『現代日本語書き言葉均衡コーパス』の小説サンプルに対する分類情報付与

国立国語研究所論集 (NINJAL Research Papers) 25: 59–74 (2023)
ISSN 2186-1358

『現代日本語書き言葉均衡コーパス』の小説サンプルに対する分類情報付与

加藤　祥[a]　　浅原正幸[b]

[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 研究系

要旨

我々は『現代日本語書き言葉均衡コーパス』の書籍サンプルに含まれるすべての小説サンプルについて，小説の内容に関するジャンルや舞台設定等の分類情報（「推理」「SF」「アドベンチャー」「ロマンス」など）を付与した。分類情報の策定にあたっては，小説サンプルの取得された各書籍について，書店や出版社の分類情報をはじめ，小説の内容を表すと複数作業者が判断した特徴語句を広く収集し，結果を整理した。各小説サンプルには様々な分類項目を重複して付与した。本稿の作業により，これまで分類されていなかった小説の分類情報が付与された。新たに付与された分類情報により，分類別の語彙分布や文体特徴が確認できるようになった。本稿では，作業手順と情報付与結果を報告する *。

キーワード：現代日本語書き言葉均衡コーパス，書籍サンプル，小説サンプル，分類，アノテーション

1. はじめに

『現代日本語書き言葉均衡コーパス』（以降 BCCWJ）の書籍サンプルの約四分の一は小説のサンプルである。書籍サンプルのジャンル情報としては，日本十進分類法（以降 NDC）に基づく NDC 番号と，日本図書コードの分類コード（C コード）が付与されているため，NDC9 番台「文学」以下の「9X3 小説，物語」（地域をここでは X と示した。X は 1–9 の任意の数字で言語区分を表すに該当する）もしくは C コードの第 3–4 桁が「93 日本文学，小説・物語」「97 外国文学，小説」であることにより，当該書籍サンプルが小説であると確認される。しかし，小説の内容に関する分類は付与されていない。そのため，一般書籍では情報の付与されているジャンル分類や時代分類等はなく，BCCWJ 書籍サブコーパスの四分の一のサンプルが「小説・物語」というフィクションに大別されているのみということになる。小説サンプルの語彙は，たとえば江戸時代を舞台とした推理小説，中世ヨーロッパを舞台とした恋愛小説，現代の高校を舞台とした部活動の青春小説，第二次世界大戦を舞台とした戦争シミュレーション小説などでそれぞれ大きく異なる

* 本稿は国立国語研究所の共同研究プロジェクト「実証的な理論・対照言語学の推進」（プロジェクトリーダー：浅原正幸）のサブプロジェクト「アノテーションデータを用いた実証的計算心理言語学」（プロジェクトリーダー：浅原正幸）および科研費 JP18K00634 によるものである。また，2022 年 8 月 30 日にオンラインで実施された国立国語研究所言語資源開発センターの「言語資源ワークショップ 2022」における発表「『現代日本語書き言葉均衡コーパス』書籍サブコーパスの小説サンプルに対するジャンル情報付与」に対し，頂戴したコメントを反映させた内容を含む。

ことが明らかであり，一まとめに「小説・物語」と扱うことが難しい場合がある。現代日本語ではあまり使わない語彙が頻出すると判断された結果，時代小説が対象外とされる調査も見られる（柏野 2013，石川 2015 など）。そこで，すべての小説サンプルに対し，分類情報を付与することを試みた。但し，日本の小説の分類に関する明確な基準は存在しない。本稿では，複数作業者が各小説の特徴を示すと考えられるキーワードを収集し，分類項目として整理を試みた結果を報告する。本稿の作業の結果，BCCWJ の小説サンプルには，ジャンルや時代に関わる複数の分類項目が付与された。

2. データ

本稿の作業対象は，BCCWJ の書籍サンプルに含まれる小説サンプルである。

2.1 BCCWJ の書籍サンプル

書籍サンプル[1]は，PB: 出版サブコーパス（以降，出版 SC），LB: 図書館サブコーパス SC（以降，図書館 SC），OB: 特定目的サブコーパス（以降，特定目的 SC）の「ベストセラー」である。書籍サンプルは，22,058 サンプル（PB：10,117，LB：10,551，OB：1,390）により構成される。

出版 SC は，2001 年から 2005 年までに国内で発行された書籍・雑誌・新聞を対象とし，そこに含まれる総文字数（推計 65,471,677,099 文字）によって母集団が定義されており，このうちの「書籍」（PB）は，「2001 年から 2005 年までの 5 年間に日本国内で出版されたすべての書籍」の冊数・ページ数を調査して定義された。また，図書館 SC は，「1986 年から 2005 年までの 20 年間に発行された書籍のうち，東京都内のより多くの公共図書館で共通に所蔵されている書籍」を定義するため，東京都立中央図書館で取りまとめられている「ISBN 総合目録」を集計し，総文字数（推計 47,877,656,072 文字）によって母集団が定義された。特定目的 SC の「ベストセラー」（OB）は，あらゆる書籍の中で特に多くの人に読まれたものであり，出版の実態を反映する「出版 SC」の書籍，流通の実態を反映する「図書館 SC」の書籍に対して，一般読者に受容された実態を反映する資料として設計された。1976 年から 2005 年までの 30 年間において，各年のベストセラーとして『出版年鑑』（出版ニュース社）および『出版指標年報』（全国出版協会出版科学研究所）のどちらかで 20 位までに挙げられた書籍（951 冊）を対象として定義された。

