『日本語歴史コーパス 平安時代編』出現形容詞に対する古典分類語彙表番号アノテーション

池上尚
埼玉大学 教育学部
nikegami@mail.saitama-u.ac.jp

１．はじめに

国立国語研究所で進められている『日本語歴史コーパス』の構築は、現在までに「平安時代編」・「鎌倉時代編I説話・随筆」・「室町時代編I狂言」・「江戸時代編I洒落本／II人情本」（試作版）・「明治・大正編I雑誌」としてその成果が公開されており、日本語史研究における今後の利用が大いに期待される。しかし、コーパスを通時的なものへと発展させていく一方で、共時的なものとして、より多角的な観点で様々なアノテーションを付与していくことも求められる。後者の例として、係り結びなどの統語情報や、文脈レベルでの意味情報といったアノテーションの付与が挙げられる。

本稿では、『日本語歴史コーパス』に対する意味情報アノテーションの試行として進めている、「平安時代編」出現形容詞に対する古典分類語彙表番号の付与について作業状況を報告する。

２．『日本語歴史コーパス　平安時代編』

アノテーションの対象となる『日本語歴史コーパス平安時代編』（以下、『CHJ 平安編』）にはすでに、形態論情報や本文種別（地の文・会話・和歌の種別）、話者情報（一部）などが付与されている。本文は小学館『新編 日本古典文学全集』（以下、新編全集）に基づき、「中納言」での検索結果である用例リストからジャパンナレッジ新編全集の当該ページを直接参照し、本文・現代語訳・注釈を確認することができる。古典語に対して分類語彙表番号を付与する場合、一般的解釈の反映と考えられる現代語訳がすでに用意されていることは作業の大きな足掛かりとなる。試行として『CHJ 平安編』を選定した理由もここにある。

このコーパスの作品別語数、ならびに、本研究に関連する形容詞の内数は表 1 の通りである。

表 1 作品別語数（括弧は形容詞内数）

作品名	延べ語数	
竹取物語	10316	(285)
古今和歌集	31260	(691)
伊勢物語	13825	(344)
土佐日記	6685	(197)
大和物語	23091	(693)
平中物語	12403	(325)
蜻蛉日記	47264	(1859)
落窪物語	54586	(2471)
枕草子	66037	(3682)
源氏物語	445715	(23255)
和泉式部日記	10891	(451)
紫式部日記	17442	(815)
堤中納言物語	15696	(640)
更級日記	14660	(765)
大鏡	71267	(1795)
讃岐典侍日記	15544	(492)
計	856682	(38760)

※空白・記号・補助記号は含まない

３．『日本古典対照分類語彙表』

宮島編 1971 の改訂増補版である『日本古典対照分類語彙表』は、奈良時代成立の『万葉集』から鎌倉時代成立の『徒然草』まで、計 17 作品における語の頻度・分類語彙表番号を対照できる表である。現代語の分類語彙表番号（国立国語研究所編 2004）に対して、古典語特有の語義に対する分類語彙表番号を追加した構成になっており（図 1）、付属の CD には本書の内容が Excel ファイルで保存されている（図 2）。

3.50(自然) ――――――――――――――――――――――――――――――――――――
3.5000(自然) ……該当なし
3.5010(光) あかし=明・赤 あきらか=明 あざあざと=鮮鮮 あざやか=鮮
ありありと おほ=凡 おぼおぼし おぼろ=朧 かくやく=赫奕 かそけし
きはやか=際 きらきらし=煌煌 きらきらと=煌煌 くらし=暗
けざけざと けざやか けちえん=掲焉 こぐらし=木暗 さやか=清明
さやけし=清明 しろし=著 せいめい=清明 たどたどし=迚迚
なまぐらし=生暗 はなやか=華 ひたてり=直照 ひらひらと ほのか=仄
ほのぐらし=仄暗 ほのぼの=仄仄 まさやか みやうみやう=冥冥
ものあざやか=鮮 をぐらし=小暗
3.5020(色) あかし=明・赤 あからか=赤 あさし=浅 あざやか=鮮
あさら=浅 あめまだら=飴斑 あをし=青 あをじろ=青白 あをやか=青
いろぶかし=色深 うすらか=薄 かあをし=青 かうかう=皓皓
かぐろし=黒 き=黄 くろぐろ=黒黒 くろし=黒 こまやか=細・濃
さうさう=蒼蒼 しろし=白 しろじろと=白白 つきよらし=着宜
にほひやか=匂 ひたあを=真青 ひたくろ=黯 ふしぐろ=節黒
まくろ=真黒

