特集

現代日本語書き言葉均衡コーパス

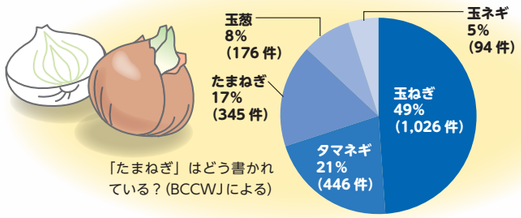
# BCCWJ開発秘話

**現代日本語の書き言葉の全体像を把握する――そのために国立国語研究所（国語研）が構築し、2011年より公開したのが、「現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese：BCCWJ）」です。約1億語からなる、日本語書き言葉の全体をバランスよく反映した2024年現在唯一のコーパス。その開発は試行と挑戦の連続でした。**

## コーパスとは？

コーパスとは、実際に使われた言葉を大量かつ体系的に集め、品詞情報など研究用の情報を付加してさまざまな検索ができるようにした、言葉のデータベースです。

野菜の「たまねぎ」は、どう書かれている？「○○的」という言葉には、どのようなものがある？　コーパスを使うと、そうした言葉の使われ方を調べることができます。




「たまねぎ」はどう書かれている？（BCCWJによる）

| | 出現数トップ 10 | 出現数 10 の例 | 出現数 1 の例 |
|---|---|---|---|
| 1 | 具体的（10,547） | アイドル的 | BGM 的 |
| 2 | 基本的（9,965） | スター的 | アウトドア的 |
| 3 | 積極的（7,585） | タイミング的 | スーパーマン的 |
| 4 | 一般的（6,395） | あたし的 | あたくし的 |
| 5 | 社会的（5,988） | 味的 | 当たり前的 |
| 6 | 比較的（4,141） | 色彩的 | あら探し的 |
| 7 | 経済的（4,049） | 観光的 | おぬし的 |
| 8 | 個人的（3,416） | 推論的 | 漢字的 |
| 9 | 総合的（3,343） | 立場的 | 惣菜的 |
| 10 | 効果的（3,297） | 実感的 | 俗人的 |

「○○的」という言葉にはどのようなものがある？（BCCWJ の元データによる「／○○／的」の検索結果）

## 世界各地でコーパスが次々に誕生

言葉を研究するには、言葉を集めて分析する必要があります。研究者はそれぞれの目的や関心に従って言葉を集めますが、個人では集めることができる量や範囲は限られます。そこで、多様な目的に利用できて言語研究の共通基盤となるように、言葉を大量かつ体系的に集めたコーパスを、大学や国が中心となってつくり始めました。

世界初の大規模なコーパスは、イギリスの「Survey of English Usage Corpus（SEU）」です。1959年から話し言葉と書き言葉それぞれ50万語の収集を始め、紙のカードに言葉を書き取って整理する方法でつくられました。

1964年にはアメリカで、100万語の書き言葉を収集した「Brown Corpus」が完成しました。コンピュータで使えるようにした、世界初の電子コーパスです。その後、世界各地でコーパスが開発され、1994年にはイギリスで話し言葉と書き言葉を合わせて1億語を集めた「British National Corpus（BNC）」が完成しました。

## 1億語の日本語書き言葉コーパスをつくろう！

文庫本1,700冊分に相当

国語研では1950年代から、雑誌や新聞などを対象に、言葉の使われ方や頻度を調べる語彙調査を行ってきました。それらは紙のカードを用いたものでした。2004年には国語研初の電子コーパスとして「日本語話し言葉コーパス」が完成しました。しかし、日本語の書き言葉の全体をバランスよく反映したコーパスが、まだありませんでした。

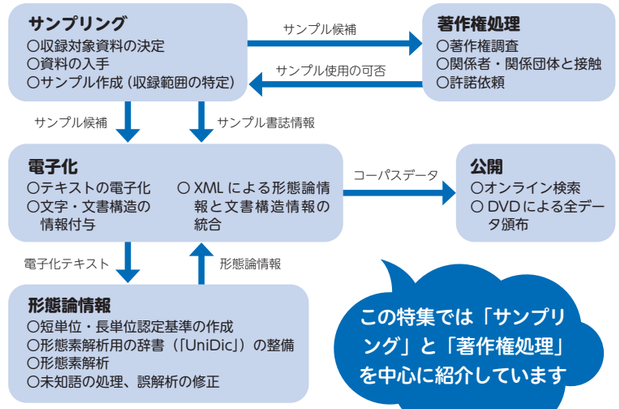
そこで、BNCの日本語版を目指し、「1億語の書き言葉コーパスをつくろう！」というプロジェクトが文部科学省科学研究費補助金（特定領域）の助成を受けて2006年に始まりました。

