Enlargement of Nippon Decimal Classification Metadata of Book Samples in the “Balanced Corpus of Contemporary Written Japanese”: Extraction of Essays from Book Samples According to NDC Metadata and Writing Style Analysis

国立国語研究所論集 (NINJAL Research Papers) 21: 65–84 (2021)
ISSN 2186-1358

『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——

加藤　祥[a]　　森山奈々美[b]　　浅原正幸[c]

[a] 目白大学／国立国語研究所 共同研究員
[b] 国立国語研究所 コーパス開発センター 技術補佐員［–2020.03］
[c] 国立国語研究所 コーパス開発センター

要旨

本研究では『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。作業は，国立国会図書館サーチの NDC 情報を参照し，人手によって分類の確認と追加を行った。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。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)」「教科書 (-078)」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。本研究では，情報付与作業の方法と基礎情報を報告し，分類例を示す。本データを用いた研究事例として，NDC 情報を用いた随筆の抽出と随筆の文体調査結果を報告する。本データは「中納言」の検索で利用できる *。

キーワード：『現代日本語書き言葉均衡コーパス』，日本十進分類法，文体，対数尤度比

1. はじめに

『現代日本語書き言葉均衡コーパス』（以降 BCCWJ）（Maekawa et al. 2014）を検索する際，「中納言」（https://chunagon.ninjal.ac.jp/）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をすることが可能である。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 2013，国立国語研究所 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，いわゆる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能や産業などに分類され，適切に収集し

* この論文は，2018 年 9 月 5 日（水）に開催された「言語資源活用ワークショップ 2018」（於：国立国語研究所）にて行った発表「『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」と，2019 年 9 月 2 日（月）に開催された「言語資源活用ワークショップ 2019」（於：国立国語研究所）にて行った発表「『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」をもとにしているが，データの追加及び大幅な改訂を行っている。また，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」（プロジェクトリーダー：浅原正幸）と，科研費基盤研究（C）（課題番号 18K00634）「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」（代表：加藤祥）の研究成果である。

難かった。そこで，BCCWJ に付与された NDC 記号を拡張し，下位分類を用いて BCCWJ サンプルを「随筆」や「理論」などのジャンルで分類することを可能とした。あわせて，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。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。また，本作業で付与した情報を用いて「随筆」サンプルを抽出し，文体調査を試みた。

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（22,058 サンプル）を対象として，NDC 情報の増補作業を行った。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すべてのサンプルを扱う。

NDC 分類記号がなかったサンプルの場合は，新たに番号を付与する。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する。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た。本データの下位区分の扱いは，確認時（2019 年 5 月）の国立国会図書館サーチ API（以下 NDL サーチ https://iss.ndl.go.jp/）の付与済み NDC 情報に依拠する。

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）の書籍サンプルは，NDC 分類記号として（1）（2）（3）（4）のような 3 桁の番号が付与されており，ジャンルによる分類が可能である。「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）。なお，2021 年 3 月の「中納言」の更新より，旧 NDC 情報（BCCWJ-1.1）と新 NDC 情報（本データ）のいずれの情報によってもジャンル指定が可能になる。旧 NDC 情報は，BCCWJ 構築時のサンプリング[1]時に用いられたものであり，サンプリング比が重要である場合はこの情報を用いるべきである。一方，ジャンルごとの比較を行う場合には，新 NDC 情報を用いることにより，より詳細な分析が可能となる。

（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）

[1] サンプリングについては，https://ccd.ninjal.ac.jp/bccwj/sampling.html を参照。

(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）

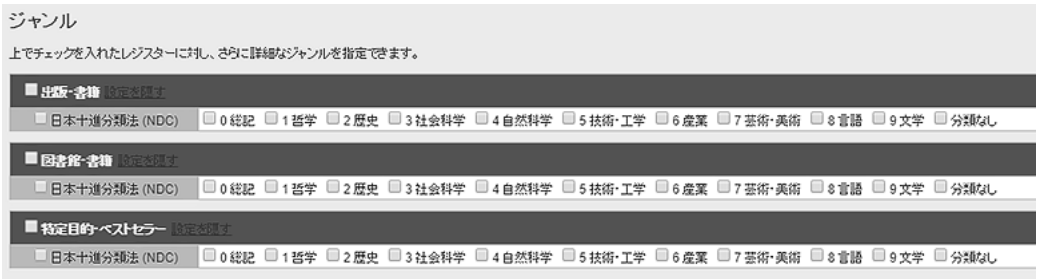

図 1　書籍サンプルを指定した際の「中納言」ジャンル指定画面例

2.3 本作業により増補される情報

NDC 分類記号は，当初より BCCWJ に付与されていた 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

BCCWJ-1.1 の NDC 分類情報は，NDL サーチが提供する NDC 上位 3 桁とほぼ合致しているため, NDL サーチの NDC 分類情報を参照した。BCCWJ において 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かったサンプル（「分類なし」）についても，NDL サーチで該当書籍に 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。なお，作業中，NDL サーチの 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。

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，ISBN で書籍の同定ができなかった場合には，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。

