BCCWJ-WLSP-LUW: 『現代日本語書き言葉均衡コーパス』に対する長単位語義情報アノテーション

加藤祥 [1]　浅原正幸 [2]
[1] 目白大学　[2] 国立国語研究所・総合研究大学院大学
masayu-a@ninjal.ac.jp

概要

本稿では，『現代日本語書き言葉均衡コーパス』（BCCWJ）の長単位への分類語彙表番号アノテーション作業とその結果について報告する．今回構築した BCCWJ-WLSP-LUW は，BCCWJ の書籍・新聞・雑誌データに対し，『分類語彙表増補改訂版』を基に分類語彙表番号を付与し，長単位の文脈的な意味分類を可能にしたアノテーションデータである．短単位への分類語彙表番号付与作業を基盤としつつ，未定義の長単位語義については一部手作業による対応を行った．これにより，長単位においても短単位と同様に分類語彙表を用いた詳細な意味分析が可能となった．

1　はじめに

本稿は，『現代日本語書き言葉均衡コーパス』[1]（以下 BCCWJ）に対する長単位への分類語彙表番号アノテーション作業と作業結果である BCCWJ-WLSP-LUW について報告する．

BCCWJ-WLSP [2] は，BCCWJ の書籍・新聞・雑誌データに『分類語彙表増補改訂版』[3] の分類語彙表番号を付与し，文脈的な意味分類によってコーパスを調査することを可能にしたアノテーションデータである．BCCWJ コアデータに含まれる書籍サンプル（PB），新聞サンプル（PN），雑誌サンプル（PM）のそれぞれの一部である 347,094 短単位を対象とし，自立語 の 182,166 語に分類語彙表番号を付与している．

今回，BCCWJ-WLSP の長単位についても同様の手続きによるアノテーション作業を行った．複数短単位については「分類語彙表番号-UniDic」対応表 [4] が整備されているが，対応のない長単位への分類語彙表番号付与は，手作業による．作業の結果，BCCWJ の長単位の分類語彙表番号の整備が完了し，BCCWJ-WLSP-LUW として公開した．長単位についても，短単位同様の意味分類の利用が可能となった．また，短単位の付与作業結果を確認することで BCCWJ-WLSP-LUW の意味分類作業を確認した．

2　BCCWJ-WLSP

加藤他 [2] では，短単位における自立語（のべ 182,166 語）について，文脈上適切な分類語彙表番号を人手により付与した．なお，このうちのべ 19,438 語（自立語の 10.7%）については，分類語彙表にない語に対して新たに分類番号を付している．

表 1　分類番号の構造　（例：この（分類番号：3.1010））

類	部門	中項目	分類項目
相（3）	関係（.1）	真偽（.10）	こそあど（.1010）

作業は，まず，UniDic 語彙素番号 [5] と分類語彙表番号を人手で対応させたデータ [6] を用い，自立語について BCCWJ の短単位ごとに，対応可能性のある分類語彙表番号の 5 桁目まで（表 1）を列挙し，作業者が分類語彙表番号の選択肢から，手作業で文脈的に該当する意味分類を選択した．

また，加藤他 [7] は，分類語彙表番号の付与されたコーパスの助動詞にも用法情報を付与する作業を行った．作業対象は，『現代語の助詞・助動詞』[8] の第二部（助動詞）に掲載された 27 種類とし，分類（用法の詳細説明）や用例の確認は同書に拠った．作業対象とした助動詞には，たとえば「う」「よう」のように UniDic では動詞の意志推量形となるなど，他辞書では助動詞扱いではないものも比較的広く含めている．作業者は，自立語の分類番号付与作業と同様に，助動詞が出現する文脈を読み，該当語がいずれの分類に該当するのかを判定した．これらの助動詞の用法は，小松原 [9] が『現代語の助詞・助動詞』を電子化したことにより，分類語彙表番号との

対応が進められている．

3 長単位への分類語彙表番号付与作業

長単位の分類語彙表番号付与は，短単位の分類語彙表番号を用いて行われる．77249 件については，短単位と長単位の形態素単位が同じであるためにそのまま流用する．短単位と長単位の形態素単位が異なる自立語を中心に 50240 件に対して新たに分類語彙表番号を付与した．既存の BCCWJ-WLSP における短単位に付与された意味分類を用いることで，長単位の分類語彙表番号をその構成語情報に基づき付与した．

(1) における長単位は，「家庭訪問」「てくる」「5 日制」「廃止する」「懇談会」であるが，複合名詞の場合，「家庭訪問」は「訪問（1.3520: 体-活動-交わり-応援・送迎)」，「5 日制」は「制（1.3082: 体-活動-心-制度・慣例)」，「懇談会」は「会（1.3510: 体-活動-交わり-集会)」の分類番号が適用可能である．サ変動詞「廃止する」では，「廃止（1.1503: 体-関係-作用-終了・中止・停止)」の類が異なるのみであり，「廃止する（2.1503: 用-関係-作用-終了・中止・停止)」の分類番号となる．しかし，「てくる」のような複合語は，短単位が「て」「くる（2.1527: 用-関係-作用-往復)」であるが，複合語となるために意味分類が異なり，分類語彙表上も長単位として「てくる（2.1500: 用-関係-作用-作用・変化)」がある．分類語彙表に掲載のある長単位は，該当の番号を付与することとなる．

