大学共同利用機関法人 人間文化研究機構
国立国語研究所
National Institute for Japanese Language and Linguistics

文化庁
Agency for Cultural Affairs, Government of Japan

2026年5月11日
※解禁日時はありません

高品質日本語データベース一般公開開始
2億語規模(完成時)の言語資源を研究・産業界へ

国立国語研究所(大学共同利用機関法人 人間文化研究機構)は、2026年3月、『現代日本語書き言葉均衡コーパス 第2部』(BCCWJ2)の一般公開を開始しました。

文化庁委託事業により構築するBCCWJ2は、従来のBCCWJ1と統合することで、高品質な2億語規模の書き言葉のデータを提供するものです。言語研究に加え、生成AI開発、言語教育、辞書編纂、言語政策など、幅広い分野での活用が期待されます。

コーパスとは？
書籍やウェブなどで実際に使われた言葉を大量に収集、整理したデータベースです。言葉の使われ方や意味の変化を客観的に調査・分析することができ、研究や産業分野で広く活用されています。

BCCWJ2の特徴① 21世紀の日本語の変化を可視化

BCCWJ1と合わせて利用することで、2001～2025年の日本語の変化を分析できます。BCCWJ1(2011年公開)では見られなかった以下の語の用例も確認できるようになりました。

スマートフォン／ツイッター／iPad／LCC(ローコストキャリア) など

BCCWJ2の特徴② 生成AI時代以前の日本語を収録した最後のコーパス

BCCWJ2には、生成AIによる文章生成が普及する以前の書き言葉が収録されています。

そのため、「人によって書かれた日本語」を体系的に収録した最後の大規模コーパスと位置づけられ、高品質な日本語AI開発を支える基盤データとして、大規模言語モデルの再学習や性能評価への活用が期待されます。

紹介・デモンストレーションのご案内

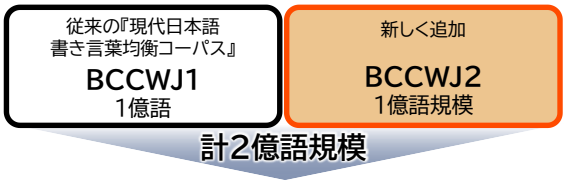
BCCWJ2シンポジウム―中間成果の報告と展望―
日時:2026年5月30日(土)13:00～17:00
会場:対面(コモレ四谷 タワーコンファレンス)＋オンライン(Zoom)
主催:国立国語研究所　共催:文化庁
詳細:https://www.ninjal.ac.jp/events jp/20260530a/
※当日もご取材可能です。シンポ終了後に、個別にお応えいたします。

BCCWJ2の概要

BCCWJ2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」により構築する、1億語規模の現代日本語書き言葉のコーパスです。

2011年公開の『現代日本語書き言葉均衡コーパス』(BCCWJ1、文部科学省科研費特定領域研究「日本語コーパス」の助成により構築)は主に2005年までの日本語データ約1億語分を収録していました。BCCWJ2はそれを拡張する形で構築し、主に2006～2025年のデータを収録します。

2026年3月に、その一部として2006～2010年の書籍データ約2300万語を一般公開しました。今後、段階的に公開範囲を拡大し、2028年度末までに全体を公開する予定です。



収録資料

BCCWJ2では、以下の3種類の資料を収録予定です。

1. 書籍　（2006～2025年、計約1億語）
2. 教科書　（小学校・中学校・高等学校使用の検定教科書、使用年度3か年分）
3. SNS　（BCCWJ2で新規収録、データ設計は現在検討中）

利用方法

BCCWJ2は、以下の2つのウェブツールから無料で利用できます。また、研究者・事業者向けに、データセットの有償提供も予定しています。

- 「少納言」(https://shonagon.ninjal.ac.jp/)
  利用登録不要。文字列検索に対応。
- 「中納言」(https://chunagon.ninjal.ac.jp/)
  利用登録必要。語単位で付与された情報(形態論情報)を使った、高度な検索に対応。