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 938 件の半数以上に対し, NDC 情報を付与することができた（表 1）。なお，「分類なし」のままとなった書籍は，概ねムック本などであった。すなわち，NDL サーチでは雑誌扱いとされていたため，NDC 情報の付与がなかったものである。

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（42.4%）	938（100.0%）

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 に付与されていた番号（J-BISC）と現在の NDL サーチが異なっていた 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 分類が更新された。また，84.1% 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	8,682（82.3%）	1,869（17.7%）	10,551（100.0%）
OB	1,137（81.8%）	253（18.2%）	1,390（100.0%）
PB	8,728（86.3%）	1,389（13.7%）	10,117（100.0%）
総計	18,547（84.1%）	3,511（15.9%）	22,058（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。以下ではまず，形式区分を用いた BCCWJ 書籍サンプルに含まれる随筆の分布を報告する（4.1 節）。なお,「随筆」については, 高崎ら（2007）が『文藝春秋』から収集した「随筆」500 篇を用いた文体調査を行っており，同著者らをはじめとする「随筆」を調査対象とした研究がある（立川 2014，高崎 2012）。高崎ら（2007）の一連の研究は，雑誌の「随筆」特集などから独自に収集したデータを用いた分析である。そこで，本稿では，BCCWJ 書籍サンプルに含まれる「随筆」の分析を試みることで,「随筆」の特徴語と文体特徴を調査した例を示す（5 節）。

また，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，細区分が活用できる。均衡性を有した書き言葉コーパスにおいて，また出版や図書館などのサブコーパスにおいて分類されているサンプルの抽出を可能とすることにより，ジャンルとの関連や他形式文章との文体対照が可能である。BCCWJ を用いて特定表現を収集するに際しても，当該サンプルの形式区分等を用いた文体分析への活用を図ることが可能となる。以下で，日本の小説の時代区分を用いた分布（4.2 節）と形式区分を用いた分布（4.3 節）を例示する。

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。

表 3　BCCWJ の随筆サンプル（類目別件数）

サブコーパス	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
LB	3	2	9	45	16	5	1	17	6	301	405
OB	0	6	0	0	2	2	2	2	2	60	76
PB	3	6	4	43	22	16	2	8	5	114	223
総計	6	14	13	88	40	23	5	27	13	475	704

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 229 サンプル確認される。なお，本分析においては，「-049」（40 サンプル確認）のみならず，「-049 「雑著」」以外で，途中に「-049」が補助区分として埋め込まれているものすべてを随筆サンプルとした。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 2018），形式区分を利用した分析を行う（5 節）。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，反対に随

筆サンプルを除外した傾向分析も可能である。

4.2 BCCWJ の日本の小説サンプル内訳

NDC 分類「913」は，日本の小説であるが，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 4 に「913」の時代区分を用いた分類別サンプル数を示した。

表 4　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）	3,948
913.68	近代：複数作家の作品集	130
913.7	講談・落語	12
913.8	童話	10
	総計	4,322

4.3 BCCWJ の書籍サンプル内の教科書の分布

NDC 分類において，形式区分「-078」はその分野の教科書を表す。BCCWJ において，小中高等学校の教科書は「教科書サンプル」（OT）に収録されているが，一般向けの教科書がどの程度収録されているかは不明であった。表 5 に形式区分「-078」を用い要目別サンプル数を LB サンプルと PB サンプルごとに計数し，示す。なお OB サンプル内には形式区分「-078」に適合するサンプルは含まれなかった。

社会科学・自然科学においては，社会学・動物学・医学のサンプルが確認できた。芸術においては，主に器楽合奏のサンプルが確認でき，LB サンプルよりも PB サンプルのほうが多いこと

も確認できた。言語・文学の教科書においては，韓国語・フィリピン語・ペルシャ語・オランダ語の教科書（会話も含む）が言語の分類で確認された一方，中国語・英語・ドイツ語・フランス語などは文学の分類の教科書として確認された。

表 5　BCCWJ の書籍サンプル内の教科書の分布

NDC（要目）	要目名	LB	PB	総計
361	社会科学・社会学	9	4	13
481	自然科学・動物学・一般動物学	3	2	5
493	自然科学・医学・内科学	1	0	1
494	自然科学・医学・外科学	1	5	6
764	芸術・音楽・器楽合奏	5	9	14
829	言語・その他の東洋の諸言語	2	3	5
849	言語・その他のゲルマン諸語	0	1	1
912	文学・日本文学・戯曲	0	1	1
920	文学・中国文学	0	1	1
923	文学・中国文学・小説，物語	1	0	1
930	文学・英米文学	4	5	9
933	文学・英米文学・小説，物語	5	7	12
934	文学・英米文学・評論，随筆	1	0	1
940	文学・ドイツ文学	1	1	2
943	文学・ドイツ文学・小説，物語	0	1	1
948	文学・ドイツ文学・作品集	0	1	1
950	文学・フランス文学	0	1	1
	合計	33	42	75

