

『BCCWJ 図書館サブコーパスの文体情報』（2015 年公開第 1 版）

○概要

『BCCWJ 図書館サブコーパスの文体情報』は、『現代日本語書き言葉均衡コーパス』（BCCWJ）¹の図書館サブコーパス（LB）²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、表現に関わる文体の特徴を捉えるための分類指標を設計し、②図書館サブコーパス収録のサンプルすべてに文体情報のアノテーションを行った成果を公開用にまとめたデータです。

書籍サンプルへその文体分類情報をアノテーションした作業概要は次のとおりです。

- 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。（→分類①）
 - 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。（→分類②）
- ※一部、分類①と分類②の両方を付与しているものがある。
- 分類①：
 - (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
- 分類②：
 - [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）
否/対談・座談/インタビュー/対話形式/往復書簡/シナリオ/そのほか対話系
 - (b) 引用系（Q&A 形式、投稿形式、その他引用編集形式）
否/Q&A 形式/投稿形式/そのほか引用が多いなど特徴あり
 - (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）
否/紙面にコマ割が多い/図解/そのほか写真やイラストが主
 - (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）
否/用語解説/辞書・辞典形式/見本・カタログ形式/そのほかリスト系
 - [内容や表現上の特定の特徴]
 - (e) 前書きや後書きである 否/該当
 - (f) 明治時代より以前の古い言葉が多い 否/該当
 - (g) 外国語が多い 否/該当
 - (h) 数式やプログラミング言語などが多い 否/該当
 - (i) 法律文が多い 否/該当

¹ 『現代日本語書き言葉均衡コーパス』（国立国語研究所）http://www.ninjal.ac.jp/corpus_center/bccwj/

² 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。

- (j) 教育現場で使いがたそうである 否/該当
(k) その他一定量の「本文」が認めがたい 否/該当

※この文体情報は、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](#)」，「[コーパスアノテーションの基礎研究](#)」による成果の一部です。

※『BCCWJ 図書館サブコーパスの文体情報』の説明は、次の文書をご覧ください。

- ・ 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.[\(pdf\)](#)

○利用方法

『BCCWJ 図書館サブコーパスの文体情報』は次の2種類のデータ形式にて公開します。

- ・ **エクセルファイルデータ** LB_all.xlsx (5.9MB) [Microsoft®2010]にて作成したファイル
- ・ **テキストファイルデータ** LB_all.csv (2.1MB) CSV形式で作成したテキストファイル

※このコーパスは[クリエイティブ・コモンズ 表示 - 非営利 - 継承 2.1 日本 \(CC BY-NC-SA 2.1JP\)](#)



で公開します。

○利用条件

『BCCWJ 図書館サブコーパスの文体情報』を利用した研究成果を発表される場合は、下記の情報を明記してください。

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）

また、下記の文献ほか、参考にした関連文献を明記してください。

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.

○関連文献

[Wakako Kashino, Manabu Okumura \(2010\) An Approach toward Register Classification of Book Samples in the Balanced Corpus of Contemporary Written Japanese, Proc. of PACLIC24, pp.433-438.\(pdf\)](#)

浅原正幸・加藤祥・立花幸子・柏野和佳子(2015)「文体指標と語彙系列の対応分析」『第7回コーパス日本語学ワークショップ予稿集』pp.7-16.

[浅原正幸・加藤祥・立花幸子・柏野和佳子\(2014\)「文体指標と語彙の対応分析」『6回コーパス日本語学ワークショップ予稿集』pp.39-48. \(pdf\)](#)

[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一』『第1回コーパス日本語学ワークショップ予稿集』pp.131-138. \(pdf\)](#)

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回年次大会論文集』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」『ことば工学会』41, pp.21-29.

[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「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. \(pdf\)](#)

[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』 pp.63-70. \(pdf\)](#)

[柏野和佳子・中村壮範\(2014\)「BCCWJ図書館サブコーパスの文体情報検索ツールによるテキスト分析」『第5回コーパス日本語学ワークショップ予稿集』 pp.171-180. \(pdf\)](#)

[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.\(pdf\)](#)

[保田祥・柏野和佳子・立花幸子・丸山岳彦\(2012\)「「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.\(pdf\)](#)

[保田祥・柏野和佳子・立花幸子・丸山岳彦\(2012\)「「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. \(pdf\)](#)

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」『ことば工学会』41, pp.3-10.

[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」は何のために用いられるのか」『第3回コーパス日本語学ワークショップ』 pp.143-152. \(pdf\)](#)

[保田祥・立花幸子・柏野和佳子・丸山岳彦\(2013\)「ベテランは足を保護する」が語りかけるとき」『第4回コーパス日本語学ワークショップ予稿集』 pp.345-354. \(pdf\)](#)

保田祥・柏野和佳子・立花幸子・丸山岳彦(2013)「アノテーターコメントを用いた「語りかけ性」分析の試み - 頻度情報から捉え難いテキスト性質の解明に向けて -」『言語処理学会 第19回年次大会 発表論文集』 pp.358-361.

○開発スタッフ

担当者： 柏野和佳子（専任研究者）

協力者： 丸山岳彦 浅原正幸（専任研究者） 加藤祥（プロジェクト研究員）
田嶋明日香 立花幸子 平本智弥（技術補佐員）

○問合せ先

『BCCWJ 図書館サブコーパスの文体情報』に関するお問い合わせ、ご意見などは、柏野和佳子 (waka@ninjal.ac.jp)まで電子メールにてお寄せください。

Copyright (c) 2015 国立国語研究所