言語資源小委(R8.6.30)
資　料　2

文化審議会国語分科会言語資源小委員会（第14回）R8.6.30

BCCWJ2の進捗状況について

山崎誠（国立国語研究所・客員教授）

目次

- はじめに
- 令和7年度の進捗
  - データ（書籍、教科書、SNS）
  - ツール（形態素解析用辞書、長単位解析ツール）
  - その他（公開、形態論情報の見直し、学会発表）

はじめに

- 本報告は日本語学会2026年度春季大会において実施したワークショップ「BCCWJ2の構築を通してコーパスの新展開を考える」における以下の4つの発表等に基づいています。

- 出版書籍データの設計方針（山崎誠・呉寧真）
- 教科書レジスターの設計：複数時期の教科書の比較分析の可能性（近藤明日子）
- SNS データの収集とコーパス化：データの特徴と整備をめぐる論点（落合哉人）
- BCCWJ1形態論情報の更新（小木曽智信）

コーパス名の整理

- BCCWJ1
- 2011年に公開した「現代日本語書き言葉均衡コーパス」（従来「BCCWJ」と呼んでいたもの）

- BCCWJ2
- BCCWJ1の拡張部分（2024年度から構築が始まった部分）

- BCCWJ
- BCCWJ1とBCCWJ2の両者を合わせた名称

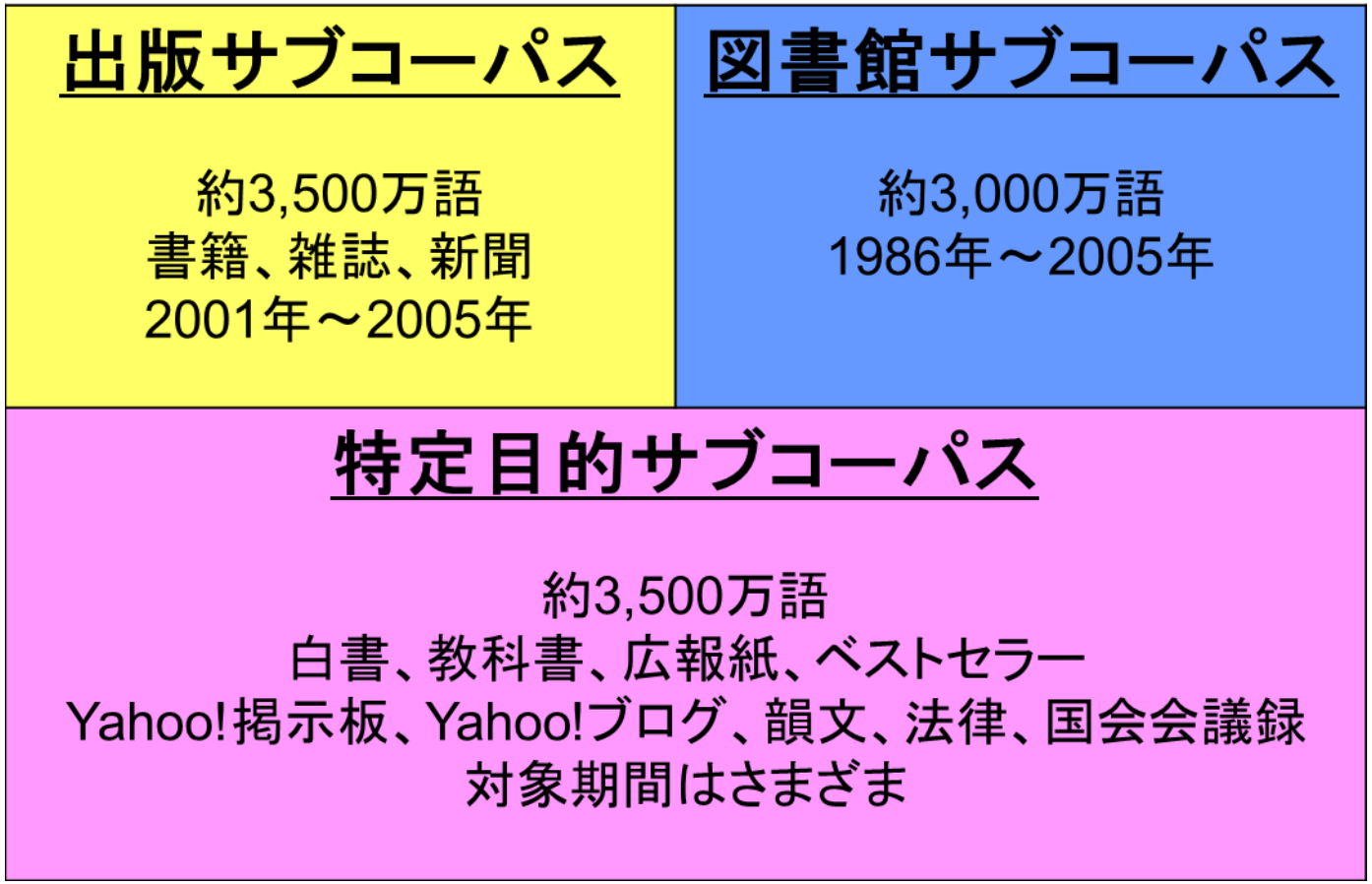
BCCWJ1について

- 『現代日本語書き言葉均衡コーパス』
- Balanced Corpus of Contemporary Written Japanese