5. BCCWJ の随筆の検討分析

「随筆」は，「筆のおもむくままに筆者の考えを端的に表現した短編」とされ，「自己の見聞した事物のありようを描く作品から，そういった経験に対する見解や批評を述べる作品まで内容や特徴には幅がある」ため，「様々なタイプの文章を内包するがゆえに，構造的に特定の形式が認められていない」ことが指摘され，日本語の文章に特有のジャンルであるといわれる（高崎ら 2007）。そのため，話し言葉とも書き言葉とも，語彙や表現などにおける対照が行われやすい種類の文章であると考えられる。本稿の NDC 増補により，BCCWJ に含まれる「随筆」と分類された文章が抽出可能になり，様々なコーパスとの対照分析が可能となった。

高崎ら（2007）は，雑誌「文芸春秋」の巻頭随筆 500 篇を用い，文章特性を調査している。ここでは，特性として指摘されている表現の検証を行うとともに，BCCWJ 書籍サンプルに含まれる随筆（704 編）に見られる特性を確かめる。

5.1 随筆の特徴語彙の検討

本節では, BCCWJ の「随筆」に特有な語彙を調査する。具体的には, 随筆とそれ以外の 2 群（A 群:随筆と B 群:それ以外）に分け，それぞれの群における語彙素の頻度をもとに，どちらに偏っているかを対数尤度比（log-likelihood ratio，以下 LLR と呼ぶ）により数値化し，調査を行う。LLR は，コーパス言語学で特徴語彙を取り出すために用いられる指標で，次式によって定義する：

$$\begin{aligned}\mathrm{LLR}(w) = 2\,(&a\log_e a + b\log_e b + c\log_e c + d\log_e d\\ &-(a+b)\log_e(a+b) - (a+c)\log_e(a+c)\\ &-(b+d)\log_e(b+d)\\ &-(c+d)\log_e(c+d)\\ &+(a+b+c+d)\log_e(a+b+c+d))\end{aligned}$$

ここで a：A 群に出現する語彙素 w の出現頻度，b：B 群に出現する語彙素 w の出現頻度，c：A 群の延べ語数 – a，d：B 群の延べ語数 – b とする。LLR (w) 自体は偏りしか評価しないために，どちらの群に偏っているかを示さない。この問題を扱うために，w の A 群における使用率（a / a+c）が，B 群における使用率（b / b+d）よりも小さい時に -1 を乗ずる。これを修正 LLR と呼ぶ。

5.1.1 随筆の特徴語彙

本稿で行った BCCWJ への NDC 増補によって，これまでも取得が可能であった文学の類目にあたる「9X4」分類の随筆に加え，文学以外に含まれる「-049」分類の随筆が取得可能となった。

A 群を書籍サンプル内の随筆（9X4 および -049）とし，B 群を随筆以外とした場合の修正 LLR 上位語（記号や固有名詞を除く）を表 6 に示す。随筆（A 群）の特徴的な語彙として，一人称代名詞の「私」をはじめ,「ね」「か」のような読み手に語りかける終助詞や敬体の「です」, 接続詞として「けれど」のように話し言葉的なくだけた語などが得られている。表外でも,「僕（書籍全体修正 LLR 538.6 以下同様）」「自分（661.7）」などの一人称に関する語,「思う（872.0）」「好き（359.8）」「面白い（321.0）」のような判断や評価に関する語が上位語として散見される。これらの語彙は，随筆の文体的な特徴として，主観性や語りかけ性，硬度やくだけ度などと関わる可能性が考えられる。また,「って」「と」「言う」のように引用に関わる語も見つかっている。随筆における引用は，客観的な論拠というよりも，一般論や同意を求めるための表現と考えられるため，専門性などとの関わりが考えられる。高崎ら（2007）においても，文末形式として「という」が上位頻度であることが指摘され，体験や思考など,「何らかの意味で異質の次元が持ち込まれている」とされている。なお，分析の手法上，表外上位に固有名詞が散見される。

表 6　書籍サンプルの随筆データの特徴語彙（文学（9X4）と文学以外（-049）をあわせて検討）

順位	全体	修正 LLR	LB	修正 LLR	PB	修正 LLR	OB	修正 LLR
1	私	3466.8	私	1948.3	フォント	1487.6	です	772.9
2	言う	2491.5	ね	1281.5	私	973.6	ます	763.1
3	です	2200.7	言う	1180.1	言う	551.8	裏技	581.3
4	だ	1686.1	です	1023.5	喜	524.2	夢	549.6
5	ね	1432.7	けれど	941.0	だ	510.0	言う	340.6
6	書く	1428.4	だ	878.3	書く	471.6	と	266.6
7	も	1420.3	書く	796.9	積層	435.0	エクスタシー	247.5
8	小説	1354.2	小説	786.3	点検	432.8	患者	208.2
9	と	1143.0	笑い	786.0	小説	377.0	歌	199.3
10	か	1141.1	って	762.1	も	353.7	浮気	196.2

5.1.2 随筆の特徴語彙：文学と文学以外の対照

これまでも取得が可能であった文学の類目にあたる「9X4」分類の随筆と，本稿で行った BCCWJ への NDC 増補によって取得可能となった「-049」分類の随筆の異同を確かめておく。BCCWJ に含まれる随筆は，文学に分類される随筆が三分の二にあたる（表 3）が，文学以外のジャンルに横断的に含まれる随筆とは異なる傾向が見られる可能性がある。表 7 に，文学（9X4）と文学以外（-049）の分類について特徴語彙を示す。

