『日本語歴史コーパス』に対する分類語彙表番号アノテーションとその利用

《資料・情報》

『日本語歴史コーパス』に対する 分類語彙表番号アノテーションとその利用

浅原正幸・池上尚・鈴木泰・市村太郎
近藤明日子・加藤祥・山崎誠

キーワード：日本語歴史コーパス、分類語彙表、語義、対応分析

1. はじめに

『日本語歴史コーパス』（CHJ）（国立国語研究所 2023）は、奈良時代編から明治・大正編まで様々な資料を電子化し、形態論情報を付与したコーパスである。我々は CHJ に対して、『分類語彙表増補改訂版』（国立国語研究所 2004）および『日本古典対照分類語彙表』（宮島ほか 2014）の分類語彙表番号付与作業を進めた。

『分類語彙表増補改訂版』は、語を意味によって分類・整理した類義語集である。同梱されている CD-ROM には CSV 形式のデータが格納されており、同データは『分類語彙表──増補改訂版データベース──』として公開されている。分類語彙表では、語を分類番号・段落番号・小段落番号・語番号の順に階層的に体系化している。分類番号は、ピリオドを挟んで左に統語分類（品詞などで決まる文法的な機能）である類を示す 1 桁の数字と、ピリオドを挟んで右に意味分類である分類項目を示す 4 桁の数字からなる。類は、体言である「1. 体」、用言である「2. 用」、形容詞・副詞・連体詞などの「3. 相」、接続詞・感動詞などの「4. 他」の 4 種類のラベルからなる。現在のところ、助詞・助動詞に対する分類番号は限定的であり、さらに固有名詞は分類語彙表にあまり登録されていないために、分類番号が未定義である要素も出現する。分類項目の 1 桁目を部門と呼ぶ。部門は「.1 関係」「.2 主体」「.3 活動」「4. 生産物」「5. 自然」の 5 種類のラベルからなる。また、分類項目の 2 桁目までを中項目と呼び、意味分類は、部門－中項目－分類項目の順で細かくなる階層構造をなす。『日本古典対照分類語彙表』（宮島ほか 2014）は、「竹取物語」・「徒然草」など 17 資料に出現する語彙の頻度とともに対応する分類語彙表番号を付与したものである。同梱されている CD-ROM には頻度表のエクセルファイルが収録されている。現代語と古典語と枕詞などを除いてほぼ同じ分類体系ではあるが、古典語の分類番号はピリオドを用いない。

表 1 に「今昔物語集」のタグ付け例を示す。「今」の例では現代語の分類番号「1.1641」が wlsp 列に付与され、類「1.」が「体」、部門「.1」が「関係」を表す。さらに、中項目「.16」が「時間」、分類項目「.1641」が「現在」を意味する。「越後」の例では古典語の分類番号「12590」が付与され、類「1」が「体」、部門「2」が「主体」を表す。さ

表 1 「今昔物語集」のタグ付け例

pSampleID	pStart	orthToken	lemma	wlsp	類ラベル	分類項目
30-今昔 1100_12001	160	今	今	1.1641	体	関係-時間-現在
30-今昔 1100_12001	170	は	は			
30-今昔 1100_12001	180	昔	昔	1.1642	体	関係-時間-過去
30-今昔 1100_12001	190	、	、			
30-今昔 1100_12001	200	越後	エチゴ	12590	体	主体-公私-固有地名
30-今昔 1100_12001	220	の	の			
30-今昔 1100_12001	230	国	国	1.2550	体	主体-公私-政治的区画
30-今昔 1100_12001	240	に	に			
30-今昔 1100_12001	250	聖人	聖人	12410	体	主体-成員-専門的・技術的職業
30-今昔 1100_12001	270	有	有る	2.1200	用	関係-存在-存在
30-今昔 1100_12001	280	けり	けり			
30-今昔 1100_12001	300	。	。			

らに、中項目「25」が「公私」、分類項目「2590」が「固有地名」を表す。タグ付け作業は、平安時代編、鎌倉時代編、明治・大正編の資料を主な対象として進め、一部であるが、室町時代編の資料にも作業を行った。