- 現代日本語の書き言葉の全体像を把握するために構築したコーパスであり、現在、日本語について入手可能な唯一の均衡コーパスです。（https://clrd.ninjal.ac.jp/bccwj/ ）

- 国立国語研究所を中心に開発
- 2006年構築開始、2011年公開

ＢＣＣＷＪ１の構成



合計約1億語

ＢＣＣＷＪ2の構築

- 出版サブコーパス
- 書籍：BCCWJ1の方法が継承できる
- 雑誌：典拠となる資料が廃刊
- 新聞：著作権者の利益を考慮

- 特定目的サブコーパス
- 教科書
- ＳＮＳ

ＢＣＣＷＪ２の設計方針

- 基本的にBCCWJ1の設計を踏襲し、継続性を保つ（→公平な比較ができることを担保）
- 円滑に構築を進めるために効率化を図る
- 2018年の著作権法改正（柔軟な権利制限規定）に準拠

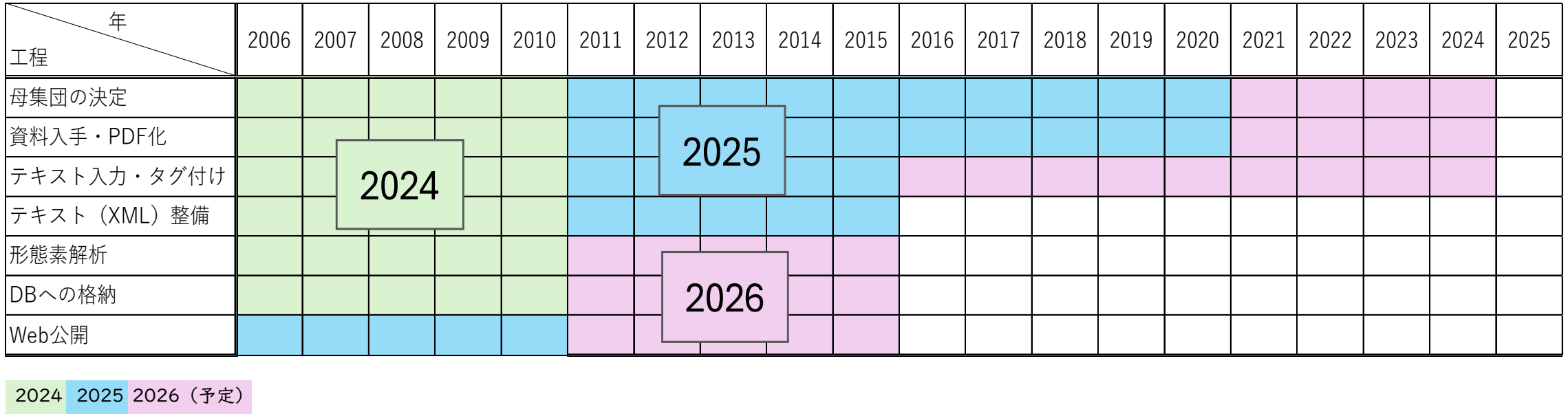
令和7（2025）年度の進捗

- データ
  - 書籍
  - 教科書
  - ＳＮＳ
- ツール
  - 形態素解析用辞書
  - 長単位解析ツール
- その他
  - 公開
  - 形態論情報の見直し
  - 学会発表

書籍

- 2011～2015年刊行書籍のコーパス化：書籍の入手、サンプル部分のPDF化、テキスト入力
- 2016～2020年刊行書籍のコーパス化：サンプル部分のPDF化のみ
- 2006～2009年刊行書籍のコーパス化（補填）：書籍の入手、サンプル部分のPDF化

