国語研プロジェクトレビュー
NINJAL Project Review　Vol.4 No.1 pp.43–53（June 2013）

書籍サンプルの文体を分類する

Classifying Writing Styles of Book Samples

柏野 和佳子（KASHINO Wakako）

1. はじめに

筆者は，『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）の構築に携わり，その書籍サンプルを収録する過程において，書籍テキストには，実に多種多様な形式，内容，表現方法がとられていることを経験的に知った。それらを類型で整理することによって，書籍サンプル全体を正確に把握でき，かつ，より緻密な文体研究が実現するに違いないと考えた。

そこで，「テキストの多様性を捉える分類指標の策定」プロジェクトを実施した。本プロジェクトの目的は，書籍コーパスをより有効に活用し文体研究を進めるために，①書籍サンプルの多種多様な形式，内容，表現に関わる特徴を捉えるための分類指標を設計する，②コーパス収録サンプルすべてにアノテーションを行う，③その結果の分析，検証を行う，というものである（柏野・奥村 2012，保田ほか 2012a，2012b，2013，柏野ほか 2012a，2012b，2013）。

コーパスへ文体情報を付与することの重要性は，EAGLES（1996）等で早くから議論され，Lee（2001）によって The British National Corpus（BNC）への付与が実現されている。また，BCCWJ に収録されるサンプルテキストの文体を計量的に考察する試みはすでにいくつか行われている（間淵ほか 2010，小磯ほか 2011）。しかしながら，約 1 万という一定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になる。

本稿では，文体を分類するために設計した分類指標について説明し，それらを付与した典型例のサンプルを示すことで，本プロジェクト成果の概要を伝える。

2. 文体分類のアノテーション

2.1 アノテーション作業の概要

書籍サンプルへその文体分類情報をアノテーションする作業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] 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が母集団とされ，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。

● 1 サンプルの範囲と長さ：サンプルの一部を字数を揃えて抽出することはせず，1 サンプル全体を範囲とする。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）なものを内容・表現の文体的特徴の印象判定により細分類する。（→ 2.2 節）
②①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）

2.2　内容・表現の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，発売形態，また，著者情報，形態論情報などが付与されており，それらを利用して，半自動的に種々の観点から分類することは可能である。しかしながら，EAGLES（1996）がコーパスへ付与することが望ましいとあげる，（A）対象読者に想定される読解レベル（難易度），（B）テキストの作成意図，（C）さまざまな文体情報の 3 種に関する情報は C コード以外には与えられておらず，それらの観点によるサンプルの分類や抽出は困難である。そこで，(A）を補う「専門度」，(B）を補う「客観度」，(C）を補う「硬度」「くだけ度」「語りかけ性度」という，あわせて 5 つの分類指標を新たに設計した。EAGLES（1996）でコーパスに備えることが望ましいと議論されている「文体情報」とは，形式性，親疎性，口語性に関わる文体情報だと言える。よって，その形式性，親疎性を問うものとして「硬度」と「くだけ度」の指標を，口語性を問うものとして「語りかけ性度」という指標を設けた。

(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　形式・内容・表現に文体判断が単純にいかない特徴をもつものの分類指標

先の 2.2 節にあげた文体の分類指標の，付与がしがたいと考えた一群の書籍サンプルがある。一つは，テキスト構造・紙面形式に目立つ特徴をもつものである。もう一つは，内容や表現に特定の特徴をもつものである。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。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。なお，いずれの分類指標も該当サンプルには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）

［内容や表現上の特定の特徴］
(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 [2]
(k) その他一定量の「本文」が認めがたい