出版 SC および図書館 SC は，国会図書館の書誌データ「J-BISC」を用いて層別されており，NDC（一次区分＋「記載なし」）による分類を基準としている。「ベストセラー」は層別が実施されていない。

サブコーパスによって母集団の定義が異なるため，サブコーパスによって書籍の分布も異なっている。出版 SC と図書館 SC やベストセラーでは小説の割合は当然異なり（次の 2.2 節を参照。OB では SC 全体の 46% に上る），さらには小説の分類やその分布も異なることが予測される。

[1] 本稿の BCCWJ のサンプル情報については，丸山ほか（2011）の報告データに基づく。

2.2 BCCWJ の小説サンプル

小説サンプルは，NDC の一次区分で「9 文学」（6,354 サンプル，加藤ら 2019 の増補版；BCCWJ-NDC による）の層に含まれ，「小説・物語」は「9X3」である。NDC「9X3」サンプルは，LB が 2,668 サンプル，OB が 632 サンプル，PB が 1,771 サンプルで，計 5,071 サンプルである（表 1 参照）。

BCCWJ の書籍サンプルにおいて，小説サンプルは約四分の一に該当する 23%，「9 文学」においては 80% を占める。小説サンプルは BCCWJ において多くの割合を占めるものの，その分類以下の細区分は存在しなかった。

柏野ら（2012）は，BCCWJ の LB に付与した文体情報を用い，テキストの硬軟を分析している。硬軟の印象の考察においてテキストの特徴として「親密度の低そうな語」「難解な内容や説明」（硬い）と「平易な語」「平易な内容や説明」（軟らかい）が指摘されている。柏野（2015）の「9X3」1,724 サンプルにおいては，「硬い」が 341 サンプル，「軟らかい」が 1250 サンプルという分布が見られる。小説サンプルは「軟らかい」と判断される割合が高いといえようが，「硬い」と判断される小説がどのような種類であるのかという確認はしにくい。どのような小説が「硬い」のか，「親密度の低そうな語」「難解な内容や説明」を含む小説がどのように分類されるのかなどの疑問に対し，多数の小説を大まかに分類する基準が求められる。

3. 小説の下位分類例とマルチラベルの必要性

BCCWJ では小説の下位分類は付与されていないが，Brown Corpus [2] は，500 サンプル中「Imaginative Prose」を 126 サンプル含んでおり，これらが NDC の「小説・物語」に該当する分類と考えられ，内訳として以下の 6 分類が設定されている。

K. General Fiction（29 サンプル）
L. Mystery and Detective Fiction（24 サンプル）
M. Science Fiction（6 サンプル）
N. Adventure and Western Fiction（15 サンプル）
P. Romance and Love Story（29 サンプル）
R. Humor（9 サンプル）

この分類については，「Adventure and western fiction often includes short stories from fantasy fiction magazines and anthologies. (McEnery & Hardie 2012: 98)」のような指摘や，「It is tempting to believe that this is the case because the corpus compilers felt that these were the most useful, salient, or interesting categories — perhaps these are basic-level genres, or prototypical sub-genres (especially those which keep appearing in different corpora). (Lee 2001)」のような不明瞭であることの指摘が確認されるが，小説の分類が求められることに対応する試みと考えられる。

[2] 本稿での Brown Corpus のサンプル情報については，Francis & Kucera（1979; Revised Edition）に示されたデータに基づく。

石川（2015）は，ほぼ同一の標本抽出基準を共有する一連のコーパスが，Brown Corpus 同様に「小説」以下の内容に関する 6 分類を持つことに着目し，対照を目的とした BCCWJ の小説データセット（K：一般小説（29 種），L：ミステリ（24 種），M：科学小説（6 種），N：冒険活劇小説（29 種），P:恋愛小説（29 種），R:ユーモア小説（9 種）の 6 ジャンル）を選定している。石川（2015）の抽出対象は，対照対象の FROWN/FLOB Corpus [3] にあわせるため，基本的に 1991 年〜 1992 年刊行のサンプルとされ，LB と OB が対象となっており，（a）時代小説など現代以前を主たる舞台とした作品を除き（登場人物のせりふなどに，現代日本語ではない言い回しが多く見られるため），かつ，（b）その他の各ジャンルにあてはまらない作品と定義された。

よって，たとえば，石川（2015）の分類では，ホラー小説，企業小説，社会小説，経済小説，スポーツ小説，青春小説，アクション要素を含まないファンタジー小説など，様々な小説が「K:一般小説」に分類される。また，ヒロイックファンタジー，バイオレンス小説，アクションサスペンスなどが「N：冒険活劇小説」に分類される。この分類にあたり，石川（2015）も「恋愛小説が推理や謎解きの要素を含むことは珍しくないし，同様に，冒険活劇小説に恋愛や推理の要素が混在することもふつうである。また，ある作品が何らかの度合いで恋愛のテーマを扱っている場合，それを，一般小説とするか，恋愛小説とするかの線引きもむつかしい」（p. 8）と指摘している。