「少納言」検索画面

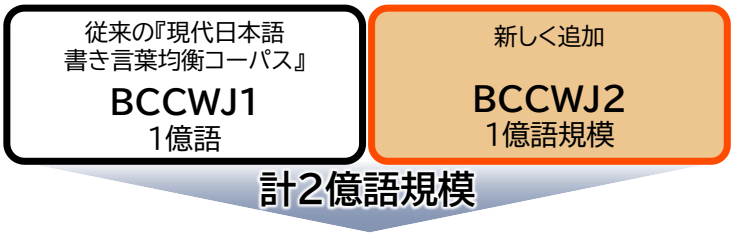
「中納言」検索画面

お問い合わせ先

人間文化研究機構　国立国語研究所　BCCWJ2事務局
TEL: 042-540-4539　E-mail: bccwj2@ninjal.ac.jp
BCCWJ2ウェブサイト: https://www2.ninjal.ac.jp/BCCWJ2/

BCCWJ2
『現代日本語書き言葉均衡コーパス』の拡張

概要



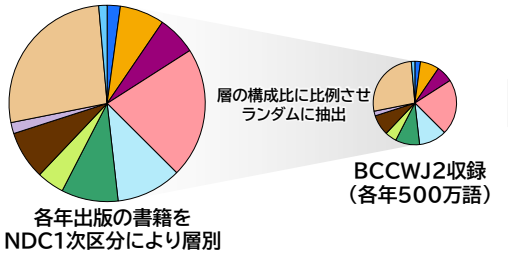
- 国立国語研究所では、**文化庁の委託事業**として、2024年度より5か年計画で「信頼できる言語資源としての現代日本語の保存・活用のためのデジタル基盤整備事業」を実施している。
- 本事業の一環として、従来の『現代日本語書き言葉均衡コーパス』(BCCWJ1、文部科学省科研費特定領域研究「日本語コーパス」の助成により2011年公開)を拡張し、**『現代日本語書き言葉均衡コーパス第2部』(BCCWJ2)**を構築する。これにより、コーパスの規模を従来の1億語から**2億語規模**に拡張する。
- 全体の公開は**2028年度末**を予定しており、2026年3月に、その一部として2006～2010年の書籍データ(約2,300万語)を公開した。

収録資料

以下の3種類の資料の収録を予定している。

① **書籍**

BCCWJ1に収録されている2001～2005年刊行の書籍データの拡張として、2006～2025年に日本で刊行された書籍を対象に、日本十進分類法(NDC)の1次区分の構成比に基づき、書籍の章・節の単位でテキストをランダムに抽出する。各年500万語、合計1億語を収録する。



② **教科書**

小学校・中学校・高等学校で使用された検定教科書を対象に、直近の3回の学習指導要領改訂ごとに使用年度1か年分(各学年・各教科1種ずつ)、計3か年分を選定し、その全文を収録する。

直近3回の学習指導要領改訂ごとに
1か年分、計3か年分の教科書を収録
① 平成10・11年改訂
② 平成20・21年改訂
(平成27年一部改正)
③ 平成29・30年改訂

③ **SNS**

近年、SNSの利用が拡大、そこで使用される日本語も現代日本語の一側面として重要である。その実態を把握できるよう、BCCWJ1にはないSNSデータを新たに収録する。データ設計については現在検討中である。

特徴① 21世紀の日本語の変化を可視化

書籍データをBCCWJ1と併せて利用することで、2001年から2025年までの日本語の変化を通時的に分析できる。BCCWJ2では、「**スマートフォン**」「**ツイッター**」「**iPad**」「**LCC**(ローコストキャリア)」など、BCCWJ1には見られなかった語の使用例も確認でき、**最新の日本語の変化**をより詳細に観察することが可能である。

