表ではこの原文 KWIC 表示機能について述べる。

1.『日本語歴史コーパス』における「原文」

過去の時代の日本語を研究するにあたっては、当時使われた用例がほとんど唯一の手がかりであり、それがどのように書かれているのかは日本語研究者にとってきわめて重要な情報である。用例の確認は、根本的には一次資料である原本そのものやその写真・画像にあたることができれば良いが、その一方で、原本のままでは現代人には読みづらく検索ができないため、現行の活字にそのまま直した翻字本文や、表記を読みやすく改め誤りを正した校訂本文が必要とされる。校訂済みの本文は、現代人にとって読みやすいだけでなく、『日本語歴史コーパス』のように形態素解析を施して単語の情報等を付与する際にも適している。このように、一つの日本語史研究資料であっても、原文画像、翻字本文、校訂本文とさまざまな段階があり、それぞれが研究上で必要とされる価値を持っている。

『日本語歴史コーパス』（小木曽 2016）の「平安時代編」では、小学館『新編日本古典文学全集』（新編全集）の校訂本文を底本としている。ここでは、コーパスにとっての「原文」は校訂本文が唯一のものである。ところが、「鎌倉時代編 I　説話・随筆」の『今昔物語集』では、本文が漢字カタカナ交じりであるだけでなく、部分的に漢文の語順で書かれ、返り点が表示されている。そのためこうした作品をコーパス化するにあたっては、底本である新編全集の本文をさらに改変し、形態素解析が可能な通常の語順の漢字ひらがな交じり文に直す必要があった（冨士池・田中 2012，冨士池ほか 2013）。また、「明治・大正編 I 雑誌」では、当時出版された雑誌そのものを底本としたため、自ら本文校訂を行う必要があり、そのために本文を修正したほか、漢字カタカナ交じり文の記事はひらがなに直した上でコーパス化を行っている（近藤 2016）。

『日本語歴史コーパス』のテキストは XML で構造化・タグ付けされており、以上のよう

な本文修正については元の様態を保存し再現できるようにタグ付けがなされている。このタグによって再現される元のテキストを「原文文字列」と読んでいる。

2. これまでの「中納言」と原文表示

『日本語歴史コーパス』は Web 上のコーパス検索アプリケーション「中納言」を通じて提供されているが、ここでの検索結果として表示される本文（KWIC の前後文脈）は、形態素解析の対象となった校訂済みの本文である。したがって、修正前の本文については文脈からは確認できない。しかし、日本語研究のための資料として、調査対象の用例についてはできるかぎり原態を示したいため、検索結果の表に「原文文字列」という列を設け、ここでキーとなった語の「原文」の様態を確認できるようになっている（中納言バージョン 2.2.2.2, 図１）。

サンプルID	開始位置	連番	コア	前文脈	キー	後文脈	語彙素読み	語彙素	品詞	原文文字列	振り仮名
60M明六1875_39001	10800	7490	1	法律|の|條數|其|幾|千萬|なる|を|知ら|ず#其|東洲|に|在|て|は|唐律|を|以|て|	粗	|備はれ|る|者|と|し|我が|國|往昔|の|律法|亦|之|に|本づく#而て|明律|あり#清律|あり	ホボ	略	副詞	粗	
60M明六1874_17001	5500	3750	1	其|詳|を|論ぜ|ん|に|民選|議院|の|事|は|世上|に|公論|あり|て|其|制|も|亦|	畧ぼ	|分明なれ|ば|今|又|贅論|せ|ず#衒|に|は|只|其|財務|に|預かる|要件|のみ|を|掲ぐ	ホボ	略	副詞	畧ボ	
60M明六1874_18006	4010	2730	1	従|て|議|を|作り|譜|を|案じ|て|調|を|爲す|の|法|なり#此|法|支那|に|は|	畧ぼ	|これ|有り#歐米|諸國|に|は|殆ど|精妙|を|極む#只|我|邦|に|未だ|開け|ず#今|之|を	ホボ	略	副詞	畧ホ	
60M明六1875_33004	8960	5790	1	に|來入し|其|勢|を|變ぜ|ずし|て|若干|歳月|を|経|ば|終|に|は|金紙|の|間|	畧ぼ	|平均|を|成す|に|至る|べし#一たび|平均|を|成す|に|至ら|ば|是より|紙幣|を|燒却|する|て	ホボ	略	副詞	畧ホ	
60M明六1875_41003	5250	3510	1	ん|や#勇|を|養ふ|は|武|に|あり#我|邦|風習|太古|より|武|を|重んじ|名|を|惜む#	畧と	|日耳曼|に|似|たり#王室|中古|の|文弱|は|上|たる|者|武|を|卹む|に|より|而して|武士|の	ホボ	略	副詞	畧と	
60M明六1874_05004	14210	9460	1	を|壓服|する|極めて|難く|人民|常|に|自由|の|氣質|を|保持|せ|ん#亞非利加|も|亦|其|氣候|	畧々	|亞細亞|の|南方|と|似たる|を|以て|其|人民|の|隷從|する|概ね|亦|相|同じく|又|亞米利加	ホボ	略	副詞	畧ノ\	
60M明六1874_04001	10480	7220	1	歐羅巴|J|に|於|て|は|氣候|既に|漸|を|逐|て|強|と|強|と|相|隣し|其|力|	畧	|相|比敵|する|を|以|て|甲國|より|乙國|を|一擧し|て|壓服|する|て|極めて|難し#是れ|即ち	ホボ	略	副詞	畧	