表 7　書籍サンプル随筆データの特徴語彙（文学（9X4）と文学以外（-049）を分けて検討）

順位	9X4（文学）	修正 LLR	-049（文学以外）	修正 LLR
1	た	919.9	ます	-977.8
2	私	755.0	為る	-463.6
3	小説	380.9	上司	-404.6
4	だ	330.7	相続	-338.6
5	男	316.9	会社	-316.3
6	書く	280.7	裁判	-316.1
7	女	270.8	ストレス	-315.2
8	言う	263.7	条	-295.2
9	文学	241.4	コミュニケーション	-289.3
10	彼	212.6	点検	-264.1

随筆サンプルの中でも，文学類目に分類される「9X4」と様々な類目に分散する「-049」とでは，それぞれ特徴語に違いが見られる。「9X4（文学）」では，随筆一般に特に多く見られた「私」のほか，「男」「女」「彼」のような名詞，「小説」「書く」「文学」のような文学類目ゆえと考えられる語が特徴的に現れている。これに対し，「-049（文学以外）」では，「ます」が特徴的であり，そのほかには表に見る「相続」「ストレス」「上司」「コミュニケーション」「点検」など，様々な

ジャンルの語彙であると推測される内容語が見られる。すなわち，「9X4」分類においては敬体が特徴とはいえないが，様々なジャンルの「随筆」では，ジャンルに関わる内容語のほか，特に敬体が特徴だといえる。「9X4（文学）」分類の分析では，文学類目としての偏りや文学類目の特徴語彙が取得されるが，「-049（文学以外）」として各類目に分散していた「随筆」テキストを加えることにより，敬体のような特徴語が取得できると考えられる。次節でも，「随筆」としての総計に加え，「9X4（文学）」と「-049（文学以外）」との異同についても確かめておくこととする。

5.1.3 随筆の機能語の特徴語彙（品詞別）

表 7 に見た特徴語には内容語が目立ったが，敬体のほかにもジャンル内容に関わらない特徴語の存在が考えられる。内容に関わらない特徴語を調査するため，機能語の特徴語について確認したい。また，話し言葉的か書き言葉的かという特徴について確認するにあたっては，「中納言」の中のコーパスを 1 度にまとめて検索できる「まとめて検索 KOTONOHA」（https://chunagon.ninjal.ac.jp/integrated/）を用い，書き言葉として『現代日本語書き言葉均衡コーパス』『国語研日本語ウェブコーパス』，話し言葉として『日本語話し言葉コーパス』『日本語日常会話コーパス』『名大会話コーパス』『現日研・職場談話コーパス』の検索結果について，調整頻度 per million word（pmw）を示す。

以下の表 8 は，随筆一般に確認された特徴語から，助動詞のみの修正 LLR 上位語を示したものである。随筆では，敬体の「です」「ます」のほか，特に，「てる（話し言葉:6702，書き言葉:2591）」「つう（話し言葉：639，書き言葉：160）」などの話し言葉的な表現が特徴語となっていることがわかる。このような話し言葉的な表現は，「くだけている」という印象に関わる可能性が考えられる。また，「たい」による主観性との関わりや，「らしい」伝聞などの影響もあり得るだろう。

表 8　書籍サンプル随筆データの特徴語彙：助動詞（修正 LLR 上位 10 位）

順位	語彙素	随筆頻度	（内）文学 9X4	（内）文学以外 -049	随筆以外頻度	修正 LLR
1	です	12,171	7,844	4,327	275,293	2,200.7
2	だ	65,631	48,116	17,515	2,003,681	1,686.1
3	てる	1,951	1,447	504	46,283	287.4
4	し	1,275	939	336	29,513	208.9
5	ず	5,151	3,613	1,538	154,083	161.5
6	たい	2,020	1,395	625	54,233	153.2
7	らしい	724	574	150	17,581	95.7
8	な	1,644	1,192	452	46,960	79.1
9	ます	11,041	6,193	4,848	367,855	62.7
10	つう	322	258	64	7,470	52.2

表 9 には，助詞の上位語を示す。随筆の特徴語として，副助詞（表中太字部）が特に目立つ。

助動詞同様，「って（話し言葉：8054，書き言葉：1850）」のように話し言葉的な表現が取得されている。また，「たり」「くらい」「ほど」のように婉曲表現が見られるほか，「なんて」「だけ」「ばかり」「しか」のような限定表現が見られ，書き手の判断が特徴的に表現されることが，主観的な印象に関わる可能性がある。このほか，終助詞の「ね」「よ」などは，読み手に語りかける印象との関係が考えられよう。なお，副助詞が特徴語に多く見られることから，表 10 には副詞を示す。

