キリシタン資料のローマ字原文対応和文テキストの作成

片山　久留美・小木曽　智信・中村　壮範（人間文化研究機構　国立国語研究所）

国立国語研究所で構築中の『日本語歴史コーパス』に『室町時代編Ⅱキリシタン資料』として『天草版平家物語』『天草版伊曽保物語』の２作品が追加された．これらの資料は原本がポルトガル式ローマ字により表記されており，当時の発音を知ることができる資料として日本語研究上重要な位置を占める．コーパス化にあたっては，ローマ字テキストと和文テキストを用意し両者をアラインメントにより対応付けることで二つのテキストを同時に参照可能にした．その際，仮の和文テキストを作成して形態素解析を行い，付与された形態論情報を利用することによって，均質性の高い独自の和文テキストを自動で出力したほか，ローマ字テキストと和文テキストのアラインメント作業も効率よく行うことが可能となった．

Creating parallel texts of "Christian materials" in Muromachi era: the original Romanized Text and its transliteration to Japanese characters

Kurumi Katayama / Toshinobu Ogiso / Takenori Nakamura
(National Institute for Japanese Language and Linguistics)

The National Institute for Japanese Language and Linguistics created a morphologically annotated corpus of the “Christian Materials” (Kirishitan Shiryō) written in the Muromachi era. The original texts of the Christian Materials “*Feiqe no Monogatari*” and “*Esopo no Fabulas*” were written in the Roman alphabet with Portuguese spellings that were current in the 16th century. For the study of the phonetics and morphology of the Japanese language at that time, we had to make it possible to align the original Romanized texts with corresponding texts transliterated into Japanese characters. To conduct this task, we prepared intermediate texts with Japanese orthography in conformance with previous research, and performed a morphological analysis on those texts with a parser using information from our dictionary for Late Middle Japanese. In this way we were able to both automatically generate Japanese texts with a high level of regularity and also align the original Roman alphabet texts and their Japanese character counterparts into parallel texts in an efficient manner.

１．まえがき

国立国語研究所では，古代語から近代語までの資料を網羅する『日本語歴史コーパス』の構築を進めている．2018 年 3 月，『室町時代編Ⅱキリシタン資料』として新たに『天草版平家物語』『天草版伊曽保物語』（以下，天草版平家・伊曽保）の二作品，約 14 万語を追加した．これらの作品は，原本がローマ字表記されているというこれまでの『日本語歴史コーパス』所収の資料にはない特徴を持っている．ローマ字テキストは日本語研究上重要な意味を持つものであるため，これを参照できる形でのコーパス化が求められる．

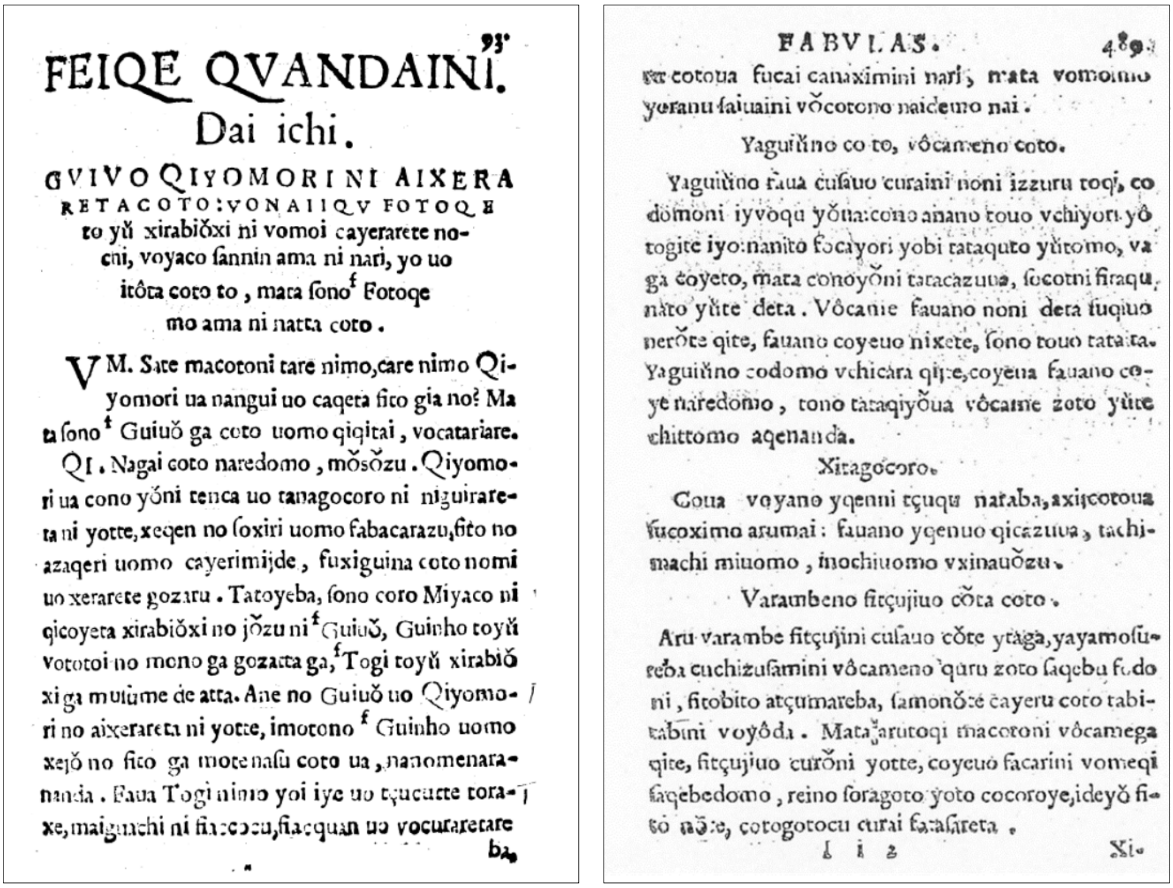
そこで本発表では，ローマ字テキストとそれに対応する和文テキストの作成手順，および双方を効率的に関連づける手法を報告し，この手法がローマ字テキスト以外の本文統制の必要な資料にも応用可能であることを述べる．