書籍



➢ 大まかな工程を示す

2011～2015年の集計（暫定）

年	サンプル数	延べ字数（全体）	延べ字数（記号等なし）	延べ語数（全体）	延べ語数（記号等なし）
2011	1378	10,687,897	10,037,592	6,434,099	5,500,600
2012	1404	10,805,563	10,175,959	6,522,791	5,576,426
2013	1299	10,007,925	9,390,758	6,019,476	5,146,135
2014	1184	10,052,689	9,491,273	6,083,907	5,201,218
2015	1226	10,615,198	10,006,509	6,414,173	5,483,567
計	**6,491**	**52,169,272**	**49,102,091**	**31,474,445**	**26,907,946**

- 延べ語数（短単位）は推定値

2016～2020年刊行書籍

- 2016～2020年刊行書籍のコーパス化（サンプル部分のPDF化のみ）
- 各年の目標語数（500万語）に達するに十分なサンプル数を確保

年	サンプル数
2016	1,422
2017	1,400
2018	1,401
2019	1,421
2020	1,400

2006～2009年刊行書籍のコーパス化（補填）

- 書籍の入手、サンプル部分のPDF化まで実施
- テキスト入力は令和8年度に実施

年	サンプル数
2006	232
2007	122
2008	163
2009	88

教科書

- BCCWJ1 における「教科書」レジスターは、「教科書コーパス」（田中・近藤・平山 2011）の一部

- 「教科書コーパス」とは、「1998～1999年改訂の学習指導要領のもとで2005～2007年度に使用された小・中・高校の教科書から各教科・各学年につき占有率の最も高い1 種ずつ、計144 冊の全文」（近藤 2026: 204)を収録したもの。

- 当時の著作権法のもとでは全文の公開が不可能であった。

教科書

- 2018年の著作権法改正により、「教科書コーパス」全体の公開が可能に

- 「BCCWJ2では「教科書コーパス」全体を収録するとともに、その後の2回の学習指導要領改訂期に使用された教科書についても、「教科書コーパス」の設計方針にならいコーパス化し、併せて3期分を「教科書レジスター」として公開する計画とした。」（近藤 2026: 204）

教科書

表１　BCCWJ2 教科書レジスターに収録する教科書の使用年度および冊数

学習指導要領 改訂期	使用年度（上段）/ 冊数（下段）		
	小学校	中学校	高等学校
1998～1999年	2006年度 59冊	2005年度 28冊	2006年度 57冊
2008～2009年 （2015年一部改正）	2011，2018年度 63冊	2014，2019年度 29冊	2012～2015年度 55冊
2017～2018年	2025年度 63冊	2025年度 29冊	2025年度 57冊

- 近藤（2026: 205）より

教科書

- BCCWJ2の教科書レジスターは完成時には「延べ語数約2000万語（記号類を含む推計値）の規模となる見込みである。」（近藤 2026: 204-205）

- 「複数の校種・教科・時期にわたる検定教科書の全文を収録したコーパスの公開は前例がない。」（近藤 2026: 205）

教科書

- 1998-1999年改訂期
- 特定領域研究「言語政策に役立つ,コーパスを用いた語彙表・漢字表等の作成と活用」（研究代表者：田中牧郎、課題番号： 18061008 、期間：2006-2010年度）で作成済み

- 2008-2009年改訂期
- 「2008～2009 年改訂期の中学校教科書については、河内昭浩氏（群馬大学）が作成されたデータ（JSPS 科研費JP25381226 助成）を利用してコーパス化するご許可をいただいた。」（近藤 2026: 204）
- 教科書のコピー入手にあたって文化庁・国語課の協力を得た

教科書

- 2017-2018年改訂期

- BCCWJ1と同じ方針で教科書を選定することにした
- 教科書のコピー入手にあたって文化庁・国語課の協力を得た

令和7年度の進捗

- 2017-2018年改訂期
- 高等学校教科書（52冊）　42冊まで入力完了
- 小学校・中学校教科書（92冊）　PDF化、入力指示作成完了

- 2008-2009年改訂期
- 小学校・中学校・高等学校教科書　PDF化完了

- 1998-1999年改訂期
- 小学校教科書　59冊　PDF化完了

複数時期の教科書の語彙の比較

- 1998～1999年改訂期と2008～2009年改訂期の中学校の教科書の語彙の比較
- 2008～2009年改訂は、いわゆる「ゆとり教育」の見直しが行われた時期