3. 分類の典型例

各分類指標が付与されたサンプルより，典型例と認められる例を示す。サンプルの出典は，BCCWJ のサンプル ID と書名とで示す。

3.1 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』）

> ＥＦ以外のＲＢ結合タンパク質としては，転写因子ＲＡＸ，Ｔ細胞が活性化するときに誘導されるＩＬ－２，ＧＭ－ＣＳＦ，ＨＩＶ－２などの転写を活性化する転写因子ＥｌＦ－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，ＥｌＦ－１やサイクリンＤのＲＢ結合ドメインにはｌａｒｇｅＴ抗原やＥ１Ａタンパク質と同じようにＬＸＣＸＥというアミノ酸配列が存在する．また，ＲＢタンパク質は骨格筋分化を支配する重要な遺伝子群ＭｙｏＤファミリー（ＭｙｏＤ，ｍｙｏｇｅｎｉｎ，ＭＲＦ４，ｍｙｆ－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…ァァン、シャラ…ァァン…。

[2] 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。

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ぅ…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）客観度：1 とても客観的（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当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らるべきであり、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を一緒にしました。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、中生代の三畳紀のはじめにいた、「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。

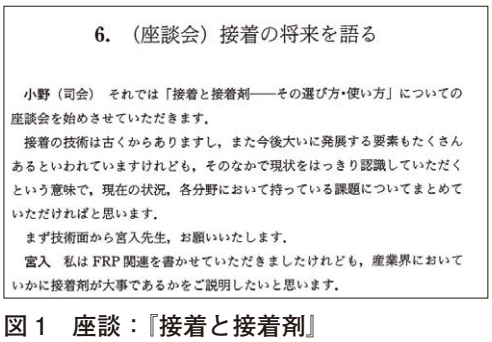
（7）くだけ度：1 とてもくだけている（LBf9_00067『男はオイ！女はハイ…』）

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、こいつひとついってやれとばかりすぐ電話にとびついた。
「ハイ、こちら―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。

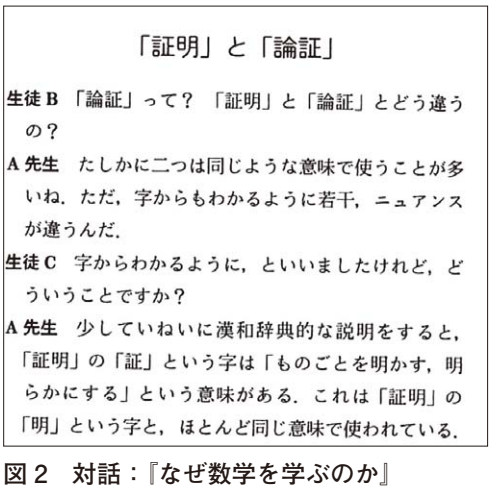
3.2　文体判断が単純にいかない特徴をもつものの分類指標の付与サンプル例



6.（座談会）接着の将来を語る

小野（司会） それでは「接着と接着剤――その選び方・使い方」についての座談会を始めさせていただきます．
接着の技術は古くからありますし，また今後大いに発展する要素もたくさんあるといわれていますけれども，そのなかで現状をはっきり認識していただくという意味で，現在の状況，各分野において持っている課題についてまとめていただければと思います．
まず技術面から宮入先生，お願いいたします．
宮入 私は FRP 関連を書かせていただきましたけれども，産業界においていかに接着剤が大事であるかをご説明したいと思います．

図 1　座談：『接着と接着剤』
（LBd5_00012）



「証明」と「論証」

生徒 B 「論証」って？ 「証明」と「論証」とどう違うの？
A 先生 たしかに二つは同じような意味で使うことが多いね．ただ，字からもわかるように若干，ニュアンスが違うんだ．
生徒 C 字からわかるように，といいましたけれど，どういうことですか？
A 先生 少していねいに漢和辞典的な説明をすると，「証明」の「証」という字は「ものごとを明かす，明らかにする」という意味がある．これは「証明」の「明」という字と，ほとんど同じ意味で使われている．

