

BCCWJ2の設計と構築



前川喜久雄・小木曾智信・山崎 誠・落合哉人
(国立国語研究所)

話の構成

- BCCWJとBCCWJ2
- BCCWJ
 - 開発の経緯
 - コーパスの構成
 - 公開実績
- BCCWJ2
 - 開発の経緯
 - 対象レジスター
 - 書籍
 - 教科書
 - SNS/CMC
 - アノテーション
 - 公開方式
 - ロードマップ
- まとめ

参考文献

BCCWJとBCCWJ2

- **BCCWJ**(Balanced Corpus of Contemporary Written Japanese)
 - 科研費特定領域研究「日本語コーパス」(2006-2010)で構築
 - 各種レジスター総計 1 億語規模
 - 我国初の公開を前提とした均衡コーパス ← 構築過程で種々の試行錯誤
 - Google Scholarでの引用件数 > 4,000
- **BCCWJ 2** : BCCWJの後継開発計画
 - 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」(2024-2028)で構築
 - 書籍 1 億語規模 + α
 - 従来の経験を活かして設計と構築の両面を合理化

BCCWJ (=BCCWJ1)

BCCWJ開発の経緯

- ボトムアップの開発計画
 - 『日本語話し言葉コーパス』の経験に立脚して自発的に立案
 - 国語研内部でもコーパスの有用性・優先度・実行可能性について疑義
 - 「外部資金でやるならどうぞ」
- 我国初の均衡コーパス開発につき様々な試行錯誤
 - 均衡させる(balanceをとる) 範囲
 - 厳密なサンプリングと言語研究上の要請のすりあわせ
 - アノテーション・メタデータの設計と範囲
 - 文字集合・エンコーディングの選択
 - 公開用データフォーマット(XML等)
 - 文末の認定
 - 著作権処理・個人情報保護の実践

BCCWJの構成

出版（生産実態）サブコーパス 約3,500万語 書籍、雑誌、新聞 2001年～2005年	図書館（流通実態）サブコーパス 約3,000万語 書籍 1986年～2005年
特定目的サブコーパス 約3,500万語 白書、教科書、広報紙、ベストセラー Web掲示板、ブログ、韻文、法律、国会会議録 対象期間はさまざま	

図2-1: BCCWJの構成

2 種類のサンプル

- 固定長サンプル

- 1000文字固定（約590短単位）。母集団の特性を確率論的に推定することが目的

- 可変長サンプル

- 文章の構造を反映したサンプル（節・章など）。上限1万字

表2-2: レジスターとサンプルのタイプ

サブコーパス	レジスター	サンプルのタイプ
出版サブコーパス	書籍(PB)	固定長、可変長
	雑誌(PM)	固定長、可変長
	新聞(PN)	固定長、可変長
図書館サブコーパス	書籍(LB)	固定長、可変長
特定目的サブコーパス	白書(OW)	固定長、可変長
	教科書(OT)	可変長
	広報紙(OP)	可変長
	ベストセラー(OB)	可変長
	Yahoo!知恵袋(OC)	可変長
	Yahoo!ブログ(OY)	可変長
	韻文(OV)	可変長
	法律(OL)	可変長
	国会会議録(OM)	可変長

2 種類の形態論情報

- **短単位 (SUW)**

- 語の形態的側面に注目した単位 (～国語辞典の見出し語)

