BCCWJ 図書館サブコーパス全テキストへの文体情報付与結果の分析

柏野 和佳子*（国立国語研究所 言語資源研究系）
立花 幸子（国立国語研究所 コーパス開発センター）
保田 祥（国立国語研究所 コーパス開発センター）
飯田 龍（東京工業大学 大学院情報理工学研究科）
丸山 岳彦（国立国語研究所 言語資源研究系）
奥村 学（東京工業大学 精密工学研究所）
佐藤 理史（名古屋大学 大学院工学研究科）
徳永 健伸（東京工業大学 大学院情報理工学研究科）
大塚 裕子（はこだて未来大学 メタ学習センター）
佐渡島 紗織（早稲田大学 留学センター）
椿本 弥生（はこだて未来大学 メタ学習センター）
沼田 寛（はこだて未来大学 メタ学習センター）

Writing Style Annotation for the Library Subcorpus of the Balanced Corpus of Contemporary Written Japanese

Wakako Kashino (Dept. Corpus Studies, NINJAL)
Sachiko Tachibana (Center for Corpus Development, NINJAL)
Sachi Yasuda (Center for Corpus Development, NINJAL)
Ryu Iida (Dept. Computer Science Graduate School of Information Science and Engineering, Tokyo Instiute of Technology)
Takehiko Maruyama (Dept. Corpus Studies, NINJAL)
Manabu Okumura (Precision and Intelligence Laboratory, Tokyo Institute of Technology)
Satoshi Sato (Graduate School of Engineering, Nagoya University)
Takenobu Tokunaga (Dept. Computer Science Graduate School of Information Science and Engineering, Tokyo Instiute of Technology)
Hiroko Otsuka (Center for Meta-Learning, Future University Hakodate)
Saori Sadoshima (Center for International Education, Waseda University)
Mio Tsubakimoto (Center for Meta-Learning, Future University Hakodate)
Hiroshi Numata (Center for Meta-Learning, Future University Hakodate)

１．はじめに

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」（平成 21～24 年度）の成果報告である。『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には，10,551 の書籍サンプルが収録されている。本研究ではそのコーパスをより有効に活用し，テキスト研究を進めるために，書籍テキストの多種多様な形式，内容，表現に関わる特徴を捉えるための分類指標の設計と付与，検証とを行ってきた（柏野・奥村 2012，柏野ほか 2012，柏野ほか 2012a，柏野ほか 2012b，保田ほか 2012，保田ほか 2012a，2012b）。

コーパスへ文体情報を付与することの重要性は，EAGLES（1996）等より議論され，例えば Lee（2001）によって，British National Corpus（BNC）への付与が実現されている。また，BCCWJ に収録されるテキストの文体を計量的に考察する試みがすでにいくつか行われている（小磯ほか 2008，2011，間淵ほか 2010）。しかしながら，サブコーパスに収録される約 1

* waka@ninjal.ac.jp

万という大量の書籍サンプルすべてを精査し，体系的に文体情報を付与するような試みは，本プロジェクトの実践がはじめてのことである。

これまで，柏野ほか（2012），柏野ほか（2012a，2012b）において述べてきたとおり，アノテーション作業は次の二段階で行った。

① 主に形式による判定を行う。構造的に単純なテキストタイプ（例：章節構造）であれば②の細分類の対象とする。
② 内容・表現の細分類をする。「専門度（幼児・小学生～専門家：5段階），客観度（とても客観的～とても主観的：4段階），硬度（とても硬い～とても軟らかい：4段階），くだけ度（とても・どちらかといえば・くだけていない：3段階），語りかけ性度（とてもある・どちらかといえば・特にない：3段階）」の分類指標を付与する。

上記①の段階で図書館サブコーパスの 10,551 の書籍サンプルのうち，8,887（84%）を「構造的に単純なテキストタイプ」と判断し，上記②の内容・表現の細分類の対象とした。柏野ほか（2012），柏野ほか（2012a，2012b）では，その②のアノテーション作業結果に関する報告を重ねてきた。本稿では，これまで取り上げなかった，②の細分類の対象外としたものを取り上げ，それらの類型とアノテーション作業結果について報告する。該当サンプルは，全部で 1,664（16%）である。これまで対象外としたサンプルの分類結果まで分析することにより，図書館サブコーパスに収録される書籍サンプルの全体像と特徴とをより正確に把握することが狙いである。本稿で取り上げる一群のテキストを，以降「特徴的な類型のテキスト」と呼ぶこととする。