(1) この時期、恒例なのが「家庭訪問」だ。昔は一律に行われてきたが、最近は、学校 5 日制や共働きが増えたことで、廃止したり面談や懇談会に切り替えたりするケースもある。
（PN1a_00002, 句読点は原文ママ，下線は著者による（以降同様））

また，(2) の「デビューする（2.1210: 用-関係-存在-出没，2.3833: 用-活動-事業-興行)」のように長単位の意味分類に複数の可能性がある場合や，(3) の「立ち遅れる」のように，分類語彙表に掲載のない複合語もある．文脈的な意味を確認し，意味分類（2.1584: 用-関係-作用-限定・優劣）を付与した．

(2) 大学在学中にデビューし、キャスターとして活躍するその後の進路を思うにつけ、「本当によく見ていてくれた」と感謝する。
（PN1c_00001）

(3) 西海岸に比べて立ち遅れていた東海岸も、今やハイテク企業の輩出地。
（PN1c_00004）

当該長単位が分類語彙表にない場合，分類項目については，類語を確認して該当すると判断された番号を付与した．形状詞や副詞の指示があっても分類語彙表番号には体の類しかないなど，類のみ該当番号がない場合は，該当すると判断された類の番号を作成して付与した．特に，機能語を含めたすべての外国語が，形態素解析結果において「名詞-普通名詞-一般」とされるため，個別の確認が必要となった．

また，分類語彙表に対応する意味がある場合は，該当する分類番号を付与するが，分類語彙表と対応のない場合は，以下のように最小単位毎に分類語彙表番号を付与し，「&」で接続して示した．

- 分類語彙表番号に意味が対応していない場合
  例）一度：1.1960&1.1962
  体-関係-量-数記号（一二三）&体-関係-量-助数接辞
- 分類語彙表の分類で対応できない場合
  例）重厚：1.1914&1.1911
  体-関係-量-軽重&体-関係-量-長短・高低・深浅・厚薄・遠近

(4) において，短単位「紫外（1.5020&1.1770: 体-自然-自然-色&体-関係-空間-内外)」「線（1.1711: 体-関係-空間-線)」「B（1.1961: 体-関係-量-順位記号（甲乙丙)」，「紫外（1.5020&1.1770: 体-自然-自然-色&体-関係-空間-内外)」「線（1.1711: 体-関係-空間-線)」「量（1.1900: 体-関係-量-量)」は，それぞれに分類語彙表が付与されたが，長単位は「紫外線B（1.5010: 体-関係-作用-動き)」「紫外線量（1.1900: 体-関係-量-量)」が付与された．

(4) 日焼けの原因となる「紫外線B」の年間照射量は、沖縄県が北海道の約 2 倍。（中略）だが、札幌市の子があびた 1 平方メートルあたりの紫外線量は年 3 万 9117 ジュールだったのに、那覇市は年 3 万 4568 ジュールと少なかった。
（PN5a_00003）

4 構築したデータの基礎統計

本節では構築した長単位の分類語彙表付与結果を確認する．長単位 (BCCWJ-WLSP-LUW) の類・部

表 2　長単位の分類語彙表番号付与結果 (BCCWJ-WLSP-LUW)

類・部門	PB		PM		PN		全体	
	頻度	割合 (%)	頻度	割合 (%)	頻度	割合 (%)	頻度	割合 (%)
1.1 体-関係	7,900	8.19	8,894	9.25	10,438	11.77	27,232	9.68
1.2 体-主体	5,424	5.63	4,816	5.01	6,643	7.49	16,883	6.00
1.3 体-動作	4,547	4.72	5,188	5.39	7,135	8.05	16,870	6.00
1.4 体-生産物	2,106	2.18	2,354	2.45	1,388	1.57	5,848	2.08
1.5 体-自然	1,946	2.02	1,777	1.85	1,035	1.17	4,758	1.69
2.1 用-関係	7,766	8.06	6,745	7.01	5,041	5.68	19,552	6.95
2.3 用-動作	6,480	6.72	5,264	5.47	4,902	5.53	16,646	5.92
2.5 用-自然	281	0.29	222	0.23	163	0.18	666	0.24
3.1 相-関係	5,598	5.81	4,632	4.82	2,723	3.07	12,953	4.61
3.2 相-主体	1	0.00	0	0.00	0	0.00	1	0.00
3.3 相-動作	1,047	1.09	1,027	1.07	623	0.70	2,697	0.96
3.5 相-生産物	207	0.21	163	0.17	74	0.08	444	0.16
4 他	1,257	1.30	882	0.92	402	0.45	1,664	0.59
全形態素	96,406	100.00	96,167	100.00	88,686	100.00	281,259	100.00

表 3　短単位の分類語彙表番号付与結果 (BCCWJ-WLSP)