- **長単位 (LUW)**

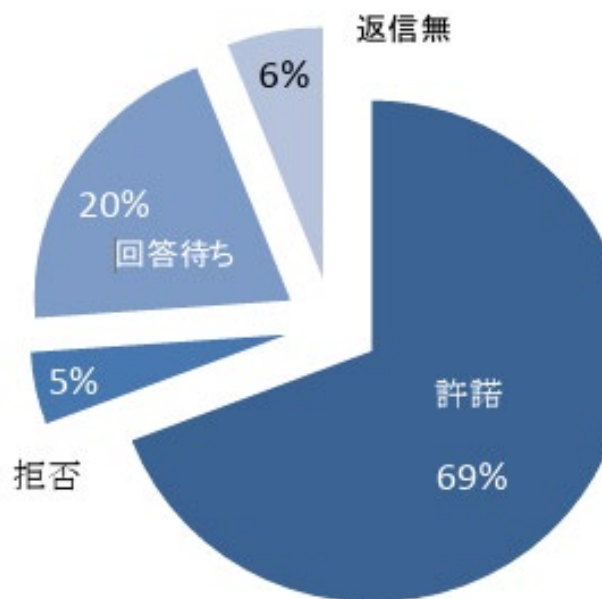
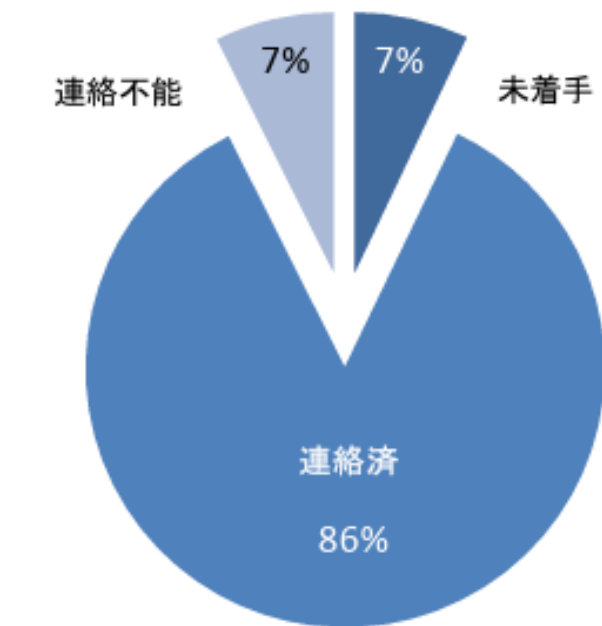
- 語の構文的側面に注目した単位 (～複合語、複合辞)

文節	日本語コーパスについて						解説した		
長単位	日本語コーパス			について			解説し		た
短単位	日本	語	コーパス	に	つい	て	解説	し	た

全体は自動解析。コア（100万短単位）は人手で精度向上

権利処理

- 現代語コーパスでは不可避
- 社会情勢による変化が速い
 - 著作権法がしばしば改正
 - 法曹家の解釈も不安定
- BCCWJでは「誰にも文句をいわれない」「馬鹿正直な」処理を実施
- 書籍を中心に約24,000件を処理



特定領域研究「日本語 コーパス」

総括班

項目A
コーパス
の構築

データ班

ツール班

電子化辞書班

項目B
コーパス
の評価

日本語学班

日本語教育班

言語政策班

辞書編集班

言語処理班

データ班内部構成

サンプリンググループ

形態論グループ

著作権グループ

電子化グループ

国語研に設置
常勤と非常勤で総勢25名程度

公開方式・利用実績

- テキスト・形態論情報の無償公開
 - 『少納言』 文字列検索のみ 登録不要
 - 『中納言』 短単位・長単位検索 要登録
- データ全体の有償公開（含商業利用）
 - DVD により頒布
- アノテーション・メタ情報の追加公開
 - 国語研学術情報レポジトリで随時公開（現在23件）

https://repository.ninjal.ac.jp/search?page=1&size=20&sort=custom_sort&search_type=2&q=472