出現語に対して語義情報（分類語彙表番号）を付与することにより、語彙の通時的な変化を捉えられるようになる。分類語彙表番号の分布から、各時代の表現としての文体の移り変わりや、資料ごとの内容を代表する主要語義情報を得ることができる。語義情報を用いることにより、資料に対する索引としての利用のほか、割り当てられる語彙素の選択から語彙の通時的変化を捉えられる可能性がある。さらには、初学者向けに本文とともに語義情報を示すことで可読性をあげ、古典に親しみやすい環境整備に資する。

本稿では、大規模に語義情報を CHJ に付与するために行った工夫の概要を示す。さらに事例として「今昔物語集」の語義情報の統計分析結果を示す。

2. 作業の概要

本作業では、分類語彙表番号付与対象を、CHJ すべてのサンプルに定義されている国語研短単位とする。資料によっては、長単位など別単位の分類語彙表番号も付与した。分類語彙表番号アノテーション作業手法を確立するために、まずは現代語に対する分類語彙表番号付与手法について検討した。研究所内で整備途中であった「分類語彙表番号－UniDic 語彙素番号対応表」の整備を行い WLSP2UniDic として公開した（近藤・田中 2020）。WLSP2UniDic は形態論情報に基づき可能な語義を枚挙した対応表である。ChaKi.NET とともに配布されているスタンドアロン版 ChaMame には WLSP2UniDic

が含まれており、これを用いることでテキストに出現する形態素に対して対応表に含まれる全ての語義を枚挙することができる。この環境を用いて、『現代日本語書き言葉均衡コーパス』（BCCWJ）（Maekawa et al. 2014）に対する分類語彙表番号付与を進めた。先の WLSP2UniDic を BCCWJ の語彙素番号に対して突合し、枚挙した分類語彙表番号を Microsoft Excel 上に表層形・語彙素とともに展開し、正しい語義を人手で選択する方法によりアノテーションを行った。対応表に出てこない語義については、『分類語彙表増補改訂版データベース』を参照しながら、分類語彙表番号を手入力した。その過程で作業用のリファレンスとして、分類語彙表版 CradleExpress を整備した。結果、BCCWJ の書籍・新聞・雑誌コアデータの半分約 33 万語に対して分類語彙表番号を付与した。なお、BCCWJ に対しては「現代語の助詞・助動詞」（国立国語研究所 1951）にあげられている助動詞の用法も付与した。

BCCWJ に対する分類語彙表アノテーション作業手法が確立した状況で、CHJ に対する分類語彙表番号アノテーションの作業手法について検討した。池上により既に実施していた中古和文サンプルに出現する形容詞に対する分類語彙表番号アノテーション（池上 2017）の作業手法を参考に、主として奈良～平安～鎌倉時代の資料を元に構築された『古典対照分類語彙表』（宮島ほか 2014）の番号を参照しながら、「竹取物語」「土佐日記」「方丈記」「徒然草」の 4 資料について小規模なアノテーションを実施した。作業者は小学館『新編 日本古典文学全集』の現代語訳を参照しながら、適切な語義を Microsoft Excel 上で選択する作業を実施した。その過程で「古典対照分類語彙表番号 -UniDic 語彙素番号対応表」の基礎となる表が構築された。最終的に同対応表を整理し「古典対照分類語彙表番号 -UniDic 語彙素番号対応表」WLSP2UniDic_historical として公開した。