図 1 古典分類語彙表

見出し	順漢字	語種	品詞	注記	作品	合計	徒然	平家	宇治	方丈	新古	大鏡	更級	紫	源氏	枕	蜻蛉	後撰	土左	古今	伊勢	竹取	万葉	意味分類
あかし	1灯				7	11		1	1				1	1	3	1	3							14600(灯火)
あかし	2証				1	1						1												11113(理由・目的・証拠)
あかし	3明石				7	71		9			4	1			48						1		1	7 12590(固有地名)
あかし	4明・赤		形		15	152	2	5	28	1		11	15	3	38	33	8	2	2	1	1			2 33420(人柄)/35010(光)/35020(色)
あかしう	明得		動下二		1	1																		1 21635(朝晩)
あかしおほと	明石大門				1	1																		1 12580(固有地名)
あかしがた	明石潟				2	2					1													1 12580(固有地名)
あかしがたし	明難		形		1	2									2									31348(難易・安危)
あかしかぬ	明不堪		動下二		3	9		4							4									1 21635(朝晩)
あかしくらしわぶ	明暮侘		動上二		1	1									1									23330(生活・起臥)
あかしくらす	明暮		動四		7	37	2	8	1					1	21		3					1		23330(生活・起臥)
あかしつる	明釣		動四		1	1																		1 23811(牧畜・漁業・鉱業)
あかしはつ	明果		動下二		2	4									3	1								21635(朝晩)
あかしぶみ	明文				1	1									1									13154(文章)

図2 Excel ファイル

4. 形容詞（短単位）に対するアノテーション試行

4.1 データ

今回は、古典分類語彙表番号「3.50(自然)」で始まる形容詞（短単位）のうち、文脈レベルで番号の曖昧性が問題となる以下の 22 語を対象とした。

表2 現在までの作業対象

語彙素読み	語彙素	延べ語数
アカイ	赤い	30
アカイ	明い	110
アサイ	浅い	126
オドロオドロシイ	おどろおどろしい	100
オビタダシイ	夥しい	13
オボオボシイ	おぼおぼしい	5
カタイ	固い	15
カタイ	難い	183
カライ	辛い	53
キヨイ	清い	205
キラキラシイ	きらきらしい	33
クライ	暗い	122
コワイ	強い	12
コワゴワシイ	強々しい	7
サヤケイ	清けい	7
サワガシイ	騒がしい	86
タカイ	高い	209
タドタドシイ	たどたどしい	37
ナゴイ	和い	5
ニガイ	苦い	1
ハヤイ	早い	82
モロイ	脆い	6
	計	1447

4.2 手順

番号付与にあたり、その根拠となる「現代語訳」や「対象」といった情報も記録した。これらを踏まえ該当する番号をプルダウンメニューから選択する（図3）。

「現代語訳」列

新編全集で当てられている当該形容詞の現代語訳を「現代語訳」列にそのまま引用し入力する。なお、現代語訳が意訳で当該箇所と完全に一致しない場合は、その部分を括弧で括って入力する（図4）。形容詞が否定を伴い表現全体で意訳されている場合（「高からず」を「低い」など）は、現代語訳全体を括弧で括って入力する。

「対象」列

当該形容詞の叙述対象を本文から引用し「対象」列に入力する。ここでいう「対象」とは、形容詞の表す属性が帰属する主体（主語・被連体修飾語・被連用修飾語）を指す。主体に着目した形容詞の意味記述は早くから行われており、語の統語情報と意味とは密接に関連し合っていると言われる（西尾 1972・田中 2000）。特に、内省のきかない古典語においては、この「対象」把握が文脈レベルでの番号付与の助けになると考える。また、「対象」を記録しておくことで、語認定の検討データをすぐに取り出すことができ、有益である（後述）。

なお、対象が非明示的な場合は、該当すると考えられる対象を適宜補い括弧で括って入力する（図4）。「対象」は基本的に短単位のレベルで抽出するが、接頭辞「御」が添加した名詞などは長単位のレベルで抽出せざるを得ない（「御物の怪」など）。形容詞への番号付与に直接関係するわけではないが、「対象」列の整合性については今後方針を検討していきたい。

「分類語彙表番号」列

上記2つの情報を踏まえ該当する番号を決定する。

5. 課題

試行を通して明らかになった課題について述べる。

5.1 長単位へのアノテーション

『CHJ 平安編』構築に使用された形態素解析辞書「中古和文 UniDic」における短単位と、古典分類語彙表に

現代語訳	対象	分類語彙表番号	前文脈	キー	後文脈
ごわごわした	紙	3506(材質)	と」とて、いと	こはく	、すくよかなる紙
手ごわい	物の怪	31400(力) 33300(文化・歴史・風俗) 33420(人柄) 35060(材質) 35230(地)	いふ中にも、	こはき	物の怪にとりこ
強く	まもり		り。#まもりの	こはく	やおはしけむ、
険しい	坂		どにて、坂の	こはき	を登りはべりした
固く	烏帽子		押し入れて、	こはから	ぬ烏帽子振りや
かたくるしく	言葉	3330(文化・歴史・風俗	言葉、いとうたて	強く	憎げなるさまを、