複数時期の教科書の語彙の比較

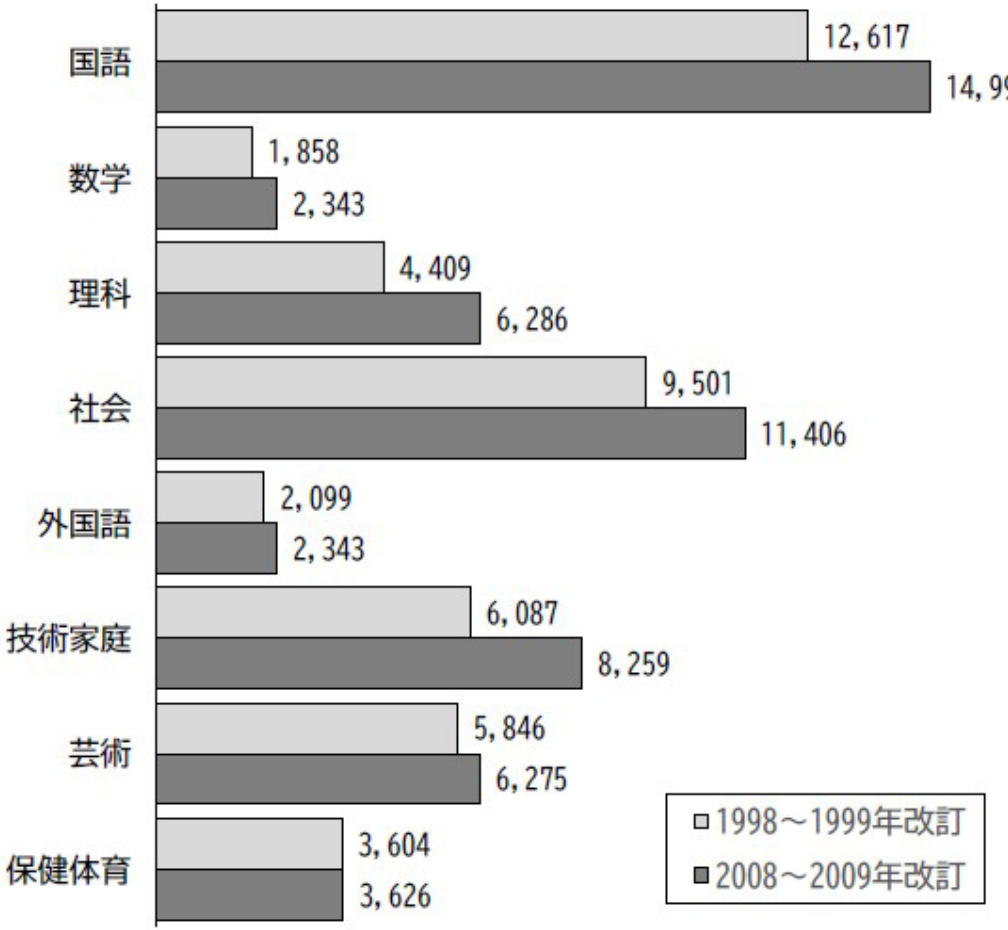


図 1　1998–1999 年改訂期と 2008–2009 年改訂期における中学校教科書の異なり語数の比較

- 近藤（2026: 205）より

複数時期の教科書の語彙の比較

- 「すべての教科において2008～2009年改訂期のほうが異なり語数が多い。増加幅は、理科（1.43倍）が最も大きく、技術家庭（1.36倍）、数学（1.26倍）がこれに続く。」（近藤 2026: 205-206）

- 「この増加は、学習指導要領改訂による授業時間数の増加に伴い、学習内容が拡充されたことに対応するものと考えられる」。（近藤 2026: 206）

ＳＮＳ

- BCCWJ1の構築はSNSが本格的に普及する前に行われた
- BCCWJ1におけるインターネット上の書き言葉のデータは「Yahoo!知恵袋」と「Yahoo!ブログ」

- ツイッター（2008年）
- フェイスブック（2008年）
- LINE（2011年）
- Instagram（2014年）

ＳＮＳ

- 「直近15 年あまりにおいてそのようなインターネット上の書きことばは、主にスマートフォンやSNSで産出されるものへと変化している。」（落合 2026: 207）

- 「そこでBCCWJ2では新たに、Blueskyをはじめとする公開型SNSの投稿と、LINEにおける非公開型の発信の収録を行う。」（落合 2026: 207）

対象と収集方法

- Blueskyについては、「日本語による投稿として判定された（lang:ja の属性を持つ）最新の投稿をAPI（Firehose）を通じて継続的に取得し、10 分刻みで保存」（落合 2026: 207）

