

BCCWJ 図書館サブコーパスの 文体情報検索ツールによるテキスト分析

柏野 和佳子* (国立国語研究所 言語資源研究系)
中村 壮範 (国立国語研究所 コーパス開発センター)

Text Analysis of the Library Subcorpus of the BCCWJ Using a Text Search Tool

Wakako Kashino (Dept. Corpus Studies, NINJAL)
Takenori Nakamura (Center for Corpus Development, NINJAL)

1. はじめに

国立国語研究所の共同研究プロジェクト「テキストの多様性を捉える分類指標の策定」(2009-2012年度)において、『現代日本語書き言葉均衡コーパス』(BCCWJ) の図書館サブコーパス(10,551サンプル)に収録されている全ての書籍サンプルを対象に、文体情報を付与した(柏野ほか2012a, 柏野ほか2012b)。文体情報とは、テキスト構造が単純(例: 章節構造)なもの(全体の約8割)に付与した「専門度、客観度、硬度、くだけ度、および語りかけ性度」といった評定値や、テキスト構造・紙面形式などの点で上記分類になじまないもの(約2割)にその特徴を表すものとして付与した「対談、Q&A形式、図解、用語解説」等の分類情報である。今回、我々は、それらの付与した文体情報をもとに、特定の文体的特徴をもつテキストを検索し、形態論情報等を一覧できる文体情報検索ツールを構築した。本ツールを用いて概観できる、図書館サブコーパスの文体的特徴について報告する。

2. 付与した文体情報

2.1 アノテーション作業の概要

文体情報のアノテーション作業の概要は次のとおりである。

- 対象: BCCWJに収録されている図書館サブコーパス(10,551サンプル)の書籍サンプル¹。
- 1サンプルの範囲と長さ: サンプル全体の印象判定を行うため、1サンプル全体を範囲とする。
サンプルの一部より一定量の字数を揃えて抽出するようなことはあえてしない。1サンプルの平均はおよそ3,000語。
- 作業者: 言語データ構築経験有のおおよそ20~50代の女性、延べ9名。
- 内容:
 - ① 文体判断が可能と判断されるもの、即ち、テキスト構造が単純(例: 章節構造)などを内容・表現の文体的特徴の印象判定により細分類する。(→2.2節)
 - ② ①以外、文体判断が単純にいかないと判断されるもの、即ち、テキスト構造・紙面形式に特徴をもつものを選別、分類する。(→2.3節)
- 態勢:
 - ① 典型例抽出を目的とする作業(Kashino and Okumura 2010)
 - ・ 判断のゆれを検証するために同一サンプルを作業者3人で判定。最終的には3人の結果を見直

* waka@nijal.ac.jp

¹ 1986年から2005年までの20年間に発行された書籍のうち、東京都内の13自治体以上の公共図書館で共通に所蔵されていた書籍を母集団とし、そこから抽出したサンプルから成るサブコーパスである。

し判定値を一つに決定。3,324サンプルまでこの態勢。（全体の31.5%）

②全体付与を目的とする作業（柏野ほか2012b）

- ・1サンプル作業者1人で判定。残りの①以外を作業し、全10,551サンプルを完了。

2.2 文体的特徴を表す分類指標

BCCWJ に収録されている書籍サンプルには、NDC（日本十進分類法）によるジャンルや、C コード（日本図書コード）による販売対象²、また、著者情報、形態論情報などに加え、テキストの内容・表現の文体的特徴を表す指標として、次のものを設定した。そのうえで、印象判定を行い、その結果を付与した。

- (a) 専門度 1 専門家向き、2 やや専門的な一般向き、3 一般向き、4 中高生向き、5 小学生・幼児向き
- (b) 客観度 1 とても客観的、2 どちらかといえば客観的、3 どちらかといえば主観的、4 とても主観的
- (c) 硬度 1 とても硬い、2 どちらかといえば硬い、3 どちらかといえば軟らかい、4 とても軟らかい
- (d) くだけ度 1 とてもくだけている、2 どちらかといえばくだけている、3 くだけていない
- (e) 語りかけ性度 1 とても語りかけ性がある、2 どちらかといえば語りかけ性がある、3 特に語りかけ性はない

2.3 文体判断が単純にいかないと判断されるもの分類指標

文体判断が単純にいかないと判断されるもの、即ち、テキスト構造・紙面形式に特定の特徴をもつものがある。そこで、次のような分類指標を設け、該当するサンプルの選別、分類を行った。理由は、それらは、必要に応じて、その特徴による類型で整理分類をし、別途、その文体を吟味すべきと考えたからである。なお、いずれの分類指標も該当サンプルすべてに付与したため、複数の分類指標が付与されたサンプルも存在する。

[テキスト構造・紙面形式上の特徴]