２．キリシタン資料とは

キリシタン資料とは，16 世紀から 17 世紀にかけて，主にカトリックの宣教師がキリスト教布教のためにつくった一連の日本語資料群のことを指す．布教に用いるキリスト教の教義書を日本語で書いたものや，宣教師が日本語を学ぶための辞書・文法書・読み物などがある．今回コーパス化の対象とした二作品は，宣教師が日本語を学ぶためのリーダーとして使用されたものである．『天草版平家物語』は，『平家物語』の内容を喜一検校という語り手が右馬の允という聞き手に語って聞かせる対話形式で書かれている．『天草版伊曽保物語』は，イソップ寓話を平易な語り口で描いたものである．いずれも当時の口語体で書かれており，中世の重要な口語資料となっている．

天草版平家・伊曽保は，全文がポルトガル式ローマ字によって表記されている（図 1）．これにより，漢字や仮名による表記ではわからない当時の発音や音韻が明らかになる．たとえば（1）では「逸物（イチモツ）」が「ychimot」と語末を t 入声音で表記されている．また(2)の「rŏdô」のように，オ段長音の開合の別なども明示されており，当時の発音を窺い知ることができる．

口語体かつローマ字表記であるという資料は，天草版平家・伊曽保の他にない．そのため数あるキリシタン資料の中でも，この二作品が日本語研究上特に重要な資料と位置付けられているのである．



FEIQE QVANDAINI. 93.
Dai ichi.
GVIVO QIYOMORI NI AIXERA RETACOTO:VONAIIQV FOTOQE
to yŭ xirabiŏxi ni vomoi cayerarete no-
chi, voyaco ſannin ama ni nari, yo uo
itôta coto to, mata ſono[f] Fotoqe
mo ama ni natta coto.

VM. Sate macotoni tare nimo,care nimo Qi-
yomori ua nangui uo caqeta fito gia no? Ma
ta ſono[f] Guiuŏ ga coto uomo qiqitai, vocatariare.
QI. Nagai coto naredomo, mŏsŏzu. Qiyomo-
ri ua cono yŏni tenca uo tanagocoro ni niguirare-
ta ni yotte,xeqen no ſoxiri uomo fabacarazu,fito no
azaqeri uomo cayerimijde, fuxiguina coto nomi
uo xerarete gozaru. Tatoyeba, ſono coro Miyaco ni
qicoyeta xirabiŏxi no jŏzu ni[f] Guiuŏ, Guinho toyŭ
vototoi no mono ga gozatta ga,[f]Togi toyŭ xirabiŏ
xi ga mutŭme de atta. Ane no Guiuŏ uo Qiyomo-
ri no aixerareta ni yotte, imotono[f] Guinho uomo
xejŏ no fito ga motenaſu coto ua, nanomenara-
nanda. Faua Togi nimo yoi iye uo tçucutte tora-
xe,maiguchi ni fiacocu,fiacquan uo vocuraretare
ba,

FABVLAS. 489.
ga cotoua ſucai canaximini nari, mata vomoimo
yoranu ſaiuaini vŏcotono naideuo nai.

Yaguiŭno co to, vôcameno coto.

Yaguiŭno faua cuſauo curaini noni izzuru toqi, co
dŏmoni iyvoqu yŏnacono anano touo vchiyori yô
togite iyo:nanito ſocayori yobi tataquto yŭtomo, va
ga coyeto, mata conoyŏni tatacazuua, ſucotni firaqu,
nato yŭte deta. Vôcame fauano noni deta tuqiuo
nerŏte qite, fauano coyeuo nixete, ſono touo tataita.
Yaguiŭno codomo vchicara qite,coyeua fauano co-
ye naredomo, tono tataqiyŏua vôcame zoto yŭte
chittomo aqenanda.

