『日本語歴史コーパス 奈良時代編Ⅱ宣命』形態論情報の概要

2020 年 3 月 31 日　呉寧真

１．言語単位

『日本語歴史コーパス 奈良時代編Ⅱ宣命』（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『日本語歴史コーパス　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　奈良時代編Ｉ万葉集』（以下「『奈良時代編Ｉ万葉集』」と呼ぶ）と同様の規程に基づいているが、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定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』『奈良時代編Ｉ万葉集』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：／掛け／まく／も／畏き／天皇（スメラ）／が／御／世／御／世／仕へ／奉り／

漢　語：／三／宝／　／謀／反／　／禅／師／　／辺／戍／
外来語：／盧舎那／　／袈裟／　／舎利／　／王（ユニキシ）／
記　号：／、／　／。／
人　名：／日並所知／皇太子／　／阿倍／朝臣／東人／　／葛城／曽豆比古／　／押勝／
地　名：／岡宮／　／唐国（モロコシ）／　／大八嶋／国／　／高天原／

● 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　：　現御神　朕　さだか　いそしい　集る … 漢　語　：　最　勝　王　経 … 外来語　：　菩薩　袈裟　菩提 …
付属要素		接頭的要素　：　相（あい）　御（み）　大（おお）　従（ひろき）… 接尾的要素　：　らま　じもの　重しみ　等（たち、ら）…
その他	記　号	、　。
	数	一　三　五　百　千 …
	固有名	人　名　：　藤原　麻呂　真備　大鷦鷯 … 地　名　：　葦北　飛鳥　淡路　因幡　須伎　春日…
	助詞・助動詞	の　を　そ　ゆ　ゆり　い　ず　ごとし　なり　たり　む　ましじ…

（２）短単位

● 短単位の認定規定、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】｜天｜つ｜日=嗣｜　｜御=世｜　｜中=今｜　｜辞=立つ｜
　　｜内=相｜　｜護=法｜梵=王｜　｜乾=政｜官｜　｜最=勝｜王｜経｜　｜大｜法=師｜

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】｜柵戸（きのへ）｜　｜公民（おほみたから）｜　｜詔旨（おほみこと）｜

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】｜鈴｜印｜契｜

《外来語》
1 最小単位を 1 短単位とする。
【例】｜観世音｜菩薩｜　｜盧舎那｜如来｜　｜御｜袈裟｜

［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。なお、「万」「億」等は、従来単独で 1 短単位とするが、本コーパスには例がない。
【例】｜八｜年｜　｜十｜五｜日｜　｜廿｜三｜日｜　｜五千｜戸｜

［3］その他
1 最小単位を 1 短単位とする。

付属要素	｜大｜御｜舎利｜　｜皇子｜等｜　｜労し｜み｜重し｜み｜
助詞・助動詞	｜朝庭｜に｜侍へ｜奉ら｜む｜を｜ば｜必ず｜治め｜賜は｜む｜
人名	｜藤原｜朝臣｜麻呂｜　｜県｜犬養｜橘｜夫人｜　｜百済王｜敬福｜
地名	｜武蔵｜国｜　｜近江｜大津｜宮｜　｜高天原｜　｜佐保｜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。形態素解析処理においては、形態素解析器に「MeCab」、解析用辞書に「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、CHJ の他の時代編のサブコーパスおよび 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】｜天の｜下の｜公民｜諸｜聞き=食へと｜詔り=たまふ。｜
｜現御神と｜大八嶋=国｜所知す｜天皇が｜大命=らまと｜詔り=たまふ｜大命を、｜

（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】｜同じ｜事｜ぞ｜と｜勅りたまひ｜て｜、｜治め賜ひ｜慈び賜ひ｜けり｜。｜

［2］付属語は 1 長単位とする。

【例】｜大八嶋国｜知ら｜さ｜む｜次｜と｜、｜天｜つ｜神｜の｜御子｜ながら｜も｜

［3］主語・主題、連用修飾成分、連体修飾成分の後ろで切る。

【例】貴き｜高き｜広き｜厚き｜大命｜を｜受け賜はり｜

［4］体言に形式的な意味の「為る」「奉る」が直接続く場合、体言と切り離さない。

【例】｜順=する｜　｜助け=奉る｜

［5］連用形で続く動詞連接は、全体を 1 長単位とする。

【例】｜護り助けまつれ｜　｜進退ひ匍匐ひ廻ほり白し賜ひ受け賜らく｜

ただし、上記形式中に「奉る」「たまふ」が複数含まれる場合、切り離す。

【例】｜上げ奉り｜治め奉る｜　｜教へ賜ひ｜おもぶけ賜ひ｜答へ賜ひ｜宣り賜ふ｜随｜

［6］同格の関係にある体言連続は切り離さない。

【例】｜祖父=太政大臣｜　｜みまし=大臣｜

［7］並列された語は切り離さない。

【例】｜君臣=祖子｜　｜由紀=須伎｜　｜鈴=印=契｜　｜上=中=下｜

［8］表記上切り離せない語は 1 長単位とする。

【例】｜卿｜（※｜まえ｜つ｜きみ｜）
　　｜大赦し＝たまふ｜（※｜ひろく｜つみ｜ゆるしたまふ｜）

従って、表記を重視し、表記の違いによって、同じ意味を表す同じ語の組み合わせでも、1 長単位でとらえる場合ととらえない場合がある。

【例】｜大八洲｜所知す｜（おおやしま｜しらしめす）
　　｜御大八洲す｜（おおやしま=しらしめす）

４．他のコーパスと異なる処理・特殊な処理

宣命は和文資料でありながら、漢文・漢語に由来する特殊な表記の語が多い。既存のサブコーパスにおける処理では対処できない箇所について、独自の処理を行った。以下にはそのうち、特に注意を要したものを挙げる。なお挙例は短単位による。

