『現代日本語書き言葉均衡コーパス』出版書籍サンプルのNDC別語彙分布

著者	加藤 祥, 浅原 正幸
雑誌名	言語資源活用ワークショップ発表論文集
巻	6
ページ	218-225
発行年	2021
URL	http://doi.org/10.15084/00003496

『現代日本語書き言葉均衡コーパス』出版書籍サンプルの NDC 別語彙分布

加藤 祥（目白大学）†
浅原 正幸（国立国語研究所）

Lexical distribution of the Balanced Corpus of Contemporary Written Japanese by NDC

Sachi Kato (Mejiro University)
Masayuki Asahara (National Institute for Japanese Language and Linguistics)

要旨

『現代日本語書き言葉均衡コーパス』の書籍サンプルには NDC 情報が付与されており，構築当時に情報のなかった書籍などへの増補も行われた（加藤ほか 2021）。また，コーパスに付与された NDC を利用することで，ジャンル別の特徴語の抽出などが試みられてきた（内田・藤井 2015）。しかし，一般動詞など，多義的あるいは補助的に使用される語は，語義情報なしでは語彙としての分布傾向が見られにくく，ジャンル横断的な分布となる。そこで，本稿は，増補した NDC（加藤ほか前掲）を用いてジャンルの語彙分布を再確認するとともに，分類語彙表番号の付与された BCCWJ-WLSP（加藤ほか 2019）と重ね合わせることにより，語義分布に内容別の傾向が見られることを確認する。

１．はじめに

内田・藤井（2015）は，『現代日本語書き言葉均衡コーパス』（以降 BCCWJ）の書籍等サンプルの語彙頻度をジャンル別に調査し，上位頻度語に着目したジャンルの特徴抽出を試み，また，ジャンル内の語に観察される類似性は，同一または類似のフレームを喚起することであると示す。同ジャンルの文章は，類似文脈であることが期待され，上位頻度語における何らかの類似性が推察される。但し，意味的にジャンル別の特徴があることが示唆されるが，ジャンル横断的に出現する上位頻度語の調査にあたっては，類似文脈上どのような意味で用いられているかという意味的な情報が必要となる。よって，類似文脈における語の類似性をどのように計るかという点，類似文脈において特徴的でないジャンル横断的な多義語類は意味的な情報が求められるという点により，文脈の類似性は，意味的な観点で検証することが求められよう。

BCCWJ-WLSP では，書籍サンプルの一部（PB）に人手で意味的な情報が付与されている。意味的な観点での検証が可能となっている。また，加藤ら（2021）は，『現代日本語書き言葉均衡コーパス』（BCCWJ）の書籍全サンプル 22,058 サンプル（PB（出版）10,117 サンプル・LB（図書館）10,551 サンプル・OB（ベストセラー）1,390 サンプル）に付与された日本十進分類法（NDC）分類記号の補助分類を拡張した。NDC 分類が BCCWJ 構築当時に収集できておらず「分類なし」となっていた 938 サンプルについても NDC 分類を確認し，540 サンプルについて増補を行った。現在，増補された BCCWJ の NDC 情報を用いた語彙

† s.kato@mejiro.ac.jp

分布調査が可能となった。

本稿は，BCCWJ の増補された NDC 情報をジャンル情報として用い，ジャンルによる語彙の分布傾向を確かめるとともに，BCCWJ-WLSP を重ね合わせることにより，文章内容と語義の分布を調査する。また，ジャンル横断的と考えられる語についても，語義的な分布傾向のあることを確かめる。

２．調査に使用したデータ

本稿では，BCCWJ-WLSP と重ね合わせることから，PB を使用した集計を行う。

２.１　NDC 情報

加藤ほか（2021）の増補した NDC 情報を使用する。BCCWJ 構築時に NDC 分類記号がなかったサンプルにも，新たに番号を付与しているためである。

増補データにおいては，付与されていた NDC 分類記号（第一次区分：類目表・第二次区分：綱目表・第三次区分：要目表）に下位区分が確認された場合は，該当する番号を追加している。NDC 新訂 9 版（日本図書館協会分類委員会 1995）では，6 区分（形式区分・地理区分・海洋区分・言語区分・言語共通区分・文学共通区分）が一般補助表にあたり，類の一部分に固有補助表（細区分表）がある。なお，新訂 10 版（日本図書館協会分類委員会 2018）では言語共通区分・文学共通区分が固有補助表となっている。これらのデータの扱いは国立国会図書館サーチ API（以下 NDL サーチ）の付与済み NDC 情報に依拠したものである。

