Refinement of NDC Annotation on the Balanced Corpus of Contemporary Written Japanese and Writing Style Analysis of Essays Based on the NDC Auxiliary Table

『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

加藤 祥（国立国語研究所コーパス開発センター）†
櫻井 芽衣子（日本工業大学）
森山 奈々美（津田塾大学大学院）
浅原 正幸（国立国語研究所コーパス開発センター）

Refinement of NDC annotation on the Balanced Corpus of Contemporary Written Japanese and writing style analysis of essays based on the NDC auxiliary table

Sachi Kato (National Institute for Japanese Language and Linguistics)
Meiko Sakurai (Nippon Institute of Technology)
Nanami Moriyama (Tsuda University)
Masayuki Asahara (National Institute for Japanese Language and Linguistics)

要旨

現在,『現代日本語書き言葉均衡コーパス』に含まれる書籍サンプルに付与された日本十進分類法（NDC）分類記号に，補助分類を拡張する作業を進めている。国立国会図書館の NDC 情報（8 版・9 版）を参照し，人手によって補助分類の確認と追加を行う。本発表は，現在までに作業の完了した図書館サブコーパス 10,551 サンプルについて，情報付与作業方法とその結果を報告する。本作業により，たとえば形式区分を利用し，ジャンルの分散する「随筆」「理論」「研究法」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る。そこで，付与した形式区分「随筆」サンプル群を例とし，語彙特徴から文体的な傾向を調査した。さらに，柏野（2015）の文体指標を用い，「随筆」の文体特徴として考えられてきた「主観的」で「軟らかく」「くだけている」傾向などを確認する。

１．はじめに

あるテキストの文体的な特徴を分析するときには，著者に関する情報のほか，新聞記事か書籍か, あるいは Web 上のブログの文章かなど, ジャンルの違いに着目することが多い。文体分析にコーパスを活用し，ジャンルごとの文体特徴を明らかにすることが期待される。現在,『現代日本語書き言葉均衡コーパス』（以降 BCCWJ）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。しかし，ジャンル分類は主に媒体と内容に基づき，書籍の形式については分類されていない。すなわち，現状，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るため，文体分析対象の「随筆」は適切に収集し難い。反対に，「自然科学」ジャンルの典型例であろう「方法論」の文体を調査したい場合では，随筆の混在が分析結果へ影響を及ぼすともいえる。そのため，BCCWJ 公開後にも，たとえば「随筆」を調査対象とした研究は多々見られるが（立川 2014，高崎 2012 など），いずれも雑誌の「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り，大規模データの調査は難しい状態にあると考えられ

† yasuda-s@ninjal.ac.jp

る。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」などのジャンルでの分類も可能とすることを目指す。本稿は，現在までに付与の完了した図書館サブコーパス 10,551 サンプルについて，情報付与作業方法とその結果を報告する。また，実際に形式区分「随筆」を用い，「随筆」サンプル群の語彙的・文体的な傾向を調査する。

２．BCCWJ 書籍サンプルに対する NDC 記号アノテーション

２.１　アノテーション対象

BCCWJ には，出版・書籍（PB：7,482 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：405 サンプル）の 3 種類の書籍サブコーパスがある。本研究は，これらすべての書籍サンプルに対し，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）があれば，該当する番号を付与する。また，本稿で報告する図書館サブコーパスから開始し，順次アノテーションを進める。

BCCWJ 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となっている。「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。

⑴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）

⑵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）

⑶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）

⑷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）

本アノテーション作業により，既存の 3 桁に「.」以降の番号（下位区分）が追加される場合（(1)’(2)’(3)’(4)’に例示する）がある。すなわち，(1)では文学共通区分で時代情報が追加され，(2)では歴史の小項目が追加されるというように，さらに詳細な書籍サンプルの分類が可能となる。また，(3)(4)のように，形式区分が追加される場合は，内容ではない「事典」「随筆」などの分類が可能となる。