２．特徴的な類型のテキストのアノテーション作業

２.１　特徴的な類型のテキストの分類指標

柏野ほか（2009）では，BCCWJ 構築のサンプリングの過程で観察されたサンプルの多様性を報告した。その際に，文章形式に特徴のあるサンプルとして，Q&A 形式（例 1），会話形式（例 2），引用編集形式（例 3）を取り上げた。例 3 は，講義のあまった時間に学生に書かせたものを集めたものであるらしい。編者がそれらを引用して編集しているものとして，「引用編集方式」と呼ぶこととする。さらに，紙面形式に特徴のあるサンプルとして，コマ割りや図，イラストなどの視覚的表現を多用する一群（例 4）を取り上げた。以下，例を示す（サンプルの出典は，BCCWJ のサンプル ID と書名とで記す）。

例 1：Q&A 形式（PB33_00111『環境経営なるほど Q&A　環境先進企業へのヒント』）

Q3－7　マネジメントのための環境会計

マネジメントのための環境会計にはどんなものがありますか？　それぞれの特徴を教えて下さい。

A

■内部環境会計の意義

環境会計は，その目的により，外部報告目的の環境会計と内部管理目的の環境会計とに分類されています。わが国では環境省のガイドラインも推進力となって，多数の企業が環境会計を外部に公表するようになってきた一方，企業の意思決定に役立つ内部管理目的の環境会計の研究も進められています。

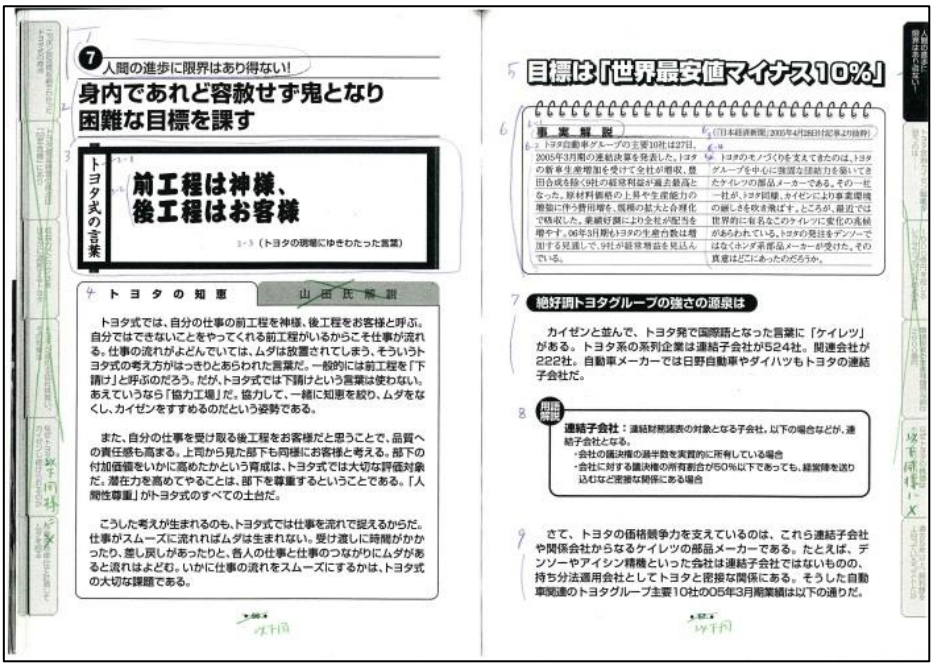
例 2：会話形式（PB53_00480『感性ちゃんと頭脳君の対話』）

感性　そういうことか。分かったわ。つまり，「肌の表面に何を塗っても，その物質がバリアゾーンを通過して有棘細胞層や基底細胞層にまで到達するわけがない」ってことなのね？
頭脳　そうだよ。そんなことは不可能なんだよ。もしもそれが可能だとしたら，肌の防衛網が機能していないことになるから，おそらくそういう人は生きていけないだろうね。
感性　物質のサイズを小さく，細かくしてもダメなの？
頭脳　ダメだよ。無理だね。バリアゾーンが健全な場合には，水の分子一個ですら通さないんだ。