PB においては，106 の「分類なし」サンプルに NDC が付与されている。表 1 に本稿で調査対象とする PB の NDC 別語数を示す。

表 1　PB の NDC 別語数

NDC	0 総記	1 哲学	2 歴史	3 社会科学	4 自然科学	5 技術工学	6 産業	7 芸術美術	8 言語	9 文学	分類なし
全体	951540	1758372	2951728	8570302	3206332	2791835	1454079	2050221	610483	8343115	1003574
固定長	242367	413380	670390	1921190	762655	672686	328787	489380	142616	1660215	232743

２.２　語義情報

BCCWJ の短単位に分類語彙表の意味分類が付与された BCCWJ-WLSP（加藤ほか 2019a）を意味情報として使用する。BCCWJ-WLSP は読み手が文脈上どのような意味として読んだのか，手作業で意味分類が付与されており，読み手が当該文脈において読み取った意味情報が付与されたデータである。BCCWJ-WLSP の PB データ は 111,983 短単位であり，付属語を除く 52,812 短単位に語義（加藤ほか 2019a）が，助動詞 10,321 短単位に用法（加藤ほか 2019b）が付与されている。

表 2 に本稿で調査対象とする BCCWJ-WLSP の PB データの NDC 別語数を示す。BCCWJ-WLSP の PB データは，2 番台のサンプルが長く，3 番台のサンプルが短いなどの影響により語数分布にばらつきが見られるため，以降，固定長を用いた調査を行う。

表 2　BCCWJ-WLSP の PB データの NDC 別語数

NDC	0 総記	1 哲学	2 歴史	3 社会科学	4 自然科学	5 技術工学	6 産業	7 芸術美術	8 言語	9 文学
全体	5469	7038	22434	4802	9633	6917	6283	7554	4236	37617
固定長	2213	3663	4525	2217	2121	1402	2903	1474	1610	9681

３．NDC 別語彙分布

３.１　NDC 別上位頻度語

NDC 別に上位頻度語を確認しておく。表 3 に固定長を用いた相対頻度を示す。ジャンルに特有の分布が明らかとはいい難い。記号の「「」「」」が 9 番台に頻出する傾向が確認できるものの，頻度の高い語は助詞や記号，補助動詞，名詞であっても数詞や形式的な名詞などであり，ジャンルに特徴的とはいえまい。高頻度の語はジャンル横断的に出現する傾向にあるといえる。

ジャンルに限定的な名詞を見ることで，ジャンルの特徴語を抽出することは可能であるが，いずれも低頻度という問題がある。なお，名詞のみを見ても，「製造」「市民」「株主」のようなジャンルに明確な分布のある語の頻度は 1000 位以降の集計結果となる。

