『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

著者	加藤 祥，森山 奈々美，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	4
ページ	155-160
発行年	2019
URL	http://doi.org/10.15084/00002564

『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

加藤 祥（国立国語研究所コーパス開発センター）†
森山 奈々美（津田塾大学・国立国語研究所コーパス開発センター）
浅原 正幸（国立国語研究所コーパス開発センター）

Enlargement of NDC metadata on the Book samples in the Balanced Corpus of Contemporary Written Japanese

Sachi Kato (National Institute for Japanese Language and Linguistics)
Nanami Moriyama (Tsuda University / National Institute for Japanese Language and Linguistics)
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（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。また，開発当時 NDC 分類記号が付与されていなかったサンプル（「分類なし」）などの見直しもあわせて行った。作業は，国立国会図書館の NDC 情報を参照し，人手によって分類の確認と追加を進めた。本作業結果により，たとえば形式区分を利用し，ジャンルの分散する「随筆(-049)」「理論(-01)」「研究法(-07)」などのカテゴリで BCCWJ サンプルを分類することが可能となった。このほか，時代情報や小項目が追加されたサンプルもあり，今まで以上に詳細な分類が可能となった。本発表では，情報付与作業の方法と基礎情報を報告し，分類例を示す。本作業結果データは「中納言」の検索結果として利用可能となる。

１．はじめに

『現代日本語書き言葉均衡コーパス』（以降 BCCWJ）を検索する際，「中納言」ではサブコーパスを指定し，新聞，雑誌，書籍，Web ブログなどのテキスト属性の分類が可能である。さらに，書籍は日本十進分類法（NDC）分類記号による主題の分類や，図書分類コード（C コード）による販売対象と発行形態の分類が付与されているほか，図書館書籍には人手によって文体情報が付与されている（柏野，2015）。

しかし，ジャンル分類情報は主に媒体と内容に基づいているため，ジャンル横断的な基準による分析が困難であった。たとえば「随筆」の文体分析を行いたい場合，芸能人やアスリート，料理人などによる随筆は，その内容からそれぞれ芸術や産業などに分類されるため，文体分析対象の「随筆」は適切に収集し難かった（加藤ら，2018）。そこで，BCCWJ に付与された NDC 記号を拡張し，下位分類（「随筆」や「理論」などのジャンル）を用いた BCCWJ サンプルの分類を可能とした。さらに，NDC 分類が BCCWJ 構築当時に収集できておらず「分類なし」となっていた約 1000 サンプルについて，NDC 分類を確認し，増補を行った。本稿は，アノテーション方法と作業の結果，本作業による「中納言」データの更新について報告する。

† yasuda-s@ninjal.ac.jp

２．BCCWJ 書籍サンプルの NDC 情報増補作業

２.１　NDC 情報付与作業の概要

BCCWJ の書籍サンプルすべて（22,058 サンプル）を対象とした。よって，出版・書籍（PB：10,117 サンプル），図書館・書籍（LB：10,551 サンプル），特定目的・ベストセラー（OB：1,390 サンプル）の 3 種類の書籍サブコーパスに含まれるサンプルを扱う。

NDC 分類記号がないサンプルの場合は新たに番号を付与する。現在付与されている 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（小項目，補助表[1]の形式区分・地理区分など）が確認された場合は，該当する番号を追加する。

２.２　BCCWJ データバージョン 1.1 の NDC 情報

BCCWJ データバージョン 1.1（以下 BCCWJ-1.1）書籍サンプルの NDC 分類記号としては，(1)(2)(3)(4)のような 3 桁が付与されており，主題による分類が可能である。「少納言」や「中納言」による検索では，NDC の類目を用いたジャンル指定も可能である（図 1）。

(1)サンプル ID：LBl9_00056『伊達政宗』……913
9：文学（類目），91：日本文学（綱目），913：小説・物語（要目）

(2)サンプル ID：LBl2_00027『縄文人・弥生人 101 の謎』……210
2：歴史（類目），21：日本史（綱目）

(3)サンプル ID：LBq5_00062『いい音が聴きたい』……547
5：技術（類目），54：電気工学（綱目），547：通信工学・電気通信（要目）

(4)サンプル ID：LBg4_00014『お天気博士の季節へのラブレター』……451
4：自然科学（類目），45：地学（綱目），451：気象学（要目）

図 1　書籍レジスターを指定した際の「中納言」ジャンル指定画面例

２.３　本作業により増補される情報

NDC 分類記号は，現在までの BCCWJ に付与されている 3 桁に「.」以降の番号（下位区分）が追記されている場合（(1)’(2)’(3)’(4)’に例示する）がある。そこで，本作業は，3 桁の NDC 分類記号に加え，(1)’では文学共通区分で時代情報，(2)’では歴史の小項目というよう