例 3：引用編集形式（PB23_00427『ほろっと本音キラッと青春』）

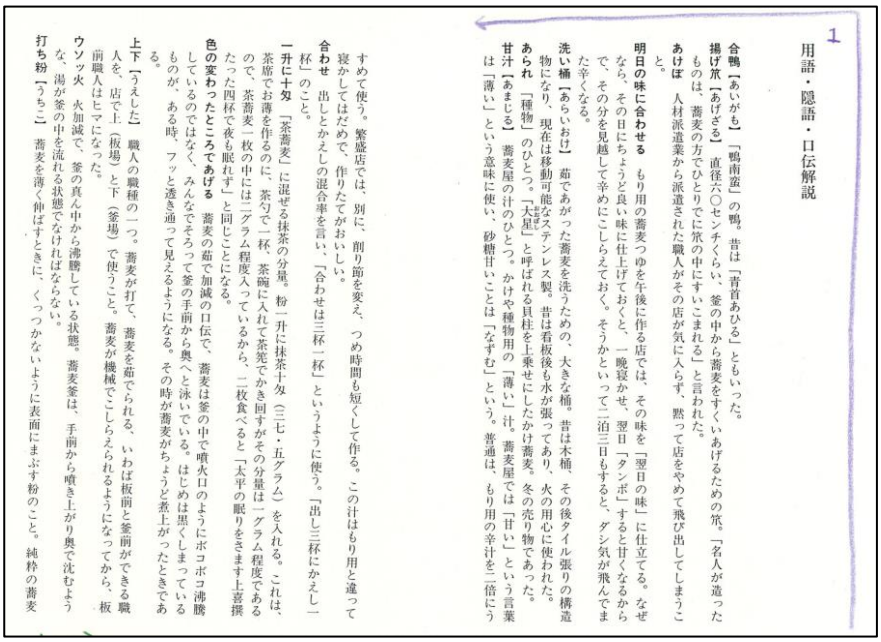
一八歳ってこんなものかなあ。ちょっと予定とはちがう。
なんだか毎日平凡。だけど，毎日平凡に過ごせていることを幸せだと思う。何も特別じゃなくていいと思いながら，毎日を平凡に頑張ってます。
友よ！　おまえらみんなさめすぎや。もっと毎日，感動的に生きろよ。

例 4：コマ割り（PB5n_00141『トヨタだけがなぜ儲かるのか！？』）



以上の観察に加え，辞書形式やカタログ形式をもつテキスト（例 5，6）も文章形式に特徴のあるものと考えられる。

例 5：辞書形式（LBp6_00009『蕎麦屋のしきたり』）



用語・隠語・口伝解説

合鴨【あいがも】「鴨南蛮」の鴨。昔は「青首あひる」ともいった。

揚げ笊【あげざる】直径六〇センチくらい、釜の中から蕎麦をすくいあげるための笊。「名人が造ったものは、蕎麦の方でひとりでに笊の中にすいこまれる」と言われた。

あけぼ　人材派遣業から派遣された職人がその店が気に入らず、黙って店をやめて飛び出してしまうこと。

明日の味に合わせる　もり用の蕎麦つゆを午後に作る店では、その味を「翌日の味」に仕立てる。なぜなら、その日にちょうど良い味に仕上げておくと、一晩寝かせ、翌日「タンポ」すると甘くなるからで、その分を見越して辛めにこしらえておく。そうかといって二泊三日もすると、ダシ気が飛んでまた辛くなる。

洗い桶【あらいおけ】茹であがった蕎麦を洗うための、大きな桶。昔は木桶、その後タイル張りの構造物になり、現在は移動可能なステンレス製。昔は看板後も水が張ってあり、火の用心に使われた。