⑴’ 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））

⑵’ サンプル ID：LBl2_00027『縄文人・弥生人 101 の謎』……210.025

[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。

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））

２.２　アノテーション方法

BCCWJ の書籍サンプルについて，国立国会図書館の NDC 情報（8 版・9 版が付与されている）を参照した。各サンプルの候補となる書籍情報を収集し，人手により BCCWJ サンプルの書籍タイトル・著者・出版社・発行年を確認し，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，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。また，BCCWJ 構築時に 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。本稿で報告する LB サンプルに対しては，作業者 2 名が NDC 情報付与作業を行った。

２.２　LB サンプルへのアノテーション結果

LB サンプル（総数：10,551 サンプル）への情報付与結果を表 1 に示す。3 桁の番号が付与されたサンプルは 10,092 サンプルあり，「分類なし」となっていた番号のないサンプルが 459 サンプルあった。本作業により，8,690 サンプルに補助分類が追加され，83%のサンプルにおいて NDC 番号が拡張された。番号の追加がなかったサンプルは，「9X3」（要目：小説）が 18%を占めるほか，「X04」（要目：論文集）が 5%など，国立国会図書館データにおいて補助分類の付与がない書籍である。

また，BCCWJ 構築当時に「分類なし」であったサンプルのうち，410 サンプルにも NDC 番号を付与することができた。なお，NDC 番号の確認できなかった書籍（49 サンプル分）は，いずれも国立国会図書館では雑誌扱い（ムックや特集版など）とされていたものである。また，現在の BCCWJ とは番号の異なる書籍が 10 サンプル分見つかった。これらの違いの原因には，NDC の版の違いや，後の修正などが考えられる。

表 1. LB サンプルへの NDC 番号拡張アノテーション結果（いずれもサンプル数）

LB サンプル	番号付サンプル	拡張サンプル	新規番号付与	番号違い
10,551	10,092	8,690	410	10

３．NDC 情報を用いた随筆サンプルの抽出

本稿では，拡張した NDC 番号の補助区分を用い，特に文体研究に活用が可能と考えられる「随筆」サンプルを BCCWJ の LB から抽出することを試みる。

３.１　NDC による「随筆」の抽出

NDC では，文学者が書いた随筆，あるいは多数主題を扱った随筆が「9X4」（文学）に分類されるが，特定主題を扱った随筆・エッセイは「.049」の形式区分が付与される。ゆえに，BCCWJ サンプルから「随筆」テキストを収集するためには，既存の「9X4」のほか，形式区分「.049」の付与されたサンプルを抽出する必要がある。LB において「9X4」は 300 サンプルある。

３.２　LB サンプルの形式区分による「随筆」の抽出

まず，LB サンプルに付与された形式区分について整理しておく。表 2 は，本作業によって付与された NDC 補助分類の形式区分が，どのように分布しているのか整理したものである。形式区分は，全サンプルの約 10%に付与されていた。

表 2 では，「.04」が 3.6%を占めている。「.04」は「論文集」であるが，このうち，「.049」は「特定主題随筆」にあたる。「特定主題随筆」は，各ジャンル（内容による分類：NDC の類目にあたる）に分散した随筆・エッセイである。よって，「.049」の付与されたサンプルを抽出すれば，料理人やアスリートなどによって記された随筆についても調査対象とすることが可能になる。「.049」が付与されたサンプルは 81 サンプルあった。

表 2．LB サンプルの NDC 形式区分

形式区分	サンプル数	LB サンプル全体における割合
.01 理論．哲学	87	0.8%
.02 歴史的・地理的論述	275	2.6%
.03 参考図書	61	0.6%
.04 論文集．法論集．講演集．会議録	382	3.6%
.05 逐次刊行物	62	0.6%
.06 団体	51	0.5%
.07 研究法．指導法．教育	116	1.1%
.08 叢書．全集．選集	35	0.3%
計	1069	10.1%