続いて鎌倉時代以降のデータに対して大規模なアノテーションを実施した。前述の「分類語彙表番号 -UniDic 語彙素番号対応表」と「古典対照分類語彙表番号 -UniDic 語彙素番号対応表」により、自動で可能な語義ラベルを付与し、人手で曖昧性解消を行った。「古典対照分類語彙表番号 -UniDic 語彙素番号対応表」は形態論情報に基づき可能な古典の語義ラベルを追加で枚挙したもので、これにより辞書引きでコーパス中の出現語彙に対する古典語義の分類語彙表番号の自動入力が行われ、作業者は現代語の作業と同様に文脈を見たうえで正しい語義を選択することでアノテーションを行えるようになった。しかしながら、いずれの対応表にも出現しない語義（未登録語義）については、分類語彙表の分類体系から適切な分類番号を手入力で記述した。鎌倉時代の資料として主に説話を扱い「今昔物語集」「宇治拾遺物語」「十訓抄」に対して作業を実施した。この時代の資料のみで 30 万語規模のデータを構築した。室町時代語の資料として「虎明本狂言集」の一部のデータに対して作業した。近代の資料として「国定国語科教科書（第 1 期）」「国定国語科教科書（第 2 期）」「太陽（第 1 巻第 1 号-第 12 号）」に対して作業した。

3. 基礎統計

3.1 基礎統計：類の分布

表 2 に類の頻度表を示す。年代ごとに作業量にばらつきがあるが、全体で 647,751 語に対して作業を行った。このうち 353,890 語が助詞・助動詞・記号・固有名詞などの分類語彙表番号が割り当てられない（分類番号未定義）ものであった。

表 2　統語分類：類の頻度表

	体	用	相	他	未定義	総計
0900 竹取	2,318	2,252	706	72	7,409	12,757
0934 土佐	1,710	1,272	453	45	4,728	8,208
1100 今昔	40,687	29,498	8,518	1,189	95,706	175,598
1212 方丈	1,433	792	342	100	2,735	5,402
1220 宇治	24,214	21,336	6,290	716	68,149	120,705
1252 十訓	19,808	13,039	3,974	460	52,896	90,177
1336 徒然	8,876	6,138	2,688	213	22,919	40,834
1642 虎明	1,255	811	369	88	2,925	5,448
1895 太陽	13,256	6,131	3,116	853	23,038	46,394
1904 小読	10,846	5,312	2,620	794	25,762	45,334
1910 小読	28,915	12,833	6,388	1,135	47,623	96,894
総計	**153,318**	**99,414**	**35,464**	**5,665**	**353,890**	**647,751**

3.2 基礎統計：部門の分布

表 3 に部門の頻度表を示す。部門「.2 主体」と「.4 生産物」は、類「1. 体」にしか定義されていない分類であるために相対的に頻度が低い。

表 3　意味分類：部門の頻度表

	関係	主体	活動	生産物	自然	未定義	総計
0900 竹取	2,305	577	1,680	229	485	7,481	12,757
0934 土佐	1,507	360	990	167	411	4,773	8,208
1100 今昔	36,706	12,221	21,591	3,615	4,570	96,895	175,598
1212 方丈	1,264	255	609	147	292	2,835	5,402
1220 宇治	24,549	6,726	14,516	2,661	3,388	68,865	120,705
1252 十訓	16,012	5,461	11,640	1,540	2,168	53,356	90,177
1336 徒然	8,150	2,011	5,626	714	1,201	23,132	40,834
1642 虎明	1,103	389	598	113	232	3,013	5,448
1895 太陽	10,936	3,324	6,502	609	1,132	23,891	46,394

1904 小読	8,603	2,960	3,644	1,363	2,208	26,556	45,334
1910 小読	23,051	6,426	10,243	3,025	5,391	48,758	96,894
総計	**134,186**	**40,710**	**77,639**	**14,183**	**21,478**	**359,541**	**647,751**

4. 事例研究：「今昔物語集」の巻別・説話別の分布の検討

以下では「今昔物語集」の巻別・説話別の分類番号の分布の検討を行う。コアデータと呼ばれる形態論情報の人手による修正が行われている部分にのみ分類番号を付与したために、分析対象を巻第十二・第十七・第二十・第二十七・第二十九とする。

