「現代日本語書き言葉均衡コーパス」の拡張 —BCCWJ2 の構築—

山崎誠[1] 高橋雄太[1] 小木曽智信[1]

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

概要

国立国語研究所では，2024 年度より文化庁からの委託事業「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」を開始した．この事業は「現代日本語書き言葉均衡コーパス」（BCCWJ）の拡張として企画されているものである．構築の中心となるのは，BCCWJ の出版サブコーパスの書籍部分の拡張で，2006 年～2025 年の書籍サンプル約 1 億語を現在の BCCWJ に追加するものである.本発表ではその設計について報告する．

1 はじめに

「現代日本語書き言葉均衡コーパス」(BCCWJ)は，2011 年の公開以来，人文系日本語研究を中心として幅広く利用されてきた．しかし，公開から 10 年以上が経過し，いくつかの問題点が出てきた．小木曽(2023)[1]では以下の 3 点が指摘されている．

1. 収録資料が現代語としては古くなっている
2. 経年的な調査ができない
3. 規模が必ずしも十分でない

このようなことを背景に文化庁では 2024 年度から BCCWJ の拡張を目指した「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」を開始し，国立国語研究所がその業務を請け負うこととなった．

なお，便宜上，拡張する部分を BCCWJ2，旧来の BCCWJ を BCCWJ1 と呼ぶことにする．

2 計画・設計の基本方針

2.1 簡素化

基本的に BCCWJ1 を踏襲するが，設計や仕様をかなり簡素化し，円滑に作業が進むようにする（作業コストの軽減）．具体的には，

(1) タグの種類を減らす．「中納言」での表示に関するものを優先し，研究にあまり使われていなそうなタグは採用しない．

(2)(出版点数にかかわらず)毎年一定量の語数(500 万語）を取得する．後述の 3.1 の図に示すように，対象となる母集団の書籍の数は，2015 年ごろからゆるやかに減少している．しかし，対象となる 20 年分の書籍の数が確定するのを待っていると作業が進まないので，各年の差は無視した．

(3) サンプルサイズを大きくする．1 つのサンプルから効率的にデータを確保できるようにする．ランダムに決めた基準ページの前後 10 ページずつの範囲から可変長サンプルを取得する．また，固定長サンプルは必ず可変長サンプルに含まれているようにすることで設計を分かりやすくした．

2.2 アウトソーシング

BCCWJ1 の構築と違って，単純な作業は基本的に外注する．サンプリング，テキスト入力は外注し，形態素解析以降の作業を国語研で行う．

著作権処理も（行うとしたら）外注だが，今回は後述のように実施しない方向で検討中である．

2.3 コアデータの作成

BCCWJ1 と同様に解析精度を高めたコアデータを作成する．

2.4 公開の予定

2025 度末に約 3000 万語を公開予定，以降順次追加し，2028 年度末までに全体の公開を終える．

2.5 提供方法

BCCWJ1 と同様，少納言，中納言，全データ公開（有償契約）の 3 種類とする．

3 書籍のサンプリング

3.1 母集団の決定

BCCWJ1 と同様に国立国会図書館提供の書誌データ（JAPAN/MARC）を利用する．初期データとして 2023 年 12 月 17 日時点でのデータの提供を受けた．レコード数は 6,465,267 件である．上記の日付以降のデータは別途国立国会図書から個別にダウンロードすることになっている．

母集団の決定に当たって BCCWJ1 同様，以下のような書籍は対象外とした．

・外国語で書かれた書籍

・マンガ・絵本，学習参考書・試験問題集，要覧・名簿・電話番号簿，写真集・月報・小冊子・大型本・豆本，官庁資料，外国の教科書，脚本，手稿，書誌・目録，百科事典・一般年鑑，博士論文，地図，ピアノ譜．これらは国立国会図書館の書誌分類である NDCL を元に除外した．

・ページ数不明の書籍

・価格が 2 万円以上の書籍

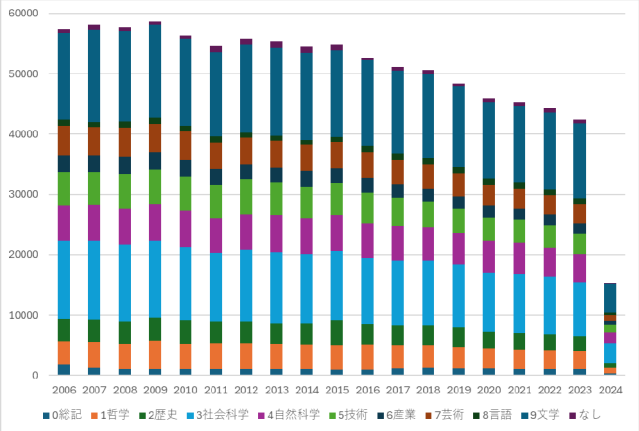

図 1　各年の書籍タイトル数（NDC 別）

また，著作権的に問題が出てきそうな短い言語作品である俳句・短歌・詩なども対象外とした．具体的には書名に「句集・歌集・詩集」が入っているものである．これらの過程を経て 95925 冊を抽出した．