あられ「種物」のひとつ。「大星」と呼ばれる貝柱を上乗せにしたかけ蕎麦。冬の売り物であった。

甘汁【あまじる】蕎麦屋の汁のひとつ。かけや種物用の「薄い」汁。蕎麦屋では「甘い」という言葉は「薄い」という意味に使い、砂糖甘いことは「なずむ」という。普通は、もり用の辛汁を二倍にうすめて使う。繁盛店では、別に、削り節を変え、つめ時間も短くして作る。この汁はもり用と違って寝かしてはだめで、作りたてがおいしい。

合わせ　出しとかえしの混合率を言い、「合わせは三杯一杯」というように使う。「出し三杯にかえし一杯」のこと。

一升に十匁　「蒸蕎麦」に混ぜる抹茶の分量。粉一升に抹茶十匁（三七・五グラム）を入れる。これは、茶席でお薄を作るのに、茶匁で一杯、茶碗に入れて茶筅でかき回すがその分量は一グラム程度であるので、茶蕎麦一枚の中には二グラム程度入っているから、二枚食べると「太平の眠りをさます上喜撰たった四杯で夜も眠れず」と同じことになる。

色の変わったところであげる　蕎麦の茹で加減の口伝で、蕎麦は釜の中で噴火口のようにボコボコ沸騰しているのではなく、みんなでそろって釜の手前から奥へと泳いでいる。はじめは黒くしまっているものが、ある時、フッと透き通って見えるようになる。その時が蕎麦がちょうど煮上がったときである。

上下【うえした】職人の職種の一つ。蕎麦が打て、蕎麦を茹でられる、いわば板前と釜前ができる職人を、店で上（板場）と下（釜場）で使うこと。蕎麦が機械でこしらえられるようになってから、板前職人はヒマになった。

ウソッ火　火加減で、釜の真ん中から沸騰している状態。蕎麦釜は、手前から噴き上がり奥で沈むような、湯が釜の中を流れる状態でなければならない。

打ち粉【うちこ】蕎麦を薄く伸ばすときに、くっつかないように表面にまぶす粉のこと。純粋の蕎麦

例 6：カタログ形式（LBj6_00025『熱帯魚・水草カタログ』）

その他の仲間の魚たち

以下のページで「その他の仲間」として紹介している魚たちは、本来、別々のグループに属している種類を、編集の都合上、「その他の仲間」として一まとめにしたものだ。したがって、ハゼの仲間やハタの仲間、そしてカワハギの仲間など、色々なグループの魚が登場してくる。多種多様な魚が含まれるということは、個々の魚の飼育方法や、飼育上の注意点などもすべて異なってくることを意味するので注意してほしい。

その他の仲間の魚の中で、特に人気の高い種類は、パープルファイヤー・ゴビー、ディープウォーター・アンティアス、そして、ロイヤルグラマなどである。これらの魚は、最近特に人気が高まりつつあるサンゴなどの無脊椎動物のレイアウト水槽での飼育に適した小型美魚である。この種の水槽では、自然のサンゴ礁に近い環境が再現されるので、こうしたレイアウト水槽に適した小型の美しい魚たちに注目が集まりつつあるのだ。

346

キイロサンゴハゼ

Gobiodon okinawae　~3 分布/西部太平洋

大きさ 3 cm
水温 24~27℃
難易度 普通
水槽 45cm以上
混泳 同大の温和な魚と
入荷 普通
価格 1,000~2,000円
エサ 冷凍エサ各種、フレークフード、顆粒状乾燥餌、アサリのミンチ、クリル、魚肉、魚卵

●全身が鮮やかな黄色一色に染まっているかわいい熱帯産のハゼの仲間である。あまり活発な魚ではなく、普段は岩の窪みや海藻の葉の上などにちょこんと体を乗せて一休みしていることが多い。輸入量はわりと多く、入手は容易だ。とても小さな魚なので、あまり大きな水槽に泳がせるとどこにいるかわからなくなってしまう。むしろ、60cm前後の小型水槽に無脊椎動物のレイアウトを作り、温和な魚とだけ混泳させて飼育するとよい。