表 3　ジャンル別上位頻度語（PB 全体・固定長・相対頻度）

	品詞	PB	0 総記	1 哲学	2 歴史	3 社会科学	4 自然科学	5 技術工学	6 産業	7 芸術美術	8 言語	9 文学
、	補助記号-読点	44269	44152	54691	50381	39841	29502	38813	41051	48833	40788	53638
の	助詞-格助詞	43632	42790	43418	47396	48029	44345	43751	45017	43426	39624	37733
。	補助記号-句点	31467	32566	33333	31095	26569	22697	28615	28912	33898	29204	41503
に	助詞-格助詞	30615	28618	32573	32196	32030	29766	29850	29420	29613	28124	30277
て	助詞-接続助詞	28027	26938	31818	26969	27034	25542	25611	25862	27931	25642	32435
を	助詞-格助詞	27772	30202	29665	26131	27169	26104	29449	25545	26141	24626	29177
は	助詞-係助詞	27615	26262	30093	27645	26741	26569	24475	26181	27177	28068	31213
だ	助動詞	26155	26060	29982	24299	24395	23385	22336	23419	26787	26680	31833
為る	動詞-非自立可能	23283	28321	22868	21819	27472	25414	27099	24600	20330	19942	16399
た	助動詞	23169	19178	19239	28624	17877	14442	16074	17336	24006	16667	38235
が	助詞-格助詞	20170	19244	20400	18726	19899	22714	19299	20582	20064	18084	20702
と	助詞-格助詞	18987	18575	23680	19812	20428	17961	17148	18298	19647	21891	17601
で	助詞-格助詞	10420	12143	9630	9873	9779	10758	11457	10925	12530	10770	9684
居る	動詞-非自立可能	9685	8995	10738	9742	9044	9024	7851	9054	9714	8435	11973
も	助詞-係助詞	9471	9354	10750	9037	8120	8352	7753	8744	10863	9894	11874
有る	動詞-非自立可能	8798	8380	11725	10193	10287	10289	8014	9261	7973	10483	6059
「	補助記号-括弧開	8146	8108	8460	6186	6896	4053	5280	6433	6619	13603	14005
」	補助記号-括弧閉	8136	8120	8435	6168	6886	4053	5274	6421	6606	13596	13991
の	助詞-準体助詞	7626	8091	9432	6971	5856	6053	5640	6433	8627	6893	11186
一	名詞-数詞	7473	6527	5503	8650	8742	8712	10428	9803	8423	8470	2803
,	補助記号-読点	7001	5904	1570	1168	11424	21071	10281	9045	928	8414	45
事	名詞-普通名詞-一般	6720	6729	9572	5670	7877	7265	5808	6335	5973	7608	5781
言う	動詞-一般	6046	6614	8764	5719	5950	4929	4305	4912	6584	8358	6961
れる	助動詞	5766	6573	6147	6592	6531	6357	6052	5520	5190	5504	4269

３.２　NDC 別体用相分布

まず，ジャンルによって品詞分布に差の現れることが期待される。そこで，表 4 に体用相の分布を示す。体は UniDic の 品詞における名詞と代名詞，用は動詞，相は形容詞・形状詞・副詞・連体詞として集計している。

3 番台，5 番台，6 番台で体の類が高く，1 番台や 9 番台で用の類が低く相の類が高い。3 番台では相の類が低いというジャンル別の分布傾向が確認できる。また，MVR を見ると，3 番台で動作，7 番台と 9 番台で様態が描写される傾向にあるという特徴がわかる。

表 4　ジャンル別体用相分布（PB 全体・固定長・相対頻度）

品詞	相対頻度	0 総記	1 哲学	2 歴史	3 社会科学	4 自然科学	5 技術工学	6 産業	7 芸術美術	8 言語	9 文学
体	312250	307043	278860	334635	340754	333929	347510	347146	312918	288509	241789
用	116873	116431	131051	111571	114522	111892	111129	108788	113726	116053	127944
相	45118	44763	50847	40918	39124	41053	40099	40379	47419	46418	57233
MVR	(100 万語当たり)	384457	387991	366744	341630	366895	360832	371170	416962	399976	447329

３.３　NDC 別語義分布

NDC は各書籍の内容で分類されていることから，ジャンル別の内容語は文脈的に類似していることが期待され，意味的に異なる分布を示すと考えられる。語義別の分布を見てみたい。はじめに，表 5 に分類語彙表の部門による語義分布を示す。表 6 では各語義（部門）がどのような記事で出現するかを検討したカイ二乗検定結果を示す。品詞間のジャンルごとの頻度を係数し，カイ二乗検定を行い、標準化残差の値により検討する。標準化残差は ±1.96 より外側の場合 $p<0.05$ 水準で有意（表中⇗⇘で示す）、±2.56 より外側の場合 $p<0.01$ 水準で有意（表中⇑⇓で示す）とされる。