図 1　これまでの「中納言」の「原文文字列」表示

前後文脈まで含めた原文は提供されておらず、機能は限定的である。それでも、これまでに公開してきた資料については機能的に十分であったと言える。

3.「万葉集」「キリシタン資料」と原文

今回、新たに『日本語歴史コーパス』に『万葉集』とキリシタン資料『天草版平家物語』『エソポのハブラス（伊曽保物語)』を追加することとなった。これらの資料は、原文が漢字仮名交じり文ではないため、原文と書き下した校訂本文との差が甚だしく、従来の枠組みでは扱いきれない。

「奈良時代編 I 万葉集」として収録される『万葉集』の原文は周知の通り漢字だけの万葉仮名で書かれており、通常はこれを漢字仮名交じり文に書き下したものを読んでいる。次に例を示す。

金野乃 美草苅葺 屋杼礼里之 兎道乃宮子能 借五百礒所念
秋の野の み草刈り葺き 宿れりし 宇治のみやこの 仮廬し思ほゆ

（7 番歌）

熟田津尓 船乗世武登 月待者 潮毛可奈比沼 今者許芸乞菜
熟田津に 船乗りせむと 月待てば 潮もかなひぬ 今は漕ぎ出でな

（8 番歌）

許等〻波奴 樹尓波安里等母 宇流波之吉 伎美我手奈礼能 許等尓之安流倍志
言とはぬ 木にはありとも 愛しき 君が手馴れの 琴にしあるべし
（811 番歌）

許等騰波奴 紀尓茂安理等毛 和何世古我 多那礼乃美巨騰 都地尓意加米移母
言とはぬ 木にもありとも 我が背子が 手馴れの御琴 地に置かめやも
（812 番歌）

また、「室町時代編 II キリシタン資料」として収録される予定の『天草版平家物語』『エソポのハブラス』の原文は当時のポルトガル語ローマ字で書かれており、これも漢字仮名交じり文に書き下したものとともに利用されている。次に例を示す。

VManojô. Qẽgueônobŏ, Feiqe no yurai ga qiqitai fodoni, ara ara riacu xite vo catari are.
QIICHI. Yaſui coto de gozaru : vôcata catari maraxôzu.
右馬の允．検校の坊，平家の由来が聞きたいほどに，あらあら略してお語りあれ．
喜一．やすいことでござる：おほかた語りまらせうず．
（平家物語　巻第一）

EVROPA no vchi Phrigiatoyŭ cunino Troia toyŭ jŏrino qinpenni Amoniato yŭ ſatoga vogiaru. Sono ſatoni nauoba Eſopoto yŭte, yguiŏ fuxiguina jintaiga vogiattaga, ſono jidai Europano tencani cono fitoni maſatte minicui monomo vorinacattato qicoyeta.
エウロパの中ヒリジヤといふ国のトロヤといふ城裡の近辺にアモニヤといふ里がおぢゃる。その里に名をばエソポというて、異形不思議な仁体がおぢゃったが、その時代エウロパの天下に、この人にまさって醜い者もおりなかったと聞えた。
（エソポのハブラス　エソポが生涯の物語略）

これらの資料において、上段に示した「原文」と下段に示した形態素解析対象となる読み下し本文は、既存のサブコーパスのように形態素解析等のために本文を校訂してカタカナをひらがなに直したといったレベルではなく、全く異なる文字種によるテキストとなっている。個々における「原文」は、研究上の利用価値が高く、漢字仮名交じり文では落ちてしまう貴重な情報を含んでいる。

