「日本語歴史コーパス　平安時代編」　形態論情報の概要

2014 年 3 月 31 日

１．２種類の言語単位

（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**

「日本語歴史コーパス　平安時代編」で採用したこの 2 種類の言語単位は、「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中古和文用に設計したものである。基となっている BCCWJ の言語単位は「日本語話し言葉コーパス（CSJ）」との互換性の保持を図り、国立国語研究所が行った語彙調査の単位を基に設計された。

「日本語歴史コーパス　平安時代編」の言語単位は通時的な日本語研究で利用するために、現代語のコーパスとの互換性の保持を図っている。これまでに国立国語研究所が実施してきた語彙調査における言語単位のうち、短い単位の系列に属するものが「短単位」、長い単位の系列に属するものが「長単位」である。なお、長単位・短単位認定規程は、BCCWJ の規程をそのまま用いるのではなく、中古和文用に修正・拡張を行っている。

短単位・長単位とも、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た漢字等の表記に相当するものである。

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。中古和文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、使用実態に基づき個別の判断をすることがある。語種等により、次のように認定する。

  ／　は最小単位の分割位置を表す。

  和語：／雲／の／あなた／は／春／に／や／ある／らむ／
  漢語：／関／白／　／加／持／　／大／納／言／
  外来語：／迦陵頻伽／　／菩薩／　／瑠璃／色／（「瑠璃」のみ外来語）
  記号：／。／　／・／

人名：／紀／貫之／　／王／昭君／
地名：／大和／の／国／十市／の／郡／

- 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。

表：最小単位の分類

分　類		例
一　般		和　語　：　春　花　あはれ　言ふ　言葉　…
		漢　語　：　関　白　加　持　…
		外来語　：　阿闍梨　菩薩　瑠璃　…
付属要素		接頭的要素　：　相　御（おおん、ご、み）　打ち　なま　…
		接尾的要素　：　君（ぎみ）　難し　気（げ）　様（さま）　…
その他	記　号	、　・　。　「　」　…
	数	一　二　十　百　千　…　幾　数　何
	固有名	人　名　：　源　貫之　伊勢　あこぎ　…
		地　名　：　大和　土佐　入間　住吉　吉野　逢坂　…
	助詞・助動詞	の　を　ぞ　こそ　し　る・らる　ず　まじ　まほし　なり　…

（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ている。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》

最小単位二つの結合までを１短単位とする。

【例】｜母｜　｜宮｜　｜母=宮｜　｜あいだち=なし｜　｜心=のどか｜
　　｜法=師｜　｜右｜大=将｜

例外：複合動詞は原則として分割する。

【例】｜聞き｜渡る｜　　｜出で｜来｜

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】｜大殿籠る｜　　｜返り申し｜

例外：最小単位が三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】｜仏｜法｜僧｜

《外来語》

１最小単位を１短単位とする。

【例】｜菩提｜樹｜　｜瑠璃｜色｜

［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】｜二十｜四｜日｜　　｜十｜万｜億｜　　｜二三十｜束｜

［３］その他

１最小単位を１短単位とする。

付属要素	｜相｜乗る｜　｜者｜ども｜　｜得｜がたし｜
助詞・助動詞	｜雲｜の｜あなた｜は｜春｜に｜や｜ある｜らむ｜　｜そ｜の｜
記号	｜、｜　｜。｜　｜「｜
人名	｜紀｜貫之｜　｜王｜昭君｜　｜院源｜僧都｜
地名	｜大和｜の｜国｜十市｜の｜郡｜　｜あさか｜山｜

- 短単位データの作成は自動形態素解析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「中古和文 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、中古和文では、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。複合辞は付属語として認めない。
- 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、副助詞が挿入された複合動詞がある。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。

｜　は文節の分割位置を、　＝　は文節を切らないことを表す。

｜小野=小町｜　｜雁=が=音｜　｜わた=つ=うみ｜　｜天=の=川｜　｜尚侍（ないし=の=かみ）｜
｜一宮（いち=の=みや）｜　｜うち=も=臥されず｜　｜鳴き=こそ=渡れ｜

（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、‖　は注目している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】｜春｜は｜あけぼの｜。｜
｜「｜雀｜の｜子｜を｜犬君｜が｜逃がし｜つる｜、｜伏籠｜の｜中｜に｜籠め｜たり｜つる｜もの｜を｜」｜とて｜、｜

［２］付属語は１長単位とする。

【例】｜「｜雀｜の｜子｜を｜犬君｜が｜逃がし｜つる｜、｜伏籠｜の｜中｜に｜籠め｜たり｜つる｜もの｜を｜」｜とて｜、｜

［３］主語・主題、連用修飾成分、連体修飾成分の後ろで切る。

【例】｜いと｜はしたなき｜こと‖多かれど｜
｜長雨｜晴れ間‖なき｜ころ｜

［４］体言に形式的な意味の「す」「きこゆ」「はべり」「まゐる」「つかうまつる」が直接続く場合、切り離さない。

【例】｜御曹司=し｜て｜　　｜返りごと=したまへ｜

［５］「御（おほん・お・み・ご）～す・きこゆ」「～おはす・おはします・きこゆ・さぶらふ・たてまつる・たまふ・つかうまつる・はべり・もうす」という形式の敬語表現は、全体を１長単位とする。

【例】｜御曹司=し｜て｜　　｜思う=たまへ=忍び｜つれ｜ど｜

上記形式中に付属語が含まれる場合、切り離さない。

【例】｜御覧ぜ=させ=たまふ｜

［６］同格の関係にある体言連続は切り離さない。

【例】｜母=北の方｜　｜出雲権守=時方朝臣｜

［７］並列された語は切り離さない。

【例】｜女神=男神｜　｜ありさま=心ばへ｜

［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】｜知ら=ず=読み｜　｜我=は=顔｜

- 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．品詞付与方針

短単位と長単位の品詞体系は共通であるが、品詞付与方針が異なる。短単位では可能性を考慮した品詞を付与しており、「名詞-普通名詞-形状詞可能」等がある。これに対して長単位では文脈に即して品詞を付与する方針をとり、名詞-普通名詞-〇〇可能といった品詞は設けない。例えば、「哀れ」は短単位では「名詞-普通名詞-形状詞可能」であるが、長単位では文脈に則し「もののあはれ知りすぐし、」の場合は名詞を、「皇子もいとあはれなる句を作りたまへるを」の場合は形状詞を付与する。

参考文献

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
http://dl.dropboxusercontent.com/u/73297026/report/unidic-EMJ_rulebook2012.pdf
冨士池優美（2012）「中古和文における長単位の概要」第 2 回コーパス日本語学ワークショップ予稿集

参考 URL

中古和文 UniDic　http://www2.ninjal.ac.jp/lrc/index.php?UniDic
中・長単位解析器 Comainu　https://maro.ninjal.ac.jp/Comainu/
http://comainu.org/