表 5　ジャンル別語義分布（BCCWJ-WLSP 部門・固定長・相対頻度）

部門	相対頻度	0 総記	1 哲学	2 歴史	3 社会科学	4 自然科学	5 技術工学	6 産業	7 芸術美術	8 言語	9 文学
なし	535132	496611	543817	527956	531349	505422	515692	476404	501357	549068	574631
1:関係	222547	230908	211575	230718	205683	254125	245364	258353	241520	192547	205970
3:活動	143827	179394	142506	142541	176816	142857	131241	182914	154003	132298	119926
2:主体	52847	61907	64428	58785	55931	29703	14265	38236	71913	65839	52887
4:生産物	23012	20786	12285	23646	4962	9430	59914	32036	13569	34161	25927
5:自然	22635	10393	25389	16354	25259	58463	33524	12056	17639	26087	20659

表 6　ジャンル別語義分布のカイ二乗検定結果（BCCWJ-WLSP 部門別標準化残差）

	0総記	1哲学	2歴史	3社会科学	4自然科学	5技術工学	6産業	7芸術美術	8言語	9文学
-	-3.77	1.12	-1.04	-0.37	-2.84	-1.49	-6.66	-2.66	1.15	9.34
1:関係	0.98	-1.70	1.43	-1.98	3.62	2.10	4.87	1.79	-2.97	-4.70
3:活動	4.94	-0.24	-0.27	4.59	-0.13	-1.37	6.30	1.14	-1.35	-8.04
2:主体	1.98	3.33	1.93	0.67	-4.93	-6.60	-3.69	3.35	2.39	0.02
4:生産物	-0.72	-4.60	0.31	-5.88	-4.32	9.43	3.40	-2.48	3.06	2.29
5:自然	-4.01	1.19	-3.07	0.86	11.48	2.80	-4.02	-1.32	0.96	-1.57

表 5・6 から，特に 4 番台では「.5：自然」が突出し，5 番台で「.4：生産物」が高頻度と

いうジャンル特徴が明らかとなっている。そのほかのジャンル別の傾向として，6 番台と 4 番台で「.1：関係」が多く，6 番台では「.3：活動」も多いことや，「.2：主体」は 7 番台で高頻度となるなども確認される。1 番台でも「.2：主体」の出現率が高い。

次に，表 7 に中項目の上位頻度を示す。たとえば，表 5 で見た部門「.3：活動」中でも，「.30：心」が 3 番台，「.31：言語」は 8 番台，「.35：事業」は 5 番台に頻出しているというジャンル別の特徴が明らかとなっている。同様に，ジャンルに大差がないように見えた部門「.1：関係」であっても，「.19：量」は 7 番台，「.15：作用」は 5 番台，「.17：空間」が 4 番台と 5 番台に頻出するような分布が明確である。NDC の分類ごとに，語義分布が確認でき，ジャンル別に語義の分布をみることが可能といえよう。

但し，全体的な語義分布を見るにあたっては，特徴語となりやすい名詞の影響が強く考えられるという問題が残る。よって，ジャンル横断的な品詞についても，語義分布に特徴が見られるのかを確認しておきたい。