公開方式・利用実績

- テキス

- 『少紙』

- 『中紙』

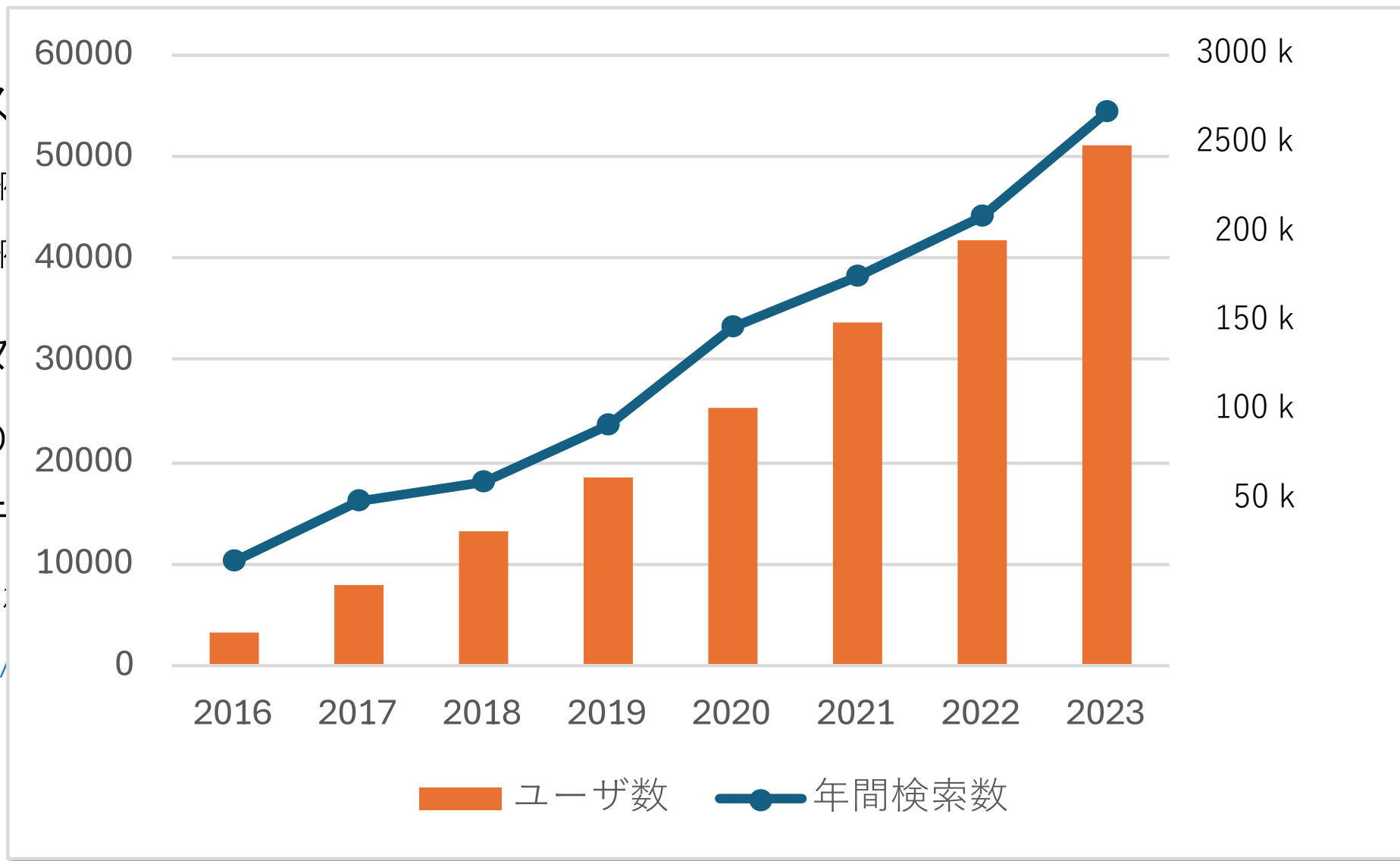
- データ

- DVD

- アノテ

- 国語

<https://>



&q=472

BCCWJ2

開発の動機

- 問題点

- BCCWJが現代語コーパスとしては古くなった
- 経年的な調査研究ができない
- 規模的にも十分でない

- 文化庁国語課による概算要求提案から

国（独立行政法人国立国語研究所）が2005年までのデータで整備したBCCWJに、2006年から**2025年までの20年分**の日本語データを追加し、1億語規模から**2億語規模**の現代日本語コーパスに拡充する。書籍、新聞等から、現代日本語の縮図となるように統計的に適切な文のサンプルを選択・特定し、著作権処理をした上で、日本語の品詞、意味、文構造等の情報を付与し、電子データ化を図る。1年間に2千万語を整備し、5年間で1億語を追加する。