表 9　書籍サンプル随筆データの特徴語彙：助詞（修正 LLR 上位 16 位）

順位	語彙素	随筆頻度	（内）文学 9X4	（内）文学以外 -049	随筆以外頻度	修正 LLR	品詞
1	ね	3,968	2950	1,018	72,335	1,432.7	終助詞
2	も	26,117	18,954	7,163	735,753	1,420.3	係助詞
3	と	50,001	36,155	13,846	1,540,328	1,143.0	格助詞
4	か	13,669	9,832	3,837	361,726	1,141.1	終助詞
5	って	2,155	1,659	496	40,858	699.0	**副助詞**
6	たり	2,278	1,609	669	51,856	400.0	**副助詞**
7	の	109,808	78,954	30,854	3,755,894	287.7	格助詞
8	くらい	1,005	781	224	21,455	224.6	**副助詞**
9	ば	4,709	3,188	1,521	142,568	129.5	接続助詞
10	なんて	582	455	127	12,572	124.7	**副助詞**
11	よ	2,948	2,177	771	85,633	122.7	終助詞
12	だけ	2,549	1,823	726	77,235	69.4	**副助詞**
13	ほど	1,343	1,000	343	38,069	68.8	**副助詞**
14	こそ	502	365	137	12,264	64.5	係助詞
15	ばかり	696	542	154	18,202	62.1	**副助詞**
16	しか	669	484	185	18,057	49.0	**副助詞**

文体的な特徴が現れやすいとされる副詞であるが，表 10 から，「矢張り（話し言葉：1592，書き言葉：333）」「まあ（話し言葉：588，書き言葉：221）」「兎に角（話し言葉：143，書き言葉：68）」「随分（話し言葉：89，書き言葉：34）」「結構（話し言葉：563，書き言葉：111）」などの話し言葉的な表現が取得される傾向にあるとわかる。頻度は低いが，「ギシ」のようなオノマトペも見られている。また，「そう」「こう」など指示的な副詞が特徴語となっていることが特筆される。表 11 には，指示詞を示す。

表 10　書籍サンプル随筆データの特徴語彙：副詞（修正 LLR 上位 10 位）

順位	語彙素	随筆頻度	（内）文学 9X4	（内）文学以外 -049	随筆以外頻度	修正 LLR
1	そう	4,356	3,345	1,011	108,822	501.1
2	ギシ	68	1	67	9	436.0
3	矢張り	715	588	127	13,046	257.3
4	こう	1,302	955	347	32,518	149.8
5	まあ	374	299	75	6,707	140.7
6	どう	2,195	1,553	642	62,559	107.6
7	兎に角	261	209	52	4,725	95.8
8	随分	192	156	36	3,168	88.3
9	結構	184	134	50	3,013	86.0
10	もう	1355	1085	270	37,461	84.0

指示詞は一般に話し言葉で多く見られ，代名詞（話し言葉：24988，書き言葉：12877），「- レ」（此れ（話し言葉：4188，書き言葉：1716），其れ（話し言葉：6093，書き言葉：1950），彼れ（話し言葉：631，書き言葉：127））で目立つ。随筆でも特徴語となることが考えられる。また，「こんな」「そんな」も特徴語となっている。

表 11　書籍サンプル随筆データの特徴語彙：指示詞（修正 LLR 上位 5 位＋「此れ」）

順位	語彙素	随筆頻度	（内）文学 9X4	（内）文学以外 -049	随筆以外頻度	修正 LLR
1	其れ	6,854	5,316	1,538	188,400	443.9
2	彼れ	477	406	71	7,585	237.8
3	こんな	745	535	210	16,890	132.9
4	其の	8,324	6,196	2,128	263,418	128.9
5	そんな	1,105	836	269	29,276	91.1
9	**此れ**	3,425	2,370	1,055	114,712	16.9

しかし，随筆において「此れ」は特徴語ではない（修正 LLR：16.9）という結果になった。「其れ」「彼れ」は上位語として確認できる。特に，「其れ」は文学（9X4）ジャンルにおける特徴語である。

高崎ら（2007）の指示語の調査でも，「これ」が 864 件に対し，「それ」が 1,713 件という結果が示されている（p. 23）。表 11 においても，「此れ」（表中太字）は随筆において「其れ」よりも頻度が低い。「まとめて検索 KOTONOHA」の調査結果でも，書き言葉において「此れ」が 1,716pmw に対し，「其れ」が 1,950pmw とさほどの差が見られないことから，「此れ」が用いられにくいことも，書き言葉としての随筆の特性であると考えられる。「此れ」の頻度は，レジスター別に見てもほとんど差がないが，新聞（そのほか広報紙・韻文）において頻度の低い傾向が見られるほか，国会議事録にのみ突出するという特徴がある（表 12）。対して，「其れ」は，「此れ」

同様に新聞（広報紙）において頻度が低く，国会議事録で突出する傾向が見られるが，白書と法律でも低頻度であること，ベストセラーにおいて高頻度であるという点（表 13）において「此れ」とは違いがある。すなわち，指示詞は話し言葉的なレジスターに多く用いられやすい傾向があるものの，「此れ」は白書や法律などでも用いられやすく，国会議事録のような話し言葉の書き起こしでなければ，文体の違いが頻度差に関わるのではないと考えられる。よって，随筆に「此れ」が特徴的ではないことは，話し言葉的な語彙が用いられるとしても，あくまでも書き言葉であるという特質の表れといえよう。