- 「比較対象（ないしは第2 候補）として同じく公開型SNS であるMisskey についても同様のデータ収集の仕組みを構築した。」（落合 2026: 207）

- 「いずれも2025年4月起点として継続的に運用を進めており、2026年1月末日時点で約1.2億投稿が保存できている。」（落合 2026: 207）

対象と収集方法

- 「LINE のデータは、個人間のやりとりを対象とするため、「首都圏・20代」「首都圏・30代」「関西圏・20代」の3 つのグループについて提供者を募り、すべての参加者の研究利用・公開の同意を得た上で収集を進めている。」（落合 2026: 207）

- 「2026年3月時点で、3グループ合わせて計120談話、約5万発信が収集できており、」（落合 2026: 207）

未知語の問題

- 「2025年4月に取得したBlueskyの投稿データから2ファイル（2025年4月1日22時10分台および2025年4月28日22 時0分台）をランダムに取り上げて解析したところ、未知語である可能性がある文字列はことなりで約7,500 件抽出された。」（落合 2026: 208）

- 「その多くは（固有）名詞が中心であり、短期間で消滅することから、すべてを辞書に登録することが現実的ではない可能性もある。」（落合 2026: 208）

未知語の問題

表 1：未知語候補として抽出されたもの
のうち、頻出する形式上位 9 件

形式	出現数
ブルスカ	14
リヨ	14
スパコミ	13
ですー	11
いいい	10
イファ	9
タオバ	9
フォロバ	9
みこち	9

- 落合（2026: 208）より

検討課題

- BCCWJ2に収録する規模（cf. BCCWJ1のYahoo!知恵袋、Yahoo!ブログはそれぞれ約1000万語）

- 投稿の種類（例：ポスト、リプライ、リポスト）
- 投稿の時間帯
- ユーザーのバランス
- ユーザーのIDや添付された画像・動画に関して、どの程度の情報を公開するか

- 以上、落合（2026: 208）より

ツール

- 形態素解析用辞書
- 長単位解析ツール

形態素解析用辞書

- 2025.12.31
- 現代語用, 古文用, 方言用のUniDic 14種（ver.2025.12）を更新・新規公開
- https://clrd.ninjal.ac.jp/unidic/

長単位解析ツール

- 長単位解析器であるMonakaの開発（尾崎他 2026.3）
- https://github.com/komiya-lab/monaka

- 2026.4 Web茶まめで長単位の出力が可能に
- https://chamame.ninjal.ac.jp/index.html

Web茶まめ


Web茶まめ
各種UniDicを使用した形態素解析ツール
更新履歴
Web茶まめについて
国立国語研究所
使い方
• テキストボックスに入力した文章か、ファイル選択ボタンからアップロードしたテキストデータに対してMeCabによる形態素解析を行います。
• テキストを入力後、解析前処理の有無／形態素解析に使用する辞書／出力項目／出力形式を選択し一番下にある「解析する」ボタンを押すと形態素解析が実行されます。
テキストを入力
クリア
解析したいテキストを入力
テキストファイルから解析
ファイルを選択
✔複数のファイルを選択してアップロードできます。
▲CSV形式でダウンロードする場合のテキストデータの容量は、1ファイルにつき5MBまでです。
▲辞書、出力項目を増やすと出力ファイルのデータサイズが大幅に増加しますので、ご注意ください。
▲容量が5MB以上のテキストデータは、5MBずつに分けてアップロードしてください。合計ファイルサイズが大きすぎると、失敗する場合があります。
▲Excel形式でダウンロードする場合、テキストデータの容量は1ファイル100KBまでです。
解析前処理
HTMLタグ・《》タグを削除
半角→全角変換
踊り字を展開
カタカナひらがな反転
数字処理
改行処理
辞書バージョン選択
最新版（2025年12月版）
辞書選択 2辞書まで同時に解析・比較することができます。解析を行いたい辞書を最大で2つ選んでください。
現代語
現代語話し言葉
近現代口語小説
旧仮名口語
近代文語
近世江戸口語
近世上方口語
近世文語
中世口語
中世文語
中古和文
上代語
和歌
関西方言
IPAdic(現代語)
出力項目 （短単位）
語彙素ID
語彙素
語彙素読み
語形
品詞
活用型
活用形
書字形(基本形)
発音形出現形
仮名形出現形
語種
発音形(基本形)
仮名形(基本形)
LexID
語彙素ID
出力項目 （長単位）
文節境界
長単位境界
長単位品詞
長単位活用型
長単位活用形
長単位読み

