『日本語歴史コーパス 室町時代編Ⅱキリシタン資料』形態論情報の概要

2018 年 3 月 30 日　片山久留美

１．言語単位

『日本語歴史コーパス 室町時代編Ⅱキリシタン資料』（以下「本コーパス」と呼ぶ）には以下の２種類の言語単位が用意されている。

（１）用例収集を目的とした「短単位」

（２）言語的特徴の解明を目的とした「長単位」

短単位・長単位ともに、代表形（語彙素読み）・代表表記（語彙素）・品詞・活用型・活用形を与える。代表形は国語辞典の見出しに、代表表記はその見出しに与えられた漢字等の表記に相当する。

これらは『現代日本語書き言葉均衡コーパス（BCCWJ）』で採用した単位を基に設計したものである。本コーパスの言語単位は、通時的な日本語研究での利用を可能にするため、『BCCWJ』をはじめとする現代語のコーパスや『日本語歴史コーパス（CHJ）』の他の時代のコーパスとの互換性の保持を図っている。

その一方で、『BCCWJ』や『CHJ』平安時代編等の規程をそのまま用いるのではなく、本コーパス用に単位認定規程の修正・拡張を行った。基本的には同時代資料のコーパスとして先行して公開されている『CHJ』室町時代編Ｉ狂言（以下「『室町時代編Ｉ狂言』」と呼ぶ）と同様の規定に基づいているが、資料の特性上一部本コーパス独自の処理を行った箇所がある。

本文書では短単位・長単位の単位認定規定を概説しつつ、本コーパス独自の処理をした箇所についてもコーパス使用の際に留意が必要な点を中心に説明する。

２．短単位の概要

短単位は、言語の形態的側面に着目して規定した言語単位である。短単位の認定にあたっては、まず意味を持つ最小の単位（最小単位）を規定し、その最小単位を文節の範囲内で短単位認定規程に基づいて結合させる（もしくは結合させない）ことで認定する。

（１）最小単位

● 最小単位は現代語において意味を持つ最小の単位である。本コーパスにおける最小単位については、現代語との関連を重視して、原則として現代語を対象とした最小単位認定を行うが、必要に応じて使用実態や平安時代編・近代語コーパスの状況に基づき個別の判断をすることがある。語種等の違いにより、それぞれ次のように認定する。

※「／」 は最小単位の分割位置を表す。

和　語：／これ／ほど／まで／は／無かっ／た／物／を／

漢　語：／読／誦／　／熟／柿／　／知／恵／　／帝／王／
外来語：／Ｃｏｌｌｅｇｉｏ／　／Ｐａｓｔｏｒ／　／伽藍／　／修羅／
記　号：／．／　／，／　／：／　／？／
人　名：／Ｅｓｏｐｏ／　／清盛／の／三男／知盛／　／梶原／源太／　／喜一／
地　名：／Ｇｒｅçｉａ／　／胡／国／　／阿波／の／国／　／一の谷／

● 上記のように認定した最小単位を、短単位認定のために下表のとおりに分類する。

分類		例
一般		和 語 ： 獣 哀れ いとおしい 遥か 畏まる 流す 語る 為る …
		漢 語 ： 悪 行 威 勢 一 門 御 所 …
		外来語 ： Historia Latin 閼伽 弥陀 菩提 卒塔婆 …
付属要素		接頭的要素 ： 大(たい、だい) 御(お、ご、み、おん) 新(しん) 打ち(うち) 故(こ)…
		接尾的要素 ： 兼ねる 立て さ 位(い) 共(ども) 卿(きょう) 殿(どの) 寺(じ)…
その他	記号	， ． ； ： ！ ？ （ ） Ｄ Ｌ ア オ …
	数	０ １ ９ 一 五 十 百 千 万 余 数 …
	固有名	人 名 ： 平 源 熊谷 北条 義経 与一 俊寛 Xantho …
		地 名 ： Egypto Babilonia 日本 粟田口 宇治 鎌倉 屋島 関東 熊野 七条 …
	助詞･助動詞	の は や こそ んば さえ まで いで が て い かし た る･らる むず ごとし じゃ なんだ ろう …