表 12　語彙素「此れ」の pmw 分布（一部）

BCCWJ 全体	新聞	広報紙	韻文	国会議事録
1907.0	861.5	547.7	461.5	5468.4

表 13　語彙素「其れ」の pmw 分布（一部）

BCCWJ	新聞	白書	広報紙	ベストセラー	法律	国会議事録
2598.8	774.6	346.2	199.4	4026.0	21.3	4320.3

機能語を確認することにより，「随筆」の特徴語として，音変化を含む話し言葉的な語や，書き言葉よりも話し言葉において頻度の高い語が取得される傾向が確認された。敬体の「です」「ます」が特徴語であることも含め，「随筆」の文体はくだけた印象や軟らかい印象となっている可能性が考えられる。また，婉曲表現や程度の判断に関する表現が特徴的であることから，主観性が高いことや専門性が低いことなどの印象に影響する可能性もある。次節では，文体指標を用いた検証を行う。

5.2 随筆の文体の検討

一般に，随筆には特徴的な文体傾向が現れると考えられている。前節で見たように，特徴語として音変化を含む語をはじめとする話し言葉的な語，話し言葉に高頻度な語が現れており，文体特徴と関わりのある可能性がある。ジャンル別の調査を行う際には，「随筆」が特徴的な文体を有する文章群として着目されることが多い。しかし，これまで「随筆」という文章群の大規模かつ主題横断的な調査は困難であったため，文体傾向についても，客観的な対照を可能とする指標を用いた分析結果は示しにくかったといえる。そこで，本節では，BCCWJ の LB に含まれる全随筆サンプルについて，柏野（2013）の示す文体指標を参照し，随筆の文体傾向分析を行う。

国立国語研究所（2015）では，BCCWJ の LB サンプルについて人手で文体分類を行い，以下の情報を付与している。

（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

以下，（a）から（e）の 5 つの指標について，LB 内の随筆サンプルと随筆サンプル以外とを対照し，随筆の文体に特徴が見られるのかを検証する。なお，サンプルにより，文体指標の付与されていない場合（文体情報付与対象外）がある [2]。

5.2.1 随筆の専門度

表 14 に専門度分布を示す。随筆サンプルの 8 割程度が「一般向き」と判定されており，随筆の対象読者は，概ね「一般」であると考えられる。文学に分類される随筆（「9X4」）と各ジャンルに分散していた（形式分類「-049」）随筆に差異は見られない。

前節で見た特徴語彙として，敬体が現れていた（5.1 節の表 6・表 7 参照）ことは，類目によっては「やや専門的な一般向き」のような内容であったとしても，敬体である「です」「ます」を使用することによって（表 6・表 7 参照）やや専門度をやわらげ，「一般向き」のテキストであるという印象を読み手に与えることに役立っている可能性がある。話し言葉に高頻度な語彙や，話し言葉的な語彙の影響もあり得る。

[2] 国立国語研究所（2015）は，「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）なもの」を分類①とし，内容・表現の文体的特徴の印象判定により細分類を行っている。印象判定の付与されない分類②「文体判断が単純にいかないと判断されるもの」とされた場合には，「テキスト構造・紙面形式上の特徴」情報（対話系・引用系・視覚表現多用系・データベースやリスト系），「内容や表現上の特定の特徴」情報（前書きや後書き・明治時代より以前の古い言葉が多い・外国語が多い・数式やプログラミング言語などが多い・法律文が多い・教育現場で使いがたそう・一定量の「本文」が認めがたい）が付与されている。

表 14　LB における専門度分布

	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	文体情報付与対象外	総計
文学（9X4）	0	8	246	1	0	45	300
	0.0%	2.7%	82.0%	0.3%	0.0%	15.0%	100.0%
文学以外（-049）	0	3	81	0	0	21	105
	0.0%	2.9%	77.1%	0.0%	0.0%	20.0%	100.0%
随筆計	0	11	327	1	0	66	405
	0.0%	2.7%	80.7%	0.2%	0.0%	16.3%	100.0%
随筆以外	141	918	6,738	383	302	1,664	10,146
	1.4%	9.0%	66.4%	3.8%	3.0%	16.4%	100.0%
LB 全体	141	929	7,065	384	302	1,730	10,551
	1.3%	8.8%	67.0%	3.6%	2.9%	16.4%	100.0%

5.2.2 随筆の客観度

表 15 に客観度分布を示す。いわゆる随筆は，主観的であることが予想される。そして，本稿の調査結果を見ても，「とても主観的」が半数近く，「どちらかといえば主観的」をあわせた主観的傾向は，7 割程度に見られる。このことは，特徴語彙として評価や判断に関わる語（5.1 節）が現れていたこととも関連性があろう。また，書籍全体に含まれる「とても主観的」なサンプルの割合が随筆の影響により高くなっているといえる。