[1] NDC 新訂 9 版では，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（2017 年以降）では言語共通区分・文学共通区分が固有補助表となったが，国立国会図書館サーチ API の付与済み NDC 情報に依拠する。

に，さらに詳細な分類を付与する。また，(3)’(4)’のように，形式区分（内容ではない「事典」「随筆」などの分類）を付与する場合もある。

(1)’ サンプル ID：LBl9_00056『伊達政宗』……913.6
913（日本文学小説）.6（文学共通区分（明治以降））

(2)’ サンプル ID：LBl2_00027『縄文人・弥生人 101 の謎』……210.025
210（日本史）.025（小項目（考古学））

(3)’サンプル ID：LBq5_00062『いい音が聴きたい』……547.033
547（通信工学・電気通信）.033（形式区分（事典））

(4)’サンプル ID：LBg4_00014『お天気博士の季節へのラブレター』……451.049
451（気象学）.049（形式区分（随筆））

２.４　アノテーション方法

これまでの BCCWJ の NDC 分類情報は，国立国会図書館が提供する NDC 上位 3 桁とほぼ[2]合致しているため，国立国会図書館の NDC 分類情報を参照した。現状 NDC 分類番号が確認できず，3 桁の NDC 番号が付与されていないサンプル（「分類なし」）についても，国立国会図書館データで該当書籍に NDC 情報が付与されている場合は，新規に番号を取得することとした。また，補助分類（(1)’(2)’(3)’(4)’に見られる「.」以降の番号）があれば追加を行う。

ISBN で書籍の同定が可能な場合は ISBN を確認したが，ISBN がデータ上付与されていない書籍も多い。よって，各サンプルの候補となる書籍情報を収集し，人手（作業者 4 名）により BCCWJ サンプルの書籍タイトル・著者・出版社・発行年を確認し，該当書籍情報を得た。データの確認には，国立国会図書館サーチ　API（http://iss.ndl.go.jp/information/api/）を用いた。

３．BCCWJ 書籍サンプルの NDC 情報増補結果

本作業により，BCCWJ-1.1 において「分類なし」として NDC を用いたジャンル分類の対象外となっていた書籍サンプルの半数以上に NDC 情報を付与することができた（表 1）。なお，「分類なし」のままであった書籍は，概ねムック本[3]などであり，国立国会図書館では雑誌扱いとされていた。

表 1　「分類なし」への NDC 情報付与（BCCWJ-1.1「分類なし」の 57.6%）

サブコーパス	新規追加	分類なし	BCCWJ-1.1「分類なし」数
LB	410（89.3%）	49（10.7%）	459（100.0%）
OB	24（80.0%）	6（20.0%）	30（100.0%）
PB	106（23.6%）	343（76.4%）	449（100.0%）
総計	540（57.6%）	398（32.4%）	938（100.0%）

[2] 作業中，国立国会図書館の NDC 情報が誤りである可能性が見られたため，他公的機関の多数において付与されていた NDC 情報を取得した例が 2 例ある。

[3] ISBN は有するが，雑誌コードも有している。このほか，一部の大型絵本やリーフレットは NDC 情報が付与されていなかった。

本作業で新規に追加された 540 件（BCCWJ-1.1「分類なし」からいずれかの NDC 分類に変更）のほか，BCCWJ-1.1 と国立国会図書館で付与されていた番号が異なった 27 件（これまでの NDC 分類から変更の生じるサンプルが 16 件含まれる）については，NDC 分類が更新されることになる。

また，8 割以上の書籍サンプルで，補助分類を追加できた（表 2）。

表 2　補助分類の追加（書籍サンプルの 84.1%）

サブコーパス	下位分類追加	追加なし	サンプル数
LB	8682（82.3%）	1869（17.3%）	10551（100.0%）
OB	1137（81.8%）	253（18.2%）	1390（100.0%）
PB	8728（86.3%）	1389（13.7%）	10117（100.0%）
総計	18547（84.1%）	3511（15.9%）	22058（100.0%）

４．　BCCWJ 書籍サンプルの NDC 補助区分を用いた分類

本作業で付与した補助区分によって，随筆や論文のようなジャンル横断的な分類のサンプルを，形式区分を用いて調査対象とすることが可能となる。また，「中納言」のジャンル分類（NDC の第 1 次区分：類目に該当する）ごとの補助区分を参照することで，詳細なデータ整理を行うことができる。以下では，形式区分を用いた BCCWJ 書籍サンプルに含まれる随筆と論文の分布を報告するほか，時代区分を用いた日本の小説の時代分布を例として示す。このほかのジャンルにおいても，ジャンルごとの固有補助表を参照し，細区分が活用可能である。