- (a) 対話系（対話、対談・座談、インタビュー、往復書簡、シナリオ、その他対話形式）
- (b) 引用系（Q&A 形式、投稿形式、その他引用編集形式）
- (c) 視覚表現多用系（コマ割多用、図解、その他写真やイラストの多用）
- (d) データベースやリスト系（用語解説、辞書形式、見本・カタログ形式、その他リスト形式）

[内容や表現上の特徴]

- (e) 前書きや後書きである
- (f) 明治時代より以前の古い言葉が多い
- (g) 外国語が多い
- (h) 数式やプログラミング言語などが多い
- (i) 法律文が多い
- (j) 教育現場で使いがたそうである³
- (k) その他一定量の「本文」が認めがたい

² C コードの第 1 桁は販売対象を示す。0:一般 1:教養 2:実用 3:専門 5:婦人 6:学参 I (小中) 7:学参 II (高校) 8:児童 9:雑誌扱い

³ 例えば、暴力的な描写や性的な描写を含むものを区別するための指標である。文体情報付与のための指標というよりは、コーパス活用のためのテキスト整理の指標として設けたものである。

2.4 アノテーション文体情報のリスト

アノテーション文体情報はリスト形式でまとめている。次の表1はじめの10件分を例として示す⁴。このリストを用いることにより、特定の文体的特徴をもつテキスト群の、該当するテキスト数や、コーパス収録時に付与されているサンプルID、書名、NDC（日本十進分類法）、Cコード（図書分類コード）といった書誌情報を得ることまでは簡単にできるようになっている。しかしながら、このようなリストだけでは、絞り込んでいったテキストの内容分析にすぐにとりかかることができない。テキストを分析したい場合は、ここで得られたサンプルIDをもとに、別途、BCCWJの収録DVDより該当テキストを抽出するという大きな手間が必要になる。そこで、アノテーション文体情報をもっと簡便に活用できるように、特定の文体的特徴をもつテキストを検索し、形態論情報等をすぐさま一覧できる文体情報検索ツールを構築することとした。

表1 アノテーション文体情報のリストの例

番号	書誌情報				文体的特徴					備考	非対象とする該当理由										その他、一定量の「本文」が認められたか	
	Sample ID	NDC	Cコード	タイトル	専門度	客観度	硬度	くだけ度	語りかけ度		小説の主人公あるいは語り手の人称	対話系か	引用系か	視覚表現多用系か	テータベースやスクリプト系か	前書きや後書きか	明治時代以前の古い言葉が多いか	外国語が多いか	数式やブログ用語などが多いか	法律文書などが多いか	教育現場で使ったそなであるか	
1 LBa2_00007	288	1323	光明皇后	3一般向き	2どちらかといえは 客観的	2どちらかといえは 硬い	3くだけていない	3特に語り性はない	判定不能	否	否	否	否	否	否	否	否	否	否	否	否	否
2 LBa2_00027	292		NHK大黄河	3一般向き	3どちらかといえは 主観的	2どちらかといえは 硬い	3くだけている	2どちらかといえは 語り性がある	[私+達]	否	否	否	否	否	否	否	否	否	否	否	否	否
3 LBa5_00005	596		日本の郷土料理	3一般向き	3どちらかといえは 主観的	2どちらかといえは 硬い	3くだけていない	3特に語り性はない	私	否	否	否	否	否	否	否	否	否	否	否	否	否
4 LBa7_00003	762		フルトヴェングラー	3一般向き	3どちらかといえは 主観的	3どちらかといえは 軟かい	3くだけている	3特に語り性はない	判定不能	否	否	否	否	否	否	否	否	否	否	否	否	否
5 LBa9_00002	933		呪われたブルー・エラー	3一般向き		3どちらかといえは 軟かい	3どちらかといえは 軟かい	3特に語り性はない	わたし	否	否	否	否	否	否	否	否	否	否	否	否	否
6 LBa9_00062	913	0193	汚名	3一般向き		2どちらかといえは 硬い	2どちらかといえは くだけている	3特に語り性はない	判定不能	否	否	否	否	否	否	否	否	否	否	否	否	否
7 LBa9_00068	910	3391	日本文芸史	2やや専門的な一般向き	3どちらかといえは 主観的	2どちらかといえは 硬い	3くだけていない	2どちらかといえは 語り性がある	わたくし	否	否	否	否	否	否	否	否	否	否	否	否	否
8 LBa9_00107	913		アローン・アゲイン	3一般向き		3どちらかといえは 軟かい	2どちらかといえは くだけている	3特に語り性はない	わたし	否	否	否	否	否	否	否	否	否	否	否	否	否
9 LBan_00008			時間と空間の冒険	5小学生・幼児向き		3どちらかといえは 軟かい	2どちらかといえは くだけている	3特に語り性はない	わし	否	否	否	否	否	否	否	否	否	否	否	否	否
10 LBan_00022		8093	ちょっと気になる駄菓子	5小学生・幼児向き		4とても軟らかい	1とてもくだけている	3特に語り性はない	おれ	否	否	否	否	否	否	否	否	否	否	否	否	否

