中古和文における長単位の概要

冨士池　優美（国立国語研究所　コーパス開発センター）†

Outline of Long-unit-word in the Early Middle Japanese

Yumi Fujiike　(Center for Corpus Development, NINJAL)

１．はじめに

国立国語研究所では「通時コーパスの設計」プロジェクトを中心に、歴史的な日本語のコーパス構築の準備が進められている。「通時コーパス」に格納される資料の一つに、源氏物語を中心とした中古和文がある。通時コーパスの形態論情報については、言語単位として、コーパスからの用例収集に適した「短単位」と、格納したサンプルの言語的特徴の解明に適した「長単位」の 2 種類を採用する。この 2 種類の言語単位について代表形・品詞等の情報を与える。

中古和文については、長短 2 種類の言語単位のうち短単位は中古和文 UniDic[1]で形態素解析後、人手修正をする形でデータ整備が進められており、認定基準については既に小椋・須永（2012）にまとめられている。これに対して長単位の認定基準については現在検討中の段階である。本稿では長単位の認定基準について、その概要と課題について述べる。また、長単位解析の現状について、あわせて報告する。

２．長単位の概要

「通時コーパス」中古和文の長単位は、『現代日本語書き言葉均衡コーパス』（Balanced Corpus of Contemporary Written Japanese、以下 BCCWJ）で採用した単位を中古和文用に修正・拡張する方向で検討を進めている。この BCCWJ の長単位は、『日本語話し言葉コーパス』（Corpus of Spontaneous Japanese、以下 CSJ）で採用した単位を書き言葉用に修正・拡張したものであり、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、長い単位の系列[2]に属するものである。

長単位は文節を基にした単位である。長単位の認定は、文節の認定を行った上で、各文節の内部を規則に従って自立語部分と付属語部分に分割していくという手順で行う。そのため、長単位の認定基準は、文節と長単位、二つの認定基準から成る。

本節では、文節の認定基準、長単位の認定基準、長単位の認定基準に関する BCCWJ からの変更点、長単位の長所及び認定基準に関する検討課題について述べる。以下、例文中の

† yfujiike@ninjal.ac.jp

1 中古和文 UniDic については小木曽ほか（2012）を参照。

2 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調査単位の概要については、小椋ほか（2011）pp.1-4（第 1 章『現代日本語書き言葉均衡コーパス』の言語単位）を参照。

文節の境界を「┃」、長単位の境界を「｜」とし、注目している境界を「‖」、切らないことを示す場合には「-」を、中でも注目している部分には「=」を用いる。また、注目している単位には下線を付す場合がある。

２．１　文節の認定基準

長単位の認定にあたっては、まず文節の認定を行う。現代語では文節は一般に付属語又は付属語連続の後で切れるが、中古和文においては、文節の切れ目に当たる付属語がないことが多いため、付属語を伴わない自立語については特に構文的な機能に着目して文節を認定することになる。文節を認定する上で問題となることの一つに、固有名、付属語を含む敬語表現、「―が～」「―つ～」「―の～」で１短単位と認める体言句、表記上分割不可能なものがある。これらについては、内部にある助詞・助動詞の後では切らないこととする。

┃源=頼朝┃　┃小野=小町┃　┃壇=ノ=浦┃　┃造ら=せ=たまふ┃
┃雁=が=音┃　┃わた=つ=うみ┃　┃天=の=川┃　┃尋常（よ=の=つね）┃

２．２　長単位の認定基準

長単位は、規定に基づいて文節を分割する、あるいはしないことによって得られた要素を１単位とする形式であり、文節を超えることはない。文節・長単位・短単位の関係を次ページの表１に示す。