347

以上述べたようなものを分類するために，次のような指標を設けた。

(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）

さらに，文体を吟味する際，「本文」であるのか「前書き」や「後書き」であるのかは区別すべきと考えた。また，「内容」や「表現」の文体判断が困難になるようなものもそれぞれ別扱いすべきと考えた。その結果設けた指標は次のものである。

(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[1]
(k) その他一定量の「本文」が認めがたい

なお, 収録サンプルの中には,「後書き」が「本文」であるテキストが存在する（LBr9_00086『あとがき大全』）。この場合は（e）ではない。引用編集形式であるため，（b）の指標が付与されている。

２.２　アノテーション作業の概要

作業対象と内容は次のとおりである。

●対象テキスト：BCCWJに収録されている図書館サブコーパス（10,551サンプル）の書籍テキスト。
●1テキストの範囲と長さ：コーパス収録テキストの分類指標とするため，その一部を字数を揃えて抽出することはせず，1サンプル全体（平均3,000語）を範囲とする。
●作業ファイル：サンプルを取得した書籍の紙面コピーを参照する。
●作業量：１セット約400～500の書籍テキストに対する指標付与を延べ約10日で行う。
●内容：
下記に該当する場合に指標を付与する。排他的ではなく該当するものすべてを付与する。
(a)対話系，(b)引用系，(c)視覚表現多用系，(d)データベースやリスト系，(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，(k)その他一定量の「本文」が認めがたい

３．アノテーション作業結果

３.１　分類指標の付与結果

今回の対象データである 1,664 テキストに対する NDC 別分類指標の付与結果を表 1 に示す。分類指標は排他的ではないため合計は 1,664 を超える。図書館サブコーパス収録サンプルの NDC 別の数と比率は，図 1 に示すとおり「9.文学」と「5.社会科学」が多い。よって，表 1 で「9.文学」「5.社会科学」が全体的に多いのは，もともとのサンプル数の比率の大きさに寄るところがある。しかしながら，図 2 の NDC 別分類指標の付与比率をみてみると，収録サンプル比率とは異なる次のような特徴を確認することができる。

[1] 厳密には，(j)は文体判断が困難な類型ではない。小中学校の教育現場等において用例表示をする際に避けた方が無難だと思われるような，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目的からは外れるが，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。田野村（2009）は，そういったテキストに対し「日本語の学術的研究という観点からそれらを排除すべき理由は本来ない」が，「危うい内容のデータは排除ないし隔離するという処置を講じる必要があるように筆者には思われる」と述べている。この分類はその試みの一つになると考える。

表 1　NDC 別分類指標の付与結果（1,664 テキスト）

NDC	サンプル数	(a)対話系	(b)引用系	(C)視覚表現多用系	(d)データベースやリスト系	(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	(k)その他一定量の「本文」が認めがたい
0.総記	46	9	12	5	9	8	0	0	2	1	1	1
1.哲学	75	17	20	3	10	21	1	0	0	0	1	7
2.歴史	143	32	20	20	48	26	5	0	0	0	0	6
3.社会科学	355	112	68	10	66	54	3	0	0	13	17	31
4.自然科学	120	30	18	16	35	15	0	3	4	1	1	2
5.技術	180	18	22	57	71	13	0	1	1	3	0	11
6.産業	54	8	2	13	25	5	0	0	0	1	0	3
7.芸術	177	45	18	59	35	12	0	0	0	0	3	11
8.言語	86	11	14	1	39	7	0	16	1	0	0	5
9.文学	339	77	25	1	16	55	5	0	0	0	115	50
n.なし	89	9	10	30	26	5	1	1	0	0	3	5
計	1664	368	229	215	380	221	15	21	8	19	141	132