その他

- 公開
- 形態論情報の見直し
- 学会発表

公開

- 2026.3.24
- 2006～2010年刊行の書籍のサンプル約2300万語を公開

- 少納言

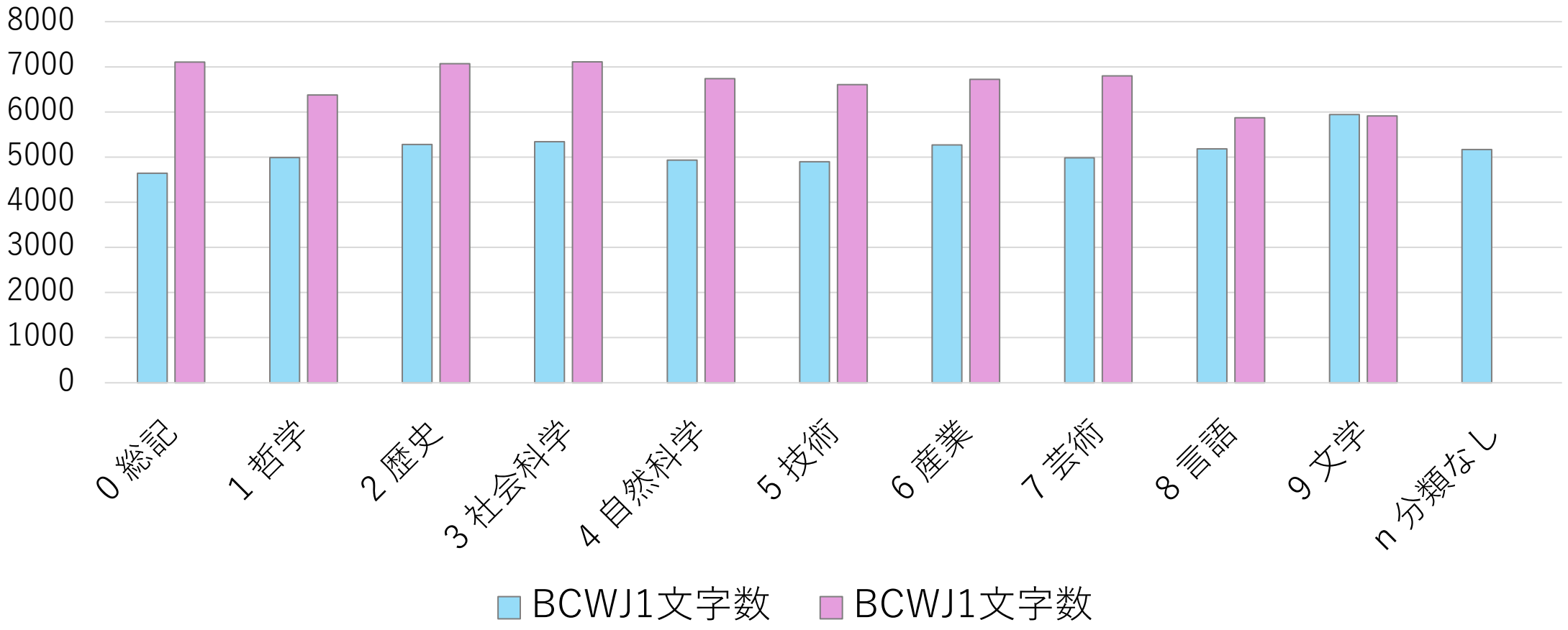
中納言

2006～2010年公開部分の集計

年	サンプル数	延べ字数（全体）	延べ字数（記号等なし）	延べ語数（全体）	延べ語数（記号等なし）
2006	1,018	7,899,470	7,130,319	5,025,851	4,257,160
2007	1,001	8,270,863	7,486,300	5,243,045	4,459,321
2008	1,402	8,664,522	7,829,349	5,509,513	4,675,425
2009	1,233	8,905,477	8,069,401	5,657,744	4,822,167
2010	1,289	9,557,858	8,639,849	6,085,655	5,168,416
計	**5,943**	**43,298,190**	**39,155,218**	**27,521,808**	**23,382,489**