以下、長単位認定基準の概要を示す。

〔１〕区切り符号は１長単位とする。

｜時めきたまふ｜あり｜けり｜。｜

〔２〕付属語は１長単位とする。

｜あら｜ぬ｜が｜、｜

ただし、「～おはす・おはします・きこゆ・たてまつる・たまふ・はべり」という形式の敬語表現の中に現れる付属語の後ろでは切らない。

｜弾か=せ=たまふ｜

〔３〕助詞・助動詞を伴わない自立語は、主語・主題、連用修飾、連体修飾の各成分の後ろで切る。

｜いと‖はしたなき‖こと‖多かれど｜

〔４〕体言に形式的な意味の「す」が直接続く場合、体言と「す」とを切り離さない。

｜心づかひ=し｜て｜　｜消息=せよ｜

〔５〕並列の関係にある語は切り離す。

｜四位‖五位｜こきまぜ｜に｜　｜あまたたび｜傾き‖あやしぶ｜

〔６〕同格の関係にある体言連続は切り離さない。

｜母=北の方｜なむ｜いにしへ｜の｜人｜の｜

〔７〕数を表す要素を含む自立語は、以下のように長単位を認定する。

(1)数を表す要素は、単位の変わり目の後ろで切る。

｜三尺‖六寸｜ばかり｜

(2)数を表す要素の前で切る。

｜長さ‖二十丈｜、｜広さ‖五丈｜

(3)数を表す要素とそれに続く体言・接辞とは切り離さない。

｜ひとつ=后腹｜

表１　文節・長単位・短単位の関係

文節	長単位	長単位 語彙素	長単位 語彙素読み	長単位品詞	短単位
B				空白	
	いづれ	何れ	イズレ	代名詞	いづれ
	の	の	ノ	助詞-格助詞	の
B	御時	御時	オオントキ	名詞-普通名詞-一般	御
					時
	に	なり	ナリ	助動詞	に
	か	か	カ	助詞-係助詞	か
	、	、		補助記号-読点	、
B	女御	女御	ニョウゴ	名詞-普通名詞-一般	女御
	、	、		補助記号-読点	、
B	更衣	更衣	コウイ	名詞-普通名詞-一般	更衣
B	あまた	数多	アマタ	副詞	あまた
B	さぶらひたまひ	侍ひ給う	サブライタマウ	動詞-一般	さぶらひ
					たまひ
	ける	けり	ケリ	助動詞	ける
B	中	中	ナカ	名詞-普通名詞-一般	中
	に	に	ニ	助詞-格助詞	に
	、	、		補助記号-読点	、
B	いと	いと	イト	副詞	いと
B	やむごとなき	やんごとなし	ヤンゴトナシ	形容詞-一般	やむごとなき
B	際	際	キワ	名詞-普通名詞-一般	際
	に	なり	ナリ	助動詞	に
	は	は	ハ	助詞-係助詞	は
B	あら	有り	アリ	動詞-一般	あら
	ぬ	ず	ズ	助動詞	ぬ
	が	が	ガ	助詞-格助詞	が
	、	、		補助記号-読点	、
B	すぐれ	優る	スグル	動詞-一般	すぐれ
	て	て	テ	助詞-接続助詞	て
B	時めきたまふ	時めき給う	トキメキタマウ	動詞-一般	時めき
					たまふ
B	あり	有り	アリ	動詞-一般	あり
	けり	けり	ケリ	助動詞	けり
	。	。		補助記号-句点	。

２．３　BCCWJ からの変更点

（１）文節認定基準における主語・主題に関する変更

助詞・助動詞を伴わない自立語に関しては、文節を「主語・主題の後ろで切る」という規定が CSJ にあった。これに対して、BCCWJ は書き言葉を対象としておりこの規定が適用されることが少なく、「持続可能な」「センス抜群の」のような漢語形状詞を述部に持つものの文節認定の判断が難しいこともあり、規定を削除していた。しかし、中古和文においては、助詞を伴わずに主語・主題を示すことが多いため、この規定が再び必要となった。

（2）並列に関する変更

CSJ では並列を切り離していたのに対して、BCCWJ では、漢語の並列を中心に、並列と見るか 1 語と見るかの判断が困難な例が頻出したため、切り離さないこととしていた。