各分類における典型例を選定するという目的や，特定の分類項目に該当するテキストを抽出する目的であれば，各サンプルを一つの範疇に分類することは有用であると考えられる。しかし，各サンプルの内容において分類を試みるにあたっては，ある程度多様かつ重複した要素を考慮しなければならない。また，現代語コーパスであるとはいえ，様々な時代や「Western」のような舞台設定は語彙や文体に大きな影響を及ぼす可能性が考えられる。マルチラベルを設定する必要があろう。

4. 本稿における分類情報付与

本稿では，各小説サンプルの特徴および分類に関する情報収集（4.1）と，収集した情報を整理した項目設定（4.2）の 2 段階の作業を行った。作業の結果，各小説サンプルの分類情報として，書店（4.1.1）と出版社（4.1.2）の分類が付与された。また，作業者 3 名が自由に収集した書籍内容に関する特徴的語句（4.1.3）を整理した分類項目（4.2.3）も付与された。なお，情報収集は，主にオンライン上で行い，書店および出版社 Web サイトにおける分類見出しや書籍概要等の説明文，書籍（カバー・帯等）写真を確認したほか，BCCWJ 本文を確認した。

4.1 分類設定のための情報収集

そもそも，小説の内容分類は，書店や出版社で各々行われている何らかの基準があり，一般的

[3]「the Brown family of corpora」と呼ばれる 4 種類（Leech & Smith 2006; アメリカ英語（書き言葉）の Brown Corpus（1961 年），同イギリス英語版の LOB Corpus（1961 年），1990 年代初頭のアメリカ英語（書き言葉）の FROWN Corpus（1991 年），イギリス英語版の FLOB Corpus（1992 年））のうち，1990 年代のコーパスを示す。

にも曖昧な認識があると考えられるが，統一的な分類や明確な定義が存在するのではない。よって，複数作業者が，各小説サンプル（5,071 サンプル）の掲載された書籍について，以下の手順で内容の分類に関連する情報の収集を試みた。

作業者は Web 上で収集可能な小説の分類情報を取得したが，サンプル該当部分の内容を確認するにあたっては，BCCWJ に収められたテキストを参照した場合がある。また，LB に関しては，別プロジェクトで実際の書籍から取得した情報の備考欄に内容に関して記載されていた場合があったため，参照資料とした。

書籍内容の分類に関する情報として，書店分類（4.1.1），出版社分類（4.1.2），書籍内容に関する特徴的語句（4.1.3）の 3 つを以下の手順で収集した。作業は 3 名の言語学を専攻する大学院生が行い，情報が取得されなかったサンプルについては 1 名が再作業を行った。なお，BCCWJ の小説サンプルには，（1）（2）に見られるように，書籍タイトル，副題や巻数，著者名，出版社情報のほか，ISBN などが付されており，調査に利用することが可能であった。

（1）　サンプル ID：LBa9_00043　『陳舜臣全集』第 7 巻，陳舜臣，講談社（ISBN：4061926071）
（2）　サンプル ID：PB49_00642　『四季』秋，森博嗣，講談社（ISBN：4061823531）

4.1.1 書店分類の収集

書店の分類を確認した。まず，作業者は，ISBN で検索が可能な Amazon（https://www.amazon.co.jp/）を参照した。Amazon で調査時に販売がなかったなど，当該書籍の分類が確認できない場合は，紀伊国屋書店（https://www.kinokuniya.co.jp/）などの他書店を順次参照した。（1）では，Amazon から「文学・評論」，紀伊国屋書店から「日本文学全集」の分類情報が取得された。（2）は，Amazon から「日本文学」「ミステリー・サスペンス・ハードボイルド（本）」「講談社ノベルス」が取得された。

4.1.2 出版社分類の収集

作業者は，出版社の分類を確認した。

（1）（2）のサンプルでは講談社の書籍情報（https://bookclub.kodansha.co.jp/）によって検索を行い，シリーズ名情報など追加で取得可能な情報があれば取得した。（2）は「講談社ノベルス」が取得された。その他，各サンプルの書籍について出版社のシリーズの分類特徴（「ライトノベル」「ミステリー」など）が取得できる場合は取得することとした。また，「短編集」「アンソロジー」などの形式についても情報があれば取得した。

4.1.3 書籍内容を特徴的に表している語句の収集

上記の手順において，あらすじや概要などの内容紹介文面が確認できれば，書籍内容を特徴的に表していると考えられる語句を取得することとした。（2）は「ミステリー」が取得された。

この手順においては，「長屋の人情ドラマ」「トラベルミステリー」「ハートフル群像劇」「バイ

オレンス剣豪アクション」「SF スペースロマン」「法廷サスペンス」など，各作業者は，書籍内容に関すると判断した多様な説明語句を収集した。作業者が何らかの分類に有用であると判断した場合には，時代（平安，幕末，バブルなど）やジャンル（学生もの，探偵もの，政治家の半生，国際謀略など）に関わる補足的な語句についても，適宜取得した。

なお，書籍内容の情報が 4.1 の手順から全く取得できない場合には，Web 上で検索した書評や帯の説明画像等を参照して内容を確認した例があった。作業者が備考欄にその旨を付した。特に情報が取得できず（備考欄にその旨を付した），BCCWJ で該当サンプル本文を確認した例もあった。

4.2 分類の設定

次に，第一著者は作業者 3 名が自由に収集した書籍内容の特徴説明語句における重複記述と表記や類似表現を整理し，小説の分類項目として設定することを試みた。書店および出版社の分類項目は「日本文学」のように大きいため，ここでは利用しなかった。なお，「9X3」であっても小説ではないサンプルの除外（4.2.1），該当サンプル範囲の特定（4.2.2）を行った上で，整理（4.2.3）作業を実施した。