3.2 取得する文字数（語数）の算出

各年において，NDC ごとに必要な冊数を算出した．BCCWJ1 ではすべての書籍を文字数ベースで集計し，その中から 1 文字を選ぶという方式でランダムサンプリングを行っていたが，今回その方式はとらず，書籍 1 冊単位でのランダムサンプリングを実施した．表 1 は 1 冊から 10000 字（約 5000 短単位）取得できたと想定した場合の冊数である．

表 1　2006 年の場合

年	NDC	目標冊数
2006	0 総記	28
2006	1 哲学	57
2006	2 歴史	79
2006	3 社会科学	280
2006	4 自然科学	113
2006	5 技術	90
2006	6 産業	45
2006	7 芸術	71
2006	8 言語	23
2006	9 文学	206
2006	N (なし)	8
計		1000

3.3 C コードの取得

対象読者層を示すメタ情報である C コードが国立国会図書館のデータベースには含まれなくなったことから，新たに JPO 出版情報登録センターの保有しているデータベースを利用し，C コードを取得することにした．

3.4 途中経過

図 2，3 ともに 2008 年の書籍サンプル 368 個の段階での数値である．

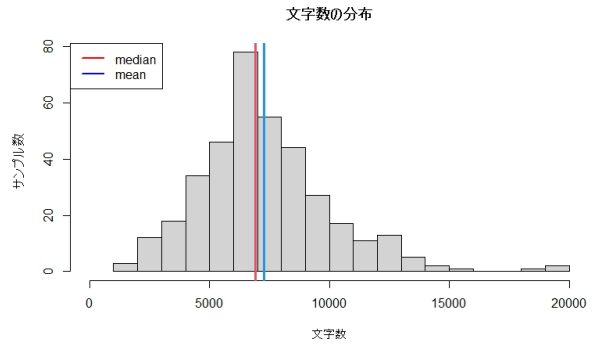

図 2　サンプルあたりの文字数

1 サンプルあたりの平均文字数は約 7300 字である．BCCWJ1 の出版書籍は 1 サンプル平均 5000 字程度なので，約 1.5 倍のサイズになっていることが分かる．

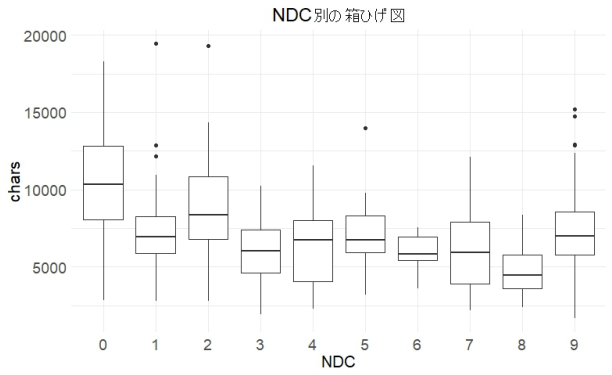

図 3　NDC 別の文字数

0 番台（総記）が多く，8 番台（言語）が少ない．

4 文字入力

4.1 基本方針

文字入力も BCCWJ1 に準拠するが複雑なルールは設けていない．

・文字符号化方式：UTF-8
・文字集合：JIS X 0213:2004
・JIS 水準外の文字は = で入力

入力文字の精度は，99.95%以上である．外注の業者には簡易なタグでルビや注記などを入力してもらい，納品後それを XML に自動変換する．

5 形態論情報

形態論情報としては，BCCWJ1 を踏襲して短単位は mecab+最新の UniDic，長単位は新規に開発した解析ツール Monaka を利用する[2].
https://github.com/komiya-lab/monaka

6 著作権処理

6.1 2018 年の著作権法改正

2018 年の著作権法の改正でコーパスの構築が非常にやりやすくなった．詳細は文化庁著作権課(2019)[3]参照．以下，コーパス構築に与える影響を小木曽(2024)[4]より抜粋する．

30 条の 4：著作物に表現された思想又は感情の享受を目的としない利用　→構築したコーパスの利用・分析等が可能に

47 条の 4：電子計算機における著作物の利用に付随する利用等　→集めた資料の電子化等が可能に

47 条の 5：電子計算機による情報処理及びその結果の提供に付随する軽微利用等　→コーパスの検索サービスが可能に

この権利制限の緩和で，BCCWJ1 のときと比べて大幅なコストダウンが見込まれる．実際に「昭和・平成書き言葉コーパス」（SHC）はこれを受けて構築された[5].