## 目指したのは現代日本語の書き言葉の縮図

新規にコーパスをつくる場合、どういう性質のコーパスとするのか、その設計方針が重要です。多くの議論を経て、次の4点を念頭に置いて設計することにしました。

**①多様な書き言葉をバランスよく反映したコーパス**
**②幅広い目的に供するコーパス**
**③公開可能なコーパス**
**④先に公開した「日本語話し言葉コーパス」の解析単位との互換性を保持**

**コーパス構築の流れ**




## 1億語をどう選ぶ？

書き言葉にはさまざまなものがありますが、活字になって刊行されたものを対象とすることにしました。具体的には、書籍・雑誌・新聞です。では、どのようにサンプリングすれば、日本語の書き言葉を代表するサンプルを収集できるのか。それが大きな課題でした。

そして導き出した方法が、「2001～2005年に出版された全ての書籍・雑誌・新聞に含まれる総文字数を推計し、その比率をもとに各媒体からサンプリングする言語量を設定する」というものです。推計する対象として、書籍は国立国会図書館の蔵書目録にあるもの、雑誌は日本雑誌協会の加盟出版社発行のもの、新聞は全国紙・ブロック紙・有力地方紙としました。


書籍　31万7117冊・7491万1520ページ
雑誌　5万5779冊・1041万4955ページ
新聞　4万9625冊・119万8189ページ
※ 2001～2005年に出版された数

**均衡とは……**
テキストのジャンルや収録する際の手続きに「偏りがない」という意味。価値判断を伴うような代表性や、全てのジャンルについて同じ量だけ収録されているという意味ではない。

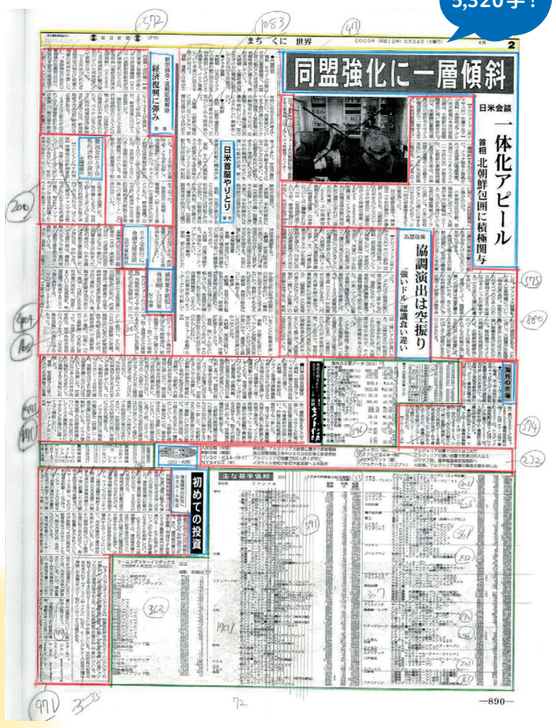
## 文字数をひたすら数える！

5年間に出版された書籍・雑誌・新聞に含まれる総文字数をどのように推計したと思いますか？

新聞については、全国紙4紙の朝夕刊の8冊・211ページを実測しました。右の写真は、実際に文字数を数えた新聞です。エリアを区切り、1文字1文字、ひたすら数えていきます。エリアごとの文字数が鉛筆で書き込まれています。このページの文字数は5,320字でした。

書籍については、日本十進分類法（NDC）の分類・判型ごとにランダムに選び出した227冊・1,135ページの文字数を、やはりひたすら数えました。雑誌は、NDCの分類・判型ごとにランダムに選び出した53冊・265ページを実測しました。

5,320字！




文字数を実測した新聞（毎日新聞2003年5月24日夕刊）

## 書き言葉の実態を把握できているのか？

5年間に出版された総文字数の推計をもとに、書籍・雑誌・新聞それぞれの種類・分類の構成比を算出しました（右上の表）。しかしこの比率でサンプリングする方法では、書き言葉が生み出された実態を把握することはできても、その言葉がどれくらい流通しているかは分かりません。また書き言葉には、書籍・雑誌・新聞以外にも多くの種類があります。