図3 作業例

現代語訳	対象	分類語彙表番号	前文脈	キー	後文脈
さわがしく	浪	3508(音)	むうきみるは浪	さわがしく	ありこそはせめ
がやがやと	(雑踏)	3508(音)	どにて、いみじう	騒がしう	おそろしきまで
はげしい	野分	3508(音)	の齢に、まだかく	騒がしき	野分にこそあは
落ち着きを失い(ました)	世の中	3331(人生・禍福)	世の中きはめて	さわがしき	に、またの年、い
世間も不穏なの(で)	(世情)	3331(人生・禍福)	物のさとししきり	騒がしき	を、いみじき御
あわただしく落ち着かぬ	(用務)	3331(人生・禍福)	御覧じつけて、	騒がしけれ	ど、#かたみにそ
おだやかならぬ(夢)	夢見	3331(人生・禍福)	しに、今宵夢見	騒がしく	見えさせたまひ
人騒がせなくらい	(宮の好色ぶり)	3331(人生・禍福)	「この宮の、いと	騒がしき	まで色におはし
大騒ぎをし	(騒動)	3331(人生・禍福)	隔てたれば、	騒がしう	、若き人をも惑

図4 「現代語訳」「対象」の括弧入力例

現代語訳	対象	分類語彙表番号	前文脈	キー	後文脈
潔白である	心	3342(人柄)	水のそこまでに	きよき	心は月ぞ照らさ
清浄な	衣	3506(材質)	#御みづからも	清き	御衣奉り、かぎ
清らかな	水・草	3506(材質)	の夕まで、水草	清き	山の末にて勤め
こざっぱりとした	筵	3506(材質)	筵、ただの筵の	清き	に敷きかへさす
まったく	(事態の程度)	3192(限度)	ぬるありさまを、	清く	知らでなどもあ
はっきりと	(発言の明瞭さ)	3192(限度)	かたはなるを、	清う	さ言はず、女房
きれいさっぱり	(行動の完全さ)	3192(限度)	た人の問ふに、	清う	忘れてやみぬる
清楚な感じの	女		たる女の、いと	清	げなる、いで来
こぎれいな感じ	さまかたち		こ、さまかたちも	清	げなりければ、
こぎれいに	顔かたち		りければ、いと	清	げに顔かたちも

図5 番号付与の単位が異なる例

おける見出し語とで単位認定にずれが生じている例が多くある。形容詞語幹に接尾的要素が付加し品詞の転成したものなどである。これらは、形容詞語幹（短単位）だけでなく、接尾的要素を含めた形（長単位）にも番号を付与していかなければならない。

しかし、短単位・長単位ともに番号が付与できそうな場合でも注意を要するものがある。典型例として、形容詞「清し」と、これの語幹に形容動詞語幹をつくる接尾辞「げ」の添加した「清げ」を挙げる（図5）。この2語は意味・用法が明らかに異なっており、古典分類語彙表でも「清し」には31921(限度)・33420(人柄)・35060(材質)、「清げ」には31345(美醜)が当てられ、2語の番号は重ならない。前者は短単位のみ、後者は長単位のみ、それぞれ番号を付与すべきだろう。

もっとも、適切な単位に番号を付与するには、『CHJ平安編』における語の共時的な在り方を検討することから始めなければならない。また、加藤・浅原・山崎2017などを参考に、通時的に見た場合のデータの整合性にも配慮しながら作業を進めていく必要がある。

5.2 長単位よりも大きな単位へのアノテーション

短単位・長単位ともに1単位として分割される形容詞が連語や慣用表現に用いられた場合、全体でまとまった意味を表すために形容詞単体に番号を付与することが難しくなる。以下のような例が見られた（「｜」は長単位の区切り）。現代語訳からも、形容詞と前接・後接要素とが一体になって特定の意味を表すことが分かる。

｜宰相の君｜と｜聞こゆる｜上臈｜の｜詠みかけたまふ｜。｜折り｜て｜見｜ば｜いとど｜にほひ｜も｜まさる｜や｜と｜すこし｜色めけ｜梅｜の｜初花｜ロ｜はやし｜と｜聞き｜て｜、｜ （源氏物語・竹河）