Xitagocoro.

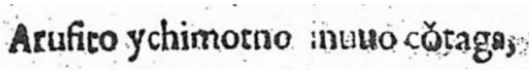
Coua voyano yqenni tçuqu nataba,axiſcotoua
ſucoximo arumai: fauano yqenuo qicazuua, tachi-
machi miuomo, mochiuomo vxinauŏzu.

Varambeno fitçujiuo cŏta coto.

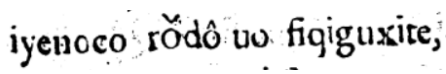
Aru varambe fitçujini cuſauo cŏte ytaga,yayamoſu-
reba cuchizuſamini vôcameno quru zoto ſaqebu ſodo
ni, fitobito atçumareba, ſamonŏte cayeru coto tabi-
tabini voyôda. Matatarutoqi nacotoni vôcamega
qite, fitçujiuo curŏni yotte, coyeuo facarini vomeqi
ſaqebedomo, reino ſoragoto yoto cocoroye,ideyŏ fi-
to naŏte, cotogotocu curai fataſareta.

Ii 2 Xi.

図 1『天草版平家』（左）『伊曽保』（右）影印

(1) 

→ Arufito ychimotno inuuo cŏtaga,
（ある人逸物の犬を飼うたが,）(『伊曽保』p.485)

(2) 

→ iyenoco rŏdô uo fiqiguxite,
（家の子郎等を引き具して,）(『平家』p.117)

３．コーパス構築上の課題

ローマ字原文は日本語研究において重要なものであるため，コーパス化にあたってはローマ字原文を利用できるようにすることが必須である．

一方でローマ字のままでは形態素解析を行うことができないだけでなく，コーパスの利用者が一見して意味を取りにくいという問題があり，漢字仮名交じりに変換したテキストも必要となる．天草版平家・伊曽保には先行研究において翻刻テキストや注釈書が出されているが，これらは，校注者それぞれが独自の方針で表記等を決定しており，形態素解析にかけにくいという問題がある．たとえば(3)では「à」を小字の「ぁ」と翻字しているが，特殊な表記であり形態素解析の際に支障を来す．また(4)や(5)のように，仮名遣いや漢字表記・仮名表記の使い分けには様々な立場が考えられ，表記揺れの原因となる．しかし天草版平家・伊曽保の原本はあくまでローマ字テキストであり，和文テキストは視認性を高めるためのものに過ぎない．そこで和文テキストには揺れや作業者の解釈を含まない，一貫した原則に基づく均質性が求められる．

(3) toxi coſo yotte gozàri tomo,
→ 年こそ寄って<u>ござぁり</u>とも， [5]
(『平家』p.172)

(4) namida uo voſayete cayetta.
→ 涙を<u>おさへ</u>て帰った． [4]
涙を<u>押さへ</u>て帰った． [5]
涙を<u>をさえ</u>て帰った． [1]
(『平家』p.102)

(5) Cono f Guendayŭ no fanguan to yŭua ,
→ この源<u>太夫</u>の判官と<u>いふは</u>， [4]
この源<u>大夫</u>の判官と<u>言うわ</u>， [1]
この源<u>太夫</u>の判官と<u>いふは</u>， [5]
この源<u>大夫</u>の判官と<u>言ふは</u>， [3]
(『平家』p.108)

つまり，キリシタン資料のコーパス化にあたっては次の４つの作業が必要となる．

Ⅰ ローマ字テキストの作成
Ⅱ 均質な和文テキストの作成
Ⅲ 和文テキストの形態素解析と解析誤りの修正
Ⅳ ローマ字テキストと和文テキストの単語レベルでの対応付け

これらの作業は，相互に関係のある内容であって，個々の作業を逐次行っていくことは効率が悪いだけでなく，不統一や矛盾の原因ともなり得る．そこで本コーパスでは解析しやすい仮の和文テキストを用意して形態素解析と修正を行った後に，辞書が持つ情報をもとに整形された和文テキストを生成し，ローマ字テキストとのアラインメントを取るという手法を取ることとした．『日本語歴史コーパス』が全文に単語の情報を付与しているという特長を活かし，上記のⅠとⅢを先に行うことで，仮のテキストに付された単語情報を用いてⅡ・Ⅳを容易に行えるようにするものである．

４．ローマ字テキストの作成

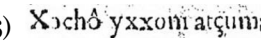
まず，原本のローマ字本文を正確に翻字したローマ字テキストを作成した．本コーパスの底本は大英図書館蔵の以下のものである．