図 2　対話：『なぜ数学を学ぶのか』
（LBpn_00038）

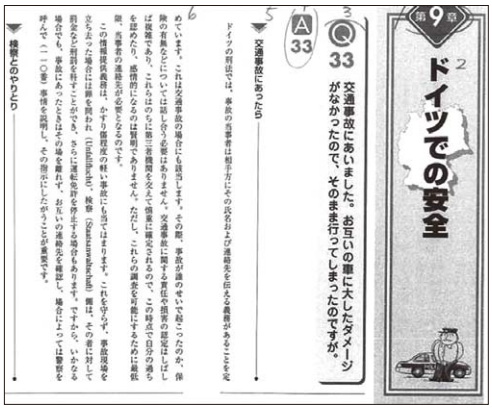


図３　Q&A：『ドイツ暮らしの法律 Q&A』
（LBr3_00037）

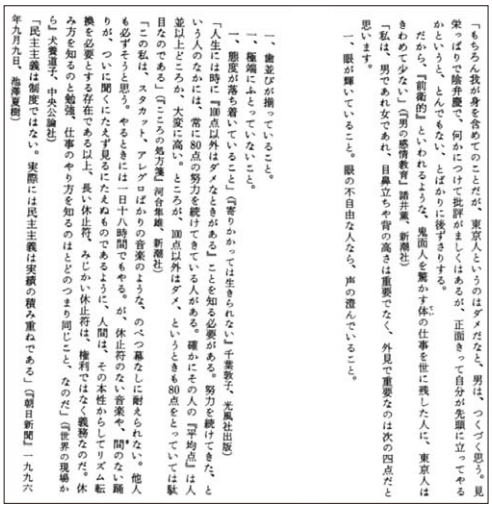

「もちろん我が身を含めてのことだが、東京人というのはダメだなと、男は、つくづく思う。見栄っぱりで陰弁慶で、何かにつけて批評がましくはあるが、正面きって自分が先頭に立ってやるかというと、とんでもない、とばかりに後ずさりする。
だから、『前衛的』といわれるような、先端の仕事を世に残した人に、東京人はきわめて少ない」（『男の感情教育』諸井薫、新潮社）
「私は、男であれ女であれ、目鼻立ちや背の高さは重要でなく、外見で重要なのは次の四点だと思います。
一、眼が輝いていること。眼の不自由な人なら、声の澄んでいること。
一、歯並びが揃っていること。
一、極端に太っていないこと。
一、態度が落ち着いていること」（『寄りかかって生きられない』千葉敦子、光風社出版）
「人生には時に『100点以外はダメなときがある』ことを知る必要がある。努力を続けてきた、という人のなかには、常に80点の努力を続けている人がある。確かにその人の『平均点』は人並以上どころか、大変に高い。ところが、100点以外はダメ、というときも80点をとっていては駄目なのである」（『こころの処方箋』河合隼雄、新潮社）
「この私は、スタカット、アレグロばかりの音楽のような、のべつ幕なしに耐えられない。他人も必ずそうと思う。やるときには一日十八時間でもやる。が、休止符のない音楽や、間のない踊りが、ついに聞くにたえず見るにたえぬものであるように、人間は、その本性からしてリズム転換を必要とする存在である以上、長い休止符は、権利ではなく義務なのだ。休み方を知るのと勉強、仕事のやり方を知るのはまったく同じこと、なのだ」（『世界の現場から』大渡順子、中央公論社）
「民主主義は制度ではない。実際には民主主義は実績の積み重ねである」（『朝日新聞』一九九六年九月九日、池澤夏樹）