たとえば、『万葉集』の例で言えば、当該例が一字一音の仮名で書かれているのか、漢字を訓読した例なのか、あるいは助詞等を補読したものなのか、という違いは、用例の価値を大きく左右するものである。また、原文が音仮名で書かれていれば上代歴史仮名遣いを確認することも可能であるが、こういった情報は漢字仮名交じりの本文では落ちてしまっている。また、キリシタン資料の原文では、ローマ字によって当時の音形が確認でき、特にオ列長音の開合の別が「ŏ」「ô」で示されていたりするが、漢字仮名交じりの本文ではこうした情報も確認できない。

このようなことから、『万葉集』やキリシタン資料にとっては前後文脈まで含めて原文が参照できることが望まれる。とくに、漢字仮名交じり文のテキストと対照する形で参照できることが望ましい。

4. 原文 KWIC 表示機能

『万葉集』とキリシタン資料のコーパスの構築にあたっては、当初より原文と形態素解析対象の本文とを別に用意し、それぞれを関連づけるアライメントを行ってきた（山田ほか 2015、Oka and Kono 2016）。原文と、形態素解析の対象となった漢字仮名交じりの本文、さらに形態素解析結果である短単位情報は、コーパスを格納した「形態論情報データベース」（小木曽・中村 2014）上でファイル頭からのオフセット値によって相互に関連付けられている。これにより、個々の単語について、前後文脈の原文テキストを出力することが可能になっている。

図 2 は、新しく公開予定の「中納言」上で前後文脈の原文を表示した例である。従来の漢字平仮名交じりの本文の KWIC（前文脈・キー・後文脈）の下段に、原文の KWIC（原文前文脈・キーの原文文字列・原文後文脈）を表示し、原文を形態素解析対象となった漢字仮名交じりテキストと対照しながら閲覧することが可能になった。検索結果のダウンロード時には、それぞれを別の列としたタブ区切りテキスト形式のデータとしてダウンロードされる。

図 2　公開予定の「中納言」の原文 KWIC 表示機能　（開発中の画面）

この機能の提供により、新しく公開される『万葉集』とキリシタン資料のデータの利用の幅が大きく広がることになるはずである。

5. 「原文」をめぐる注意点

このようにして提供される「原文」情報について、いくつか利用にあたって注意を要する点がある。

一つ目は、作品・サブコーパスごとに「原文」とされているものの実態が大きく異なることである。それぞれの中身を整理したものを表 I に示す。もともとの資料の性質が大きく異なるうえ、サブコーパスによって底本も違うためやむを得ないことであるが、利用に際しては注意が必要である。たとえば、「平安時代編」に含まれる作品の原文は、原点にまで戻れば、大部分が仮名からなる崩し字で書かれたテキストが原文であるが、底本を新編全集とする『日本語歴史コーパス』ではそこまで遡ることはできない。

二つ目は、漢字平仮名交じりの本文と原文とが、一対一に対応するとは限らないということである。たとえば、『今昔物語集』においては「未」が「未だ～ず」と読まれるような“再読文字”がある。この場合、原文の一文字が、本文中の離れた 2 箇所に対応することとなる。返り点が入るような“返読”の箇所でも、対応が 2 箇所に分かれる場合がある。また、同じよ

表 1　『日本語歴史コーパス』の「原文」

サブコーパス・資料		原文	本文（漢字ひらがな交じり）
奈良時代編 I 万葉集		新編全集（原文・万葉仮名）	新編全集（読み下し文）
平安時代編		新編全集の校訂済み本文（共通）	
鎌倉時代編 I	今昔物語集	新編全集の本文（漢字カタカナ交じり）	新編全集を日本語順に整形したもの
	その他	新編全集の校訂済み本文（共通）	
室町時代編 I 狂言		『大蔵虎明能狂言集翻刻註解』のテキスト	濁点付与など一部のみ校訂したもの
室町時代編 II キリシタン資料		原典のローマ字	ローマ字から生成した漢字仮名交じり文
明治・大正編 I 雑誌		原典をテキスト化したもの（一部漢字カタカナ交じり、踊り字あり）	原テキストを校訂した漢字仮名交じり文

うな漢文的表記で、読み下した場合に対応する読みがない“置字”がみられることがあり、この場合には原文の文字に対応する本文がないことになる。この逆の場合として、『万葉集』などで多く見られる“補読”がある。たとえば原文「金野乃」を本文「秋の野の」と読むとき、一つ目の「の」は原文に対応する文字がない。

以上のような一対一対応しないものについてもコーパスのデータベース上では問題なく格納されているが、「中納言」上での実際の利用にあたっては対応部分が見当たらなかったり複数あったりするために注意を要する。図 3 は原文 KWIC 部分を拡大したものだが、枠