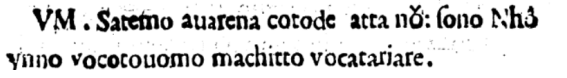
Nifon no cotoba to historia uo narai xiran to fossuru fito no tame ni xeua ni yauaraguetaru Feiqe no monogatari.
Esopo no fabulas : Latinuo uaxite Nippon no cuchito nasu mono nari.
（大英図書館蔵　請求記号 Or.59.aa.1）

影印本では判読が困難な箇所については原本を直接閲覧し確認を行っており，原本に忠実なローマ字テキストとなっている．

原本にはポルトガル式ローマ字特有の「à」「ẽ」「ô」「ŏ」「ſ」などの特殊なアルファベットや記号が用いられているが，これらについてもUnicode によって原本どおりのものを再現した．

またローマ字テキストにおいては分かち書きの有無も重要な情報となる．文字間に空白があると見られる箇所には「□」を入力することで分かち書きがされていることを示した．

(6) 
→ Xochô□yxxoni□atçumatte□fiŏgui□xite□yŭua:
（『伊曽保』p.492）

(7)

→ VM.Satemo□auarena□cotode□atta□nŏ:ſono□Nhô ynno□vocotouomo□machitto□vocatariare.
（『平家』p.394）

５．解析用本文と形態素解析

和文テキストについては，まず先行の注釈書等を参考にして作成した仮のテキストに対し，形態素解析辞書「UniDic」によって形態素解析を行った．仮のテキストは，同時代の口語資料のコーパスとして先行して公開されている『日本語歴史コーパス室町時代編 I 狂言』と同様の辞書で解析できるよう，基本的には歴史的仮名遣いに拠っている．(3)で挙げた小字の「ぁ」など，解析の支障となりうるものについては校訂を加えたが，この仮のテキストの日本語表記は十分に統一されておらず，仮名遣いや漢字・仮名表記の揺れなどが多数存在する．

形態素解析の結果，総語数は 139120 語となり，うち 8338 語に人手による形態論情報の修正を施した．解析精度は表 1 のようになっている．「L1 境界」は単語境界が正しく区切られているかを，「L2 品詞」はL1 に加え品詞認定の正しさを示す．「L3 語彙素」は UniDic の辞書見出しに相当する語彙素の認定の正しさを，「L4 発音形」は L1～L3 に加えて発音形認定の正しさを表している．たとえば「後」が「アト」か「ノチ」かを正しく認定できたかを評価したのが L3 であり，L4 は語彙素は同一ながら複数の読みのバリエーションがある場合，たとえば「狼」が「オオカミ」か「オオカメ」かといったことを正しく判定できているかを評価したものである．数字は F 値（適合率と再現率の調和平均）をパーセント表示にした値である．

表 1　仮テキストの解析精度

	L1 境界	L2 品詞	L3 語彙素	L4 発音形
精度（F 値）	99.3%	95.2%	94.0%	93.2%

仮テキストに仮名遣いや漢字・仮名等の表記揺れがあっても解析精度に大きな支障はなく，適切な解析用辞書を用いることで高い精度での解析を実現している．

誤解析はキリシタン資料特有の語彙や語形によるものが多くみられる．たとえば外国の地名や人名などの固有名詞は，仮テキストにおいても原本のアルファベット表記どおりとしたために新たに辞書への登録が必要となった．

(8) 名をば<u>Ｅｓｏｐｏ</u>と言うて（『伊曽保』p.409）
→語彙素「イソップ」発音形「エソポ」

発音形についても，ローマ字原文によって得られる読みに関する情報を最大限反映させるために修正が必要となった例がある．たとえば(1)で挙げた t 入声音を含むものは次のように発音形の該当部分を促音形としている．

(9) **逸物**　ychimot　（『伊曽保』p.485 など）
語彙素：逸物（イチモツ）発音形：**イチモッ**
(10) **末代**　matdai　（『平家』p.7 など）
語彙素：末代（マツダイ）発音形：**マッダイ**

他にも以下のようにローマ字原文の表す語形を形態論情報に反映させている．

(11) **成長**　xeigiŏ　（『伊曽保』p.415 など）
語彙素：成長（セイチョウ）発音形：**セージョー**
(12) **何**として
nattoxite（『伊曽保』p.415）
語彙素：何（ナニ）　発音形：**ナッ**
nantoxite（『伊曽保』p.425）
語彙素：何（ナニ）　発音形：**ナン**

境界認定においては，「喜うで」「及うで」のような四段動詞の連用形ウ音便の形はこれまで UniDic に書字形の登録がなく，図 2 のように誤解析となる例が見られた．

	誤解析					正解
キー	語彙素	発音形	品詞	活用型	活用形	
喜	喜ぶ	ヨロコバ	動詞-一般	文語四段-バ行	未然形-一般	発音形:ヨロコー 活用形:連用形-ウ音便
う	う	ウ	助動詞	無変化型	連体形-一般	
で	で	デ	助詞-接続助詞			語彙素:て