そこで、書籍については、読まれているという実態を反映するよう、図書館の蔵書やベストセラーを加えました。その他のさまざまな資料も加え、性質の異なる３つのサブコーパスで構成することにしました（右図）。

| | 層 | 推計文字数 | 比率 | | 層 | 推計文字数 | 比率 |
|---|---|---|---|---|---|---|---|
| 書籍 | 0. 総記 | 1,636,414,548 | 2.50% | 雑誌 | 総合 | 7,421,447,806 | 11.34% |
| | 1. 哲学 | 2,597,610,813 | 3.97% | | 教育 | 877,875,592 | 1.34% |
| | 2. 歴史 | 4,301,204,340 | 6.57% | | 政治 | 456,459,405 | 0.70% |
| | 3. 社会科学 | 12,408,321,943 | 18.95% | | 産業 | 110,640,958 | 0.17% |
| | 4. 自然科学 | 5,069,594,034 | 7.74% | | 工業 | 1,468,293,360 | 2.24% |
| | 5. 工学 | 4,615,929,967 | 7.05% | | 医療 | 180,964,513 | 0.28% |
| | 6. 産業 | 2,196,387,437 | 3.35% | | | | 16.06% |
| | 7. 芸術 | 3,258,432,447 | 4.98% | 新聞 | 全国紙 | 2,417,622,461 | 3.69% |
| | 8. 言語 | 888,800,128 | 1.36% | | ブロック紙 | 1,296,592,154 | 1.98% |
| | 9. 文学 | 9,341,275,486 | 14.27% | | 地方紙 | 2,701,855,499 | 4.13% |
| | n. 分類なし | 2,225,954,208 | 3.40% | | | | 9.80% |
| | | | 74.14% | | | | |

### 現代日本語書き言葉均衡コーパス（BCCWJ）

**出版サブコーパス**
生み出された書き言葉を幅広く収集
書籍・雑誌・新聞（2001～2005年）
約3,500万語

**図書館サブコーパス**
流通している書き言葉を幅広く収集
東京都内13自治体以上の公立図書館で共通して所蔵しているもの（1986～2005年）
約3,000万語

**特定目的サブコーパス**
上の２つでは捉えにくいもので、現代日本語書き言葉の実態を把握するのに重要な資料を収集
白書・教科書・広報紙・ベストセラー・Yahoo!知恵袋・Yahoo!ブログ・韻文・法律・国会会議録（1976～2005年）
約3,500万語

## 原本を入手するため図書館・書店・古書店を巡る

**古書店巡りの達人！？**
図書館や新刊書店で入手できない資料は、古書店を次から次へと巡って探しました。*



＊BCCWJの構築に携わったスタッフのコメントを紹介します。

構成が決まったら、サンプルを抽出する対象をコンピュータでランダムに選び、原本を入手します。図書館で借りたり、比較的安価なものは購入したりしました。館外への持ち出しが禁止されているものは、許可をいただいてスキャナを持ち込みスキャンしました。新聞は縮刷版を入手し、縮刷版がない場合は国会図書館のマイクロフィッシュから取得しました。

入手した原本は、出版サブコーパスと図書館サブコーパスの書籍と雑誌だけでも3万1000冊以上！　図書館の貸出申請・借り出し・返却、書店への発注・納品確認など、大がかりで複雑な作業でした。

## どの部分をサンプリングする？

資料の原本やスキャンデータを入手できたら、サンプリングを行います。資料ごとに、何ページのどの位置からサンプリングするか、ランダムに指定します。

１つの資料から、長さの異なる２種類のサンプルを取得します。「固定長サンプル」は基準点を始点として1,000文字目までの範囲を、「可変長サンプル」は基準点を含む章や節などの言語的なまとまり（ただし上限１万字）を抽出します。