3. Web アプリケーション「図書館コーパス検索ツール」の構築

アノテーション文体情報のより簡便な活用を目的に、特定の文体的特徴をもつテキストを検索し、形態論情報等を一覧できる文体情報検索ツールとして、Web アプリケーション「図書館コーパス検索ツール」を構築している。先行事例としては、1億語規模のBNCに対し、Lee(2001)が70種類（書き言葉46種、話し言葉24種）の言語使用域や、対象読者の性別や年齢層別、著者の属性等によって分類したものがいる。

3.1 「図書館コーパス検索ツール」の仕様

「図書館コーパス検索ツール」は、Internet Explorerなどのブラウザから利用することができるWeb アプリケーションである。インターネットが利用できる環境と標準的なブラウザがあれば、特別なソフトをインストールすることなく利用することができる。ただし、現在はまだ開発段階であるため、研究所内のネットワークからのみアクセス可能である。

BCCWJの図書館サブコーパスに付与された文体情報及び、書誌情報に基づく検索機能と、

⁴ 表の11列目の「備考：小説の主人公あるいは語り手の人称」とは、アノテーション作業時にわかる範囲で、小説の場合は主人公を、それ以外は筆者自身をどの人称で表現しているかの抽出を試みたものである。

文体情報と形態論情報(短単位)を利用して簡単な集計とグラフ表示の機能を備える。

システムの概要は以下のとおりである。

【ソフトウェア】

- ・開発ツール Microsoft Visual Studio 2012
- ・フレームワーク asp.net
- ・サーバ OS Microsoft Windows Server 2008R2 Standard
- ・Web サーバソフト Microsoft Internet Information Services ver7.5
- ・DBMS Microsoft SQLServer2012

【ハードウェア】

- ・サーバ本体 DELL PowerEdge R310
- ・CPU Intel Xeon X3430(2.40GHz, 8MB キャッシュ, 1333MHz)
- ・メモリ 24GB
- ・HDD 300GB 15000RPM(6Gbps SAS HDD/3.5 インチ)×3
※RAID5 構成で実質容量 600GB

【動作確認】

Google Chrome32, Internet Explorer10

3.2 「図書館コーパス検索ツール」の表示画面

次のように3つの表示画面から構成される。

(1)検索画面

文体情報及び、書誌情報に基づく検索が可能。

図1は、NDCの一桁目が「6」であり、文章の硬軟が「とても硬い」という指定を行っている画面である。左下の「検索」のBOXに、自動的に検索式を生成し、表示する。

(2)検索結果の表示画面

図2のように、該当サンプルのリストを表示する。該当件数が多い場合は、複数ページに分割表示する。

(3)データ表示画面

図1の左下にある、「データ表示」というリンクテキストをクリックすると、検索結果のテキストデータの各種情報を表示する。現在のところ、(a)品詞(大分類), (b)語種, (c)活用形, (d)平均文長, (e)文末の文体, (f)品詞, (g)語彙素, 以上のデータ表示が可能である。(a)～(e)までは、円グラフとリストを表示し、(f), (g)はリストのみ表示する。画面例は次の節に示す。

BCCWJ. 図書館コーパス検索ツール

① 以下から抽出条件を指定してください。
(CTRL+クリックで選択を解除できます)。

NDC
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
50
52
54
56
58
60
62
64
66
68
70
72
74
76
78
80
82
84
86
88
90
92
94
96
98
100
102
104
106
108
110
112
114
116
118
120
122
124
126
128
130
132
134
136
138
140
142
144
146
148
150
152
154
156
158
160
162
164
166
168
170
172
174
176
178
180
182
184
186
188
190
192
194
196
198
200
202
204
206
208
210
212
214
216
218
220
222
224
226
228
230
232
234
236
238
240
242
244
246
248
250
252
254
256
258
260
262
264
266
268
270
272
274
276
278
280
282
284
286
288
290
292
294
296
298
300
302
304
306
308
310
312
314
316
318
320
322
324
326
328
330
332
334
336
338
340
342
344
346
348
350
352
354
356
358
360
362
364
366
368
370
372
374
376
378
380
382
384
386
388
390
392
394
396
398
400
402
404
406
408
410
412
414
416
418
420
422
424
426
428
430
432
434
436
438
440
442
444
446
448
450
452
454
456
458
460
462
464
466
468
470
472
474
476
478
480
482
484
486
488
490
492
494
496
498
500
502
504
506
508
510
512
514
516
518
520
522
524
526
528
530
532
534
536
538
540
542
544
546
548
550
552
554
556
558
560
562
564
566
568
570
572
574
576
578
580
582
584
586
588
590
592
594
596
598
600
602
604
606
608
610
612
614
616
618
620
622
624
626
628
630
632
634
636
638
640
642
644
646
648
650
652
654
656
658
660
662
664
666
668
670
672
674
676
678
680
682
684
686
688
690
692
694
696
698
700
702
704
706
708
710
712
714
716
718
720
722
724
726
728
730
732
734
736
738
740
742
744
746
748
750
752
754
756
758
760
762
764
766
768
770
772
774
776
778
780
782
784
786
788
790
792
794
796
798
800
802
804
806
808
810
812
814
816
818
820
822
824
826
828
830
832
834
836
838
840
842
844
846
848
850
852
854
856
858
860
862
864
866
868
870
872
874
876
878
880
882
884
886
888
890
892
894
896
898
900
902
904
906
908
910
912
914
916
918
920
922
924
926
928
930
932
934
936
938
940
942
944
946
948
950
952
954
956
958
960
962
964
966
968
970
972
974
976
978
980
982
984
986
988
990
992
994
996
998
1000
1002
1004
1006
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2119