図 2　「喜うで」の解析結果

未知の固有名詞や特殊な活用形・発音形を新たに辞書に登録し人手による修正を加えることで，原本のローマ字表記を生かした形態論情報の付与を効率よく行っている．

6．均質な和文テキストの作成

形態素解析結果の修正が終わった後，付与された形態論情報を利用して新しい和文テキストを出力した．ここでは UniDic の「語形代表表記」の情報をもとにしたテキスト整形を行っている．

UniDic の各見出し語は，一般的な辞書の見出し語に相当する「語彙素」，異語形を区別する「語形」，異表記を区別する「書字形」，発音を区別する「発音形」という階層構造を持っている．「書字形」はこれまで国語研で開発してきた『日本語歴史コーパス』や『現代日本語書き言葉均衡コーパス』等のコーパスに実際に出現した表記形が登録されていくため，一つの「語形」に対し多くの「書字形」が登録されることになる．「語形代表表記」とは，そうした各語形の下に登録されている書字形のうち，最も代表的な表記と考えられるもののことを指す．主に『日本国語大辞典第二版』の表記を参考にして決定している．

語形代表表記による本文整形の例を見てみよう．たとえば仮の和文テキストには以下のような表記揺れが見られる．

(13) されどもそのことを聞きなほいた**僻こと**であれば（『平家』p.52）
(14) わが身の勅勘を許されうずと申さばこそ**僻言**でもあらうずれ（『平家』p.146）
(15) いや，それは**僻事**であらうずと言ひながら（『平家』p.180）

原本のローマ字はいずれも「figacoto」であり，「ヒガコト」という語形であることがわかるが，仮の和文テキストの表記は統一されていない．これらに図 3 のように形態論情報を付与する．

前文脈	キー	原文	後文脈	語彙素読み	語彙素	出現発音形	品詞
されどもそのことを聞きなほいた	僻こと	figacoto	であれば、とうとう帰れとて	ヒガゴト	僻事	ヒガコト	名詞-普通名詞-一般
勅勘を許されうずと申さばこそ	僻言	figacoto	でもあらうずれ:	ヒガゴト	僻事	ヒガコト	名詞-普通名詞-一般
宗盛いや、それは	僻事	figacoto	であらうずと言ひながら	ヒガゴト	僻事	ヒガコト	名詞-普通名詞-一般

図 3　語彙素「僻事」の形態論情報の付与

語彙素「ヒガゴト」の UniDic における階層構造を示したのが図 4 である．語彙素「僻事」の下に語形「ヒガゴト」「ヒガコト」があり，それぞれに複数の書字形を持っている．★を付した「僻事」が各語形の代表表記である．この階層構造を利用し，語彙素「僻事」の語形「ヒガコト」という形態論情報が付された語には，その語形代表表記である「僻事」を書字形として新たに出力する．

この手法を取ることで，同じ形態論情報を持つ語は常に同じ表記で出力されることになり，同語間での表記の揺れが生じない．仮名遣いの揺れといった問題も発生しなくなる．作業者の判断を差し挟む余地がなく，表記の揺れを排除した均質なテキストを自動で組み上げることが可能となる．

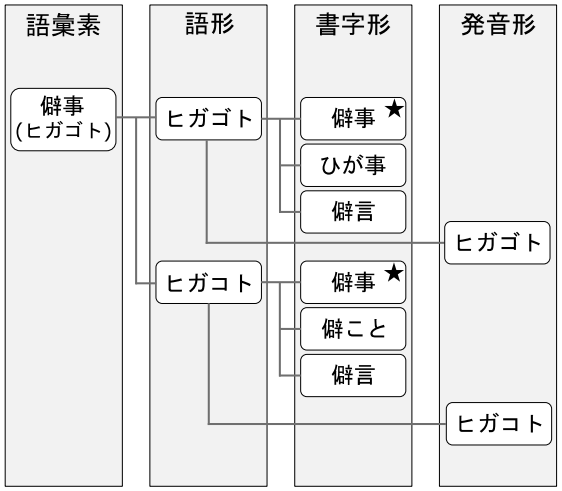


図 4　語彙素「僻事」の UniDic 階層構造

活用語についても同様に, UniDic の活用展開を利用することで出力が可能である. たとえば３節(4)で挙げた例文では,「voʃaye」の部分の表記が揺れていた. 仮の和文テキストでは以下のように歴史的仮名遣いによる表記になっている.

(16)（(4)再掲） 涙を**おさへ**て帰った.
（『平家』p.102）

天草版平家の中には「(涙を) オサエル」という例が複数現れるが, 仮のテキストでは以下のように表記揺れが見られる.
(17) 重盛涙を**抑へ**て申さるるは：(『平家』p.45)
(18) 涙を**押へ**かねさせられた. (『平家』p.179)
(19) 涙を**押さへ**出られた. (『平家』p.295)

(16)の活用語部分に形態論情報を付与すると, 図 5 のようになる.