（２）短単位

● 短単位の認定規定は、上表の分類ごとに適用すべき規定が定められる。その規定に基づき、最小単位を結合させる（又は結合させない）ことによって、短単位を認定する。以下、「一般」･「数」･「その他」に分けて、短単位認定規定の概要を示す。

※「｜」は短単位の分割位置を、「=」は短単位を切らないことを示す。

［１］一般

《和語・漢語》

最小単位２つの結合までを 1 短単位とする。

【例】｜草｜｜木｜　｜草=木｜　｜黒=糸｜　｜大=殿｜　｜遊び=女｜
　　｜知=恵｜　｜天=下｜無=双｜　｜無｜果=報｜　｜女=房｜衆｜　｜白｜拍=子｜

例外：切る位置が明確でないもの、あるいは切った場合と一まとめにした場合とで意味にずれがあるものは、3 最小単位以上の結合であっても 1 短単位とする。

【例】｜殊の外｜　｜夜もすがら｜

例外：最小単位が３つ以上並列した場合、それぞれの最小単位を 1 短単位とする。

【例】｜眼｜耳｜鼻｜舌｜

《外来語》

1 最小単位を 1 短単位とする。

【例】｜補陀落｜寺｜　｜夜叉｜御前｜　｜閻浮｜愛執｜

［２］数

「数」以外の最小単位と結合させない。「数」どうしの結合は、一・十・百・千の桁ごとに 1 短単位とする。「万」「億」等は、単独で 1 短単位とする。

【例】｜四｜位｜　｜六十｜六｜箇国｜　｜四五百｜人｜　｜一千｜一｜体｜
　　｜二｜万｜余｜騎｜

［３］その他

1 最小単位を 1 短単位とする。

付属要素　　｜御｜返事｜　｜由々し｜気｜　｜改め｜難い｜
助詞・助動詞｜失せ｜させ｜られ｜た｜と｜言う｜程｜こそ｜有れ｜
人名　　｜梶原｜平三｜景時｜　｜那須｜の｜与一｜　｜建礼門院｜　｜東方｜朔｜
地名　　｜加賀｜の｜国｜　｜備前｜備中｜備後｜　｜鬼界が島｜　｜宇治｜川｜

● 短単位データの作成は自動形態素解析と人手修正によって行われている。形態素解析処理は形態素解析器に「MeCab」、解析用辞書に「近世口語 UniDic」を使用している。

３．長単位の概要

長単位は、言語の構文的な機能に着目して規定した言語単位である。長単位の認定は、文節の認定を行った上で、各文節の内部を規定に従って自立語部分と付属語部分とに分割していくという手順で行う。

（１）文節

● 長単位の認定にあたっては、まず文節の認定を行う。現代語の文節は、一般に付属語又は付属語連続の後ろで切れる。このほかに、本コーパスでは、付属語を伴わない自立語であっても、主語・主題、連用修飾、連体修飾の各成分の後ろで切るといった規定を設けた。

● 文節を認定する上で問題となることの一つに、固有名、「―が～」「―つ～」「―の～」で 1 短単位と認める体言句、複合辞がある。これらについては、内部にある付属語の後ろでは切らないこととする。