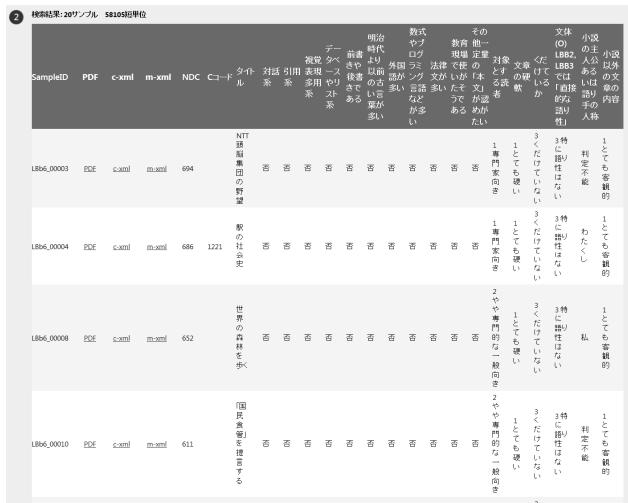


図2 検索結果の表示画面

3.3 テキスト分析例「とても硬い」と「とても軟らかい」の場合

最後に、「図書館コーパス検索ツール」を用いたテキスト分析例として、「とても硬い」と「とても軟らかい」のテキストを比較する。図は、「データ表示画面」で得られるものである（ただし、印刷の都合上、円グラフは別途作成したものと差し替えている）。