10-萬葉 0806_00006	30710	こ|の|道||にて|し|おし|てる|や|難波|の|海|と|名付け||けらし|も#士|や|も| 空しく|(ある)||	べき	|万代||に|語り|継ぐ|べき|名||は|立て|ず|し|て#我|が|背子|が|着る|衣|薄し|佐保|風
		直超乃此徑尓吾師押照哉難波乃海跡名附家良思恭士也母空		応	有万代尓語続可名者不立之而吾背子我著衣薄佐保風者疾莫吹及家左右
10-萬葉 0806_00019	11020	母|の|命|几ろか||に|心|尽くし|て|思ふ|らむ|そ|の|子|なれ|や|もますらを|や|空しく|(ある)||	べき	|梓弓|末|振り|起し|投矢|持ち|千尋|射渡し|剣大刀|腰|に|取り|佩き|あしひきの|の|つ|峰|踏み
		知智乃実乃父能美許等波播蘇業乃母能美己等於保呂可尓情尽而念良牟其子奈礼夜母大夫夜无奈之久		可	在梓弓須恵布理於許之投矢毛知千尋射和多之剣刀許思尓等理波伎安之比奇能八峯布美越左之麻久流情不障後代乃可多利都具倍久名乎多都倍志母
10-萬葉 0806_00020	62850	めづらし|も|かく|し|にそ|見し|明らめ|め|秋|立つ|ごと||に#天地|を|照らす|日月|の|極みなく|(ある)||	べき	|もの|を|何|を|か|思|は|む|乳|ざ|子|ども|狂わざ|な|せ|そ|天地|の|堅め|し|国|そ
		時花伊夜米豆良之母加久之許曾売之安伎良米晩阿伎多都其等尓天地乎呂良須日月乃極奈久阿流		倍伎	母能乎奈尓乎加於毛波牟伊射子等毛多波和射奈世曾天地能加多米之久尓曾夜麻登之麻祢波
30-今昔 1100_12014	13950	静め|て|念ひ|を|専|に|し|て|仏|を|念じ|奉ら|ば|、必ず|其|の|利益|は|(有)||	べき	|也|と|なむ|語り|伝へ|たる|と|や|。
		然レバ、人若シ急難ニ値ハム時ハ、心ヲ静メテ念ヒヲ専ニシテ仏ヲ念ジ奉ラバ、必ズ其ノ利益ハ有		べキ	也トナム語リ伝ヘタルトヤ。
30-今昔 1100_12024	17340	差せ|り|。#夏|の|事|なれ|ば|、|土葬|也|と|云へ|ども|、|少|も|香|は||(有)||	べき	|||に|、|露|其|の|臭き|香|無し|。#其|の|後|、|七日|毎||に|仏経|を|供養
		其ノ上ニ＝都婆ヲ起テ、釘抜ヲ差セリ。夏ノ事ナレバ、土葬也ト云ヘドモ、少モ香ハ		可	有｛牛｝ニ、露其ノ＝牛香無シ。其ノ後、七日毎ニ仏経ヲ供養ズ。
30-今昔 1100_12034	18510	に|抱|て|乗せ|む|事|にそ|何|なる|べき|事|に|か|有ら|む|。#極|て|恐れ|(有)||	べき	|事|かな|」#と|思ひ臥|たる||に|、|上長押|より|鼠|の|走|渡る||に|、|枕上||に|物
		若シ、健ミテ参ザラムヲ、強ニ馬ニ抱テ乗セム事コ|何ナルベキ事ニカ有ラム。極テ恐レ		可	有｛牛｝事カナ」ト思ヒ臥タルニ、上長押ヨリ鼠ノ走渡ルニ、枕上ニ物ノ撞キ落サレタルヲ見レバ、紙ノ破也。
60M明六 1874_01002	9180	其|内||に|は|学者|輩出し|て|洋字|を|以|て|和漢|の|史傳|等|を|記する|者|(ある)||	べき	|なれ|ども|要する||に|二重|の|労たる|ヿ|を|免かれ|ず#是|其|不利|の|三なり|#此
		然シナガラ其内ニハ学者輩出シテ洋字ヲ以テ和漢ノ史傳等ヲ記スル者アル		べキ	ナレトモ要スルニ二重ノ労タルヿヲ免カレズ是其不利ノ三ナリ此三不利ヲ胥シテ従来未タアラザル處ノ奇法ヲ行ハント欲スルハ三尺ノ童子ト雖トモ其至難ナルヿヲ知ルベシ
60M明六 1874_05005	17010	て|全く|同位|同等|の|者|とはなすが|故|に|又|他教|他派|を|容忍|する|制度|の|(ある)||	べき	|理|絶へて|立れ|あら|ず#臣民|論述|の|自由|禮拜|の|自由|其|他|總て|神道||關係|せ
		總テ諸教諸＝共ニ憲法上ニ於テ全ク同位同等ノ者トナス力故ニ又他教他派ヲ容忍スル制度ノアル		へキ	理絶ヘテ立レアラス臣民論述ノ自由禮拜ノ自由其他總テ神道ニ關係セル事ノ自由ハ決シテ政府ヨリ授與セラレタル特權ニアラス
60M明六 1874_20001	10480	吾人|なべて|之|を|言ふ|ヿ|を|憚る#是|人情|の|自然||に|し|て|人道|殉|に|然|(ある)||	べき	|ヿ|なり#夫|の|罪惡|を|罰する||は|政府|の|職掌|なり#然れ|ども|律法|循|舊惡|滅免|の|例|あり
		夫レ舊惡ハ吾人ナベテ之ヲ言フヿヲ憚ル是人情ノ自然ニシテ人道殉ニ然アル		べキ	ヿナリ夫ノ罪惡ヲ罰スルハ政府ノ職掌ナリ然レトモ律法循舊惡＝免ノ例アリ
60M明六 1875_31001	19280	する|者|なれ|ば|本來|の|理|に|於|て|は|人民|上||に|あり|て|政府|下||に|(ある)||	べき	|が|如く|なれ|ども|政府|は|人民|を|保護|する|の|大權|を|掌握|せ|ざる|可ら|ざる|を|以
		(人民ハ主ニシテ政府ハ人民ノ爲メニ存在スル者ナレハ本來ノ理ニ於テハ人民上ニアリテ政府下ニアル		へキ	力如クナレトモ政府ハ人民ヲ保護スルノ大權ヲ掌握セサル可ラサルヲ以テ必ス上位ヲ占ムルヲ要スルナリ