図４　引用：『試行錯誤の文章教室』
（LBl8_00017）

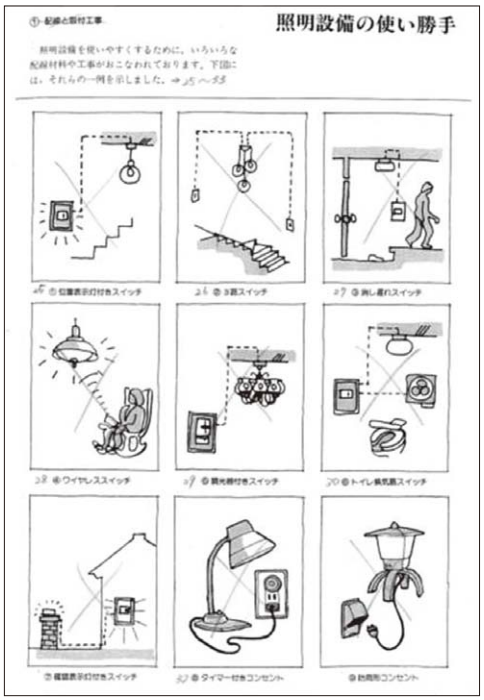


図５　イラスト：『絵ときインテリアライティングの技法早わかり』
（LBb5_00011）

図６　コマ割り：『東京で遊ぼ』
（LBsn_00022）

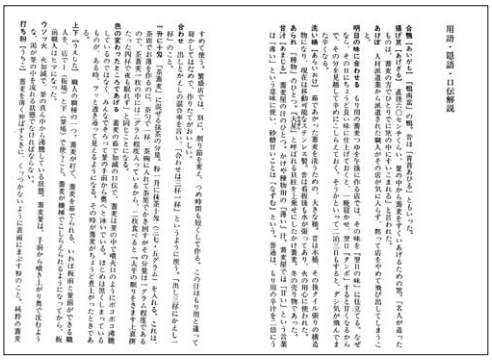

用語・題語・口伝解説

図7　辞書：『蕎麦屋のしきたり』
（LBp6_00009）

図8　カタログ：『熱帯魚・水草カタログ』
（LBj6_00025）

4.　アノテーション作業結果

図書館サブコーパスの 10,551 の書籍サンプルのうち, 2.2 節の「文体的特徴を表す分類指標」の付与対象となったのが 8,887（84%），2.3 節の「文体判断が単純にいかない特徴をもつものの分類指標」の付与対象となったのが 1,664（16%）であった。以下, 分類結果を概観する。

4.1　文体的特徴を表す分類指標の付与結果

分類指標別の付与サンプル数のグラフを図 9（次ページ）に示す（詳細は，柏野ほか 2012b）。専門度は「一般向き」が圧倒的に多い。客観度は真ん中 2 つの指標が多いが，そのうち，「どちらかといえば客観的」の方が多い。硬度はより真ん中 2 つの指標が多いが，そのうちでは「どちらかといえば軟らかい」の方が多い。くだけ度は約 3 分の 1 近くが「どちらかといえばくだけている」である。語りかけ性度は「とても」と「どちらかといえば」をあわせて約 4 分の 1 である。くだけているサンプル，語りかけ性のあるサンプルが一定数以上収録されていることが確認できる。

また，NDC 別に見ると，次のような特徴がある（柏野・奥村 2012）。

- 専門度：文学と芸術・美術以外が高い。中では哲学がもっとも高い。
- 客観度：小説類は本指標付与の対象外であるため，「文学」はエッセー類のみ対象とした。それらエッセー類が低い。歴史もまた低い。高いのは哲学，自然科学。
- 硬度とくだけ度：硬度が高ければくだけ度は低い，という相関する傾向がみえるが，その中で，自然科学だけは硬度とくだけ度がどちらも低い。硬度が低く，くだけ度が高いのは，文学，芸術・美術と技術工学。ほかはその逆。
- 語りかけ性度：自然科学や歴史が高い。文学，芸術・美術が低く，加えて，哲学がやや低めである。