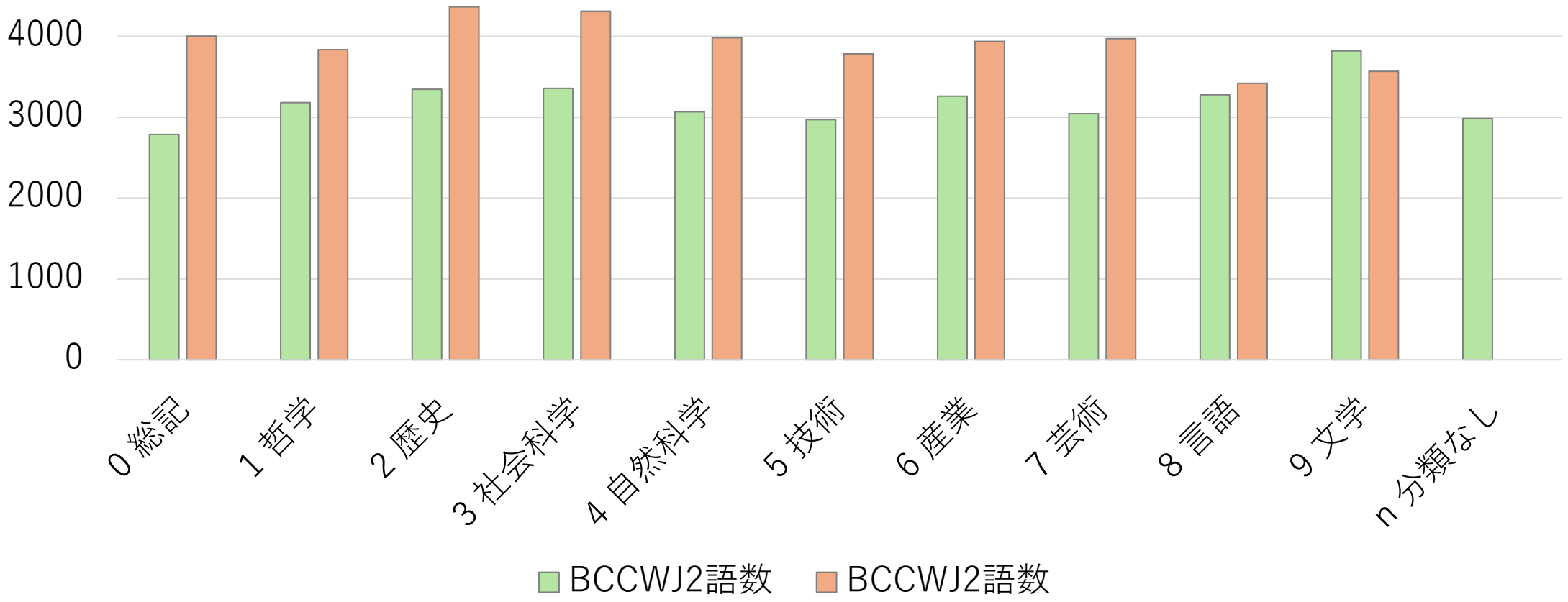
- 延べ字数、延べ語数は暫定値
- 2006～2009は補填を実施中

1サンプルあたりの文字数



- BCCWJ1は2001-2005、BCCWJ2は2006-2010。いずれも記号等を除く

1サンプルあたりの語数



- BCCWJ1は2001-2005、BCCWJ2は2006-2010。いずれも記号等を除く短単位数

形態論情報の見直し

- 既公開のBCCWJ1 の長単位・短単位解析は、「解析精度 98%以上（コアデータについては99%以上）を目標」としており、それを達成している（小椋・冨士池）。（小木曽 2026a: 210）

- しかし人手による十分な修正を経ていない非コアデータの精度は、形態素解析の性能に左右される。上記の数字はコーパス全体の精度であって、レジスターごとに差があり、サンプル単位で見ればこの精度を下回るものも少なくない。（小木曽 2026a: 210）

形態論情報規定の変更による修正

- BCCWJ2の構築にあわせて、形態論情報規程を見直し、BCCWJ1もこれにあわせて修正する。（小木曽 2026a: 211）

- ⑴ 普通名詞から固有名詞を分出
- 野球チーム名の「タイガース」「ドラゴンズ」「ドジャース」などは、「タイガー」「ドラゴン」「ドジャー」等の異語形として普通名詞として扱われてきたが、固有名詞として認定した。（小木曽 2026a: 211）

形態論情報規定の変更による修正

- ⑵ 固有名詞として一語認定
- 「ニンテン｜ドー」「Ｙｏｕ｜Ｔｕｂｅ」「ｉ｜Ｐｈｏｎｅ」など無理に構成要素等に分割していた固有名詞を一語として認定した。（小木曽 2026a: 211）