【現代語訳】「宰相の君と申し上げる上臈の女房がお詠みかけになる。折りて見ば……（折り取ってみたら、なおいっそうよいにおいもまさるかと思われますものを、少しは色めかしく咲いてくださいな。梅の初花の君よ）侍従の君が、達者なものよと感心して、」

短単位・長単位ともに分割されてしまうこうした慣用的表現は、新たな連語・複合形容詞として『CHJ平安編』の規程（国立国語研究所コーパス開発センター

現代語訳	対象	分類語彙表番号	前文脈	キー	後文脈
浅く	谷	31911(長短・高低・深浅	げば深き谷こそ	浅く	なりなめ#在原
低い官位のままで	(官位)	33410(身上)	など言ひて、いと	あさく	てやみたまひに
(薄墨色の衣)	薄墨衣	35020(色)	りあれば薄墨衣	あさけれ	ど涙ぞ袖をふち
浅(さ)	匂ひ	35040(におい)	ば、「匂ひの深さ	浅	さも、勝負の定
浅薄に	(琴の演奏)		弾くすに、はた、	浅く	なりにたるべし、
未熟(でして)	(成熟度)		何ごともまだ	浅く	て、たどり少な
(ご親切な)	とぶらひ		いとうれしう	浅から	ぬ御とぶらひ
(容易ならぬ)	絆		らのために にも	浅から	ぬ絆になむ

図6 適当な番号がない例

（池上尚）編2016）に追加すべきか検討したい。当該形容詞の主体である「対象」列の情報を合わせて整備していくことで、修飾ー被修飾関係・主述関係といった統語情報のアノテーションも進められ、結果として、先のような語（相当のもの）の検討に必要な“形容詞と対象とのコロケーション強度”を計ることも可能になる（池上2016）。

5.3 適当な番号がないもの

例えば「浅し」には、31910(多少)・31911(長短・高低・深浅・厚薄・遠近)・33410(身上)・35020(色)・35040(におい)の5つの番号が用意されているが、これらでは対応しにくい用例が見られた（図6）。

動作・行為の熟達度、人間性の成熟度などを表す場合には3.3421(才能)などが、他者との付き合いや縁の深浅を表す場合には3.3020(好悪・愛憎)などが、それぞれ適当な番号の候補となり得る。番号の曖昧性が問題とならないものも含め、用意された番号では対応が難しい場合のあることを想定して、複数人の判断を経て適当な番号を検討・付与することが望ましい。

6. おわりに

本稿では、『日本語歴史コーパス　平安時代編』に対する意味情報アノテーションの試行として進めている、形容詞への古典分類語彙表番号の付与について、その作業方針と進捗を報告した。番号を付与するレベルや、用意された番号それ自体の検討など課題は多い。現代語訳を参照できるコーパスを対象とし作業時間を短縮できる分、技術的な課題、意味論研究の在り方についても配慮したアノテーションを進めていきたい。

参考文献等

池上尚2016「中古語複合形容詞［名詞＋評価形容詞］の一語性」『国語語彙史の研究』35

加藤祥・浅原正幸・山崎誠2017「『現代日本語書き言葉均衡コーパス』に対する分類語彙表番号アノテーション」『言語処理学会第23回年次大会発表論文集』

国立国語研究所編2004『分類語彙表増補改訂版データベース』http://pj.ninjal.ac.jp/corpus_center/archive.html#bunruidb

国立国語研究所コーパス開発センター（池上尚）編2016『『日本語歴史コーパス　平安時代編』形態論情報規程集』国立国語研究所コーパス開発センター

国立国語研究所コーパス開発センター（冨士池優美・須永哲矢・池上尚ほか）編2016『日本語歴史コーパス 平安時代編』（短単位データ 1.1 / 中納言バージョン2.2.0）http://pj.ninjal.ac.jp/corpus_center/chj/heian.html

田中牧郎2000「統語的方法に基づく語の意味研究―万葉集・八代集のカナシの分析を例として―」『日本語学』19-11

西尾寅弥1972『形容詞の意味・用法の記述的研究』秀英出版

宮島達夫編1971『古典対照語い表』笠間書院

宮島達夫・鈴木泰・石井久雄・安部清哉編2014『日本古典対照分類語彙表』笠間書院

「中古和文UniDic」http://www2.ninjal.ac.jp/lrc/index.php?UniDic（20170116確認）

謝辞

本研究の一部は、平成28年度科学研究費補助金（若手研究(B) 15K16764「統語・意味情報付き形容詞を実装した通時コーパスによる中古形容詞の意味・用法研究」）、ならびに、国立国語研究所言語変化研究領域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」によるものである。