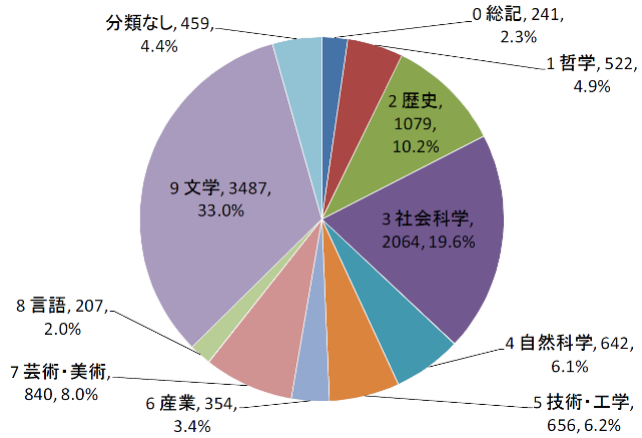


図 1　図書館サブコーパス収録サンプルの NDC 別の数と比率
（DVD 収録『『現代日本語書き言葉均衡コーパス』利用の手引第 1.0 版』（2011 年）より）

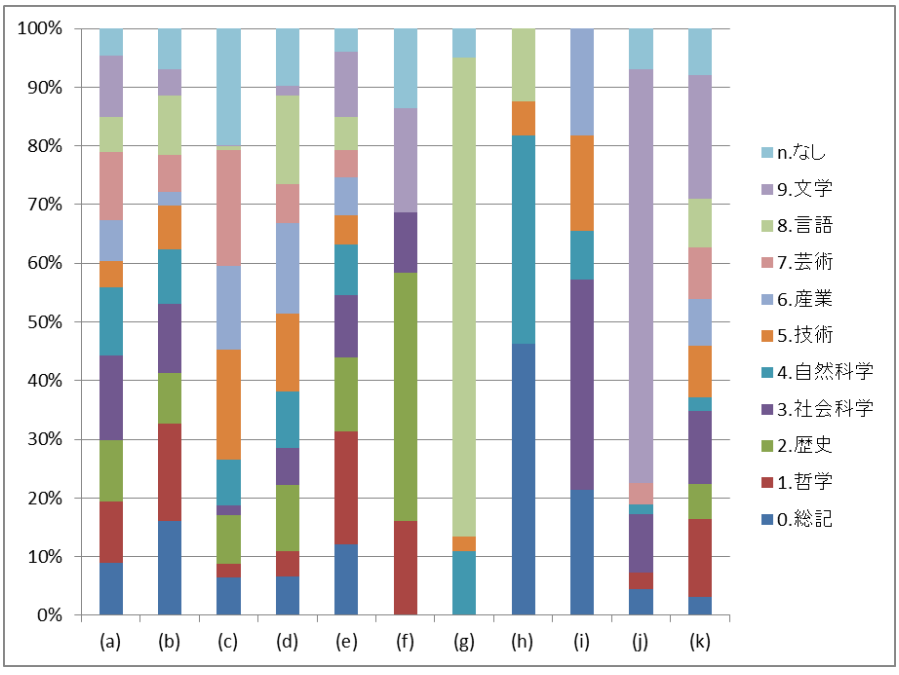


図 2　NDC 別分類指標の付与比率（1,664 テキスト）

・指標の(a)(b)(e)は，NDC の区別なく，広く用いられている形式である。
・指標の(c)は，「5.技術」「7.芸術」，「n.なし」に多い。これは「5.技術」にコンピュータのマニュアル等が多く，そこにキャプチャ画面が多用されていること，「7.芸術」に図画が多く提示されていること，「n.なし」にカタログ状の紙面が多いことに起因すると思われる。
・指標の(d)は，「6.産業」と「8.言語」に多い。「6.産業」には用語解説が，「8.言語」には辞書形式がそれぞれ多用されることによるものと考えられる。
・指標の(f)は，「3.歴史」が多くを占める。歴史を扱うテキストの中で古い言葉が多用されるからであろう。ただし，該当サンプル数はそもそも少ない。
・指標の(g)は，「8.言語」が大半を占める。外国語のテキストで、外国語が本文に入り込んでいるケースが多いためであろう。ただし, 該当サンプル数はそもそも少ない。
・指標の(h)は，「0.総記」「4.自然科学」が大半を占める。前者にはコンピュータのプログラミング言語が，後者には数式が，それぞれ多用されているためであろう。
・指標の(i)は，「3.社会科学」の比率が高い。法学を含むこの NDC で，法律が多く引用されていることがうかがえる。
・指標の(j)は「9.文学」が非常に多くを占める。暴力的な描写や性的な描写を含む小説がこの NDC に入っているためである。