特徴② 生成AI時代以前の日本語を収録した最後のコーパス

近年、生成AIによる文章生成が急速に普及しつつある。BCCWJ2には、その普及以前の日本語が収録されているため、**「人によって書かれた日本語」**を対象とした**最後の大規模コーパス**として、重要な位置づけを持つ。この特性により、今後の高品質な日本語生成AI開発において、生成AIに必要な技術である**大規模言語モデルの再学習や性能評価のための基盤データ**としての活用が期待される。

利用方法

検索ツールとして以下の2種類を提供する。

① 「**少納言**」 https://shonagon.ninjal.ac.jp/
無料、利用登録不要。文字列検索に対応し、BCCWJ1と一括検索できる。

② 「**中納言**」 https://chunagon.ninjal.ac.jp/
無料、利用登録必要。語単位で付与された形態論情報を使った高度な検索に対応。また、他のコーパスとまとめて検索も可能。

この他、研究者・事業者向けに**データセットの有償提供**も予定している。

「少納言」を使った検索の特徴

「少納言」は利用登録不要で、すぐに誰でも無料で利用できる。また、BCCWJ1と一括検索し、用例を一つの画面で確認することも可能。ただし、文字列検索のみに対応するため、検索結果に目的の語以外の用例が含まれてしまう場合がある。

↑↓ 表示番号	↑↓ 前文脈	↑↓ 検索文字列	↑↓ 後文脈
1	紙を送った。 やがて音信が途絶えた。 今年の三月、とつぜん葉書が来た。 希望の外	国語	大学に入ったという。 いまは、お兄さんと二人で暮らして、近所の中学生の家庭教師を
2	なんでもない今他所のマックで韓	国語	独学中・・・息抜きにブログ＾＾写真は何となく
3	は不意打ちにあって、まだ立ち直れないでいます。 松崎先生は私たちに、三年、四年と	国語	を教えてくださいました。先生はよく古今東西の詩を読んで聞かせてくださいました。詩
4	てしまうので工夫しました。これを覚えると買い物ができるとか、何か場面と直結した外	国語	学習というのは、身近に感じられるし、必要性、必然性が高くなるので学ぼうという気に
5	くとき、すごくうれしくって、こくごがだいすきになりました」と感想を書いた。さらに	国語	の学習に意欲的になり、その後も生き生きと文章を書いている。 今回の実践から、子ど

「国語」の検索結果（一部）。「国語」以外に「外国語」「韓国語」等の「国語」を含む別の語の用例も表示されてしまう。

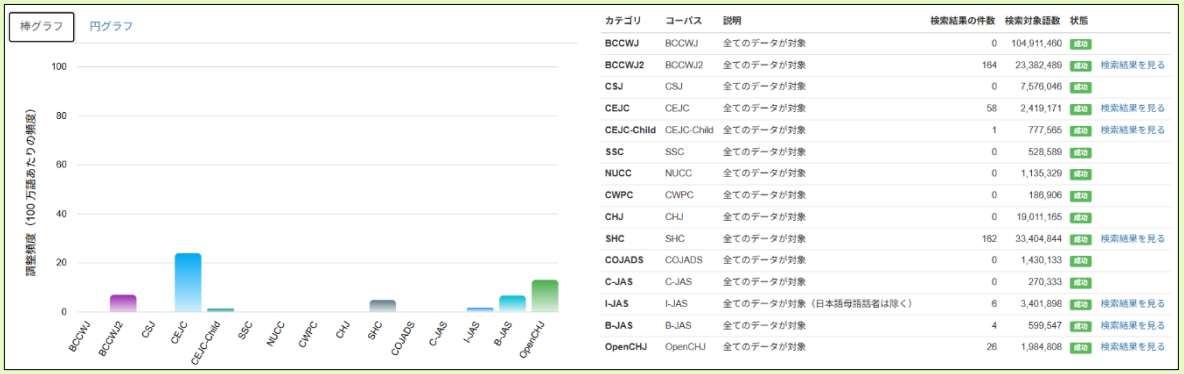
「中納言」を使った検索の特徴

「中納言」の形態論情報を使った検索では、語を指定することで、語形や表記の揺れを一括して検索できる。また、「少納言」のように検索結果に別の語が混在することも抑えられる。

