書誌情報等を利用した成人向け書籍の特定手法の検討

山崎 誠[1] 呉 寧真[1] 近藤 明日子[1] 小木曽 智信[1]

[1]国立国語研究所
{yamazaki, gons, kondo, togiso}@ninjal.ac.jp

概要

現在，国立国語研究所が構築している「現代日本語書き言葉均衡コーパス」の拡張（BCCWJ2）の一部である書籍コーパスは 2006 年～2025 年の間に日本で出版された書籍からランダムに選んだものを収録する方針である．その中にはいわゆる成人向け書籍も含まれる．これらは日本語の実態を知る上では必要なものであるが，教育現場で用例を紹介するような目的にはふさわしくない語や表現が出現する可能性がある．本稿では，成人向け書籍（主に性的な内容を含むもの）を発見する手順を経験的に求める．さらにその情報に基づいて成人向け書籍を自動的に発見するアルゴリズムを確立する方法を探る．

1 はじめに

国立国語研究所では 2024 年度より文化庁委託事業「信頼できる言語資源としての現代日本語の保存・活用のためのデジタル基盤整備事業」を実施ししている．この事業の目的は 2011 年に公開した「現代日本語書き言葉均衡コーパス」（BCCWJ）を拡張し，2 億語規模のコーパスにすることである[i]．拡張の主な部分は 2006 年～2025 年に出版された書籍からサンプルを抽出して構築する書籍コーパスである．

2 背景

BCCWJ2 に収録する予定の書籍は，出版データをもとにランダムに選ばれる．それらの中には成人向けのものが一定数含まれる．これらも日本語の実態としては必要なものであり，わざわざ遠ざける必要はないが，コーパスの利用場面として大学の授業などの教育現場を考えたとき，検索結果に性的な語や露骨な性描写が現れた場合，その扱いよっては教育的配慮を欠いた行為として問題化するおそれがある．そのようなことを防ぐためには，例えば，検索インターフェイスの機能として，検索件数としてはカウントするが，検索結果としては表示しないような工夫があれば，安心して使うことができるだろう．

「成人向け」の意味であるが，例えば「東京都青少年の健全な育成に関する条例[ii]」では，第七条の一として「青少年に対し、性的感情を刺激し、残虐性を助長し、又は自殺若しくは犯罪を誘発し、青少年の健全な成長を阻害するおそれがあるもの」を「図書類等の販売等及び興行の自主規制」の対象の一つとして挙げている．「性，残虐性，自殺，犯罪」がキーワードとして挙げられているが，この中で比較的明確に特徴を捉えやすいものは「性」であろう．「残虐性，自殺，犯罪」は判断基準の客観化が難しい．また，ポリティカルコレクトネスやコンプライアンスに反する表現も同様に問題となりうるが，これらまで対象とするのは行き過ぎと感じる．

3 成人向け書籍の特定

BCCWJ2 では，対象となる書籍からその一部を抜き出し，それをコーパスに収録している（これは BCCWJ 1 やその他の書き言葉コーパスも同じである）．抜き出した部分をサンプルと呼んでいる．そのサンプルが成人向けの内容を含んでいるかどうかを判断する方法として２つ考えられる．

１つ目は，そのサンプルに特定の語が含まれているかどうかでフィルタリングする方法である．特定の語とは人前で声に出して発言できないような，いわば NG ワードを指す．特定の語はある程度は経験的に決めることができるが，網羅的にリストアップすることは難しい．また，新しい表現や文脈に依存した比喩表現によるものは文字列の一致では見つけ

[i] 既公開の BCCWJ を「BCCWJ1」，拡張部分を「BCCWJ 2」と呼ぶ．

[ii] https://www.reiki.metro.tokyo.lg.jp/reiki/reiki_honbun/g101rg00002150.html

出すことが難しい．

２つ目は，サンプルが抜き出された書籍の書誌情報でフィルタリングするものである．書誌情報にはさまざまなものがある．一般的な図書分類のほかにウェブ上の書店のサイトなどでは，成人向けの書籍を表示しようとすると年齢確認が行われるものがある．このようなものは成人向けの書籍であり，性的な表現が出現する可能性が高いだろう．ただし，コーパスに含まれるサンプルはその書籍の一部であるから，そこには性的な表現が含まれていない可能性もある．

特定の語によるフィルタリングと書誌情報によるフィルタリングの違いとしては，前者はテキスト入力が終わってから実施可能になるのに対して，後者はテキスト入力前に実施できる点がある[iii]．

4 特定の語によるフィルタリング

2006 年発行書籍のコーパス収録サンプルを対象に語彙素「セックス-sex」を含むものを調査したところ，28 個が確認された．表 1 はそのうちの一部を示したものである．サンプル内の頻度が高いものは成人向けの内容を含む書籍である可能性が高いと思われ，実際にこのサンプルには教育現場で提示するには不向きな表現が多く見られた．

表 1　語彙素「セックス-sex」が出現するサンプル（2006 年,一部）

サンプル ID	頻度	タイトル
PB064_02432	25	誰にも聞けない女性の医学
PB065_00397	22	男の子・女の子の産み分け法
PB061_03441	8	もしも運命の人が年下だったら恋も人生もうまくいく!
PB069_14103	7	その男、魔王!?