4.2　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結果

該当した 1,664 サンプルに対する NDC 別分類指標の付与結果を表 1 に示す（柏野ほか

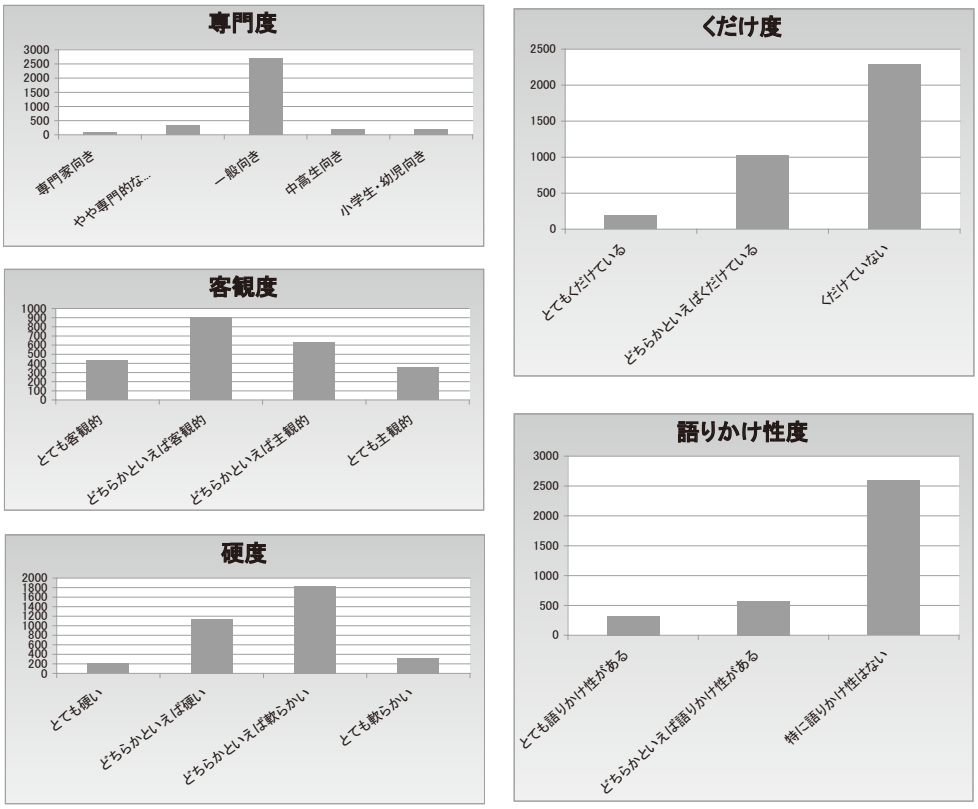


図 9　分類指標の付与結果（柏野ほか 2012b）

2013）。分類指標は複数付与を許しているため合計は 1,664 を超える。表 1（次ページ）で「9. 文学」「5. 社会科学」が全体的に多いのは，もともとの図書館サブコーパスの全体構成における比率の大きさによるところがある。

しかしながら，その影響を抜きにしたうえで，次のような特徴を確認することができる。

- 指標の（a）（b）（d）は，NDC の区別なく，広く用いられている形式である。特に（d）は合計 380 あり，多用されていることがわかる。
- 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（e）は，ランダムサンプリングの結果，こういった箇所からとられたサンプルがどの NDC においても少なくなかったことがわかる。
- 指標の（f）は，「2. 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

表 1　NDC 別分類指標の付与結果（柏野ほか 2013）

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

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」にある。前者にはコンピュータのプログラミング言語が，後者には数式が用いられるためであろう。
- 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」が多い。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。

5.　おわりに

BCCWJ に収録する図書館サブコーパスの書籍サンプルに対して，文体による特定を可能とし，文体研究などへのコーパスの有効活用を図るために，「文体的特徴を表す分類指標」及び，「文体判断が単純にいかない特徴をもつものの分類指標」を設計し，全サンプルにアノテーションを行った。その結果，多種多様な形式をもつサンプルの類型を明らかにし，かつ，各類型に該当するサンプル数がどの NDC にどの程度収録されているかを明らかにした。

分類指標を付与したアノテーション情報は公開予定である。コーパス活用の促進につなげたい。

●参照文献●

EAGLES（1996）EAGLES Preliminary recommendations on text typology, *EAGLES document EAG–TCWG–TTYP/P*, Version of Jun 1996.

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』1260−1263.

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』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』155−164.

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（2013）「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」『第 3 回コーパス日本語学ワークショップ予稿集』63−70.

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』47−52.

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）: 37−72.