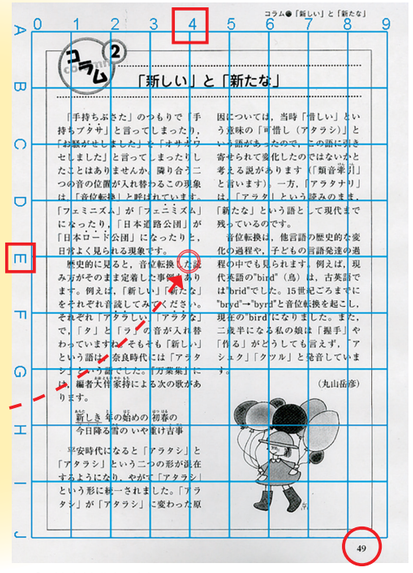
サンプル取得数
1　　SampleID：PB58_00482

NDLsize
21cm

| 優先順位 | 対象頁 | 有効 | 乱数1 | 乱数2 |
|---|---|---|---|---|
| 一位 | 49 | | 4E | 2H |
| 二位 | 25 | | 2C | 8A |
| 三位 | 78 | | 4J | 0H |
| 四位 | 20 | | 9J | 2B |

サンプル台帳の指示
「49ページの座標 4E」

マス目を印刷した透明なシートを対象ページに当て、座標により指定された交点に最も近接している文字を「サンプル抽出基準点」とする。この例では「た」。これを「キメマスシートでキメ点を決める」と言っていた。シートは判型に合わせ、さまざまなサイズを用意していた。




サンプル台帳とサンプル抽出基準点の例

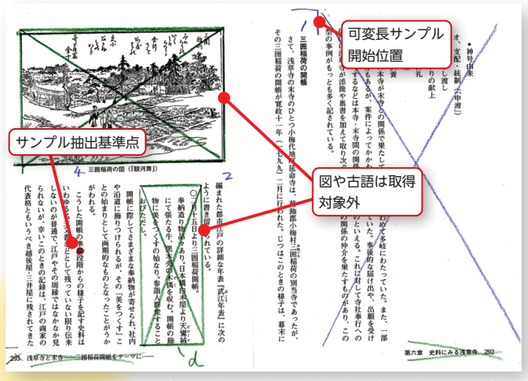
**作業者泣かせの“改行なし”**
可変長サンプルは、セリフ以外に改行がない文は、上限の１万字まで抽出することになっています。私が特に泣かされたのは、ロシア文学です。改行なしがずーっと続くんです。トルストイも、ドストエフスキーも。翻訳文も現代日本語として収録対象で、その翻訳文が素晴らしい文章なのですが、読みふけってしまっては文字数をカウントできません。ぐっとこらえて電卓をたたいていました。

## この場合はどうする？の連続

サンプル抽出基準点はランダムに指定されるため、サンプル抽出基準点に文字がない場合もあります。その場合は、サンプル抽出基準点をサンプル台帳の乱数２、３……と変えていきます。それでも抽出できない場合は、対象ページの優先順位を下げていきます。

では、古語や外国語の文章だったら？　数式の場合は？　写真の場合は？　さまざまなケースが発生します。「現代日本語で書かれた表現」が対象ですから、ひとかたまりの古語、外国語、数式や、写真、図は、原則取得対象外です。

サンプル間・作業者間で揺れが生じないように、作業規則と判断基準を細かく設定し、マニュアル化しました。そうすることで、斉一な手順で均質なサンプリングを実施しました。

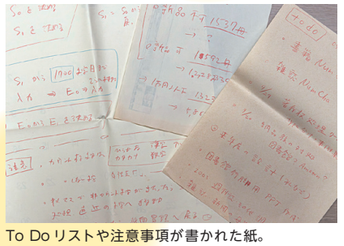



サンプル取得の例。サンプル台帳に従って、対象資料のコピーを取り、サンプル抽出基準点、固定長サンプルの範囲、可変長サンプル抽出開始位置、取得対象外などを色鉛筆で記入していく。この指示用紙に基づいて電子化作業が行われる。


ランダムに指定される資料の誌面はさまざま。作業規則と判断基準に沿って１件ずつ手作業でサンプルを取得していく。

**真夏のコピー地獄！**
暑い時期の大量コピーは、かなりの重労働でした。というのも、当時の国語研の建物の窓には、ブラインドがほとんど付いていなかったのです。しかもコピー機は窓際にありました。真夏の昼の時間帯は強い日差しが照りつけ、「そろそろ限界……」「いや、ここまで終わらせたい」という気持ちのせめぎ合いでした。今、窓という窓にブラインドが付けられているのは、この作業がきっかけでは？と勝手に考えています。