類・部門	PB		PM		PN		全体	
	頻度	割合 (%)	頻度	割合 (%)	頻度	割合 (%)	頻度	割合 (%)
1.1 体-関係	12,026	10.74	17,016	15.27	21,801	19.51	50,843	14.90
1.2 体-主体	6,573	5.87	6,678	5.99	11,006	9.85	24,257	7.11
1.3 体-動作	6,661	5.95	8,792	7.89	12,738	11.40	28,191	8.26
1.4 体-生産物	2,352	2.10	3,003	2.69	2,063	1.85	7,418	2.17
1.5 体-自然	2,361	2.11	2,702	2.42	1,685	1.51	6,748	1.98
2.1 用-関係	6,686	5.97	5,980	5.37	4,893	4.38	17,559	5.14
2.3 用-動作	7,472	6.67	6,886	6.18	5,998	5.37	20,356	5.96
2.5 用-自然	243	0.22	215	0.19	126	0.11	584	0.17
3.1 相-関係	6,023	5.38	5,729	5.14	3,873	3.47	15,625	4.58
3.2 相-主体	2	0.00	5	0.00	0	0.00	7	0.00
3.3 相-動作	964	0.86	1,070	0.96	648	0.58	2,682	0.79
3.5 相-生産物	233	0.21	229	0.21	78	0.07	540	0.16
4 他	1,206	1.08	883	0.79	408	0.37	2,497	0.73
全形態素	111,983	100.00	111,459	100.00	111,754	100.00	341,305	100.00

門（分類番号の小数点 1 桁まで）の基礎統計を表 2 に示す．対照のために短単位 (BCCWJ-WLSP) の類・部門の基礎統計を表 3 に示す．集計時には「&」を含むものなど，単一の分類語彙表番号で表現できないものは計数しなかった．

長単位の全形態素数は 281,259 であり，短単位の全形態素数 341,305 より少ない．これは主に複合語や連語が長単位として統合されることで，複数の短単位が 1 つの長単位として扱われるためである．「1.1 体-関係」の割合は，長単位（9.68%）が短単位（14.90%）より低い．一方「2.1 用-関係」の割合は，長単位（6.95%）が短単位（5.14%）より高くなっている．これは関係を表す語（短単位）の 10.6%がサ変可能名詞であり，短単位では名詞であったが複合語構成時に動詞になりやすい傾向があることが示唆される．さらに，複合名詞においても，「旧（1.1642: 体-関係-時間-過去)」「国立（1.1220: 体-関係-存在-成立)」「大量（1.1910: 体-関係-量-多少)」など，関係を表す語が複合語構成時の前件になりやすい傾向があるといえる．

5　おわりに

本稿では，『現代日本語書き言葉均衡コーパス』（BCCWJ）の長単位への分類語彙表番号アノテーション作業とその成果について報告した．今回構築した BCCWJ-WLSP-LUW は，BCCWJ の書籍・新聞・雑誌データに対し，『分類語彙表増補改訂版』を基に分類語彙表番号を付与し，長単位の文脈的な意味分類を可能にしたアノテーションデータである．短単位への分類語彙表番号付与作業を基盤としつつ，未定義の長単位語義については一部手作業で対応することで，短単位と同様に分類語彙表を用いた詳細な意味分析を長単位でも実現した．今後，BCCWJ-WLSP-LUW を公開する予定である．

謝辞

本研究は JSPS 科研費 JP22K18483 の助成を受けたものです．また国立国語研究所の共同研究プロジェクトによるものです．

参考文献

[1] 前川喜久雄 (監修), 山崎誠 (編). 書き言葉コーパス—設計と構築—. 講座日本語コーパス 2. 朝倉書店, 2014.

[2] 加藤祥, 浅原正幸, 山崎誠. 分類語彙表番号を付与した『現代日本語書き言葉均衡コーパス』の書籍・新聞・雑誌データ. 日本語の研究, Vol. 15, No. 2, p. 134, 8 2019. 資料・情報.

[3] 国立国語研究所. 分類語彙表増補改訂版データベース（ver.1.0.1）, 2018. Available at: https://github.com/masayu-a/WLSP.

[4] 片山久留美, 髙橋雄太, 菊池そのみ, 小木曽智信. 複数短単位版「分類語彙表番号-UniDic」対応表の整備と公開. 言語処理学会 第 30 回年次大会 発表論文集, pp. 165–170, 3 2024.

[5] 小木曽智信, 中村壮範. 『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用. 自然言語処理, Vol. 21, No. 2, pp. 301–332, 2014.

[6] 近藤明日子, 田中牧郎. 「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, No. 18, pp. 77–91, 2020.

[7] 加藤祥, 浅原正幸, 山崎誠. 『現代日本語書き言葉均衡コーパス』新聞・書籍・雑誌データの助動詞に対する用法情報付与. 日本語学会 2019 年度春季大会予稿集, pp. 169–174, 5 2019.

[8] 国立国語研究所. 現代語の助詞・助動詞, 国立国語研究所報告, 第 3 巻. 1951.

[9] 小松原哲太. 『現代語の助詞・助動詞』の電子化とその応用 : 直喩へのアノテーションの事例. 国立国語研究所論集, Vol. 24, pp. 45–58, 2023. 国立国語研究所.