但し，各ジャンルに分散していた（形式分類「-049」）随筆においては，随筆以外の書籍・書籍全体と同程度の「どちらかといえば客観的」なサンプルも得られている。「-049（文学以外）」と「9X4（文学）」の語彙を比較した際，「私」は「9X4（文学）」にのみ最も特徴的な語として現れていた（5.1 節表 7 参照）。「-049（文学以外）」の随筆は，内容としては各ジャンルに分類された特定主題であるため，多様な主題を扱う「9X4（文学）」分類よりも「客観的」と読み取られる可能性がある。主観的か客観的かという印象は，内容の影響が強いと考えられる。

表 15　LB における客観度分布

	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	文体情報付与対象外	総計
文学（9X4）	1	27	57	148	67	300
	0.3%	9.0%	19.0%	49.3%	22.3%	100.0%
文学以外（-049）	5	22	28	28	22	105
	4.8%	21.0%	26.7%	26.7%	21.0%	100.0%
随筆計	6	49	85	176	89	405
	1.5%	12.1%	21.0%	43.5%	22.0%	100.0%
随筆以外	944	2,474	1,481	686	4,561	10,146
	9.3%	24.4%	14.6%	6.8%	45.0%	100.0%
LB 全体	950	2,523	1,566	862	4,650	10,551
	9.0%	23.9%	14.8%	8.2%	44.0%	100.0%

5.2.3 随筆の硬度

表 16 に硬度分布を示す。随筆は，書籍全体よりも「とても軟らかい」と判断されたサンプルの割合の高いことがわかる。「どちらかといえば軟らかい」の割合では大差がないが，読み手が極端に「軟らかい」という印象を受けるテキストが，随筆の文章には高い割合で出現する可能性が考えられる。なお，特徴語彙（5.1 節）からテキストの硬軟の印象判定への直接的な影響は見えにくいが，機能語に見られる敬体や「けれど」「って」のようなくだけた語の混在と，「ね」「か」のような読み手への働きかけなどに「思う」のような主観性などが組み合わさることで，「軟らかい」印象を与える可能性は考えられよう。「-049（文学以外）」と「9X4（文学）」に差は見られない。

表 16　LB における硬度分布

	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	文体情報付与対象外	総計
文学（9X4）	1	59	164	31	45	300
	0.3%	19.7%	54.7%	10.3%	15.0%	100.0%
文学以外（-049）	0	19	49	16	21	105
	0.0%	18.1%	46.7%	15.2%	20.0%	100.0%
随筆計	1	78	213	57	66	405
	0.2%	19.3%	52.6%	14.1%	16.3%	100.0%
随筆以外	618	2,987	4,227	640	1,664	10,146
	6.1%	29.4%	41.7%	6.3%	16.4%	100.0%
LB 全体	619	3,065	4,440	697	1,730	10,551
	5.9%	29.0%	42.1%	6.6%	16.4%	100.0%

5.2.4 随筆のくだけ度

表 17 にくだけ度分布を示す。「とてもくだけている」「どちらかといえばくだけている」ともに，書籍全体よりも高い割合が明らかとなった。特徴語彙としても「けれど」「って」のような話し言葉的と考えられる表現が見られていたこと（5.1 節表 6 など）が関わっていると考えられる。反対に，「くだけていない」随筆は，随筆全体の約三分の一程度に留まる。なお，この傾向は，「-049（文学以外）」と「9X4（文学）」に大差がないため，内容別のジャンル（類目）検索を行う際には，随筆の文章の影響によって，期待しないくだけた用例の得られる可能性が考えられる。

表 17　LB におけるくだけ度分布

	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	文体情報付与対象外	総計
文学（9X4）	37	115	103	45	300
	12.3%	38.3%	34.3%	15.0%	100.0%
文学以外（-049）	15	28	41	21	105
	14.2%	26.7%	39.0%	20.0%	100.0%
随筆計	52	143	144	66	405
	12.8%	35.3%	35.6%	16.3%	100.0%
随筆以外	421	2,553	5,508	1,664	10,146
	4.1%	25.2%	54.3%	16.4%	100.0%
LB 全体	473	2,696	5,652	1,730	10,551
	4.5%	25.6%	53.6%	16.4%	100.0%

5.2.5 随筆の語りかけ性度

文章であっても，読み手が語りかけるような感じを受けるテキストは，随筆のようなテキストに現れる特徴であると考えられてきた。しかし，「語りかけ性度」に着目した調査では，いわゆるハウツー本のような教示的内容を含む書籍テキスト全般において，語りかけ性があると判断される傾向がある（加藤ら 2014）。随筆と語りかけ性度には関連性が見られるのだろうか。特徴語彙として，「ね」「か」のような直接的な語りかけと考えられる終助詞が得られた（5.1 節表 6）ことからも，随筆は語りかけ性度が非常に高いのではないかという期待があった。

本調査の結果，随筆では，書籍全般よりも「とても語りかけ性がある」「どちらかといえば語りかけ性がある」において，いくらか高い割合が示された（表 18）。もっとも，LB は小説（9X3）が全体の約 3 割（2,932 サンプル）を占めるため，小説作中における作者の顔出しや一人称小説の地の文などの影響が考えられ，大差とはいい難い。また，随筆であっても「特に語りかけ性はない」が半数以上を占め，語りかけるような文体が随筆に特有であるともいえまい。「-49（文学以外）」において「語りかけ性度」が若干高い割合となるのは，各分野における著名人などの特定主題の随筆に対し，読み手が教示を受けるような印象を持った可能性も考えられる。語りかけるような感じの文体は随筆に特徴的とまではいえない結果であった。