4.2.1 小説外書籍の特定

NDC の「9X3」（小説・物語）分類の書籍 5,071 サンプルを分類対象とする。しかしながら，NDC が「9X3」であっても厳密な意味で小説でないサンプルもある。本作業を進めるにあたって 106 サンプル（2.1%）について「小説・物語」ではないと作業者が判断し，小説外サンプルとして分類対象外とした。小説外サンプルの内訳をサブコーパス別に表 1 に示す。具体的にはエッセイ，ノンフィクション・実録，ルポタージュ・記録，解説，戯曲，体験記・日記，名言・ジョー

表 1　本作業における小説外サンプルの内訳（サブコーパス別）

	LB	OB	PB	総計
小説でない（情報付与対象外） ※網掛部は内訳を示す	64 （2.4%）	4 （0.6%）	38 （2.1%）	106 （2.1%）
エッセイ	8	2	2	12
ノンフィクション・実録	16	0	1	17
ルポルタージュ・記録	3	0	4	7
解説	9	0	13	22
戯曲	2	0	0	2
体験記・日記	6	2	1	9
名言・ジョーク	1	0	2	3
落語	7	0	2	9
論考・論評	12	0	13	25
小説（情報付与対象）	2,604 （97.6%）	628 （99.4%）	1,733 （97.9%）	4,965 （97.9%）
9X3 サンプル数	2,668 （52.6%）	632 （12.5%）	1,771 （34.9%）	5,071 （100%）

ク，落語，論考・論評を分類対象外とした。以下では，小説として認定を行った 4,965 サンプル（97.9%）を分類対象とする。

4.2.2 分類の特定

書店分類が「ミステリー・サスペンス・ハードボイルド」「SF・ホラー・ファンタジー / 童話」のような複数要素を含む場合，自由に収集された特徴的語句から「ミステリー」「ファンタジー」のように特定されていることが多い。情報収集作業時に分類が特定されている場合は，特定語句のみを分類とした。「坂本龍馬」「関ケ原の戦い」のような固有名詞など個別的な記述がある場合にも，「幕末」「江戸」のように時代に関する分類に特定した。同時に「伝記」「戦争」などの記述もそれぞれ分類項目として取得した。

また，短編集やアンソロジーについては，サンプル範囲の小説が不明瞭な場合が多いが，複数作品中の作品名とサンプル内容を照らし合わせて作品が特定できた場合のみ情報を取得した。前掲の（1）は，サンプル範囲が「水滸伝」であることが確認されたため，舞台となる時代等の内容情報が付与されており，「中世」「中国」などの情報が分類項目として取得された。

4.2.3 分類の整理

作業者が取得したキーワードが類似していると考えられた場合，たとえば「笑劇」「コミカル」などを「喜劇／コメディ」に含めるなどの整理を行った。また，具体的な動物名は「動物」，バレエや絵画などの種目名を「アート／芸術」のような上位カテゴリラベルで統合した。

作業の結果，以下のような項目に整理された。さらに統合や整理の可能な項目や，本稿の調査で取得されなかった項目も考えられるが，現段階では以下を本稿における小説の分類項目とする。

また，各サンプルに対し単一項目とはしていないため，複数の情報が付与されたサンプルもある。情報収集時に得られた「バイオレンス剣豪アクション」「江戸時代」のような情報は，「バイオレンス」「侍」「アクション」「江戸」という分類情報として 1 サンプルに付与される。また，整理のために暫定的に項目別に整理しているが，同項目内でも「サスペンスロマン」「ファンタジーアクション」「時代ミステリー」などの複数情報が付与された場合がある。

前掲の（1）は「全集」「歴史／時代」「伝奇／怪異」「中世」が，（2）は「ノベルズ」「推理／ミステリー」が付与された。

・形式に関する項目：文庫，ノベルズ，短編集／アンソロジー，絵本，全集，選書
・種類に関する項目（1）：掌編／ショートショート，パスティーシュ，ノベライズ
・種類に関する項目（2）：伝記／評伝，サーガ／物語，神話／伝説，シミュレーション，実験／トリック，民話／説話，絵巻，活劇／オペラ
・一般的な書店や出版社の分類項目（対象と見られる項目が含まれる）：SF，推理／ミステリー，歴史／時代，文学／文芸，ホラー／怪談，ライトノベル／少女小説，児童／子ども
・テーマに関する項目：恋愛，青春，友情，事件，テロ，冒険／アドベンチャー，心理／サイコ，

思想／哲学，人生／成長，成功，風刺／皮肉
・トピックに関する項目：私，社会，日常／暮らし，災害／パニック，宗教，ジェンダー／女性，幻想／ファンタジー，暗黒／闇，ヒューマン／人間，ピカレスク／犯罪，伝奇／怪異，アクション，プロレタリア，暴力／バイオレンス
・主要登場人物に関する項目：刑事／警察／捕物職，探偵，侍／武侠，スパイ／暗殺者／忍者，英雄，芸能，王侯貴族，サラリーマン，任侠／マフィア，魔法／超能力，仇討／復讐，家族／親子，群像，紀行／旅情／ロード
・ジャンル等に関する項目：戦争，経済／ビジネス，企業，金融／銀行，アート／芸術，音楽，医療／看護，病気，健康，建築／デザイン，軍事／ミリタリー，格闘／戦闘，鉄道／交通，ギャンブル／ゲーム，スポーツ，政治／外交，裁判／訴訟，サイバー／電脳，グルメ／料理，科学，動物，植物，ガーデニング，ジャーナリズム，詐欺／不正，陰謀／謀略，タイムスリップ
・時代に関する項目：紀元前，古代，中世，近世，近代，第二次世界大戦，戦後，現代，近未来