表 7　ジャンル別語義分布（BCCWJ-WLSP 中項目上位頻度・固定長・相対頻度）

中項目	相対頻度	0 総記	1 哲学	2 歴史	3 社会科学	4 自然科学	5 技術工学	6 産業	7 芸術美術	8 言語	9 文学
30:活動-心	44767	45188	55146	43536	63148	48091	33524	39614	46811	29814	41731
19:関係-量	43887	40217	28665	58122	43302	51862	37090	62005	75305	26708	35843
15:関係-作用	35116	36602	34671	30497	36085	40075	44936	34792	26459	22360	37909
12:関係-存在	34110	39313	38766	38232	24808	43847	28531	28591	31208	29814	32848
10:関係-事柄	31501	25305	36582	31602	29770	31117	29244	31002	29851	37888	31092
34:活動-行為	28640	35698	23478	31823	47361	31589	30670	37203	37313	21118	19626
16:関係-時間	26911	36150	30030	22983	31123	23102	20685	28591	33243	20497	25824
20:主体-人間	23735	15364	37128	22099	28868	16502	8559	14468	30529	27329	25101
31:活動-言語	23390	26209	25662	18343	29770	16502	14265	22046	18996	49068	22415
11:関係-類	21220	19883	20475	19227	12179	27346	36377	34792	25780	29814	15081
33:活動-生活	13581	14460	12831	14586	12630	6129	6419	16879	17639	11801	14771
13:関係-様相	13267	16268	13377	13039	14885	13201	21398	22735	10855	13043	8677
17:関係-空間	12858	16268	7098	12818	11276	17445	17118	6889	6106	8696	16527
37:活動-経済	9494	24401	7098	7514	4962	5186	4993	29625	3392	3727	6404
25:主体-公私	9431	13104	12012	15470	8570	3300	2140	1722	17639	16149	7334
38:活動-事業	8645	14912	2730	5083	4060	19331	37090	19290	2714	14286	2479
35:活動-交わり	7325	6326	4095	10387	902	13201	1427	13090	21031	1242	5578
56:自然-身体	6602	2711	8190	4199	2255	23102	713	2756	2714	1863	8780
24:主体-成員	6602	13104	7371	6409	4060	943	1427	4823	6106	9317	7644
26:主体-社会	5250	13104	4368	5746	4962	3772	713	5512	7463	1242	4855

３.４　NDC 別の動詞分布

前節では，ジャンル別に語義分布のあること，ジャンルによって語義分布に特徴のあることが確認できた。しかし，ジャンル横断的かつ多義的に用いられる語については，語義分布が語彙全般の語義分布とは異なる傾向が見られる（加藤ほか 近刊）。そのため，動詞

についてもジャンル別の特徴の有無を確かめておく。

まず，表 8 に動詞の上位頻度語を示す。「行く」や「行う」などジャンルによる頻度差が見られる語も若干あり，3・4・5・6 番台では「思う」が低頻度であるが，「考える」では 3・4 番台で高頻度，5・6 番台では低頻度となるという傾向が確認できる分布もある。しかし，「取る」のような分布が均一的で特徴が確認し難い語のあることがわかる。高頻度の動詞がジャンル横断的に出現する傾向が確認される。

表 8　ジャンル別動詞分布（PB 上位頻度・固定長・相対頻度）

動詞	相対頻度	0 総記	1 哲学	2 歴史	3 社会科学	4 自然科学	5 技術工学	6 産業	7 芸術美術	8 言語	9 文学
為る	23283	28321	22868	21819	27472	25414	27099	24600	20330	19942	16399
居る	9685	8995	10738	9742	9044	9024	7851	9054	9714	8435	11973
有る	8798	8380	11725	10193	10287	10289	8014	9261	7973	10483	6059
言う	6046	6614	8764	5719	5950	4929	4305	4912	6584	8358	6961
成る	4800	4823	5590	4390	5152	5154	4747	4918	4822	4999	4320
出来る	1855	2901	2013	1201	2157	2258	2200	1934	1661	1655	1190
来る	1805	1655	1892	1762	1460	1309	1261	1524	1833	1465	2874
因る	1688	1234	1577	1411	2578	2756	2105	2287	1099	1444	405
行く	1609	1539	2148	1505	1360	1188	1133	1363	1862	1248	2271
見る	1585	1638	1788	1574	1307	1467	1212	1457	1788	1928	2025
思う	1255	1432	1524	1143	917	868	786	985	1580	1290	2002
つく	849	689	702	689	1534	851	786	919	476	898	392
考える	839	677	1190	652	1009	1138	727	885	674	1129	649
持つ	829	747	1285	817	933	861	654	748	891	891	698
仕舞う	792	871	1050	662	573	704	679	712	1054	743	1055
行う	717	805	404	495	1148	1167	1029	903	409	365	134
於く	715	421	835	598	1359	762	785	964	323	912	105
取る	670	503	893	670	646	677	537	672	842	617	683
対する	642	553	723	516	1071	888	618	684	382	575	219