(1) 該当テキストのサンプル数と短単位数

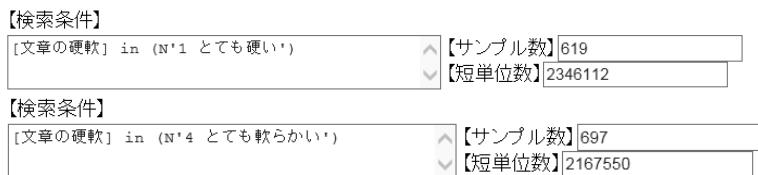


図3 「とても硬い」「とても軟らかい」のサンプル数と短単位数

(2) 品詞（大分類）、語種

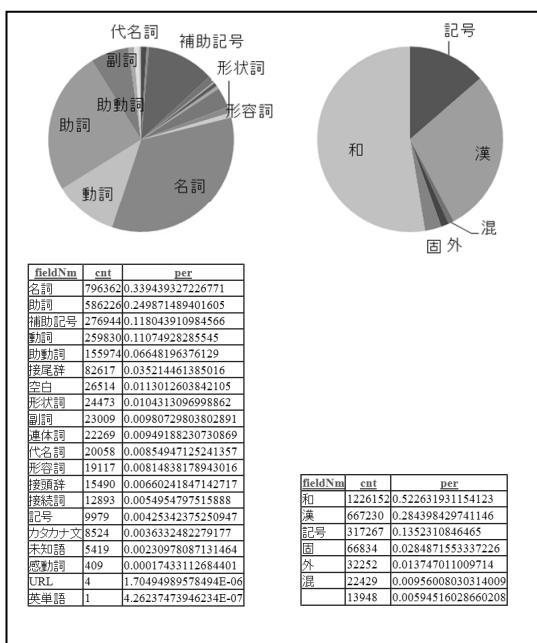


図4 「とても硬い」品詞（大分類）と語種

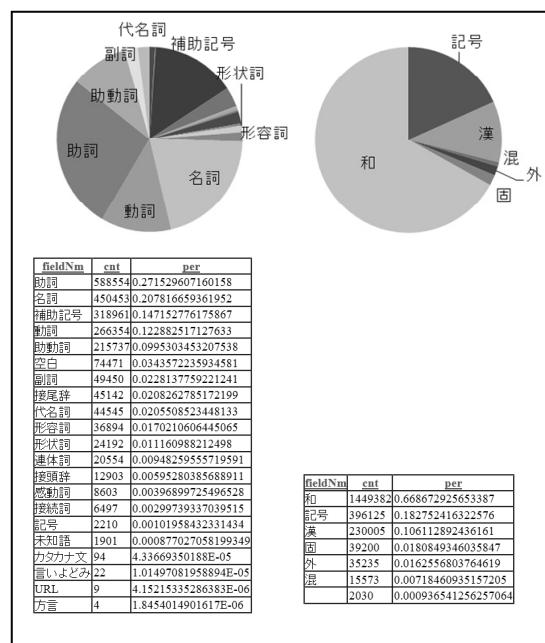


図5 「とても軟らかい」品詞（大分類）と語種

図4, 図5に示すように、このように、該当テキストに対して、BCCWJに付与されている形態論情報を利用し、品詞（大分類）や語種の円グラフと一覧を得ることができる。品詞（大分類）を比べると、「とても硬い」は名詞と助詞の二つが半分以上を占め、一方、「とても軟らかい」はその二つでは半分以下であり、代わりに、補助記号、助動詞の比率が高いことなどがわかる。語種を比べると、「とても硬い」での漢語比率、「とても軟らかい」での和語の比率の高さが顕著にみてとれる。

(3) 平均文長、文末の文体

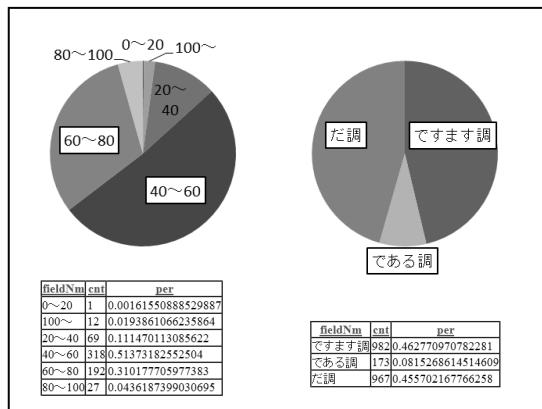


図6 「とても硬い」 平均文長と文末の文体

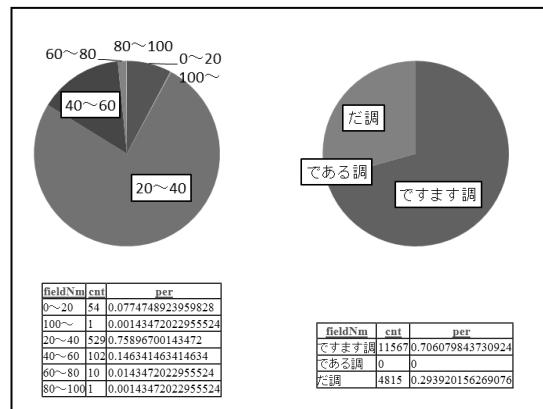


図7 「とても軟らかい」 平均文長と文末の文体

平均文長は、サンプル全体の文字数を、BCCWJに付与されている「sentence 開始タグ」の数で割って求めている。その結果、「とても硬い」の平均文長は長く、「とても軟らかい」の平均文長は短いという特徴がはっきりわかる。

BCCWJでは公開ファイル形式のひとつとしてXML形式がある。文体の文末は、会話部分に付与されているXML形式の「quote タグ」を参照し、その部分を除外し、それ以外の文末に含まれる「です/ます。」「である。」「だ。」の数を合計数で割って求めている。その結果、「とても硬い」は「だ調」と「である調」が多く、「とても軟らかい」は「ですます調」が多い、と、はっきりとした違いがみてとれる。しかしながら、文体の文末の計数機能は現在、改良中である。「です/ます。」「である。」「だ。」の終止形以外の形で終わる場合や、常体「だ・である体」の文末として、動詞、形容詞、名詞などが文末にくる場合も数え、個別の数を出し、さらに詳細に分析できるようにする予定である。

(4) 品詞と語彙素

表2 品詞一覧 (抜粋)