● 複合辞は、本コーパス内での用例数、『室町時代編 I 狂言』および BCCWJ における類例の認定状況等を基準に一部認定し、付属語として認めた。

※「｜」は文節の分割位置を、「=」は文節を切らないことを表す。

【例】｜見る｜人｜川原に｜市を｜成い=て=御座る．｜
　｜渡辺=の=源五と｜言う｜者｜熊手を｜下ろいて｜御髪に｜掛け，｜取り上げ奉る：｜

（２）長単位

● 長単位は、上記の文節を規定に基づいて分割する（又は分割しない）ことによって認定する。文節を超えることはない。以下、長単位認定規定の概要を示す。

※「｜」は長単位の分割位置を、「=」は長単位を切らないことを示す。

［１］記号は 1 長単位とする。

【例】｜詳しゅう｜奏聞すれ｜ば｜，｜斜め｜なら｜ず｜喜ば｜せ｜られ｜た｜．｜

［２］付属語は 1 長単位とする。

【例】｜詳しゅう｜奏聞すれ｜ば｜，｜斜め｜なら｜ず｜喜ば｜せ｜られ｜た｜．｜

［３］主語・主題、連用修飾成分、連体修飾成分の後ろで切る。

【例】　｜貞能｜この｜事｜をば｜如何｜思う｜ぞ｜？｜

［４］体言に形式的な意味の「為る」「遊ばす」「致す」「仕る」「為される」「参る」「召す」「申す」が直接続く場合、体言と切り離さない。

【例】｜一味同心=する｜　｜伺候=仕る｜　｜参内=致す｜

［５］「御～有る・為る・候う・仕る・為される・召す・申す」「～遊ばす・有る・致す・おわします・候う・奉る・給う・仕る・為される・参らせる・参る・申す」のような形式の敬語表現は、全体を 1 長単位とする。

【例】｜御申し=有れ｜かし｜　｜叡覧=有る｜　｜恋しがり=奉る｜

上記形式中に付属語が含まれる場合、切り離さない。　【例】｜諫め=られ=奉っ｜た

［６］同格の関係にある体言連続は切り離さない。

【例】｜次男=宗盛｜　｜讃岐の守=正盛｜

［７］並列された語は切り離さない。

【例】｜烏帽子=直衣｜　｜宮殿=楼閣｜　｜金銀=七宝｜　｜縦様=横様=蜘蛛手=十文字｜

［８］係り受けを重視し、付属語を切り出すのは不適切なものを連語として認める。

【例】｜然ら=ば｜　｜然り=ながら｜（※接続詞）

● 長単位データの作成は、人手修正済み短単位データを基に、長単位解析器 Comainu によって長単位の自動構成を行っている。

４．他のコーパスと異なる処理・特殊な処理

キリシタン資料は、狂言と同様に古代語から近代語への過渡期的な様相を示す中世語の資料として、既存の現代語・平安時代編における処理では対処できないケースが散見される。そのため可能な限り現代語・平安時代編の規定を尊重しつつ、それでは対処できない箇所については独自の処理を行った。以下にはそのうち、特に注意を要するものを挙げる。なお挙例は短単位による。

［１］ 文語活用と口語活用

現代語のコーパスおよび『日本語歴史コーパス』では、活用語について「文語」「口語（明示なし）」の二大別を行っている。ところがキリシタン資料は、前述のとおり古代語から現代語への過渡的様相を示す資料であり、文語・口語いずれかに統一することは困難である。そのため、品詞・語・出現形により方針を立てて処理を行った。

● 動詞は原則「文語」活用と見、口語活用でなければ対応できないものを「口語」とした。これは「文語」に分類される上・下二段活用動詞の一段化が進んでいないことによる。

【例】《文語》｜既に｜討手｜を｜遣わす｜．｜
→動詞-一般・文語四段-サ行・終止形-一般
｜白旗｜を｜ざっ｜と｜差し上ぐれ｜ば｜
→動詞-一般・文語下二段-ガ行・已然形-一般
《口語》｜日数｜を｜経る｜程｜に｜　（原本ローマ字「feru」）
→動詞-一般・下一段-ハ行・連体形-一般（文語下二段では処理不可）

● 形容詞型活用は原則「口語」活用と見、文語活用でなければ対応できないものを「文語」とした。ただし「―けれ」の形は文語・已然形とする。詳細は渡辺他(2015)参照。

【例】《口語》｜強い｜馬｜をば｜上手｜に｜立てよ｜
→形容詞-一般・形容詞・連体形-一般
《文語》｜前世｜の｜縁｜も｜浅から｜ず｜に｜思わ｜れ｜たれ｜ば｜
→形容詞-一般・文語形容詞-ク・未然形-補助
｜唯｜一騎｜残っ｜て｜戦｜を｜する｜こそ｜心憎けれ｜：｜
→形容詞-一般・文語形容詞-ク・已然形-一般