5 図書分類によるフィルタリング

図書分類として有効なものとして NDC（日本十進分類法）と NDLC（国立国会図書館分類表）の 2 つがある. NDC は図書分類として日本で広く普及しているものである. NDLC は BCCWJ2 の構築にあたって，国立国会図書館から書誌データの提供を受けている[1]が，その中に含まれているものである．これらを単独であるいは組み合わせて利用するフィルタリングが考えられる．

NDC では「598.2 結婚医学 : 性生活，妊娠，避妊，出産」が候補となる．NDC 第 10 版では「性に関する雑著は，ここに収める」という注記がある．

表 2 は NDC「598.2」に対応する NDLC の一覧である．対象は，2006 年の出版書籍の書誌情報データである．

表 2　NDC「598.2」に対応する NDLC（2006 年）

分類記号	分類名	サンプル数	割合 (%)
Y85	風俗本	122	92.4
Y75	家庭医学〔保健，衛生，育児，美容〕	1	0.76
EF32	家庭医学・衛生，育児，美容	9	6.82
計		132	100.00

NDC「598.2」のうち，約 9 割が NDLC の「Y85 風俗本」に対応していることが分かる．

次に，NDLC「Y85」に対応する NDC を表 3 に示した．半数以上が上記の 598.2 と対応しているが，それ以外にも「798.5　コンピュータゲーム」などが見られる. これらはアダルトゲームの解説本が多い. また，673.94 は，パチンコ店などを対象とする遊技場を対象とする分類であるが，NDC の注記に「いわゆる風俗遊戯場の経営も，ここに収める」とあり，風俗店を扱った書籍が含まれている．

表 3　NDLC「Y85」に対応する NDC[iv]（2006 年）

分類記号	分類名	サンプル数	割合 (%)
598.2	結婚医学：性生活，妊娠，避妊，出産	122	56.0
673.94	遊技場：パチンコ店，ゲームセンター，カラオケボックス，レンタルビデオショップ	5	2.3
721.8	浮世絵	3	1.4
726.1	漫画. 劇画. 諷刺画	36	16.5
726.101	漫画・劇画論. 諷刺画論	2	0.9
726.1087	漫画図集. 劇画図集. 諷刺画図集	1	0.5
748	写真集	7	3.2
759	人形. 玩具	1	0.5
778.7	各種映画：科学映画，記録映画，教育映画	2	0.9

[iii] ただし，成人向け書籍と分かったからといってそれをコーパスに収録しないわけではない．

[iv] 日本十進分類法 10 版による．

798.5	コンピュータゲーム〈一般〉：テレビゲーム，オンラインゲーム	39	17.9
計		218	100.0

このうち726.1ではじまるものは主に漫画であり，文字を主体としない書籍なのでコーパスの対象ではないが，もともとの国立国会図書館提供のデータにNDCがついていなかったためリストアップされたものである．

なお，2006年の母集団[v]には59086冊の書籍が収められているが，このうち598.2は132冊(0.22%)，Y85が218冊（0.37%）であり，両者の重なりは122冊（0.21%）であった．2006年の発行書籍でコーパスに格納されたのは，1024サンプルで，このうち598.2が2サンプル，Y85が2サンプルでこれらは同じ書籍であった．

6 特定の語と図書分類の組み合わせ

表4は語彙素「セックス-sex」が出現するサンプルにおけるNDCとNDLCの対応を示した．上位の2サンプルはNDLCのY75と対応している．Y75は，「家庭医学〔保健，衛生，育児，美容〕」という内容であり，必ずしも性的な内容であるとは限らない．PB064_05653は「冷え症・むくみ。」というタイトルの健康本であり成人向けの内容ではない．

表 4　語彙素「セックス-sex」が出現するサンプルにおける NDC と NDLC の対応（一部）

サンプル ID	頻度	NDC	NDLC
PB064_02432	25	495	Y75
PB065_00397	22	598.2	Y75
PB061_03441	8	159.6	US52
PB069_14103	7	913.6	Y81
PB069_02407	4	914.6	KH413
PB069_07162	3	933.7	KS153
PB064_05653	2	493.733	Y75
PB065_05400	2	598.2	Y85
PB067_01576	2	772.1	KD521
PB069_00346	2	913.6	Y81
PB069_02226	2	913.6	KH176
PB069_10653	2	913.6	KH626
PB069_15555	2	913.6	Y81
PB061_01075	2	143.5	SB165

[v] コーパスに収録するため，国立国会図書館の書誌データから漫画などの文字を主体としないものや古典などの現代日本語でないものを除外したリスト．

表4のPB069_02226とPB069_15555には，性的な描写が含まれる．この2つはNDLCでY81という分類が充てられている．Y81は，家庭書・娯楽書の下位分類で「小説類」となっているが，NDLCにはほかに文学という分類がある（「KG 日本文学」「KH 作品集」）．したがって，Y81のほうは，娯楽的な内容を中心にした分類であると思われる．