- ⑶ 接尾的要素の見直し
- 極めて生産性の高い「～屋」「～側」等の名詞を接尾辞（接尾的要素）として認定し、前接する要素と分割した。これは際限ない見出し語の増殖を避けるためである。（小木曽 2026a: 211）

学会発表

- （1）令和7年5月11日，日本語学会2025年度春季大会ポスター発表，「「現代日本語書き言葉均衡コーパス」の拡張とその設計」山崎誠・髙橋雄太・小木曽智信，名古屋大学 https://www.jpling.gr.jp/taikai/2025a/2025a_program/

- （2）令和7年9月11日，日本資料専門家欧州協会(EAJRS)第35回年次大会，「『現代日本語書き言葉均衡コーパス』の拡張：2025年までの2億語コーパスへ」小木曽智信・呉寧真，ハイデルベルク（ドイツ）https://www.eajrs.net/ninjal-2025

- （3）令和6年10月26日，日本語学会2025年度秋季大会ポスター発表，「BCCWJ2のメタ情報の設計―出典情報について―」小木曽智信・山崎誠，オンライン https://www.jpling.gr.jp/taikai/2025b/2025b_program/

学会発表

- （4）令和7年3月7日，第50回社会言語科学会研究大会ポスター発表，「BCCWJ2におけるSNSデータの収録―何を，どのように扱うべきか―」落合哉人，広島大学 https://www.jass.ne.jp/meeting/meeting-list/meeting50/#poster1
- （5）令和7年3月11日，言語処理学会第32回年次大会(NLP2026)ポスター発表，ライトキューブ宇都宮
- ①「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・呉寧真・近藤明日子・小木曽智信https://anlp.jp/proceedings/annual_meeting/2026/pdf_dir/Q2-19.pdf
- ②「BCCWJ2構築に向けた漢字カタカナ文抽出ツールの開発」平林照雄・呉寧真・山崎誠・小木曽智信https://anlp.jp/proceedings/annual_meeting/2026/pdf_dir/Q2-5.pdf
- ③「MonakaによるBCCWJ2 における長単位情報の付与支援」尾崎太亮・浅田宗磨・古宮嘉那子・近藤明日子・小木曽智信 https://www.anlp.jp/proceedings/annual_meeting/2026/pdf_dir/Q2-9.pdf

学会発表

- （6）令和8年3月27日，Brown Bag Workshop II: Japanese Language Corpus System Tutorial (Hamilton Library Japanese Language School Textbooks) Toshinobu Ogiso and Yuta Takahashi，ハワイ大学マノア校
  https://www.hawaii.edu/calendar/manoa/2026/03/24/45299.html

おわりに

- BCCWJ2の構築は比較的順調に進んでいる
- データを作るのがメインの仕事だが、ユーザインターフェイスをとおしてどのように届けるかが課題（小木曽 2026b）

参考文献

- 小木曽智信(2026a)「BCCWJ1形態論情報の更新」『日本語学会2026年度春季大会予稿集』 pp.210-212.
- 小木曽智信(2026b)「パネルディスカッション」, BCCWJ2シンポジウム―中間成果の公開と展望―, 2026年5月30日.
- 小椋秀樹・ 冨士池優美(2013)「4章 形態論情報」, 『『 現代日本語書き言葉均衡コーパス』利用の手引 第1.0 版』, https://doi.org/10.15084/00003227
- 尾崎太亮・浅田宗磨・古宮嘉那子・ 近藤明日子・ 小木曽智信(2026)「Monaka による BCCWJ2における長単位情報の付与支援」『言語処理学会第32回年次大会発表資料集』pp.907-912.
- 落合哉人(2026)「SNS データの収集とコーパス化：データの特徴と整備をめぐる論点」『日本語学会2026年度春季大会予稿集』 pp.207-209.

参考文献

- 近藤明日子(2026)「教科書レジスターの設計：複数時期の教科書の比較分析の可能性」『日本語学会2026年度春季大会予稿集』 pp.204-206.
- 田中牧郎・相澤正夫・斎藤達哉・棚橋尚子・近藤明日子・河内昭浩・鈴木一史・平山允子(2011)『言語政策に役立つ，コーパスを用いた語彙表・漢字表等の作成と活用』国立国語研究所. https://doi.org/10.15084/00002858
- 田中牧郎・近藤明日子・平山允子(2011)「教科書コーパス」田中・相澤・斎藤ほか(2011), pp.7-54.
- 山崎誠・呉寧真(2026)「出版書籍データの設計方針」『日本語学会2026年度春季大会予稿集』 pp.201-203.

- ご清聴ありがとうございました。