とても硬い				とても軟らかい			
fieldNm	cnt	per	fieldNm	cnt	per		
1 助詞-施動語	369312	16.8%	1 助詞-施動語	306259	14.2%		
2 名詞-普通名詞-一般	364656	16.6%	2 名詞-普通名詞-一般	266623	11.9%		
3 名詞-普通名詞-サ変可能	172170	7.3%	3 助動語	216737	10.0%		
4 助動語	166074	6.6%	4 助動語-一般	142013	6.6%		
5 疎開-主音立三點	146186	6.2%	5 喻動転写-終点	133973	6.2%		
6 疎開-終点	122386	5.2%	6 疎開-非主立三點	124541	5.7%		
7 疎開-一般	113644	4.8%	7 喻動転写-句頭	103699	4.6%		
8 名詞-数詞	91859	3.9%	8 助動語	99897	4.6%		
9 助動語-終動語	83496	3.6%	9 助動語-助動語	88138	4.1%		
10 助動語-終時動語	78106	3.3%	10 空白	74471	3.4%		
11 喻動転写-句尾	60561	2.6%	11 名詞-普通名詞-サ変可能	54406	2.5%		
12 演語辞-名詞的-一般	58893	2.4%	12 副詞	49450	2.3%		
13 名詞-普通名詞-副詞E點	47880	2.0%	13 名詞-普通名詞-副詞E點	45342	2.1%		
14 喻動転写-主語開	37632	1.6%	14 代名詞	44545	2.1%		
15 喻動転写-未語開	37335	1.6%	16 演語辞-名詞的-一般	36940	1.7%		
16 名詞-普通名詞-助動語E點	36906	1.6%	17 助詞-副動語	33326	1.5%		
17 空白	28614	1.1%	18 喻動転写-未語開	33070	1.5%		
18 副詞	23009	1.0%	19 助詞-終助詞	33066	1.5%		
19 演語辞	22269	0.9%	20 助動語-終助動語	31268	1.4%		
20 助動語-副動語	20963	0.9%	21 名詞-数詞	27867	1.3%		
21 代名詞	20058	0.9%	45 演語辞-形容開	920	0.0%		
45 助詞-終助詞	2965	0.1%					

表2は、「データ表示画面」から得られる品詞一覧をコピーし、編集しなおした表である。このように、ツールから得たデータを、別途編集し、利用することもできるようになっている。表2より、「とても硬い」では、「名詞-普通名詞-サ変可能」と「名詞-数詞」の多いことがわかる。また、「とても軟らかい」では、「副詞」「代名詞」「助詞-終助詞」の多いことがわかる。

「データ表示画面」には、頻度1に至るまでの全ての語彙素について、語彙素、語彙素読み、品詞、頻度、割合の一覧が示される。そこから、「名詞-普通名詞-サ変可能」と、「助詞-終助詞」を抽出したものが、次の表3である。両者に重なって出現する語には網掛けをして表示する。また、「助詞-終助詞」には頻度を添えて示す。表3をみると、「とても硬い」と「とても軟らかい」では重なりが少ないことがわかる。「とても硬い」の方に多く現れる語には、「研究、教育、存在・・・」といった学術的な語が多くあり、「とても軟らかい」の方に多く現れる語には、「料理、結婚、食事・・・」といった生活的な語が多くある。そして、終助詞の一覧をみると、「とても軟らかい」では多くの種類の終助詞が高頻度で出現していることがわかる。

表3 語彙素一覧より得た「名詞-普通名詞-サ変可能」と「助詞-終助詞」の語

「とても硬い」名詞-普通名詞-サ変可能（上位20語）：

物、関係、研究、意味、教育、存在、生活、労働、活動、生産、利用、戦争、運動、規定、著作、裁判、決定、影響、支配、計画

「とても軟らかい」名詞-普通名詞-サ変可能（上位20語）：

物、話、仕事、心、一緒、電話、生活、関係、意味、料理、結婚、勉強、食事、説明、涙、存在、びっくり、意識、約束、旅

「とても硬い」助詞-終助詞（頻度50以上）：

か2165、よ213、な150、ね103、わ63、の58、ぞ53

「とても軟らかい」助詞-終助詞（頻度50以上）：

か7391、よ7026、ね5442、な3730、の2109、わ1650、さ1018、ぞ608、い514、ぜ356、もの356、かしら323、や120、ねん119、け117、じゃん96、で71

4. おわりに

Web アプリケーション「図書館コーパス検索ツール」の構築について報告した。本ツールを用いることで、特定の文体的特徴をもつテキストを検索し、形態論情報等を簡単に一覧することができるようになった。その一例として、「とても硬い」「とても軟らかい」と判定されたテキストが、どのようなテキストの集合となるのか、ツールから得られる、品詞、語種、平均文長、文末文体、品詞一覧、語彙素一覧を用い、比較分析を行った。そして、それらツールから得られる情報によって、両者の文体差を支える言語的特徴の違いをはつきり確認できることを示した。本ツールを活用することで、さまざまに付与した文体的特徴を支える言語的特徴が何であるかの分析を進めていくことができる。

本ツールは、インターネットが利用できる環境と標準的なブラウザがあれば、特別なソフトをインストールすることなく利用することができる。現在は、研究所内のネットワークからのみアクセス可能であるが、広く利用してもらえる方法を検討していきたい。