４．おわりに

BCCWJ に収録する図書館サブコーパスの有効活用を可能とするために，「特徴的な類型のテキスト」に分類指標を人手付与した作業結果を報告した。多種多様な形式をもつサンプルがどの NDC にどの程度収録されているかを明らかにした。特に，テキスト形式の選択に関し，(a)対話系，(b)引用系のテキスト形式は NDC の区別なく多用されていること，(c)視覚表現多用系は，「5.技術」「7.芸術」に，(d)データベースやリスト系は，「6.産業」「8.言語」に選択的に多用されていることを確認することができた。

プロジェクト終了に際し, BCCWJ の図書館サブコーパスに収録される 10,551 サンプルの全ての分類結果についてもまとめ中である。その成果報告と分類結果を近いうちに公開する予定でいる。

本成果に基づき，さらに文体的な特徴を支える言語表現の分析を進め，辞書記述への応用を具体的に考えていきたい。

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」に基づくものです。また，BCCWJ の構築は，文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」（平成 18～22 年度，領域代表者：前川喜久雄）による補助を得たものです。

文　献

EAGLES. (1996). EAGLES Preliminary recommendation on Text Typology, *EAGLES Document EAG–TCWG–TTYP/P*, Version of Jun 1996.

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.

柏野和佳子，奥村学(2012)「書籍テキストへの分類指標人手付与の試み―『現代日本語書き言葉均衡コーパス』の収録書籍を対象に―」『言語処理学会第 18 回年次大会予稿集』pp.1260-1263.

柏野和佳子，立花幸子，保田祥(2012)「書籍テキストをその形式，内容，表現に関わる特徴によって分類する」『ことば工学研究会』41,pp.21-29.
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 ー『現代日本語書き言葉均衡コーパス』の収録書籍を対象にー」『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.
柏野和佳子・丸山岳彦・稲益佐知子・田中弥生・秋元祐哉・佐野大樹・大矢内夢子・山崎誠 (2009). 『『現代日本語書き言葉均衡コーパス』における収録テキストの抽出手順と事例』, 特定領域研究「日本語コーパス」平成 20 年度研究成果報告書 (JC-D-08-01), 特定領域研究「日本語コーパス」データ班.
小磯花絵，小木曽智信，小椋秀樹，冨士池優美，宮内佐夜香(2008)「『現代日本語書き言葉均衡コーパス』にもとづくジャンル間の文体差に関わる要因の分析」『社会言語科学会第 22 回研究大会発表論文集』pp.192-195.
小磯花絵，田中弥生，小木曽智信，近藤明日子(2011)「評定実験に基づくテキスト分類尺度の体系化の試み」『現代日本語書き言葉均衡コーパス』完成記念講演会予稿集，pp.47-52.
田野村忠温(2009)「コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発」『人工知能学会誌』24-5,pp.647-655.
間淵洋子，柏野和佳子，山口昌也，高田智和(2010)「コーパスを用いたテキスト分類指標の検討—BCCWJ の文書構造情報分析を中心に—」『言語処理学会第 16 回年次大会予稿集』pp.314-317.
保田祥，柏野和佳子，立花幸子(2012)「総体として印象を与える表現：「語りかけ性」を有すると判断する根拠」『ことば工学研究会』41,pp.3-10.
保田祥，柏野和佳子，立花幸子，丸山岳彦(2012a)「「語り性」を有する書きことばの典型例の分析」『第 1 回コーパス日本語学ワークショップ』予稿集， pp.139-146.
保田祥，柏野和佳子，立花幸子，丸山岳彦(2012b)「「語りかけ性」を有すると判断される書きことばの表現」『第 2 回コーパス日本語学ワークショップ』予稿集, pp.43-50.

関連 URL

EAGLES http://www.ilc.cnr.it/EAGLES/texttyp/texttyp.html
国立国語研究所の言語コーパス整備計画 KOTONOHA http://www.ninjal.ac.jp/kotonoha/
特定領域研究「日本語コーパス」 http://www.tokuteicorpus.jp/