［２］終止形・連体形の別

● 文語サ行変格活用等には、連体形に相当する形態で文末終止を行う場合がある。このような場合は、文末であっても終止形ではなく連体形とした。

【例】｜西国｜の｜方｜へ｜落ち行か｜う｜と｜存ずる｜：｜
→動詞-一般・文語サ行変格・連体形-一般

｜平山｜を｜打た｜す｜まい｜とて｜続い｜て｜駆くる｜．｜
→動詞-一般・文語下二段-カ行・連体形-一般

● 終助詞や助動詞に前接する場合、終止・連体形の区別が困難なケースが多い。そこで、形態的に明らかなものはその活用形とし、終止・連体同形の物は、極力平安時代編や小椋他(2011)に合わせ、終止形・連体形いずれかに統一した。

［３］助動詞「う」と意志推量形

本コーパスでは意志推量を表す形式について、基本的には『室町時代編Ⅰ狂言』での処理を踏襲し、助動詞「う」および助動詞「むず」の語形「うず」を用いて、未然形＋「う」・「うず」という形で処理している。ただし本コーパスでは、ローマ字表記の原本に即して作成した漢字仮名交じりテキストでの表記に基づいて処理を行っているため、『室町時代編Ⅰ狂言』での処理と異なる箇所がある。以下に例を示す。原文のローマ字から漢字仮名交じりテキストへの変換については別稿「『日本語歴史コーパス　室町時代編Ⅱキリシタン資料』テキストの凡例と『中納言』表示項目について」を参照されたい。

● 上一段・上二段活用以外の語
【例】｜今度｜の｜戦｜に｜命｜生き｜て｜再び｜都｜へ｜参ら｜う｜
→動詞-非自立可能・文語四段-ラ行・未然形-一般＋助動詞・無変化型・終止形-一般
｜一所｜で｜死な｜う｜事｜も｜悪しから｜うず｜
→形容詞-一般・文語形容詞-シク・未然形-補助＋助動詞・文語助動詞-ムズ・終止形-一般
【例】｜その｜成ら｜れ｜うずる｜様｜を｜見届け｜う｜とて｜，｜
→助動詞・文語下二段-ラ行・未然形-一般＋助動詞・文語助動詞-ムズ・連体形-一般
動詞-一般・文語下二段-カ行・未然形-一般＋助動詞・無変化型・終止形-一般
● 上一段・上二段活用の語
【例】｜生死｜の｜安否｜を｜試みょう｜と｜
→動詞-一般・上一段-マ行・意志推量形
｜御｜運｜の｜尽きょ｜うずる｜事｜も｜難い｜事｜で｜は｜無い｜．｜
→動詞-一般・文語上二段-カ行・未然形-一般＋助動詞・文語助動詞-ムズ・連体形-一般

上一段・上二段活用の語の場合、語形を明示するために漢字仮名交じりテキストを拗音表記としている。「試みょう」の「試みょ」のように、「う」を分割すると前接の用言を文語活用未然形と認定することが困難な場合は、『室町時代編Ⅰ狂言』での処理と同様に全体で一語とし、口語活用の意志推量形とした。ただし「うず」が後続する場合は全体で一語と認定することが難しいため、「尽きょ」のような形を文語活用の未然形として登録して処理している。

［４］「御～やる」における助動詞「やる」、「御座ある」の処理

● 「御」＋動詞連用形に後続する「やる」は、原則助動詞「やる」とする。

【例】｜何故｜に｜そなた｜は｜力｜を｜御｜添え｜やら｜ぬ｜ぞ｜

● 「おりゃる」（「お入りある」の転）や、動詞連用形と「やる」が融合した形については、動詞と助動詞とを分割するのが困難なため全体を一短単位と認定する。

【例】｜急ぎ｜我が｜方｜へ｜おりゃれ｜

→動詞-一般・文語四段-ラ行・命令形

｜この｜一｜間｜を｜我｜に｜御｜貸しゃれ｜

→動詞-一般・文語四段-ラ行・命令形

● 「オリャル」を 1 短単位としたため、意味・機能が対応する「オリナイ」「オリナシ」についても１短単位と認め、形容詞「おりない」とした。

● 存在動詞「ゴザアル」は１短単位と見、語彙素「御座る」の語形とした。また意味・機能が対応する「ゴザナイ」「ゴザナシ」も 1 短単位と認め、形容詞「御座無い」とした。