4.1 今昔物語集：類の変遷

表 4 に「今昔物語集」の巻別の類の分布に基づくカイ二乗検定結果（標準化残差）を示す。巻の番号が小さいほど、体の割合が大きく、用・相の割合が小さい傾向がみられる。巻の番号が大きくなると体の割合が小さくなり、相の割合が大きくなることが確認された。尚、サンプル数が多い場合には検定力が高まるために有意差自体は本質的なものとせず、標準化残差の大小関係を検討する。しかしながら、可読性のために有意差の閾値として $p < 0.05$ 水準（標準化残差 ±1.96 より外側）のものに * を付した。

これは、過去に指摘されてきた巻第二十を境に文体差が認められることの再確認と考える。巻第二十以前は、漢文訓読文体の要素が強く、例えば、動詞や形状詞（形容動詞の語幹）においても漢語の侵出が認められ、体の割合が大きくなることと一致する。反対に、巻第二十以降は和文体の要素が強いと言われ、巻第十二や巻第十七ほどには、動詞や形状詞において漢語が多用されないことにより、体の割合が相対的に小さくなっていると考える。

表 4　今昔物語集巻別の類：カイ二乗検定：標準化残差（未定義あり）

	体	用	相	他	未定義
巻第十二	* 14.56	* -2.96	* -6.90	1.72	* -7.42
巻第十七	* 10.69	* -2.77	* -2.52	0.99	* -6.05
巻第二十	0.07	0.57	0.89	-0.03	-0.86
巻第二十七	* -11.77	-0.20	* 3.65	-1.31	* 8.76
巻第二十九	* -13.68	* 5.23	* 4.93	-1.39	* 5.77

4.2 今昔物語集：部門の変遷

表 5 に今昔物語集の巻別の部門の分布に基づくカイ二乗検定結果（標準化残差）を示す。部門においても巻第二十を境に異なる傾向が見られた。その中でも巻第十七が主体・活動の割合が多く、生産物・自然の割合が少ないという特徴的な傾向がみられた。

表５　今昔物語集巻別の部門：カイ二乗検定：標準化残差（未定義あり）

	関係	主体	活動	生産物	自然	未定義
巻第十二	-0.90	* 2.42	* 9.72	0.11	1.35	* -7.42
巻第十七	* -2.95	* 10.53	* 11.51	* -8.94	* -6.24	* -6.05
巻第二十	* -2.40	* 2.80	0.28	-0.97	* 4.67	-0.86
巻第二十七	* 2.42	* -7.43	* -12.41	* 6.04	-1.38	* 8.76
巻第二十九	* 3.83	* -8.33	* -9.31	* 3.79	1.37	* 5.77

意味分類の中で巻別で特徴的な差異がみられた「主体」の上位頻度語を**表６**に示す。「人」は各巻に一貫して頻出するが、巻第十二・第十七・第二十と第二十七・第二十九とで傾向が異なることがわかった。前者は「我」「僧」「国」「寺」「汝」などが多いが、後者は「女」「者」「男」などが多い傾向がわかる。さらに、巻第十七の特徴として「菩薩」「地蔵」が、巻第十二の特徴として「聖人」が多い傾向がわかった。この上位頻度語の差異は、本朝仏法部（巻第十一～巻第二十）が仏教説話を、本朝世俗部（巻第二十一～巻第三十一）がその他の説話をそれぞれ収録していることの影響によると考える。このように、分類番号を用いることで、従来分析できなかった観点での定量的な評価が行えるようになった。

表６　意味分類「主体」の上位頻度語

語彙素	巻第十二	巻第十七	巻第二十	巻第二十七	巻第二十九	総計
人	233	250	256	265	207	1211
女	21	70	89	180	207	567
我	118	160	131	34	52	495
者	44	68	72	88	150	422
僧	99	164	103	9	27	402
男	11	52	39	121	158	381
国	75	77	77	39	36	304
寺	129	82	57	3	27	298
汝	44	95	68	6	1	214
菩薩	33	165	8	0	0	206
地蔵	0	191	0	0	0	191
聖人	119	11	60	0	0	190