謝 辞

本研究は、国立国語研究所の共同研究プロジェクト「テキストの多様性を捉える分類指標の策定」、「コーパスアノテーションの基礎研究」による成果を含みます。また、文部科学省科学研究費補助金基盤研究(C)「コーパス分析に基づく辞書の位相情報の精緻化」(課題番号:23520572) の助成を得たものです。

文 献

- Lee, Y. D. (2001) Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path Through the BNC Jungle, *Language Learning & Technology*, 5:3, pp.37-72.
- Wakako Kashino and Manabu Okumura(2010) An Approach toward Register Classification of Book Samples in the Balanced Corpus of Contemporary Written Japanese, *Proc. of PACLIC24*, pp.433-438.
- 柏野和佳子,立花幸子,保田祥,丸山岳彦,奥村学,佐藤理史,徳永健伸,大塚裕子,
佐渡島紗織(2012a)「テキストの硬さと軟らかさの考察－『現代日本語書き言葉均衡コーパス』の収録書籍を対象に－」『第1回コーパス日本語学ワークショップ』予稿集,pp.131-138.
- 柏野和佳子,立花幸子,保田祥,飯田龍,丸山岳彦,奥村学,佐藤理史,徳永健伸,
大塚裕子,佐渡島紗織,椿本弥生,沼田寛(2012b)「書籍テキストへの文体情報付与の試み」『第2回コーパス日本語学ワークショップ』予稿集,pp.155-164.

附録

文体的特徴を表す分類指標の付与サンプル例

(1)専門度：1 専門家向き (LBi4_00021 『がんと遺伝子』)

E 2 F 以外の R B 結合タンパク質としては、転写因子 R A X, T 細胞が活性化するときに誘導される I L - 2, G M - C S F, H I V - 2などの転写を活性化する転写因子 E 1 F - 1 や先に述べた細胞周期を制御するサイクリンDなどがある。おもしろいことに、E 1 F - 1 やサイクリンDの R B 結合ドメインには l a r g e T 抗原や E 1 A タンパク質と同じように L X C X E というアミノ酸配列が存在する。また、R B タンパク質は骨格筋分化を支配する重要な遺伝子群 M y o D ファミリー (M y o D, m y o g e n i n, M R F 4, m y f - 5) の産物とも複合体を形成し筋分化にも関与しているらしい。

(2)専門度：4 中高生向き (LBf9_00090 『超魔炎獄変』)

白く薄い空気のヴェールが、漂うように揺らめいている。
シャ…アン、シャラ…アン…。
闇を抜け、霧の中を渡る金属の響き。それは魔を覇する浄化の音。
響きに道を開けるかのようにすう…っと霧が左右に分かれた。
それは。
霧の中にたたずむそれは。
闇。
…いや。闇ではない。

(3)客觀度：1 とても客觀的 (LBo3_00158 『行政法要論』)

たとえば委員会の開催が「急施を要する場合」にあたるかどうかとか、公衆浴場の施設が「公衆衛生上不適切」かどうかは通常人の経験則によって十分判断できる事柄であるから、羈束裁量であって裁判所の終局的な判断に服すべきものとする。これに対し、外国人の在留期間の更新を適當と認めるに足る相当の理由があるかどうかは、出入国管理行政の責任者である法務大臣の政治的判断に委ねるべきであり、また、原子炉の安全性の認定は高度の科学的専門技術的知見に基づく総合的判断であるから、行政庁の便宜裁量事項であり、その当否は裁判所の審理・判断にはなじまないとする（最判昭和五三年一〇月四日民集三二巻七号一二二三頁、同平成四年一〇月二九日民集四六巻七号一一七四頁）。

(4)客観度: 4 とても主観的 (LBo3_00132『教師をめざす若者たち』)

どんなに上手な言葉を使っても、思っていないことを発すれば、子供に伝わらない。どんなに下手な言葉でも、心から伝えたいという愛情があれば、伝わるものであるということを信じることができました。この実感は日本でも通じる「教育の原則」であると思いました。

二日目、子供たちと綿花摘みと一緒にしました。敦煌の子供たちの手は「仕事をしている手」でした。

(5)硬度: 1 とても硬い (LBi3_00033『現代法社会学入門』)

取引費用がゼロである場合には、法的ルールの内容のいかんを問わず、資源配分は効率的レベルとなるというコースの定理は、法的ルールによる権利の分配のあり方のいかんを問わず、取引費用ゼロの社会では効率性が実現されることを意味する。したがって、法的ルールの選択、つまり権利の分配は、この意味のコース的世界においては、もっぱら所得分配、つまり分配的正義の観点から判断されることになる。もちろん、現実の社会では取引費用がゼロではない。

(6)硬度: 4 とても軟らかい (LBa4_00010『恐竜の世界をたずねて』)