国会図書館書庫内でのサンプリング作業

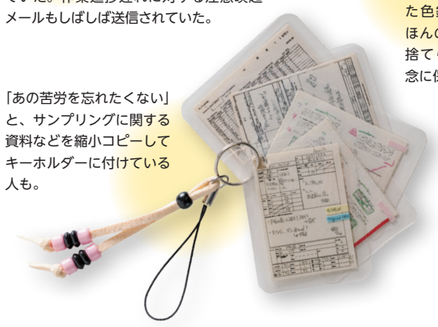


To Doリストや注意事項が書かれた紙。情報共有のためこうしたメモが、サンプリングの作業部屋の掲示板にたくさん貼られていた。作業進捗遅れに対する注意喚起メールもしばしば送信されていた。

サンプル取得に使った色鉛筆。これでもほんの一部。「とても捨てられない」と記念に保管している。

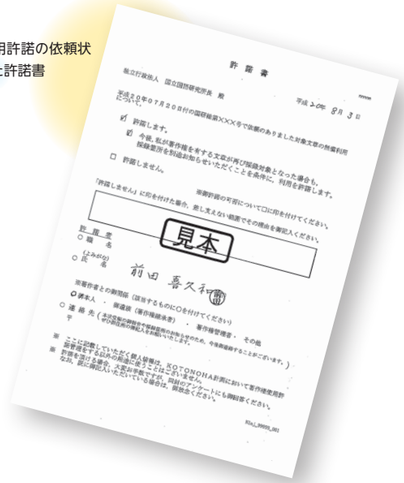

サンプリングの注意点が書かれた「魂の付箋セット」。「捨てるのがしのびない。後世まで遺していきましょう!!」とある。

サンプリングに使用した購入資料は現在も保管されている。



「あの苦労を忘れたくない」と、サンプリングに関する資料などを縮小コピーしてキーホルダーに付けている人も。

## 著作権者から許諾をいただく



著作権者に利用許諾の依頼状と共に送付した許諾書

構築したコーパスを公開することは、重要な設計方針の1つでした。そのためには、収録するサンプルの著作権者に利用許諾を得る必要があります。この著作権処理に、とても苦労しました。

著作権処理が必要なサンプルは、書籍のみで2万4421件ありました。それを1件1件、地道に処理していきました。まず、サンプルの著作権者を特定します。そして連絡先を調べ、利用許諾の依頼状を送ります。連絡先が分からず、依頼状を送ることができない場合もあります。特に雑誌の場合、連絡ができない割合が書籍よりも高くなりました。

それでも根気強く連絡先を調べ、書籍については約90%に当たる2万1986サンプルの著作権者に連絡を取ることができました。

### 日本へのお見舞い

アメリカ在住の原著者に、最初に許諾をお願いしたときには断られてしまいました。ですが東日本大震災の後、日本へのお見舞いですと言って許諾をもらうことができました。

### ようやく連絡が取れた作家さんは……

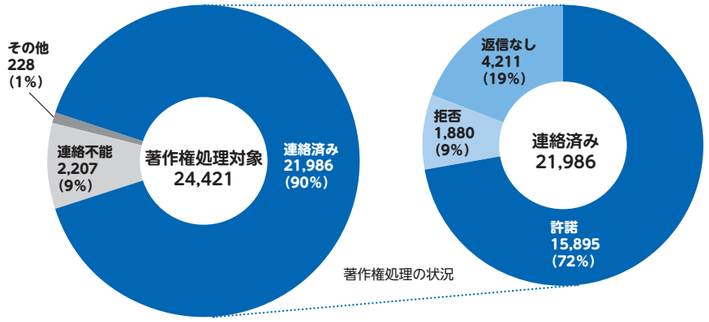
私が参加したのはプロジェクトの後期だったため、「連絡先不明」に分類された作家さんがたくさんいて、一人一人リサーチをやり直しました。そうした中、ある中国の作家さんと、幸運にも直接メールでコンタクトを取ることができました。とても丁寧なお返事と許諾もいただけて、やれやれ良かったと思っていたら……。しばらくして、ノーベル賞のニュースにその作家さんのお名前が！ びっくりしました。2012年のノーベル文学賞を受賞した莫言氏です。