キー	語彙素読み	語彙素	品詞	活用型	活用形
おさへ	オサエル	押さえる	動詞-一般	文語下二段-ハ行	連用形-一般
て	テ	て	助詞-接続助詞		
帰っ	カエル	返る	動詞-一般	文語四段-ラ行	連用形-促音便
た	タ	た	助動詞	助動詞-タ	終止形-一般

図 5「おさへて帰った」に対する形態論情報の付与

語彙素「押さえる」の語形「オサウ」の代表表記は「押さう」である. UniDic では活用語の場合, 図 6 に示したように書字形ごとに自動で活用を展開させることができる. 付与された形態論情報を用いることで, 語形「オサウ」の代表表記「押さう」を連用形-一般に活用させた「押さえ」という表記が出力される.「帰った」の部分も同様に, 語彙素「返る」の語形「カエル」に登録されている代表表記「返る」を連用形-促音便で活用させることで「返っ」という表記を出力する（図 7）. このように活用語についても, UniDic と連携することにより自動で揺れのない斉一な本文を出力することが可能になっている.

「押さえる」と「抑える」,「帰る」と「返る」のような同音異表記の語については, 漢字表記の違いが意味の違いに結びついていると考える立場もありうる. しかしこれらの語はいずれも『日本国語大辞典第二版』では同見出しとなっており, UniDic でも同語彙素として扱われている.その使い分けには様々な考え方があるが, 幅広い用途で用いられるコーパスにとって, 作業者の解釈や判断を含んだテキストを使用することは望ましくない. 本手法では付与された形態論情報に基づき一律で語形代表表記を使用することにより, 語の意味の解釈の問題には立ち入らず, コーパスにとって重要な本文の中立性・均質性を担保している.

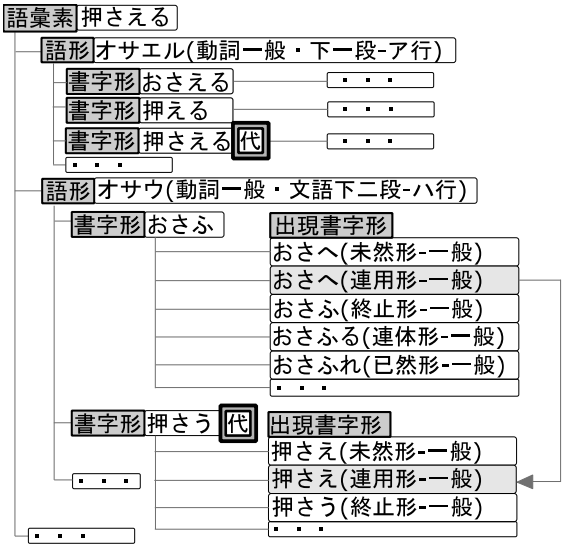


図 6 語彙素「押さえる」の活用展開

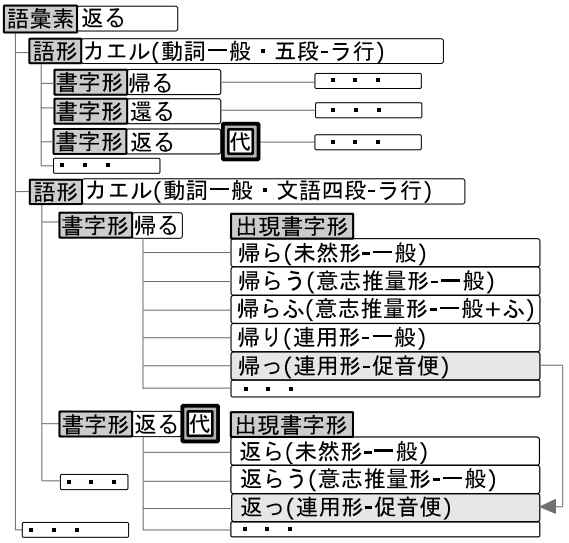


図 7 語彙素「返る」の活用展開

以上のように語形代表表記によって本文を出力していくと, (13)〜(19)の例および３節で表記揺れの例として挙げた(3)・(5)の例は以下のような表記になる. 新しい本文は歴史的仮名遣いではなく, 表音的な仮名遣いになることが特徴である.