恐竜が滅亡したわけや、恐竜たちのさいごのようすをしり、その原因をきわめるためには、恐竜の先祖のことをしらなくては、ほんとうのことがわかりません。

恐竜の先祖をしらべるには、ふるい時代につもった地層を、一枚一枚、したへしたへとしらべていかなければなりません。

このようにして恐竜の先祖をたずねていくと、中生代の三疊紀のはじめにいた、「テコドント」(図86)という、からだの長さが一メートルあまりの爬虫類にいきあたります。テコドントは、四本足であるき、走るときは二本足だったことがわかっています。恐竜の先祖は、このころから四本足または二本足の動物だったわけですね。

(7) くだけ度: 1 とてもくだけている (LBf9_00067『男はオイ！女はハイ…』)

最近流行りの通信販売。例の新聞の日曜版の裏面などに、克明にズラリと商品が写真などで広告してあるやつ。あれをば何となく眺めているうちに、どうしても欲しくなった商品があった。

よし、こいつひとつついでやれとばかりすぐ電話にとびついた。

「ハイ、こちら一です」と出たのは、耳ざわりだけでわかるアルバイトギャルの声。

「商品番号をおっしゃって下さい」

といわれて答える。

さらに「御住所と御名前、電話番号を郵便番号からどうぞ」ってんで、こいつにも律儀に返事をする。

(8) 語りかけ性度: 1 とても語りかけ性がある (LBt1_00013『5分間集中力トレーニング』)

精神的に疲れていると、「ああなったら、どうしよう」「こうなったら、どうしよう」と常に不安だらけになります。

動物病院にいらっしゃる飼い主さんには、過剰な不安を抱えている人や心配性の人がとても多い。実はそれがペットの病気をさらに悪化させることになっていますが、そういう認識をお持ちの飼い主さんは、あまりいません。詳しい説明は避けますが、不安や心配性を放っておくと、動物の具合が悪くなり、当然それが自分にも返ってきます。

それでは、どうすればいいのでしょうか。

文体判断が単純にいかない特徴をもつものの分類指標の付与サンプル例

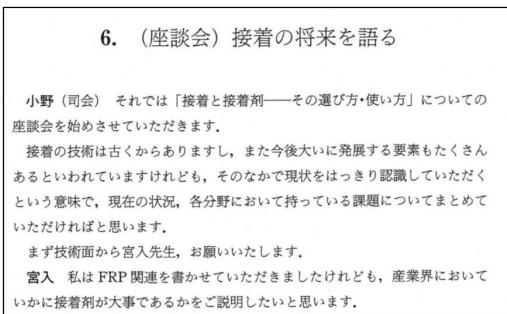


図1 座談：LBd5_00012『接着と接着剤』

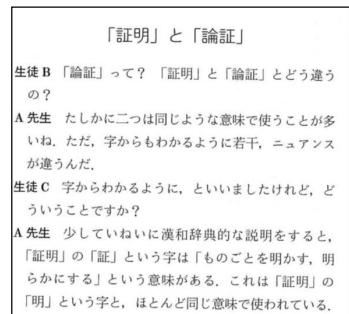


図2 対話：LBpn_00038『なぜ数学を学ぶのか』

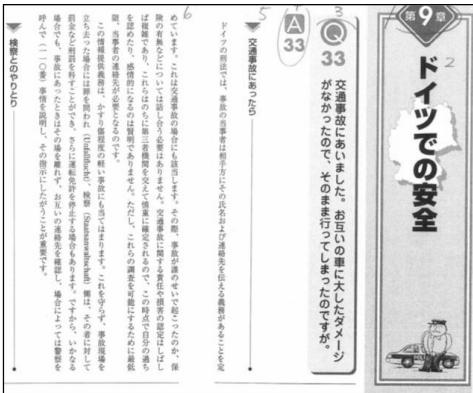


図3 Q&A：LBr3_00037『ドイツ暮らしの法律 Q&A』



図5 イラスト：LBb5_00011『絵とき インテリアライティングの技法早わかり』



図7 辞書：LBp6_00009『蕎麦屋のしきたり』

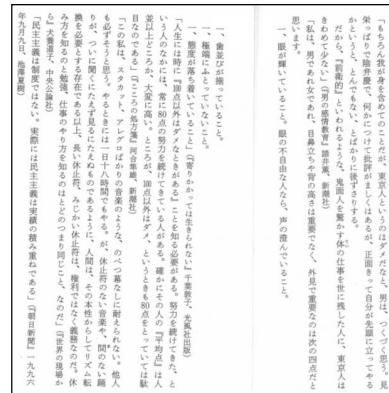


図4 引用：LB18_00017『試行錯誤の文章教室』



図6 コマ割り：LBsn_00022『東京で遊ぼ』



図8 カタログ：LBj6_00025『熱帯魚・水草カタログ』