BCCWJ：｜公正=妥当｜な｜実務慣行｜

これに対して、中古和文では漢語の並列のような問題はあまり起こらないため、並列は切り離すこととした。この規定により、複合動詞と動詞の並列が区別され、語と語との係り受けが明確になる。

中古和文：｜あまたたび｜傾き‖あやしぶ｜

２．４　長単位の長所

ここでは長単位と短単位の違いに着目して、長単位の長所を示す。短単位は、基準がわかりやすくゆれが少ないため用例収集には便利であるが、合成語を構成要素に分割してしまう場合が少なくない。短単位と比較したとき、長単位ではこのような場合にも一般的な古語辞典の見出し語に近い形式が取り出されることが大きな違いと言える。そのほか、品詞について、以下のような違いがある。

（1）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[3]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。長単位末尾に位置する短単位の品詞が名詞-普通名詞-｛副詞・形状詞・サ変形状詞可能｝の場合、その用法に基づき名詞・副詞・形状詞に判別する。「哀れ」（名詞-普通名詞-形状詞可能）を例とすると、「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。

（2）合成語の扱い

短単位では最小単位の 1 回結合を最大とするのに対して、長単位では結合回数の制限なく、合成語を認める。このとき「愛敬づく」のように、合成の結果として品詞が変わることがある。また、短単位では接辞を切り離していたのに対して、長単位では接辞を含めた形式が 1 長単位となる。接辞は品詞に影響を与える。中古和文では、漢語接頭辞がほぼ用

[3] 「形状詞可能」は、名詞としても形状詞としても使われ得ることを示す。

いられないことから接頭辞による品詞の変化は見られないが、接尾辞が付加することによって品詞が変わることが多い。

｜物語-めく｜	名詞＋接尾辞	→ 動詞
｜たへ-がたし｜	動詞＋接尾辞	→ 形容詞
｜うつくし-げ｜	形容詞＋接尾辞	→ 形状詞

上に挙げたような違いにより、より精密化された品詞情報を活用することができることが長所と言えるだろう。

２．５　認定基準に関する検討課題

（１）並列の認定

先に述べたように、中古和文では並列の扱いを BCCWJ から変更して、原則として分割することとした。ここで、問題になるのが、代名詞や副詞の並列である。例えば「こなたかなたの人々など」の「こなた」「かなた」は代名詞の並列であるが、係り受けを意識した場合には「こなたかなた」全体で「人々」にかかると見るのが妥当だろう。副詞に関しては、「と」「かく」のバリエーションが問題になる。例えば「と見かう見、見けれど、」の場合は「と」「かう」をそれぞれ副詞とすることに問題はないが、「とざまかうざまにこころみきこゆるほど」となると、「とざま」「かうざま」を並列と見てそれぞれ文節認定するのか、「とざまかうざま」全体で「こころみきこゆる」にかかると見るのか、判断が難しいところである。これらについては係り受けを考慮しながら規定を整備する必要がある。

（２）複合辞・連語の扱い

BCCWJ では複合辞を付属語（助詞・助動詞相当句）として認めていた。中古和文についても、いくつかの語が複合して助詞・助動詞のような機能を持つ形式が存在する。例えば、BCCWJ で認めていた複合辞「という」と同じ語構成の「といふ」は、中古和文でも「愛宕といふ所」のように、現代語同様の助詞相当句として用いられている。共通の形式以外に、「伏籠の中に籠めたりつるものを」「山がつになりて、いたう思ひくづほれはべりし年ごろの後、こよなく衰へにてはべるものを」の「ものを」は詠嘆・逆接を表すものとして、中古和文では助詞相当句とみることができるだろう。このように複合辞らしき形式が存在することは確かであるが、ここで問題となるのはその選定である。BCCWJ では先行研究における扱いや頻度、機能的用法の割合等に基づき複合辞を選定したが、中古和文については、個別の表現に関する検討はあるものの、複合辞となる表現を集めたリスト類がないのが現状である。複合辞を認めるかどうかがまず問題であるが、認めるとした場合には複合辞選定基準の検討が大きな課題となる。