(13)’ 然れどもその事を聞き直いた僻事で有れば
(14)’ 我が身の勅勘を許されうずと申さばこそ僻事でも有らうずれ
(15)’ 否, それは僻事で有らうずと言いながら
(16)’ 涙を押さえて返った.
(17)’ 重盛涙を押さえて申さるるは：
(18)’ 涙を押さえ兼ねさせられた.
(19)’ 涙を押さえ出られた.
(3)’ 年こそ寄って御座ありとも,
(5)’ この源太夫の判官と言うは,

7．ローマ字テキストとのアラインメント

ローマ字テキストと和文テキストの単語レベルでの対応付けについても UniDic の形態論情報を用いて効率よく行うことが可能となる．付与された形態論情報には「仮名形」や「出現発音形」がある．出現発音形は各語の発音形をカナで表したものである．この出現発音形の情報を用いて，図 8 のようにローマ字テキストとのアラインメントを行った．カナとキリシタン資料で用いられているポルトガル式ローマ字の対応表を作成し，出現発音形のカナ一字ずつとそれに対応するローマ字を学習することで，ローマ字テキストとのアラインメントを高精度で行うことが可能になっている．表 2 はカナとローマ字の対応表の一部である．サ行の「s」と「ſ」など，一つのカナに対し複数のローマ字表記が対応するものもあるが，種類は多くないため簡単な処理で対応できる．長音・促音・拗音・撥音などの特殊拍についても，「シャ：xa」「ヒョ：fio」「リョー：riŏ，reô」「ッタ：tta」などのようにパターン化することが可能なため，解析の大きな支障にはならない．

和文	涙			を	押さえ			て	返っ			た
出現発音形	ナ	ミ	ダ	オ	オ	サ	エ	テ	カ	エ	ッ	タ
カナ対応ローマ字	na	mi	da	uo vo	uo vo	ſa sa	ye	te	ca	ye	t	ta
	⇕ 対応が容易											
ローマ字テキスト	namida□uo□voſayete□cayetta											

図 8　出現発音形とローマ字テキストの対応

表 2　カナとポルトガル式ローマ字の対応表（一部）

ア	イ	ウ	エ	オ
a	i y j	u v	ye	uo vo
カ	キ	ク	ケ	コ
ca qua	qi	cu qu	qe	co
サ	シ	ス	セ	ソ
sa ſa	xi	su ſu	xe	so ſo
タ	チ	ツ	テ	ト
ta	chi	tçu	te	to
ハ	ヒ	フ	ヘ	ホ
fa	fi	fu	fe	fo
ガ	ギ	グ	ゲ	ゴ
ga gua	gui	gu	gue	go
ザ	ジ	ズ	ゼ	ゾ
za	ji gi	zu zzu	je	zo
ワ				
ua va				

（ヘボン式ローマ字と変わらない行および大文字・小文字の区別は省略）

単語レベルでのアラインメントの精度は，適合率 98.4%，再現率 98.1%で F 値は 98.2%であった．

8．おわりに

キリシタン資料のローマ字テキストと和文テキストの対照本文を作成する方法について論じた．こうして作成した本文は，『日本語歴史コーパス』の検索アプリケーション「中納言」上の原文 KWIC 機能によって，形態論情報と同時に参照できる（図 9）．「中納言」の「文字列検索」の機能を使うことで，ローマ字からの検索も可能となっている．

本研究では，解析しやすい仮の和文テキストを作成して形態素解析を行なったのち，形態論情報をもとにして新しい和文テキストを自動生成するという手法を提案した．以下の利点が挙げられる．

① 形態素解析が容易になり，解析精度が担保され形態論情報修正のコストが下がる．
② UniDic の語形代表表記の情報を用いた本文生成を行うことで，一貫した原理による均質な和文テキストをコーパス本文とすることができる．
③ UniDic の出現発音形の情報を参照することで，ローマ字テキストとのアラインメントを容易にかつ高精度で行える．
④ 直接形態素解析にかけることが難しかったローマ字テキストも，和文テキストとアラインメントを取り対応させたことにより形態論情報と関連づけることができる．

形態素解析用の仮本文を用い，UniDic による解析結果を仲介として豊富な形態論情報を利用することで，ローマ字原文に対応した和文テキストを効率的に作成・連携させることが可能になった（図 10）．

本手法は，他の資料の本文の統制や，標準化を必要とする資料などにも応用可能なものである．今回は先行の翻刻テキスト等を参考に仮テキストを作成したが，形態素解析が可能な本文があればこの手法を援用することができる．たとえば複数の人物によって音声資料を文字起こししたテキストの場合表記に揺れが生じる可能性があるが，本手法を用いれば揺れのない斉一なテキストを効率的に作成することが可能である．語形代表表記による本文出力だけでなく，UniDic に登録されている情報を用いて歴史的仮名遣い・現代仮名遣いによる本文の出力や発音形・仮名形などの出力等も可能であり，目的に応じた利用ができる（図 11）．