7 出版社から得られる成人向け情報

表5は，NDLC「Y85」を持つ書籍の出版社の情報である（頻度5以上）．1位の大洋図書は2006年に55冊を出版しており，そのうち20冊がY85の風俗本にあたる．2番目の竹書房は215冊発行しており，Y85が12冊であるが，それを上回る27冊がY81という分類である．これらはタイトルを見ると分かるが成人向けの可能性が高いものが多い．また，これら27冊のうち26冊はシリーズとして「ラヴァーズ文庫」（13冊），「竹書房ラブロマン文庫」（13冊）に該当する．これらもフィルタリングの情報になりうるだろう．

表 5　NDLC「Y85」に対応する出版社（2006 年，一部）

出版社	件数	割合(%)
大洋図書 (発売)	20	9.17
竹書房	12	5.50
コアマガジン	11	5.05
二見書房 (発売)	11	5.05
河出書房新社	9	4.13
データハウス	8	3.67
茜新社	8	3.67
ベストセラーズ	8	3.67
角川書店 (発売)	7	3.21
ぶんか社	6	2.75
イーグルパブリシン	6	2.75
宙出版	5	2.29
ジャイブ	5	2.29
近代映画社	5	2.29
メリー出版	5	2.29

8 BL 小説

7節で挙げた「ラヴァーズ文庫」はBL（ボーイズラブ）を扱うシリーズである．BL小説を成人向けとするかどうかは判断が必要であるが，性描写が登場すればそれらは成人向けの内容とみなすことがで

きる. 表 6 にコーパスに収録された 2006 年のサンプルから語彙素「セックス-sex」が出現する日本小説 12 冊をすべて示した. そのうち表 6 で網掛けをした行は BL 小説であり，約 67%に相当する．サンプル ID の先頭に「*」を付けた 5 サンプルは性描写が含まれるもので，教育現場で提示するには不向きなものである．7 節で述べたように BL 小説はすべて NDLC で Y81 という分類になっている．

なお, 網掛けをしていない 3 冊のうち, B069_00346 と PB069_00011 には男女の露骨な性描写があり，この 2 つも教育現場での提示には不向きである．

表 6　語彙素「セックス-sex」が出現する日本小説（2006 年）

サンプル ID	NDLC	タイトル
*PB069_14103	Y81	その男、魔王!?
PB069_00346	Y81	柔肌ざかり
*PB069_02226	KH176	夢に繋がれて
PB069_10653	KH626	モーニング
*PB069_15555	Y81	恥辱の檻
PB069_00011	Y81	花びらがえし
PB069_04790	Y81	短いゆびきり
PB069_08204	Y81	躰だけじゃたりねえよ。
*PB069_14586	Y81	きみがいなけりゃ息もできない
*PB069_14889	Y81	恋は淫らにしどけなく
PB069_14942	Y81	傲慢で残酷な純情
PB069_15476	KH216	吉原有情

9 その他の情報

以上挙げたもののほかにも成人向けかどうかを判断する情報として以下のようなものが考えられる．

・書籍のタイトルやサブタイトル：表 6 にその一部が垣間見えるように，扇情的な表現が使われていたりするものがある．表 6 の PB069_00346 および B069_00011 には「長編官能小説」というサブタイトルがついている．

・挿絵などの画像情報：小説の場合，サンプルとして取られた箇所に挿絵があるものがあり，その絵で内容が判断できる可能性がある．

・書影：表紙の書影がウェブサイトなどで分かる場合があり，それで判断できる可能性がある．

・年齢確認画面: ウェブ上の書店のサイトなどでは，年齢確認画面が出る場合があり，それが出ると成人向けの書籍と判断できる．

・あらすじ：一部の書籍はウェブ上で書店のサイトなどで概要が示される場合があり，それで判断できる可能性がある．

・特徴語：すでに成人向けと分かっているサンプルとそうでないサンプルを比較し，成人向け書籍の特徴語を抽出し，それで判断できる可能性がある．

・AI の活用：AI を使って成人向けの内容かどうか判断できる可能性が高い．あるいは，その他の不適切な内容も発見できる可能性もあるだろう．これも試してみる価値がある．

10 終わりに

本稿ではコーパスに収録されているサンプルから成人向けの内容を持ち，教育上の扱いに注意を要するものについて，どのような情報を用いれば特定できるかを書誌情報を中心に，経験的な知識などを基に探索的に調べた結果の一部を示した．

謝辞

本研究は文化庁委託事業「信頼できる言語資源としての現代日本語の保存・活用のためのデジタル基盤整備事業」に基づくものである。

参考文献

[1] 山崎誠，髙橋雄太，小木曽智信（2025）「「現代日本語書き言葉均衡コーパス」の拡張―BCCWJ2 の構築―」, 言語処理学会第 31 回年次大会，同発表論文集 pp.414-417，https://www.anlp.jp/proceedings/annual_meeting/2025/pdf_dir/Q1-20.pdf