サンプルID	開始位置	連番	前文脈	キー	後文脈
PB102_02334	103270	66550	たような|もの|だし|、|#|で|も|、|	Twitter	も|含め|て|、|自分|たち|の|生業|と|し
PB103_09906	65360	38640	年間|の|定量|目標|】#売上|:|二百|万|円#	twitter	|follower|:|五千|人#モチベーション|・|ワークショップ|開催|:
PB106_02414	39930	23790	クライアント# |K|i|z|n|a|では|、|	ツイッター	上|の|やりとり|を|すべて|独自|の|サーバー|に|保存
PB106_02414	30020	18230	設定|した|期間|の|終了|後|には|、|	ツイッター	を|通じ|て|期間|延長|を|アナウンス|し|、

「ツイッター」の検索結果（一部）。「Twitter」「twitter」「ツイッター」といった様々な表記も一括して検索できる。

また、「まとめて検索KOTONOHA」機能を用いて、国立国語研究所が公開する他のコーパスと一括検索を行い、出現回数を比較できる。



カテゴリ	コーパス	説明	検索結果の件数	検索対象語数	状態	
BCCWJ	BCCWJ	全てのデータが対象	0	104,911,460	成功	
BCCWJ2	BCCWJ2	全てのデータが対象	164	23,382,489	成功	検索結果を見る
CSJ	CSJ	全てのデータが対象	0	7,576,046	成功	
CEJC	CEJC	全てのデータが対象	58	2,419,171	成功	検索結果を見る
CEJC-Child	CEJC-Child	全てのデータが対象	1	777,565	成功	検索結果を見る
SSC	SSC	全てのデータが対象	0	528,589	成功	
NUCC	NUCC	全てのデータが対象	0	1,135,329	成功	
CWPC	CWPC	全てのデータが対象	0	186,906	成功	
CHJ	CHJ	全てのデータが対象	0	19,011,165	成功	
SHC	SHC	全てのデータが対象	162	33,404,844	成功	検索結果を見る
COJADS	COJADS	全てのデータが対象	0	1,430,133	成功	
C-JAS	C-JAS	全てのデータが対象	0	270,333	成功	
I-JAS	I-JAS	全てのデータが対象（日本語母語話者は除く）	6	3,401,898	成功	検索結果を見る
B-JAS	B-JAS	全てのデータが対象	4	599,547	成功	検索結果を見る
OpenCHJ	OpenCHJ	全てのデータが対象	26	1,984,808	成功	検索結果を見る

「ツイッター」の検索結果。BCCWJ2以外に、現代の日常会話で使用回数が多いことが分かる。

著作権への対応

2018年の**著作権法改正**により、著作権者の許諾なしに「資料の電子化」「コーパスの利用・分析」「検索サービスの提供」が可能になった。

BCCWJ2では、この法改正に基づき書籍・教科書データを公開する。ただし、「軽微利用」（著作権者の利益を不当に害さない範囲での著作物の一部利用）等の要件を満たすため、以下の対応を行う。

- 俳句・短歌・詩など短い著作物は収録対象外とする
- 検索ツールでは表示する文脈の長さを制限する

また、SNSデータについては、**運営会社の利用規約**も踏まえた形で公開する。

個人情報への対応

データに含まれる個人情報については、該当箇所を**伏せ字処理**した上で公開する。

学校教育での利用への対応

学校教育等でのコーパス活用の環境整備に向けて、**成人向け書籍のフィルタリング用情報**をBCCWJ1と併せて公開する。

BCCWJ2ウェブサイト
https://www2.ninjal.ac.jp/BCCWJ2/