表 9 と表 10 に動詞の語義分布を中項目で示す。表 8 で分布の不明瞭な「為る」を含む「.34：行為」や，「思う」「考える」を含めた「.30：心」のような語義的な分布が確認できる。また，表 7 と対照することにより，動詞の語義に特徴的な分布が確認できる。たとえば，表 7 で 3 番台に突出する傾向が確認された「.34：行為」が，表 9 においては同様に 7 番台と 6 番台でも頻出しているとわかる。「.17：空間」が動詞の語義として用いられていたのは，1・6・7 番台のみであるが，語義分布としては 1・6・7 番台で特に頻度の低い傾向が見られる。NDC においても，動詞の語義分布は，全体的な語義分布とは異なる傾向が確認される。

しかし，ジャンル横断的な動詞の語義においても，多義語であれば語義別の分布を確認することで，ジャンル別の特徴が確認可能であるといえる。たとえば，「.30：心」という語義で見ると，5 番台では特に低く，4 番台と 6 番台で低い傾向と，1 番台と 9 番台で高い傾向が確認できる。「.12：存在」が 1 番台と 4 番台で高頻度であり，「.31：言語」が 8 番台で

突出するが 5 番台に低いというジャンル別の傾向も明らかとなる。動詞の語義の分布としては，ジャンル別の特徴があるといえよう。

表 9　ジャンル別動詞語義分布（BCCWJ-WLSP 中項目・固定長・相対頻度）

中項目	相対頻度	0 総記	1 哲学	2 歴史	3 社会科学	4 自然科学	5 技術工学	6 産業	7 芸術美術	8 言語	9 文学
12:関係-存在	26785	29824	33033	29392	20749	33475	19971	19290	25102	27329	25824
15:関係-作用	26596	29824	27846	22320	22553	25460	32810	15846	20353	19255	33054
34:活動-行為	24333	29824	21294	26740	35183	26874	29957	31347	32564	16770	17147
30:活動-心	14870	17171	16653	15691	14885	9901	6419	9990	13569	11801	17767
31:活動-言語	12418	9941	14196	10829	15336	9430	7846	11023	10176	20497	13118
37:活動-経済	4621	6778	6279	3094	3608	1886	2140	12745	2035	3106	3615
33:活動-生活	4590	2711	4368	5083	4060	3300	2140	4134	2035	4348	6198
11:関係-類	3521	3163	3276	3315	1353	3772	7846	6545	3392	6832	2169
38:活動-事業	2767	3163	2184	1768	1353	1414	14979	3445	0	7453	1653
35:活動-交わり	2295	904	1365	3536	902	3772	0	2067	2035	0	3202
36:活動-待遇	2232	904	5460	4199	3157	471	0	2411	1357	621	1240
57:自然-生命	1886	452	3822	2873	3157	3300	0	344	2714	0	1343
16:関係-時間	975	2711	273	221	902	943	1427	2067	1357	1242	723
13:関係-様相	597	0	1365	663	451	471	713	344	1357	0	516
17:関係-空間	377	0	273	0	0	0	0	344	678	0	930
51:自然-物質	377	0	273	0	902	0	3566	0	0	0	413
19:関係-量	220	0	273	221	0	0	0	344	678	1242	103
50:自然-自然	189	0	0	0	451	0	0	0	0	0	516
32:活動-芸術	189	0	273	884	0	0	0	0	0	0	103