なお，特に国内が舞台の場合については，平安，鎌倉，南北朝，室町，戦国，安土桃山，江戸，幕末，明治，大正～昭和初期，昭和，昭和後期～平成の各時代名が判断可能な場合，国内の時代名を付与した。たとえば，「中世」に該当する場合で，国内の時代名が特定され「鎌倉」「南北朝」「室町」「戦国」「安土桃山」が付与されている場合がある。また，国内が舞台ではないが高頻度である中国の「後漢」も付与した。

・場所等に関する項目（1）：異世界，地球外，宇宙，国際などのほか，国名や地域名
・場所等に関する項目（2）：海洋，山岳，家庭，学園／学校，法廷，監獄
・傾向に関する項目：エンターテインメント，オカルト，官能／アダルト，喜劇／コメディ，教養，グロテスク／残酷，ゴシック，サスペンス，スリル／スリラー，ハード，ハードボイルド，ハートフル／ハートウォーミング，悲劇，ノスタルジー，ペーソス，メルヘン，ユーモア，ロマン，ロマンス，BL（ボーイズラブ）

これらの多様な項目の一部については，石川（2015）の分類結果と対照することが可能である。しかし，石川（2015）の「M：科学小説」分類のサンプルについては，SF ラベルまたは「科学・生物」ラベルが付与されほぼ一致した結果であったが，「R:ユーモア小説」では異なる結果となった。「ユーモア小説」は，石川（2015）が内容紹介に「ユーモア」「ユーモラス」「笑い」などの語句が含まれていることを根拠に分類を行ったのに対し，本稿の作業では作業者が小説内容を特徴的に表す語句を収集したため，手順の違いが影響した結果であると考えられる。結果として，石川（2015）の「R：ユーモア小説」分類のサンプルは，本稿の作業においては「ユーモア」が直接的に取得された場合でなければ，「企業」「日常」「ミステリー」「ライトノベル」などの分類のみが付与されており，「ユーモア」の分類は付与されなかった。

5. 結果

本稿の作業により，BCCWJ の小説サンプルに対し複数の分類情報が付与された。この結果，

サブコーパスによってサンプリングされた小説に特徴的な分布が見られた（5.1）。また，分類によって語彙の分布が異なる傾向が確かめられた（5.2）。分類による文体的な印象も異なるといえる（5.3）。

5.1 サブコーパスによる小説分類の分布

2.1 節で見たように，BCCWJ のサブコーパスは母集団の定義が異なるため，サブコーパスによって小説の種類分布が異なる可能性が高い。以下では，サブコーパス別に小説分布集計結果を例示する。

5.1.1 書店・出版社分類項目

表 2 にサブコーパス別の一般的な書店・出版社分類として収集された分類情報に関する付与結果を示す。1 サンプルに複数情報が付与されているため，「時代」「ミステリー」「ライトノベル」が同時に付与された例なども含まれる。反対に，書店・出版社で「日本の小説」のように位置付けられ，分類に関する情報が取得されなかった場合など，付与のないサンプルもある。

表 2　書店・出版社分類項目サブコーパス別分類情報付与結果（サンプル数）

	LB	OB	PB	総計
推理／ミステリー	673	112	301	1,086
歴史／時代	459	159	282	900
ラノベ／少女	279	18	181	478
文学／文芸	135	41	72	248
SF	104	20	79	203
児童／子ども	44	15	84	143
ホラー／怪談	100	8	32	140

表 2 から分布の違いが確認できる。小説サンプル数（表 1 参照）における割合を見ると，推理／ミステリーは特に LB（LB（2604 サンプル中 673 件, 以降同様）26%, OB（628 サンプル中 112 件）18%，PB（1733 サンプル中 301 件）17%），児童／子ども向けは PB（PB：5%，LB：2%，OB：2%），歴史／時代は特に OB（OB：25%，LB：18%，PB：16%）で高い傾向がある。

5.1.2 時代に関する項目

前節の表 2 から時代小説（「歴史／時代」）の割合も多いと確認されたため，以下の表 3 に時代分布をサブコーパス別に示す。時代に関する情報は，1 サンプルに 1 情報のみが付与されている。なお，複数時代に渡る場合や，海外が舞台であるなどの場合は，「中世」「近世」「近代」などの大まかな分類が付与された。PB では江戸時代の割合が高いが，OB では戦国時代の割合が高く，LB では江戸時代が多い傾向はあるものの広く分布が見られているという違いが見られる。