４．随筆の特徴語彙

本節では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: LLR）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：

$$\mathbf{LLR}(\mathbf{w}) = \mathbf{2}\ (\mathbf{a}\ \log_e \boldsymbol{a} + \mathbf{b}\ \log_e \boldsymbol{b} + \mathbf{c}\ \log_e \boldsymbol{c} + \mathbf{d}\ \log_e \boldsymbol{d} - (\mathbf{a}+\mathbf{b})\ \log_e(\boldsymbol{a}+\boldsymbol{b}) - (\mathbf{a}+\mathbf{c})\ \log_e(\boldsymbol{a}+\boldsymbol{c}) - (\mathbf{b}+\mathbf{d})\ \log_e(\boldsymbol{b}+\boldsymbol{d}) - (\mathbf{c}+\mathbf{d})\ \log_e(\boldsymbol{c}+\boldsymbol{d}) + (\mathbf{a}+\mathbf{b}+\mathbf{c}+\mathbf{d})\ \log_e(\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{c}+\boldsymbol{d}))$$

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)。

A 群を LB サンプル内の随筆(9X4 および.49)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 3 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」や動詞の「思う」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」のようにくだけた語などが得られていることがわかる。表外でも，「僕（331.66)」「自分（288.42)」などの一人称に関する語，「好き（290.88)」「面白い（273.11)」のような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。次節では，文体指標を用いた検証を行いたい。

表 3. LB サンプルの随筆(9X4 および .049) の特徴語彙

特徴語彙	修正 LLR
私	1939.13
言う	1259.50
ね	1223.40
です	1216.10
けれど	901.93
だ	853.68
笑い	832.26
書く	820.35
小説	816.71
って	712.22
も	704.44
か	678.31
と	598.10
思う	584.64

なお, 本稿の作業として新たに取得された「.049」分類の随筆と, 文学の類目にあたる「9X4」分類の随筆の異同を確かめておく。

表 4. LB サンプルの.049 と 9X4 の特徴語彙

.049 の特徴語彙	修正 LLR	9X4 の特徴語彙	修正 LLR
ます	757.11	私	-199.26
相続	510.34	た	-189.80
ストレス	443.08	の	-155.67
上司	407.72	小説	-145.90
検索	357.06	書く	-110.47
会社	307.81	男	-95.19
です	281.23	文学	-89.86
相手	266.17	女	-74.67
退職	261.88	家	-68.38
プレゼンテーション	249.08	カレー	-68.17
為る	232.21	作品	-65.78
クレーム	225.68	有る	-64.10

表 4 に，各々の分類の特徴語彙を示す。随筆サンプルの中でも，様々な類目に分散する「.049」と文学類目に分類される「9X4」では，それぞれ特徴語に違いが見られた。「9X4」では，随筆一般に特に多く見られた「私」のほか，「男」「女」「家」のような名詞，「小説」「書く」「文学」のような文学類目ゆえの語が特徴的に現れている。これに対し，「.049」では，「ます」「です」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」などをはじめ，表外でも上位語には「教育（104.67)」や「企業（98.05)」，「妃（93.39)」，

「ウイルス（70.89）」,「副腎（63.99）」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，内容語のほかで特に敬体が特徴的な文体だといえる。反対に,「9X4」分類では，敬体が特徴ではない。「9X4」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」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たと考えられる。次節でも,「随筆」としての総計に加え,「.049」と「9X4」の異同についても確かめておく。

５．随筆の文体

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。ジャンル別の調査を行う際,「随筆」が着目されることは多い。しかし，これまで「随筆」の文体傾向について，大規模かつ主題横断的な調査は困難であった。そこで，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 について人手で文体分類を行い，全てのサンプルに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，随筆サンプルと LB サンプル全体を対照し，随筆の文体に特徴が見られるのかを検証する。なお，表の各分布合計は総計と合致していないが，サンプルにより，文体指標の付与されていない場合がある（国立国語研究所，2015 の分類②にあたる場合,「客観度」は読み手に小説と判断されたサンプルに付与されていないなど）ことによる。

５.１　随筆の専門度

表 5 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。

前節で見た特徴語彙として, 敬体が現れていた（前節の表 3･表 4 参照）ことは, 特に「.049」において，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,「です」「ます」を使用することによって（表 4 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。

表 5. 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	総計
9X4	0	8	246	1	0	300
	0.0%	2.7%	82.0%	0.3%	0.0%	
.049	0	2	56	0	0	81
	0.0%	2.8%	78.9%	0.0%	0.0%	
随筆計	0	10	302	1	0	381
	0.0%	2.7%	81.4%	0.3%	0.0%	
LB 全体	141	929	7065	384	302	10551
	1.3%	8.8%	67.0%	3.6%	2.9%	