図 3　原文 KWIC が本文と一対多で対応する例

で囲んだ中の「べき」は、原文「可有キ」を本文で「有るべき」と読んでいるため、原文の２箇所に対応する。このような場合には最も前方の対応箇所を「キー」としてとり、後方の対応箇所は原文後文脈中の色つきの括弧で囲んで示す仕様となっている。

５．おわりに

以上のように、原文と読み下し本文との乖離が甚だしい日本語史資料のコーパス公開にあたって、コーパス検索アプリケーション「中納言」に原文 KWIC 表示機能を実装する。これにより、『日本語歴史コーパス』の利用の幅が一段広がることとなった。今後、こうした機能の活用により、コーパスなしでは困難であった新しい研究が実現することを期待したい。

謝　辞

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進展開」，および科研費基盤(A)「日本語歴史コーパスの多層的拡張による精密化とその活用」による成果の一部である。

文　献

冨士池優美・田中牧郎(2012).「今昔物語集の返読文字について―形態素解析の前処理を通して―」、日本語学会 2012 年度春季大会予稿集、pp.223-228

冨士池優美・河瀬彰宏・野田高広・岩崎瑠莉恵(2013).「『今昔物語集』のテキスト整形」『第 4 回コーパス日本語学ワークショップ予稿集』, pp.125-134.

小木曽智信・中村壮範(2014).『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, 自然言語処理, 21(2), pp.301-332.

山田祐実・大村舞・鴻野知暁・Kevin Duh・小木曽智信・松本裕治(2015).「万葉集を対象とした原文と読み下し文のアライメント」『第 8 回コーパス日本語学ワークショップ予稿集』, pp.243-252

小木曽智信（2016).『日本語歴史コーパス』の現状と展望, 國語と國文學, 93(5), pp.72-85.

Teruaki OKA, Tomoaki Kono (2016). Original-Transcribed Text Alignment for Manyosyu Written by Old Japanese Language, *Language Technology Resources and Tools for Digital Humanities (LT4DH),* (http://researchmap.jp/mukbfhtwa-2098193/#_2098193 より閲覧可能)

近藤明日子(2016).「『明六雑誌コーパス』『国民之友コーパス』の構築―形態論情報を付与した近代雑誌コーパスの設計―」日本語の研究, 12(4), pp.167-174

関連 URL

『日本語歴史コーパス』　http://pj.ninjal.ac.jp/corpus_center/chj/

コーパス検索アプリケーション「中納言」　https://chunagon.ninjal.ac.jp/