5.　おわりに

本研究では、『日本語歴史コーパス』に対する分類語彙表番号アノテーション作業について、その手法を関連言語資源とともに紹介した。海外ではデジタルヒューマニ

ティーズに基づく言語研究において実証性が求められており、再現性のある科学的な研究手法の導入が進んでいる。世界的にみても、本研究で構築した 60 万語規模の語義つきコーパスは類を見ない。『現代日本語書き言葉均衡コーパス』の分類語彙表番号アノテーションデータと合わせると 90 万語規模の言語資源となる。基礎統計として、統語分類（類）と意味分類（部門）の頻度情報を示した。

また、事例研究として、今昔物語集の巻別の分析をおこなった。形態論情報に基づく表現の変遷のみならず、分類語彙表（類・部門）に基づく表現の変遷の計量が可能になった。言語研究にも実証性が求められるなか、既存の研究にはない観点での定量的な分析を示した。

近年、深層学習に基づく文脈化単語埋め込み技術が発達しており、単語・句レベルの類似度が定義できるようになった。本言語資源を fine-tuning に用いることで、文脈化単語埋め込みに基づく類似度を、統語的な類似度と意味的な類似度に分類することが可能になるであろう。

引用参考文献一覧

国立国語研究所（2023）『日本語歴史コーパス』（バージョン 2023.3，中納言バージョン 2.7.1）https://clrd.ninjal.ac.jp/chj/（2023/06/01 確認）

国立国語研究所（2004）『分類語彙表増補改訂版』東京：大日本図書

近藤明日子・田中牧郎（2020）「「分類語彙表番号 -UniDic 語彙素番号対応表」の構築」国立国語研究所論集 18:77-91.

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』東京：笠間書院

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., & Den, Y. (2014). Balanced corpus of contemporary written Japanese. Language Resources and Evaluation, 48(2), 345-371.

国立国語研究所（1951）『現代語の助詞・助動詞――用法と実例――』東京：秀英出版

池上尚（2017）「『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」言語処理学会第 23 回年次大会発表論文集，pp.310-313.

参照 URL

『日本語歴史コーパス』（バージョン 2023.3）https://chunagon.ninjal.ac.jp/chj/（2023/07/21 確認）

『分類語彙表――増補改訂版データベース――』https://github.com/masayu-a/WLSP（2023/07/21 確認）

『WLSP2UniDic』https://github.com/masayu-a /WLSP2UniDic（2023/07/21 確認）

『スタンドアロン版 ChaMame』https://ja.osdn.net/projects/chaki/releases/p15635（2023/07/21 確認）

『分類語彙表版 CradleExpress』https://cradle.ninjal.ac.jp/wlsp/（2023/07/21 確認）

『WLSP2UniDic_historical』https://github.com/masayu-a /WLSP2UniDic_historical（2023/07/21 確認）

謝辞　アノテーション作業にご協力くださった皆様に感謝いたします。本研究は、JSPS 科研費 17H00917、国立国語研究所コーパス基礎研究プロジェクト「コーパスアノテーションの拡張・包括・自動化に関する基礎研究」、国立国語研究所プロジェクト「アノテーションデータを用いた実証的計算心理言語学」の支援を受けました。

付記　本データを https://github.com/masayu-a/CHJ-WLSP より公開する。

――あさはら まさゆき　国立国語研究所教授――
――いけがみ なお　埼玉大学准教授――
――すずき たい　東京大学名誉教授――
――いちむら たろう　京都府立大学准教授――
――こんどう あすこ　東京大学助教――
――かとう さち　目白大学専任講師――
――やまざき まこと　国立国語研究所客員教授――

（2023 年 1 月 30 日　第 1 稿受理）
（2023 年 7 月 31 日　最終稿受理）

Annotation of ‘Word List by Semantic Principles’ Labels for the Corpus of Historical Japanese and Its Application

Asahara Masayuki, Ikegami Nao, Suzuki Tai, Ichimura Taro,
Kondo Asuko, Kato Sachi, Yamazaki Makoto

Keywords: Corpus of Historical Japanese, Word List by Semantic Principles, Word Sense, Correspondence Analysis