［５］入声音

原本のローマ字表記によって明らかになるｔ入声音については、語形を促音にすることで入声音語形であることを示した。

【例】｜去ん｜ぬる｜治承｜三｜年｜五｜月｜の｜頃｜，｜

→原本表記「guat」　語彙素「月（ガツ）」の語形「ガッ」

｜経｜を｜書い｜て｜，｜結願｜に｜は｜大きな｜卒塔婆｜を｜立て｜て｜，｜

→原本表記「qetguan」　語彙素「結願（ケチガン）」の語形「ケッガン」

［６］原本の書き誤りと見られる箇所

原本のローマ字表記の中には明らかに書き誤りと見られる箇所がある。これらについて、本コーパスでは人名・地名などの固有名詞の場合とそれ以外のものに分け、それぞれ以下のように対処している。

● 固有名詞

原本ローマ字どおりの表記では指示対象が不明または間違った指示対象を指してしまう場合には、漢字仮名交じりテキストの該当箇所をカタカナ表記としている。語彙素は原本表記の語形をそのまま認定し、品詞は文脈から判断できる最も適切なものを用いた。

【例】｜ムネモラ｜，｜扨｜は｜力｜に｜及ば｜ぬ｜と｜言う｜て｜

（原本表記）Munemora,ſateua□chicari□ni□voyobanu□to□yŭte,

→語彙素「ムネモラ」　名詞-固有名詞-人名-名

｜シゲモリ｜如何に｜如何に｜と｜呆れ｜らるれ｜ば｜，　｜
（原本表記）Xiguemori□icani□icani□to□aqirerarureba,
※文脈上「清盛」とあるべきところ
→語彙素「シゲモリ」　名詞-固有名詞-人名-名

● 固有名詞以外

固有名詞以外で明らかに原本の書き誤りであると判断できる箇所は、漢字仮名交じりテキストを正しいと思われる形に修正した。形態論情報も漢字仮名交じりテキストに合わせて付与している。

【例】｜ムネモラ｜，　｜扨｜は｜力｜に｜及ば｜ぬ｜と｜言う｜て｜
（原本表記）Munemora,ſateua□chicari□ni□voyobanu□to□yŭte,
→語彙素「力」語形「チカラ」　名詞-普通名詞-一般
｜清盛｜公｜こそ｜過分｜の｜事｜をば｜仰せ｜らるれ｜，　｜
（原本表記）Qiyomori□cô□coſo□quabun□no□coco□uoba□vôxerarure,
→語彙素「事」語形「コト」　名詞-普通名詞-一般

参考文献

池上尚（2016）「『日本語歴史コーパス　平安時代編』形態論情報規定集」http://pj.ninjal.ac.jp/corpus_center/chj/doc/morph-heian-2016.pdf(2018 年 3 月 22 日閲覧)
市村太郎・渡辺由貴（2015・2016）「『日本語歴史コーパス　室町時代編 I 狂言』形態論情報の概要」http://pj.ninjal.ac.jp/corpus_center/chj/morph-kyogen-2016.pdf（2018 年 3 月 22 日閲覧）
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』形態論情報規定集第 4 版（下）」特定領域研究「日本語コーパス」平成 22 年度研究成果報告書
小椋秀樹・須永哲矢（2012）「中古和文 UniDic 短単位規程集」基盤研究(C)「和文系資料を対象とした形態素解析辞書の開発」研究成果報告書 2
渡辺由貴・市村太郎・鴻野知暁（2015）「『虎明本狂言集』のコーパスデータにおける短単位認定の諸問題」第 7 回コーパス日本語学ワークショップ予稿集 pp.233-240