## 利用拒否になると……

とてもありがたいことに、多くのサンプルの著作権者から利用の許諾をいただけました。しかし、利用拒否の回答が来る場合もあります。

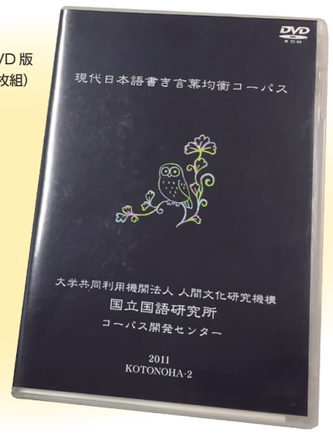
サンプリングと著作権処理は同時に並行して行っていました。利用拒否の回答が来ると、すでにサンプリングが終わっていても、そのサンプルは利用できません。対象をランダムに選んだリストから次に優先順位が高いものの原本を入手し、サンプリングをやり直さなければならず、作業計画の見直しが必要になる場合もありました。

例えば、書籍については、3つのサブコーパス合わせて1,681冊が利用不可になりました。雑誌の著作権を有する出版社に利用を拒否された例などもありました。



著作権処理の状況

## 1億語を超える日本語書き言葉のコーパス完成！



BCCWJ-DVD版
(DVD-R 4枚組)

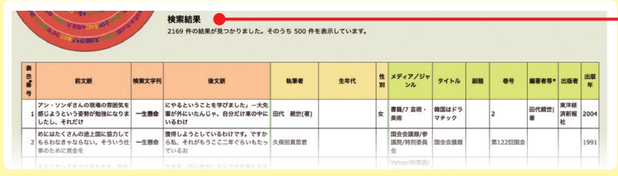
著作権処理が済んだサンプルは、電子テキスト化します。その電子テキストを用いて形態素解析を行い、品詞など研究に必要な情報を付与します。この特集では詳細を紹介できませんが、**電子化や形態論情報付与も、試行と挑戦の連続でした。**そして、全てのデータをパッケージ化。

こうして2006年から5年の歳月をかけて構築を進めてきた「現代日本語書き言葉均衡コーパス（BCCWJ）」が完成し、2011年より全てのデータを収録したDVD版の頒布（有償）とオンライン公開を開始しました。現在、Web上では、オンライン検索ツール「小納言」「中納言」「NINJAL-LWP」にて利用可能です。

## BCCWJを使ってみませんか

**国語研では、「少納言」というオンライン検索ツールを公開しています。利用条件に同意すれば、誰でも無料でBCCWJの全文検索をすることができます。「少納言」の基本的な使い方を紹介しましょう。**

https://shonagon.ninjal.ac.jp/


①**「検索文字列」の窓に調べたい語句を入力します。**例えば、「一生懸命」を検索してみます。
入力した文字列がそのまま検索されるので、その他の表記形を検索したい場合は、入力文字を変えて検索します。例えば、「一生けん命」「一生けんめい」「いっしょうけんめい」「一所懸命」「一所けんめい」「いっしょけんめい」などと入力します。

②**目的に応じて、検索対象の「メディア／ジャンル」や、刊行年代の「期間」を指定することができます。**

③**「検索」をクリックします。**
前後の文脈を指定して検索したい場合は、検索ボタンの上の「こちら」をクリックして文字列を入力します。

④**検索結果が表示されます。**検索文字列の前文脈と後文脈それぞれ40字程度と、出典情報を確認することができます。検索結果のダウンロードはできません。
「一生懸命」は2,169件見つかりました。検索結果が500件以上になった場合は、その中からランダムに選んだ500件が表示されます。
検索対象の「メディア／ジャンル」や「期間」を指定することで、検索結果を絞って表示させることが可能です。「一生懸命」の場合、「メディア／ジャンル」で「雑誌」だけにすれば検索結果は76件となり、全ての用例を確認できます。

## BCCWJの拡充が計画されています

BCCWJは、日本語研究をはじめ、日本語教育や国語教育、国語政策、辞書編集、自然言語処理など、さまざまな用途で使われています。BCCWJを用いた研究業績は2023年7月時点で1,784件でした。「少納言」は年60万回以上利用されています。

BCCWJの完成から10年以上たちました。BCCWJは、2011年の公開以降、更新がされていません。言葉は変化するものであり、またIT化により言葉が変化する速度は増していると言われることもあります。媒体の種類や文字数の比率も変わっていることでしょう。現代日本語の書き言葉の全体像を把握するためには、定期的なデータの追加・更新が必要です。

文化庁の施策の1つとして、令和6～10年度（2024～2028年度）にBCCWJを拡充する計画があります。2006年から2025年までの20年分の日本語書き言葉のデータを追加し、現在の1億語規模から2億語規模になる予定です。