また，書籍サンプル全般でも戦国時代から江戸時代，幕末から明治時代を時代設定とした小説が多いことから，語彙や文体への時代的な設定の影響が推測される。

表 3　時代項目サブコーパス別分類情報付与結果（サンプル数）

	LB	OB	PB	総計
紀元前	19	8	6	33
後漢	15	4	8	27
古代	22	0	24	46
平安	27	8	27	62
鎌倉	14	6	5	25
南北朝	6	0	2	8
中世	17	2	16	35
室町	5	0	0	5
戦国	43	58	34	135
安土桃山	14	2	10	26
江戸	157	22	112	291
近世	14	0	10	24
幕末	37	17	27	81
明治	24	19	16	59
近代	10	0	14	24
大正～昭和初期	5	0	17	22
第二次世界大戦	21	11	20	52
戦後・昭和	24	26	19	69
昭和後期～平成	6	10	3	19
近未来	13	2	9	24

5.1.3 テーマやジャンル等に関する項目

テーマやジャンル等に関する分類分布を表 4 ～表 6 に示す。なお, 以下の表では付与数の多かった項目のみとし，件数の少なかった項目は省略した。

表 4 は，テーマに関する分類情報による集計例であり，「ラブアドベンチャー」のような場合は「恋愛」「冒険」のように複数の情報が付与されている。

表 5 は，トピックに関する分類情報による集計で，「伝奇アクション」「日常を描く私小説」などは複数情報付与対象である。

表 4　メインテーマ項目サブコーパス別分類情報付与結果（サンプル数）

	LB	OB	PB	総計
恋愛	285	90	370	745
冒険／アドベンチャー	111	4	70	185
青春	79	31	46	156
人生／成長	56	30	42	128
事件	8	21	28	57
心理／サイコ	26	0	14	40
テロ	7	12	10	29
思想／哲学	8	10	9	27
風刺／皮肉	11	4	3	18
友情	9	0	7	16

表5　トピック項目サブコーパス別分類情報付与結果（サンプル数）

	LB	OB	PB	総計
幻想／ファンタジー	277	40	184	501
伝奇／怪異	68	0	63	131
ヒューマン／人間	76	22	24	122
アクション	73	6	28	107
社会	40	22	28	90
ピカレスク／犯罪	34	19	16	69
私	31	16	9	56
宗教	9	30	15	54
日常／暮らし	25	0	29	54
ジェンダー／女性	13	12	7	32
暗黒／闇	12	4	13	29
暴力／バイオレンス	13	0	6	19
災害／パニック	4	4	2	10
プロレタリア	3	0	0	3

表6は，背景とされるジャンルに関する分類情報による集計例である。表6に集計した主要ジャンルに関しては，「戦争」や「軍事」，「経済」「金融」「企業」など，類似していても統合が困難なキーワードが多く残ったため，1サンプルにおいて主なもの1つのみが付与された結果となっている。

表6　ジャンル項目（背景）サブコーパス別分類情報付与結果（サンプル数）

	LB	OB	PB	総計
戦争	64	55	79	198
企業	22	24	16	62
医療／看護	22	6	24	52
経済／ビジネス	21	16	12	49
アート／芸術	7	10	17	34
裁判／訴訟	17	5	7	29
動物	12	2	15	29
スポーツ	16	2	3	21
科学	11	6	4	21
軍事／ミリタリー	8	0	13	21
政治／外交	7	0	13	20
謀略／陰謀	4	9	7	20
サイバー／電脳	6	0	10	16
鉄道／交通	6	2	7	15
タイムスリップ	2	0	11	13
音楽	6	2	5	13
格闘／戦闘	12	0	1	13
ギャンブル／ゲーム	6	0	4	10
グルメ／料理	8	0	2	10
病気	3	2	2	7
金融／銀行	5	0	1	6

5.1.4 傾向に関する項目

傾向に関する分類情報については，作業者が直接的に情報収集した語句に基づいて付与した。「ユーモア」のような定義に揺れの生じる可能性がある項目に関しても，あらすじや内容説明，サンプルテキスト本文から何らかの判断を行って付与したのではない。そのため，出版社などの分類や記述がない小説については，一般的な定義で分類される可能性がある場合でも分類項目は付与されていない例が多いと考えられる。しかし，サブコーパス別の分布傾向が明らかな分類項目であるため，表 7 には傾向項目の情報付与結果を示す。

ロマンス小説や BL（ボーイズラブ）小説のような恋愛を主軸とした傾向の小説は PB において高い割合を占めるが，LB では少なく，OB では 1 サンプルもない。大きな分布傾向の差が見られるため，サブコーパスにおける小説分布の特性を意識したデータ利用が求められる場合もあると考えられる。

表 7　傾向項目サブコーパス別分類情報付与結果（サンプル数）

	LB	OB	PB	総計
サスペンス	240	48	137	425
ロマンス	34	0	163	197
ロマン	70	16	31	117
BL	8	0	95	103
官能／アダルト	10	17	73	100
ユーモア	51	28	18	97
ハードボイルド	60	8	25	93
喜劇／コメディ	34	4	30	68
エンターテインメント	13	0	16	29
スリル／スリラー	15	0	11	26
ハートフル／ハートウォーミング	10	0	14	24
オカルト	5	4	6	15
グロテスク／残酷	1	6	6	13
ハード	10	0	0	10
悲劇	6	0	4	10
ゴシック	6	0	3	9
教養	4	0	2	6
ペーソス	0	0	4	4
ノスタルジー	0	0	2	2
メルヘン	2	0	0	2