図 9 検索アプリケーション「中納言」による表示

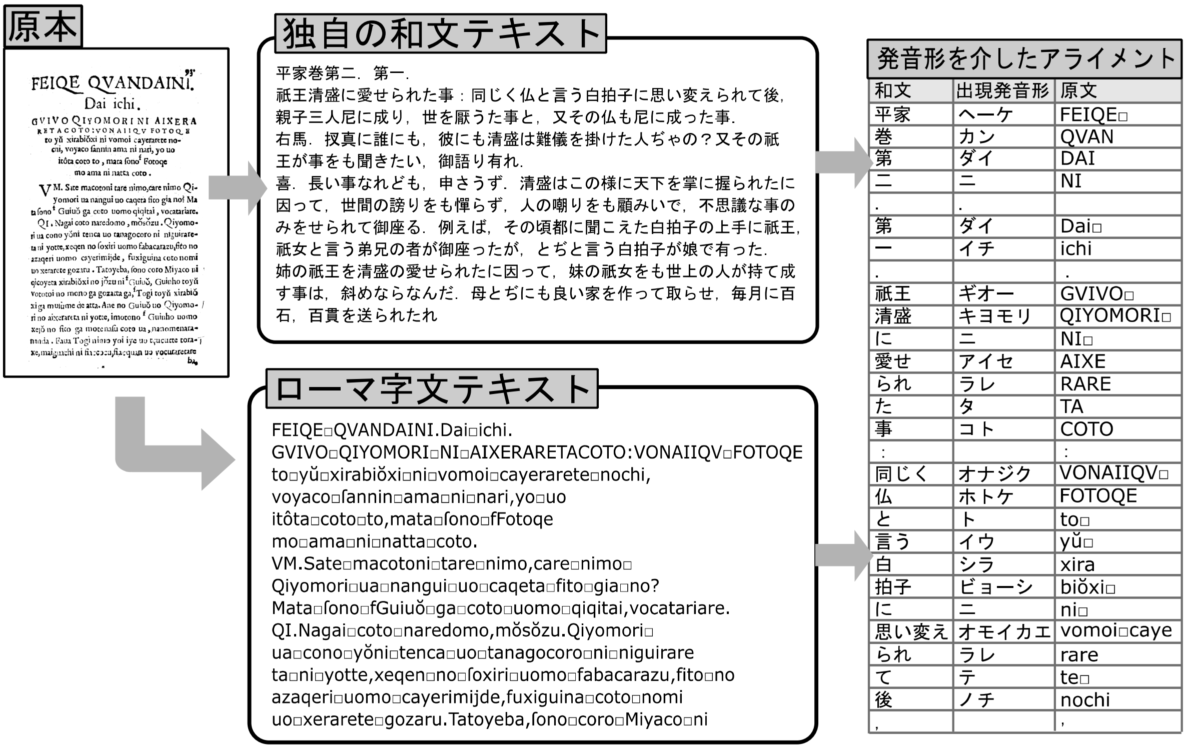


図 10 原本からのローマ字原文対応和文テキストの作成イメージ

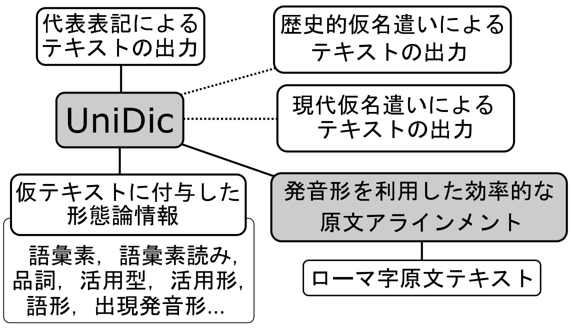


図 11 UniDic の形態論情報を介した多様な出力

付記 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新展開」の研究成果を報告したものである．

参考文献

本文中の原本画像は下記[2]および[6]の影印資料を使用した．

[1] 江口正弘：天草版平家物語対照本文及び総索引本文篇，明治書院（1986）．
[2] 江口正弘，溝口博幸（編）：天草本平家物語資料大成，尚文出版（2005）．

[3] 江口正弘（注釈）：天草版平家物語全注釈，新典社（2009）
[4]亀井高孝，阪田雪子（翻字）：平家物語：ハビヤン抄　キリシタン版，吉川弘文館（1966, 1980）
[5] 近藤政美，池村奈代美，浜千代いづみ（編）：天草版平家物語語彙用例総索引（１），勉誠出版（1999）
[6] 江口正弘（編）：天草版伊曽保物語影印及び全注釈　言葉の和らげ影印及び翻刻翻訳，新典社（2011）
[7] 大塚光信，来田隆（編）：エソポのハブラス本文と総索引，清文堂出版（1999）
[8] 国立国語研究所：日本語歴史コーパス室町時代編Ⅱキリシタン資料（短単位データ 1.0／長単位データ 1.0，中納言バージョン 2.4.2）
https://pj.ninjal.ac.jp/corpus_center/chj/muromachi.html（参照 2018-10-25）
[9] 小木曽智信，岡照晃，中村壮範，八木豊：『日本語歴史コーパス』における原文 KWIC 表示機能の実装，言語資源活用ワークショップ 2017 発表論文集，pp.252-257（2017）
[10]日本国語大辞典第二版，小学館（2000-2002）