BCCWJ では複合辞のほか、連語を認めている。これは CSJ や先行研究のほか、連濁や係り受けを考慮して、全体で 1 長単位とすることが妥当と考えられる語を人手修正作業の過程で抽出したものである。中古和文についても同様に、連語を認める方針である。「知らず読み」のような付属語を中間に持つ語等、人手修正作業の過程で抽出していく。

（３）固有名の扱い

中古和文においては、女性の呼称を中心に、「朧月夜」のような一般名詞や、「明石の御方」「藤壺」「中務」のように地名・建物名・官職名を用いて人（名）を表すため、短単位の品詞情報を基に固有名の品詞認定をすることが難しい。実例に基づき、品詞付与基準を整備していくことが今後の課題となる。

３．中古和文の長単位解析の現状

長単位は規定としては文節を分割して取り出す言語単位であるが、実際には短単位を基に自動解析する。中古和文の長単位自動解析にあたっては、短単位から長単位を自動構成する解析器 Comainu ver.0.60 [4]を用いた。解析器の学習には BCCWJ のコアデータが用いられており、中古和文への特別な対応はなされていない。

現在、源氏物語桐壺巻の人手修正済み短単位データ 6500 語[5]を基に、長単位の自動解析を行い人手修正を施した。この結果、6500 短単位は 5760 長単位となった。このうち 5166 長単位（89.7%）は短単位と境界が一致しており、長単位構成要素数（1 長単位を構成する短単位数の平均）は 1.13 である。これを BCCWJ コアデータ[6]と比較してみよう。BCCWJ コアデータにおける媒体別長単位構成要素数[7]と短長境界が一致する（1 短単位から構成される）長単位の割合[8]を表 2 に示す。源氏物語桐壺巻は長単位構成要素数が BCCWJ のどの媒体よりも少なく、短長境界が一致する長単位の割合も高い。ここから、源氏物語桐壺巻には合成語が少ない様子が見てとれる。

表２　BCCWJ コアデータにおける媒体別長単位構成要素数

媒体	書籍	雑誌	新聞	白書	Yahoo!知恵袋	Yahoo!ブログ
構成要素数	1.18	1.23	1.32	1.44	1.16	1.18
短長一致(%)	86.7	84.7	79.7	74.7	87.6	―

次に、長単位自動解析の精度について見る。源氏物語帚木～花宴巻及び伊勢物語について計 72720 短単位から記号類を除く 1000 語をサンプリングチェックしたところ、精度は語彙素認定で 95.1% [9]であった。以下、精度確認の結果から長単位境界の認定に関する典型的な誤りについて、具体的にどのような事例があるのかを見ていく。

---

4 小澤ほか（2011）を参照。

5 短単位データは中古和文 UniDic により形態素解析した後、人手修正したものを用いた。

6 「コアデータ」とは、形態素解析システムや長単位自動解析器等の学習用データとして、自動解析後に人手修正を施した高精度のデータセットである。内訳は書籍・雑誌・新聞・白書それぞれ約 20 万語と Yahoo!知恵袋・Yahoo!ブログそれぞれ約 10 万語である。

7 山崎（2011）を参照。

8 冨士池（2010）を参照。

9 精度 95.1%、つまり 1000 語のうち、49 の誤りがあったことになる。内訳は境界の誤りが 28、品詞の誤りが 9、語彙素の誤りが 0、基データの短単位の誤りが 12 であった。

（1）接辞

誤）｜うち‖具し｜　→　正）｜うち=具し｜
｜なま‖いとほし｜　→　｜なま=いとほし｜
｜頼み‖がた｜　→　｜頼み=がた｜

接辞に関しては、BCCWJ と中古和文とで認定基準上の大きな違いはない。そのため、接辞を含む形式であっても大半は正しく解析されていた。その中で誤解析となった接頭辞「うち」「なま」は中古和文で新たに接頭辞とされたものである。そのほか、現代語では比較的頻度の低い「ほの」についても同様の誤解析が見られた。「頼みがた」については、接尾辞「難い」自体は現代語で用いられるが、活用形「語幹−一般」は現代語にはない形式であった。これらの接辞の誤解析は長単位解析器の学習用データに BCCWJ、つまり現代日本語書き言葉を用いているために中古和文特有の形式に対応できていないことが原因と考えられる。