表 18　LB における語りかけ性度分布

	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	文体情報付与対象外	総計
文学（9X4）	31	54	170	45	300
	10.3%	18.0%	56.7%	15.0%	100.0%
文学以外（-049）	18	23	43	21	105
	17.1%	21.9%	41.0%	20.0%	100.0%
随筆計	49	77	213	66	405
	12.1%	19.0%	52.6%	16.3%	100.0%
随筆以外	784	1,302	6,396	1,664	10,146
	7.7%	12.8%	63.0%	16.4%	100.0%
LB 全体	833	1,379	6,609	1,730	10,551
	7.9%	13.1%	62.6%	16.4%	100.0%

5.2.6 随筆の文体特徴まとめ

文体指標との対照から，随筆（「9X4（文学）」「-049（文学以外）」）の文体は以下のような傾向が確認された。

① 一般向き
② 主観的傾向
③ 極端に軟らかい印象を受けるテキストが含まれる場合がある
④ くだけたテキストの割合が高い傾向
⑤ 語りかけるテキストの割合が高いとまではいい難い

以上により，随筆の読者対象層は広く，読者に「主観的」かつ「くだけ」た印象を与える場合が多いといえる。一般的な「随筆」に期待されると考えられる文体特徴が検証できた。また,「随筆」には，とくに軟らかい印象を与えるテキストも含まれるという可能性も見られた。「随筆」を調査の対象とするときには，これらの特徴の関わる言語現象（5.1 節参照）が得やすい可能性がある。反対に，このような「随筆」がジャンル（NDC の類目）内に分散していることで，あるジャンルを調査対象とするにあたり，これらの特徴が影響を及ぼす可能性も考えられる。特に「主観的」な文体と判断されるサンプルの割合において随筆の影響が見られているといえる。

6. おわりに

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，旧 NDC 番号とともに「中納言」上で参照できる。増補分のリストは加藤ら（2019）に掲載している。また，本データは https://github.com/masayu-a/BCCWJ-NDC にて公開している。本データにより，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。

参照文献

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): 43–53.
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』372–381.
加藤祥・柏野和佳子・立花幸子・丸山岳彦（2014）「語りかける書きことばの表現」『国立国語研究所論集』8: 85–108.
加藤祥・森山奈々美・浅原正幸（2019）「『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」『言語資源活用ワークショップ発表論文集 2019』155–160.
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）. http://doi.org/10.15084/00003109
Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced Corpus of Contemporary Written Japanese, *Language Resources and Evaluation*, 48: 345–371.
日本図書館協会分類委員会（1995）『日本十進分類法新訂 9 版』東京：日本図書館協会.
日本図書館協会分類委員会（2018）『日本十進分類法新訂 10 版』東京：日本図書館協会.
高崎みどり・新屋映子・立川和美（2007）『日本語随筆テクストの諸相』東京：ひつじ書房.
高崎みどり（2012）「美味を意味する語の使用と性差：「おいしい」を中心に」『お茶の水女子大学人文科学研究』8: 55–68.
立川和美（2014）「文章と談話における引用表現：随筆と雑談・相談を例として」『流通經濟大學論集』49(1): 31–47.

Enlargement of Nippon Decimal Classification Metadata of Book Samples in the "Balanced Corpus of Contemporary Written Japanese": Extraction of Essays from Book Samples According to NDC Metadata and Writing Style Analysis

KATO Sachi[a], MORIYAMA Nanami[b], ASAHARA Masayuki[c]

[a]Mejiro University /Project Collaborator, NINJAL
[b]Technical Staff, Center for Corpus Development, NINJAL [–2020.03]
[c]Center for Corpus Development, NINJAL

Abstract

This study presents the enlargement of Nippon Decimal Classification (NDC) metadata of book samples in the "Balanced Corpus of Contemporary Written Japanese (BCCWJ)." We revised and enhanced the NDC information about all of the book samples from the BCCWJ (22,058 samples) comprising PB (books in the publication subcorpus: 10,117 samples), LB (books in library subcorpus: 10,551 samples), and OB (books in the special-purpose subcorpus; namely, best sellers: 1,390 samples). We referred to the NDC information using the National Diet Library Search API and manually re-annotated the NDC information. In addition, we completed the empty entries of the original BCCWJ metadata. Based on these procedures, we were able to classify the BCCWJ book samples according to the genres of essay (-049), theory (-01), and textbook (-078) with the NDC supplemental tables. Furthermore, since finer-grained categories, including their chronological periods, were added to some samples, users can explore a more detailed classification of the book samples. We present the methodology of NDC information enlargement and its basic statistics. We also present experimental research on extraction essays from books and the investigation of their writing style. The compiled data can be used in the corpus query systems of "Chunagon."

Keywords: "Balanced Corpus of Contemporary Written Japanese," Nippon Decimal Classification, writing style, log-likelihood ratio