5.2 小説の分類による語彙分布

本稿の作業結果を用いた語彙分布を確認しておく。小説の語彙と BCCWJ，書籍（PB・LB・OB），NDC9 番台（「9 文学」）との異同を確かめるとともに，語彙の観点において分類が有用であることを検証する。表 8 に，語彙の頻度（百万語あたり）を示す。本稿の作業対象となった「小説」の高頻度語上位 20 位について，BCCWJ 全体，書籍（PB･LB･OB）全体，書籍における「9 文学」の調整頻度（100 万語あたりの頻度）をあわせて示した。

「9 文学」においては「小説」サンプルが 80%を占める。「小説」サンプルにおける助動詞「た」の頻度が，NDC9 番台や書籍における助動詞「た」の頻度を押し上げているものと考えられる。また，格助詞「の」は，書籍で高頻度ながら，小説では頻度が低い傾向が見られる。過去形の文末表現が多いことや，格助詞「の」による連語の用法が少ないことなどが小説の特徴であるともいえよう。

表 8　BCCWJ 全体，BCCWJ における書籍・NDC9 番台・小説の頻度上位語の調整頻度　（100 万語あたり）

lemma	POS	BCCWJ	頻度順	書籍	頻度順	9 番台	頻度順	小説	頻度順
た	助動詞	27,496	7	31,743	7	38,364	2	40,233	1
の	格助詞	48,384	1	51,230	1	38,802	1	37,361	2
て	接続助詞	33,391	3	34,807	3	32,610	3	32,655	3
は	係助詞	31,449	4	34,435	4	31,639	5	32,055	4
だ	助動詞	30,177	5	33,111	5	31,781	4	31,884	5
に	格助詞	34,189	2	36,474	2	30,546	6	30,322	6
を	格助詞	29,496	6	32,593	6	28,775	7	29,359	7
が	格助詞	22,805	9	24,166	9	20,911	8	21,246	8
と	格助詞	21,773	10	22,912	10	18,054	9	17,337	9
為る	動詞	24,508	8	25,374	8	16,373	10	15,990	10
居る	動詞	10,717	13	11,846	13	11,715	12	11,894	11
も	係助詞	12,080	12	12,191	12	11,921	11	11,864	12
の	準体助詞	10,632	15	10,254	15	11,369	13	11,520	13
で	格助詞	12,699	11	12,246	11	9,686	14	9,614	14
言う	動詞	7,677	18	8,020	16	7,367	15	7,130	15
ない	助動詞	6,090	20	6,600	18	6,751	16	6,903	16
有る	動詞	9,147	16	10,421	14	6,632	17	5,936	17
事	名詞	7,083	19	7,835	17	5,934	18	5,782	18
無い	形容詞	4,382	24	4,828	22	5,311	19	5,396	19
成る	動詞	5,392	22	5,665	21	4,374	20	4,255	20

5.3 小説の分類と文体指標分布

小説の分類は，文体と関わる可能性が考えられる。

柏野（2015）は LB の全サンプルについて，「専門度」「客観度」「硬度」「くだけ度」「語りかけ性度」の 5 つの観点の文体指標が付されたデータであり，小説サンプルについては「客観度」を除いた「専門度」「硬度」「くだけ度」「語りかけ性度」の 4 観点が付与されている（柏野 2013）。本稿では柏野（2015）の「9X3」データのうち，観点情報が付与された 1572 サンプル[4] を用い，本稿の作業で付与された小説分類と重ね合わせて文体指標の分布を確かめることとした。表 9 に，書店・出版社の小説分類項目と柏野（2015）の文体指標を重ね合わせて集計した結果を示す。専門性度の「1 専門家向き」は該当サンプルがなかったため割愛した。

専門性度は，「文学／文芸」「歴史／時代」分類で「2 やや専門的な一般向き」が見られたサン

[4] 柏野（2015）は，「歴史／時代」分類の一部や「官能／アダルト」分類などの小説を「明治時代より以前の古い言葉が多い」「教育現場で使いがたそうである」としてアノテーションの対象外としている。

プルもあったが，小説全般では「3 一般向き」が多い傾向にある。しかし，「ライトノベル」分類において「4 中高生向き」が最も多い，「児童」分類では「5 小学生・幼児向き」が最も多いという明らかな特徴が表れていた。

硬度は，「歴史／時代」分類で「1 とても硬い」の付与されているサンプルが多く見られ，「2 どちらかといえば硬い」をあわせると半数程度が硬い文体と判断されていることがわかる。また，「ライトノベル」「推理／ミステリー」分類において「3 どちらかといえば軟らかい」の付与されたサンプルの割合が高く，気軽な読み物である可能性が考えられる。

くだけ度は，小説全般で「2 どちらかといえばくだけている」の割合が高いが，「歴史／時代」分類は「3 くだけていない」サンプルが高い割合となっており，「ライトノベル」では反対に「1 とてもくだけている」サンプルの割合が高い。また，「推理／ミステリー」分類では「3 くだけていない」サンプルの割合が「ライトノベル」より高い傾向があるため，硬度において軟らかい印象であっても，くだけているほどでもないという特徴が考えられる。

語りかけ性度の分布は，「児童／子ども」分類で語りかけ性度のある（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）サンプルの割合が 40% 程度と高いものの，小説全般の傾向としては，語りかけ性が感じられるサンプルの割合は 14% 程度にとどまる。