（2）敬語表現

誤）｜ものし‖たまふ｜　→　正）｜ものし=たまふ｜
｜弾か‖せ‖たまふ｜　→　｜弾か=せ=たまふ｜

「弾かせたまふ」については付属語を含む形式であり、中古和文で新たに規定に追加されたものであるため、接辞と同様の学習用データの問題と考えられる。一方、「ものしたまふ」は動詞 2 語として解析されたが、「たまふ」の品詞は「動詞−非自立可能」であり、「動詞−非自立可能」を後項に取る複合動詞というのは現代語で一般的な形式である。「給う」の頻度が BCCWJ で低いことが影響を与えているとも考えられるが、この問題の原因は不明である。

（3）複合動詞

誤）｜具‖し｜　→　正）｜具=し｜
｜法気‖づき｜　→　｜法気=づき｜
｜消え‖のこり｜　→　｜消え=のこり｜

複合動詞に関しては現代語でも十分あり得る品詞構成であり、上に挙げた例については特に問題がないように見え、誤解析の原因は不明である。

現時点での中古和文の長単位解析においては、長単位解析器の学習に現代語を用いていることが誤解析の大きな要因になっているようである。品詞構成上は現代語・中古和文共通であっても、現代語における頻度が低いことが影響しているようにも見えるが、それだけでもないようである。上に挙げた誤解析の不明な点に関しては、解析結果の分析を継続的に行っていく中で、原因を見出していきたい。今後、長単位の認定基準とともに、人手修正済みデータを整備していく。中古和文データを用いて長単位解析器の学習を行うことについては、今後の課題としたい。

４．終わりに

本稿では、「通時コーパス」の中古和文で採用する長短 2 種類の言語単位のうち長単位の認定基準の概要及び現在検討中の課題について説明した。また、長単位解析の現状についても報告した。長単位の認定基準については、今後「通時コーパス」の準備作業を進めていく中で、適宜修正・追加を行っていく予定である。

付　記

本稿は、国立国語研究所共同研究プロジェクト「通時コーパスの設計」（プロジェクトリーダーは近藤泰弘客員教授）の成果の一部である。

文　献

小木曽智信ほか（2012）「和文系資料を対象とした形態素解析辞書の開発」、平成 21（2009）－平成 23（2011）年度科学研究費補助金基盤研究（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 1
（http://dl.dropbox.com/u/73297026/report/unidic-EMJ_report2012.pdf よりダウンロード可能）

小椋秀樹・須永哲矢（2012）「中古和文 UniDic 短単位規定集」、平成 21（2009）－平成 23（2011）年度科学研究費補助金基盤研究（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.dropbox.com/u/73297026/report/unidic-EMJ_rulebook2012.pdf よりダウンロード可能）

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』形態論情報規程集第 4 版（上）」、国立国語研究所内部報告書
（LR-CCG-10-05-01）

小澤俊介・内元清貴・ 伝康晴（2011）「BCCWJ に基づく中・長単位解析ツール」、特定領域「日本語コーパス」平成 22 年度公開ワークショップ予稿集、pp. 331-338
（https://maro.ninjal.ac.jp/Comainu/ にリンクあり）

冨士池優美（2010）「『現代日本語書き言葉均衡コーパス』における長単位の構成要素について」、日本語学会 2010 年度秋季大会予稿集、pp.237-242

山崎誠（2011）「『現代日本語書き言葉均衡コーパス』の構築と活用」、『現代日本語書き言葉均衡コーパス』完成記念講演会予稿集（JC-G-11-01）、pp.11-20

関連 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/