6.2 検索の実態

著作権法 47 条の 5 で規定されている「軽微利用」についての参考となる情報を得るために Web 検索ツール「中納言」の検索ログを取得し，利用者がどのくらいの長さの前後文脈を必要としているかを調査した．図 4，図 5 に結果を示す．

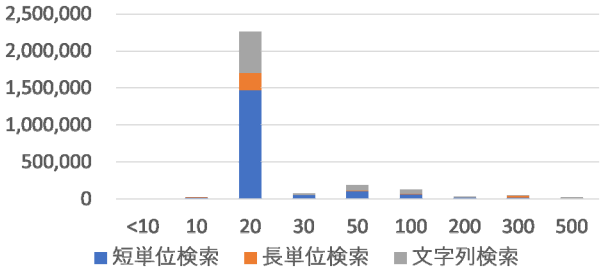

図 4　文脈長別の BCCWJ 検索回数（小木曽・山崎 2024 より）

図 4 から 80%以上の検索が標準（デフォルト設定）の 20 語のままで利用していることが分かった．

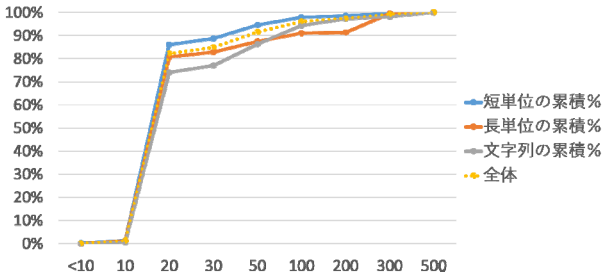

図 5　文脈長別の検索数累積構成比（小木曽・山崎 2024 より）

図 5 からは，前後 30 語で全体の 85%，50 語で全体の 92%を占めることが分かった．仮に 30 語までを軽微利用とすると，軽微利用の範囲を超える利用は 1 割程度となる．

6.3 許諾を得るかどうかの検討

2024 年秋の日本語学会で上記 6.2 の内容について来場者に意見を求めたところ，概ね 20～30 語でかまわないという反応であった．

また，許諾を得るかどうかについては，複数の弁護士事務所と相談を行っているほか，BCCWJ1 の有償契約者にアンケートを行った．

7 SNS データ

BCCWJ2 には，現代の書き言葉として重要な位置を占めている SNS も収録する予定である．現在，API 経由でデータを取得する準備中である．

8 課題・問題点

BCCWJ2 の構築にあたって順風を思われた柔軟な権利制限規定であったが，社会の AI 学習への拒否感が急速に高まってきたこともあり，むしろ許諾を得る行為そのものが効率的なコーパス構築のリスクとなった感がある．

2.2 に述べたようにアウトソーシングによる業務効率化を図ったが，仕様書の理解がまちまちであったことによる対応が多くなったことや，初年度は複数の業者が業務の担当となったため，対応が煩瑣となったことが否めない．

9 おわりに

BCCWJ2 の構築期間は 5 年間（2024～2028 年度）であり，BCCWJ1 とほぼ同じである．BCCWJ1 のノウハウが活かされるものの，今回は構築体制や環境が大きく変わったため，BCCWJ1 のとき以上に集中して取り組まなければならないように思える．今後折を見て途中経過等の発表を行っていく予定である．なお，以下の URL で BCCWJ2 の情報を随時発信していく．

https://www2.ninjal.ac.jp/BCCWJ2/

参考文献

[1] 小木曽智信（2023）「現代語の書き言葉コーパスが果たす役割―『現代日本語書き言葉均衡コーパス』の意義と今後の課題―」文化審議会国語分科会国語課題小委員会（第 60 回）資料 5 https://www.bunka.go.jp/seisaku/bunkashingikai/kokugo/kokugo_kadai/iinkai_60/pdf/93925701_01.pdf（2025 年 12 月 25 日最終閲覧）

[2] 尾崎太亮，古宮嘉那子，浅原正幸，小木曽智信(2024)「歴史的日本語資料を対象とした長単位・文節解析器 Monaka の開発と検証」日本語学会 2024 年度春季大会予稿集,(2024,06,01)

[3] 文化庁著作権課(2019)「デジタル化・ネットワーク化の進展に対応した柔軟な権利制限規定に関する基本的な考え方（著作権法第 30 条の４，第 47 条の４及び第 47 条の５関係）」https://www.bunka.go.jp/seisaku/chosakuken/hokaisei/h30_hokaisei/pdf/r1406693_17.pdf

[4] 小木曽智信・山崎誠(2024)「現代語書き言葉コーパスと著作権処理―BCCWJ2 の構築に向けて―」日本語学会 2024 年度秋季大会予稿集 pp.85-88. https://www.jpling.gr.jp/taikai/platform/2024b/2024b_all.pdf

[5] 小木曽智信・近藤明日子・髙橋雄太・間淵洋子編（2024）「『昭和・平成書き言葉コーパス』の設計・構築・公開」『情報処理学会論文誌』65(2), pp.278-291.