表 10　ジャンル別動詞語義分布のカイ二乗検定結果（BCCWJ-WLSP 中項目別標準化残差）

	0総記	1哲学	2歴史	3社会科学	4自然科学	5技術工学	6産業	7芸術美術	8言語	9文学
12:関係-存在	0.48	1.54	1.19	-1.97	2.60	-1.79	-2.36	0.13	0.72	-0.74
15:関係-作用	0.55	-0.57	-2.23	-1.30	-0.01	1.64	-3.68	-1.16	-1.59	5.27
34:活動-行為	1.38	-2.38	1.14	3.83	1.23	1.53	3.48	2.96	-1.76	-6.00
30:活動-心	0.60	0.18	0.45	0.05	-1.84	-2.82	-2.03	-0.04	-0.74	3.00
31:活動-言語	-1.43	0.33	-1.15	1.39	-1.14	-1.65	-0.37	-0.48	3.61	0.80
37:活動-経済	1.34	1.12	-1.69	-0.72	-1.85	-1.42	7.30	-1.36	-0.76	-1.77
33:活動-生活	-1.53	-0.62	0.50	-0.37	-0.80	-1.41	-0.17	-1.34	0.06	2.86
11:関係-類	-0.46	-0.62	-0.29	-1.79	0.33	2.83	3.20	0.11	2.60	-2.71
38:活動-事業	0.22	-1.01	-1.42	-1.31	-1.16	8.99	0.94	-2.00	4.01	-2.52
35:活動-交わり	-1.53	-1.50	1.87	-1.42	1.61	-1.85	-0.12	-0.06	-1.91	2.26
36:活動-待遇	-1.48	3.97	3.01	0.98	-1.73	-1.82	0.38	-0.61	-1.32	-2.49
57:自然-生命	-1.70	2.51	1.63	1.46	1.68	-1.68	-1.93	0.93	-1.73	-1.48
16:関係-時間	2.57	-1.58	-1.77	-0.10	0.01	0.56	2.15	0.61	0.46	-0.95
13:関係-様相	-1.23	1.80	0.18	-0.29	-0.20	0.18	-0.52	1.36	-0.97	-0.39
17:関係-空間	-0.98	-0.45	-1.42	-0.95	-0.91	-0.74	-0.03	0.70	-0.77	3.37
51:自然-物質	-0.98	-0.45	-1.42	1.33	-0.91	6.30	-1.06	-0.73	-0.77	0.22
19:関係-量	-0.75	0.13	0.00	-0.72	-0.69	-0.57	0.54	1.31	2.99	-0.93
50:自然-自然	-0.69	-0.93	-1.00	0.94	-0.64	-0.53	-0.75	-0.52	-0.54	2.82
32:活動-芸術	-0.69	0.30	3.66	-0.67	-0.64	-0.53	-0.75	-0.52	-0.54	-0.73

４．まとめ

BCCWJ の増補された NDC 情報を用い，ジャンルによる語彙の分布傾向を確かめるとともに，BCCWJ-WLSP を重ね合わせることにより，ジャンルと語義の分布を調査した。上位頻度語の分布特徴を見るためには，語義の分布を見ることが有用と考えられる。また，ジャンル横断的な分布が見られる一般的な動詞についても，語義的な分布傾向を確かめることができた。内容語の分布傾向が明らかではない場合でも，語義の分布によって文章内容ごとの特徴が確認できると期待される。

今後はさらに，NDC の形式分類を活用した文体と語義についても調査を進めたい。

謝　辞

本研究は，国立国語研究所コーパス開発センター共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」，科研費基盤(C)「文体分析を目的としたコーパスの文書情報拡張及びその利用」による。

文　献

加藤祥，浅原正幸，山崎誠（2019a）「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』15(2)：134-141.

加藤祥, 浅原正幸, 山崎誠　（2019b）「『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』, 161-166.

加藤祥，森山奈々美，浅原正幸（2021）　「『現代日本語書き言葉均衡コーパス』書籍サンプルの NDC 情報増補―NDC 情報を用いた随筆の抽出と文体調査― 」『国立国語研究所論集』21：65-84.

加藤祥，森山奈々美，浅原正幸（近刊）「『現代日本語書き言葉均衡コーパス』新聞記事情報を用いたジャンル別語彙分布」言語資源活用ワークショップ 2021.

国立国語研究所（2004）『分類語彙表増補改訂版』大日本図書

国立国語研究所『現代日本語書き言葉均衡コーパス』version 1.1.

内田論，藤井聖子（2015）「クラスター分析とフレーム分析による語彙のジャンル別特徴：「現代日本語書き言葉均衡コーパス」を用いて」『言語文化論究』(34)：21-34.

関連 URL

コーパス検索アプリケーション『中納言』　https://chunagon.ninjal.ac.jp/