［１］ 訓読と音読を付与

本コーパスは、地名と数詞、漢文由来の難読の熟語表記に訓読と音読を付与した。これは、掛詞や洒落などに対応するために本文の同一箇所に複数の読み・形態論情報を付与できるようにした機能によるもので、CHJ の『和歌集編』や『江戸時代編』で用いられた。なお、従来の宣命研究においては、訓読を優先させる方針があるため、本コーパスもそれに従って訓読を主本文にした。

● 副本文を付与する対象には、「二字以上の漢字熟語と思われるもの」「地名」「数詞・助数詞」の 3 種類が該当する。

【例】《漢字熟語》

「嫡子」
　訓読（主本文）→｜むかいめ｜ばら｜の｜みこ｜
　音読（副本文）→｜ちゃくし｜

「開闢け」
　訓読（主本文）→｜あめつち｜ひらけ｜
　音読（副本文）→｜かいびゃく｜

動詞の場合、漢語の部分だけに副本文を付与する。活用語尾は切り落とす。

【例】《地名》

「豊前」｜国
　訓読（主本文）→｜とよくに｜の｜みち｜の｜くち｜の｜
　音読（副本文）→｜ぶぜん｜

「越前」
　訓読（主本文）→｜こし｜の｜みち｜の｜くち｜
　音読（副本文）→｜えちぜん｜

【例】《数詞・助数詞》

「五千｜戸」

訓読（主本文）→｜い｜ち｜へ｜

音読（副本文）→｜ごせん｜こ｜

「十｜五｜日」

訓読（主本文）→｜とを｜か｜あまり｜いつ｜か｜の｜ひ｜

音読（副本文）→｜じゅう｜ご｜にち｜

● 地名、数詞・助数詞の場合、副本文の読みは『平安時代編』の規程集に従う。

漢字熟語の場合、副本文を付与するにあたって、以下の手順をとる。

【1】『日本国語大辞典　第二版』（以下『日本国語大辞典』）での項目確認

『日本国語大辞典』に該当する項目があれば、たとえ初出が上代ではなくても付与した。文字列として出現したことを重視し、検索できるように、時代を考慮せず読みを与えた。

【2】『日本国語大辞典』での意味確認

同じ書字形に二つ以上の語形がある場合、まず意味の異同があるかどうかを確認する。例えば、表記「行事」には、ギョウジとコウジの 2 語形があり、

ギョウジは

「恒例として事を執り行なうこと。また、その事柄。催し事。「年中行事」」

コウジは

「事を行なうこと。また、その人やその物事。」

と記載されている。

作中の当該用例の文脈では、「催し事」の意味ではなく、「物事」の意味のため、コウジの読みを与えた。『日本国語大辞典』によると、コウジの初出は『史記抄』（1477 年）と遅いが、本コーパスでは時代ついては考慮しない。

【3】その他

●【2】のように、同じ表記に複数の読みが想定され、かつそれらに意味の差が認められない場合、初出の順序を重視し、初出が早い語形を採用した。それでもなお決めがたい場合、『日本国語大辞典』の引用例の中に、続日本紀や漢籍を引用している語形を採用した。

●基本的に副本文は二字以上の熟語に付与したが、主本文が表記上分割できずに、2 短単位以上を付与した部分に限って、一字の語にも副本文を付与した。

【例】「像」

訓読（主本文）→｜み｜かた｜

音読（副本文）→｜ぞう｜

［２］補読の「の」

『平安時代編』の規程により、地名と人名に後接する補読の「の」は、語形で補った。

【例】「石川」朝臣　→　語彙素：｜イシカワ｜　語形：｜イシカワノ｜

「難波」大宮　→　語彙素：｜ナニワ｜　語形：｜ナニワノ｜

従来の CHJ のサブコーパスでは、地名と人名以外の「の」は補わなかったが、本コーパスでは、北川和秀（1982）『続日本紀宣命―校本・総索引』の底本の補読に従って、表記を保持した上で、語彙素の「の」を補った。

【例】「慶雲」五年　→　｜けいうん｜の｜

「陰陽」寮　→　｜おんよう｜の｜

「建内｜宿祢｜命」　→　｜たけのうち｜の｜すくね｜の｜みこと｜

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

池田幸恵、須永哲矢（2013）「「五国史」宣命のコーパス化」『第 4 回コーパス日本語学ワークショップ予稿集』

池田幸恵・須永哲矢（2015）「「五国史」宣命コーパスの設計とその利用」『訓点語と訓点資料』134

国立国語研究所（2017）『日本語歴史コーパス　奈良時代編 I 万葉集』
https://pj.ninjal.ac.jp/corpus_center/chj/nara.html#manyo

小木曽智信（2016a）「『日本語歴史コーパス』の現状と展望」『国語と国文学』93-5

小木曽智信（2016b）「多重の読みをもつテキストのコーパス化」『言語資源活用ワークショップ 2016 発表論文集』p159-162

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

村山実和子・小木曽智信・中村壮範（2017）「形態論情報の多重化による洒落本コーパスの質的拡張」 情報処理学会研究報告 Vol.2017 CH 114, No.8

松崎安子・小木曽智信・中村壮範（2019）「『日本語歴史コーパス 和歌集編 』 Ver.1.0 の公開」日本語学会 2019 年度秋季大会予稿集