表 9　書店・出版社の小説分類と文体指標（サンプル数，分類なし・分類の重複あり）

文体指標／分類	ラノベ	児童	SF	ホラー	ミステリー	文芸	時代
2 やや専門的な一般向き	0	0	0	0	0	1	2
3 一般向き	58	0	50	62	345	54	206
4 中高生向き	152	10	23	5	33	3	10
5 小学生・幼児向き	1	25	1	1	4	1	1
1 とても硬い	2	0	1	0	1	0	14
2 どちらかといえば硬い	20	1	17	14	62	9	81
3 どちらかといえば軟らかい	137	21	50	43	283	42	116
4 とても軟らかい	52	13	6	11	36	8	8
1 とてもくだけている	64	2	6	5	21	1	6
2 どちらかといえばくだけている	115	23	44	38	211	40	82
3 くだけていない	32	10	24	25	150	18	131
1 とても語りかけ性がある	7	4	0	3	5	2	2
2 どちらかといえば語りかけ性がある	20	10	5	11	46	9	21
3 特に語りかけ性はない	184	21	69	54	331	48	196
分類付与のあったサンプル数	211	35	74	68	382	59	219

以上のことから，専門性度，硬度，くだけ度の文体指標は，小説の分類と関連のあることが推測され，文体分析に小説の分類が有用といえる。

6. おわりに

BCCWJ に含まれる 4,965 の小説サンプルに対し，分類情報を付与した。書店や出版社でも小説分類は不明瞭であり，各々の小説において複数の分類が重複するという問題もあった。そこで，

複数作業者が広く収集した分類情報や特徴語句を整理することにより，分類項目を策定することとした。細かな特徴語句を整理したという性質上，さらに統合や整理の可能な項目の残ることが考えられる。本稿の調査では BCCWJ に含まれる小説サンプルを対象としているため，取得されていない分類項目もあると推察され，汎用的な小説の分類情報としては問題が残るといえる。

また，本稿の作業では，各サンプルに複数の分類情報を付与した。「江戸時代設定でユーモラスに探偵が謎を解く」という特徴の小説は，「時代小説」「江戸」「ミステリー」「ユーモア」「探偵」の各項目で分類が可能となった。

本稿の作業により，今までジャンルや時代等の区別がなくひとまとまりに「小説」とされていたサンプルについても，分類を利用した研究が可能となったといえる。そもそも BCCWJ の書籍サンプルは，PB，LB，OB の 3 種類のサブコーパスによって構成されているため，サブコーパスによって含まれる小説種類の分布が大きく異なる（5.1）。さらに，小説の分類によって語彙（5.2）や文体（5.3）の特徴も異なっていることが検証された。今後，小説の分類情報を活用した分析が求められる。

参照文献

石川慎一郎（2015）「FROWN/FLOB Corpus および BCCWJ データの再構成に基づく英日対照言語研究用小説テキストデータセットの構築の試み―English-Japanese Modern Fiction Corpus（EJ-MoFiC）の概要―」『コーパス頻度データの統計的加工』（統計数理研究所共同研究リポート 340）1–18.

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.

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

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（2012）「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.

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 2019 発表論文集』155–160.

丸山岳彦・山崎誠・柏野和佳子・佐野大樹・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子（2011）『『現代日本語書き言葉均衡コーパス』に含まれるサンプルおよび書誌情報の設計と実装』文部科学省科学研究費特定領域研究「日本語コーパス」データ班.

Lee, David. Y. W. (2001) Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology* 5(3): 37–72.

Leech, Geoffrey & Nicholas Smith (2006) Recent grammatical change in written English 1961–1992: Some preliminary findings of a comparison of American with British English. In: Antoinette Renouf and Andrew Kehoe (eds.) *The changing face of corpus linguistics*, 185–204. Amsterdam and New York: Rodopi.

McEnery, Tony & Andrew Hardie (2012) *Corpus linguistics: Method, theory, and practice*. Cambridge, UK: Cambridge University Press.

Francis, W. Nelson & Henry Kucera (1979) *BROWN CORPUS MANUAL* (Revised Edition, 1979) http://icame.uib.no/brown/bcm.html（2022 年 11 月 25 日確認）

関連 Web サイト

国立国語研究所『現代日本語書き言葉均衡コーパス』 https://clrd.ninjal.ac.jp/bccwj/（2022 年 11 月 25 日確認）

Genre Attribute-related Annotations on Fiction Samples in the Balanced Corpus of Contemporary Written Japanese

KATO Sachi[a]　ASAHARA Masayuki[b]

[a]Mejiro University / Project Collaborator, NINJAL
[b]Research Department, NINJAL

Abstract
We categorized genres and settings (e.g., "Mystery," "Science Fiction," "Adventure," "Romance," and "Historical") for all fiction works in book samples from the Balanced Corpus of Contemporary Written Japanese. To design the descriptive genre attributes, we explored the classification items of bookshops and publishers. We also newly defined the classification items by exploring characteristic words and phrases in the fiction contents. Thus, we annotated the designed classification items of genre attributes in a multi-label classification setting. The work described in this study enabled the assignment of new classification information for fiction samples in the Balanced Corpus of Contemporary Written Japanese. The genre attributes enabled us to confirm the distribution of vocabulary and stylistic features. We reported the annotation procedures and results of the classification items of the genre attributes.

Keywords: BCCWJ, book samples, fiction samples, genre attributes, annotation