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』314−317.

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』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』43−50.

保田祥・柏野和佳子・立花幸子・丸山岳彦（2013）「書きことばにおける「語りかけ」は何のため用いられるのか」『第 3 回コーパス日本語学ワークショップ予稿集』143−152.

《要旨》 文体研究などへのコーパスの有効活用を図るため，コーパスの書籍サンプルを文体によって特徴づけることを目的に，書籍サンプルの分類指標の設計と付与を行った。対象は BCCWJ 図書館サブコーパス収録の全 10,551 サンプルである。テキスト構造が単純（例：章節構造）なもの（全体の 84%）については，内容・表現の文体的特徴により，専門度，客観度，硬度，くだけ度，および語りかけ性度，という 5 観点による分類指標を定め，主観的評定によって評価値を付与した。また，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（全体の 16%）を見出し，その特徴を表す別の指標を設定した。これらを通じて，図書館サブコーパスに収録される全サンプルの多種多様な形式の類型ごとの分布や，各類型の NDC ごとの頻度が明らかになった。

Abstract: To facilitate the use of BCCWJ for writing style studies, we proposed a feature index system that characterizes the writing styles of the book samples in the corpus and then, following the proposed system, annotated all of the 10,551 samples included in its library subcorpus. For the samples with a simple text structure (84%), we chose five axes (specificity, objectivity,

formality, softness, and spokenness) and assigned a five-dimensional index to each sample based on a subjective assessment. For the remaining samples (16%), that is, the samples with a complex text structure or some specific format, we employed a different set of feature annotations. This approach allowed a systematic analysis of the diverse writing styles of the samples included in the library subcorpus. Statistics such as the number of samples with a specific style feature, and correlations between the styles and NDC (Nippon Decimal Classification) categories were obtained.

柏野 和佳子（かしの・わかこ）

国立国語研究所言語資源研究系准教授。富士通株式会社システムエンジニア，情報処理振興事業協会研究員，国立国語研究所研究員，同主任研究員を経て，2009 年 10 月より現職。

主な著書・論文：『岩波国語辞典第七版』（増補改訂，岩波書店，2009），『計量国語学事典』（共著，朝倉書店，2009），『言語処理学事典』（共著，共立出版，2009），『講座 IT と日本語研究 2　アプリケーションソフトの基礎』（共著，明治書院，2011），「和語や漢語のカタカナ表記：『現代日本語書き言葉均衡コーパス』の書籍における使用実態」（共著，『計量国語学』28(4)，2012）．

受賞：第 51 回全国大会奨励賞（情報処理学会，1996），2006 年度研究会優秀賞（人工知能学会，2007），第 12 回年次大会優秀発表賞（共著，言語処理学会，2007）．

社会活動：情報処理学会情報規格調査会学会試行標準専門委員会委員，情報処理学会情報規格調査会学会試行標準 WG3 小委員会主査．

萌芽・発掘型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」
プロジェクトリーダー　柏野和佳子
（国立国語研究所　言語資源研究系　准教授）

プロジェクトの概要

コーパスの書籍サンプルに対して文体の特定を可能とし，文体研究などへの有効活用を図るため，サンプルの分類指標の設計と付与を行った。対象は BCCWJ 図書館サブコーパス収録の全 10,551 サンプルである。テキスト構造が単純（例：章節構造）なものと判断した全体の 8 割強に「専門度，客観度，硬度，くだけ度，語りかけ性度」という 5 観点による分類指標を付与した。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と判断した全体の 2 割弱に，その特徴を表す「対談，Q&A 形式，図解，用語解説」等の分類指標を付与した。その結果，全サンプルの多種多様な形式の類型ごとの分布や，各類型の NDC ごとの頻度を明らかにした。また，特定の文体的特徴をもつサンプル抽出を実現可能にした。

さらに，機械学習による自動判定を行った。難易度について人と機械の判定精度が同程度であることを検証したうえで，BCCWJ 全サンプルに自動付与し，コーパス活用性を向上させた。