４.１　BCCWJ の随筆サンプル

NDC 分類「9X4」と一般補助表の形式区分「.049」が「随筆」に該当する。随筆にあたるサンプル数を以下の表 3 に示す。文学に分類された随筆が大半を占めるものの，その他のジャンルに分類されている随筆も 40 サンプル確認される。随筆サンプルは，文学ジャンルとその他のジャンルで語彙特徴が異なるため（加藤ら，前掲），今後，形式区分を利用した随筆分析が望まれる。

また，「社会科学」や「自然科学」などにも随筆が含まれるため，NDC のジャンル分類を利用して文体特徴などを分析する際には，特定の随筆サンプルを除外した分析も可能である。

表 3　BCCWJ の随筆サンプル（件数）

0 総記	1 哲学	2 歴史	3 社会科学	4 自然科学	5 技術	6 産業	7 芸術	8 言語	9 文学	計
1	2	10	11	4	6	1	5	1	457	498

４.２　BCCWJ の論文サンプル

NDC 分類「904」（論文集）および「908」（文学一般に関する研究・言語を特定できない作品集），一般補助表の形式区分「.04」（論文集）と「.08」（体系的な全集・論文集）は，

論文にあたると考えられる。サンプル数は表 4 に示す。なお，特に論文である可能性が高いのは「論文集（「.04」「904」）」に分類されるサンプルであろう。

表 4　BCCWJ の論文サンプル

	0 総記	1 哲学	2 歴史	3 社会科学	4 自然科学	5 技術	6 産業	7 芸術	8 言語	9 文学	計
論文集（「.04」「904」）	4	9	105	63	14	39	4	25	6	7	276
全集・論文集（「.08」「908」）	0	2	5	14	1	2	1	0	0	31	56
計	4	11	110	77	15	41	5	25	6	38	332

表 5　BCCWJ の日本の小説サンプル

NDC	分類	サンプル数
913	小説一般	190
913.2	上代	4
913.3	平安（物語文学一般）	1
913.36	平安：源氏物語	7
913.363	平安：和歌	1
913.369	平安：訳文	5
913.434	中世：平家物語	4
913.435	中世：太平記	1
913.436	中世：義経記	1
913.437	中世：曽我物語	1
913.47	中世：説話	1
913.51	近世：仮名草子	1
913.52	近世：浮世草子	3
913.56	近世：読本	1
913.57	近世：草双紙	1
913.6	近代（個人の作品集を含む）	3948
913.68	複数作家の作品集	130
913.7	講談・落語	12
913.8	童話	10
	総計	4322

４.３　BCCWJ の日本の小説サンプル内訳

NDC 分類「913」は，日本の小説[4]であるが，小説の時代区分を見ることで，時代と内容の判別が可能である。たとえば，近現代の小説に限定した分析が必要な場合などは，補助区分を「.6（近代：明治以後）」に限定することができる。表 5 に「913」の時代区分を用いた分類別サンプル数を示した。

５．　まとめ

本稿の作業により，BCCWJ の NDC 番号が増補された。増補した新たな NDC 番号は，次回の「中納言」データ更新時に公開（更新）予定である。本データに更新されることにより，これまで NDC 番号のなかったサンプルに NDC 番号が付与されたほか，補助区分を用いた詳細な書籍サンプルの分類が可能となる。随筆や論文のようなジャンル横断的な分類を利用した分析や，下位区分による詳細なデータ整理を行うことができる。今後は，本作業によって付与した NDC 番号の補助区分を用いた文体分析も進める予定である。

なお，番号に変更の生じる 567 サンプルを含め，全書籍サンプルの NDC 番号増補データは， https://github.com/masayu-a/BCCWJ-NDC/ にて配布する。特に，556 サンプルについては NDC 分類に変更が生じることになるため，集計時には注意が必要である。

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

日本図書館協会分類委員会(2018)『日本十進分類法新訂 10 版』日本図書館協会

加藤祥，櫻井芽衣子，森山奈々美，浅原正幸(2018)「『現代日本語書き言葉均衡コーパス』書籍サンプルに対する NDC 記号拡張アノテーションと NDC 形式区分を用いた「随筆」の文体分析」『言語資源活用ワークショップ発表論文集 2018』, pp. 372 -381.

国立国語研究所(2015) 『BCCWJ 図書館サブコーパスの文体情報』（第 1 版）

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Makiro Tanaka, and Yasuharu Den (2014). “Balanced Corpus of Contemporary Written Japanese”, Language Resources and Evaluation, 48, pp.345-371.

柏野和佳子(2013)「書籍サンプルの文体を分類する」『国語研プロジェクトレビュー』4(1), pp.43-53.

日本図書館協会分類委員会(1995)『日本十進分類法新訂 9 版』日本図書館協会

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/
国立国会図書館サーチ	https://iss.ndl.go.jp/

[4] 「9X3」は小説を示し，2 桁目「1」は日本の言語区分（一般補助表）を示す。