５.２　随筆の客観度

表 6 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「思う」や評価語彙（前節表 3）が現れていたこととも関連性があろう。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，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」と「9X4」の語彙を比較した際，「私」は「9X4」にのみ最も特徴的な語として現れていた（前節表 4 参照）。「.049」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」分類よりも「客観的」と読み取られる可能性がある。

表 6. 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	総計
9X4	1	27	57	148	300
	0.3%	9.0%	19.0%	49.3%	
.049	0	17	16	25	81
	0.0%	23.9%	22.5%	35.2%	
随筆計	1	44	73	173	381
	0.3%	11.9%	19.7%	46.6%	
LB 全体	950	2523	1566	862	10551
	9.0%	23.9%	14.8%	8.2%	

５.３　随筆の硬度

表 7 に硬度分布を示す。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」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（前節表 3）からテキストの硬軟の印象判定への影響は見えにくいが，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，「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。

表 7. 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	総計
9X4	1	59	164	31	300
	0.3%	19.7%	54.7%	10.3%	
.049	0	10	34	14	81
	0.0%	14.1%	47.9%	19.7%	
随筆計	1	69	198	45	381
	0.3%	18.6%	53.4%	12.1%	
LB 全体	619	3065	4440	697	10551
	5.9%	29.0%	42.1%	6.6%	

５.４　随筆のくだけ度

表 8 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」のような話しことば的と考えられる表現が見られていた（前節表 4）。反対に，「くだけていない」随筆は，随筆全体の約三分の一に留まる。なお，この傾向は，「.049」でも同様であるため，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」た用例が得られる可能性が考えられる。

表 8. 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	総計
9X4	37	115	103	300
	12.3%	38.3%	34.3%	
.049	13	23	22	81
	18.3%	32.4%	31.0%	
随筆計	50	138	125	381
	13.5%	37.2%	33.7%	
LB 全体	473	2696	5652	10551
	4.5%	25.6%	53.6%	

５.５　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とわかった（加藤ほか，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（前節表 3）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ろう。

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」ともに，いくらか高い割合が示された（表 9）。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，大差とは言い難い。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言えまい。「.049」で「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。

表 9. 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	総計
9X4	31	54	170	300
	10.3%	18.0%	56.7%	
.049	11	14	33	81
	15.5%	19.7%	46.5%	
随筆計	42	68	203	381
	11.3%	18.3%	54.7%	
LB 全体	833	1379	6609	10551
	7.9%	13.1%	62.6%	

５.５　随筆の文体特徴

文体指標との対照から，随筆（「9X4」「.049」）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言い難い

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（4 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。

６．まとめ

本研究の進める作業により，BCCWJ の書籍サンプルを，現状の内容分類（NDC の類目・綱目・要目）の小項目や形式などによって，詳細あるいは異なる基準で分類することが可能となる。本稿では，現在までに完了した LB の情報付与結果（NDC の下位区分「形式区分」）を用い，「随筆」サンプルを抽出し，大規模かつジャンル横断的な「随筆」の語彙特

徴と文体特徴の分析を試みた。文学ジャンルにとどまらないテキストを分析することで，「随筆」の特徴語彙や文体傾向が取得できた。同様に「形式区分」を用いることで，本稿で試行した「随筆」のほか，「論文集」「理論」，「研究法」「伝記」などの特定分類の抽出や分析を行うことも可能である。また，書籍の各ジャンル（NDC の類目別）の下位区分（小項目や細区分）を利用すれば，時代や地域などによる分類をはじめ，詳細に絞り込んだサンプルテキストの抽出が可能となる。PB と OB への情報付与も進め，BCCWJ 書籍サンプルのさらなる活用を図りたい。さらに，図書館情報学におけるデータの活用についても検討したい。

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

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）

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, pp.85-108

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), pp.31-47.

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.

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, pp.55-68.

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　http://iss.ndl.go.jp/