BCCWJ2開発の経緯

- BCCWJ1を前提とした拡張計画
 - 「国の事業」を受託
 - 言語資源開発の有用性に関する疑義は研究所の内外ともにほとんどなくなった
- BCCWJ1の経験を活かせる
 - 設計・仕様の簡素化
 - 外注作業の拡大
 - ツールの高性能化
 - 改正された著作権法を最大限に活用
- 国語研は独立行政法人から大学共同利用研に移管
 - 独法時代のような大規模な開発体制を組むことは困難

BCCWJ2の対象レジスター

- 新聞は各社による事業化によりデータ入手が困難
- 雑誌はBCCWJ1で母集団としていた『雑誌新聞総かたろぐ』の廃刊(2019)によりサンプリングが不可能に
 - ⇒ 新聞・雑誌はBCCWJ2の対象とはしない（できない）
 - ⇒ ただし『昭和平成書き言葉コーパス』により部分的に検索可能
- マンガ・アニメ等の日本語
 - ⇒ 現時点ではデータ入手の目途がたたない
 - ⇒ マルチメディア言語コーパスとして別途設計が必要

BCCWJ2の対象レジスター

- BCCWJ1の「出版サブコーパス」を継承・拡大
- 特定目的サブコーパスの教科書も継承・拡大
- SNS等のCMC (Computer-Mediated Communication)データを新規追加

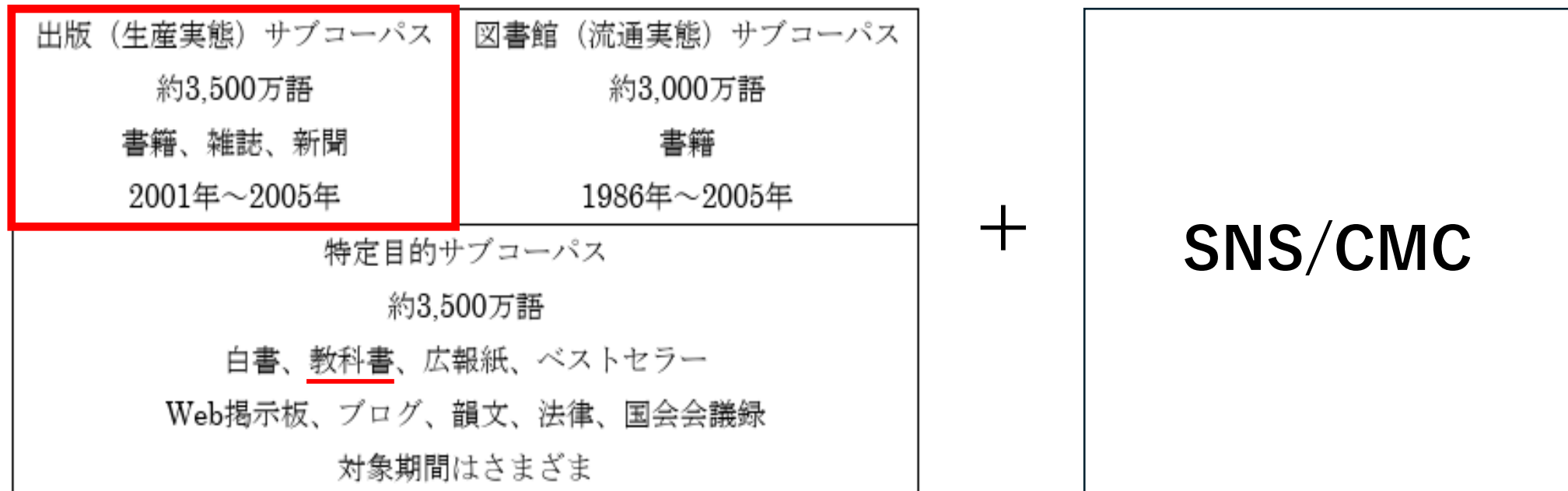
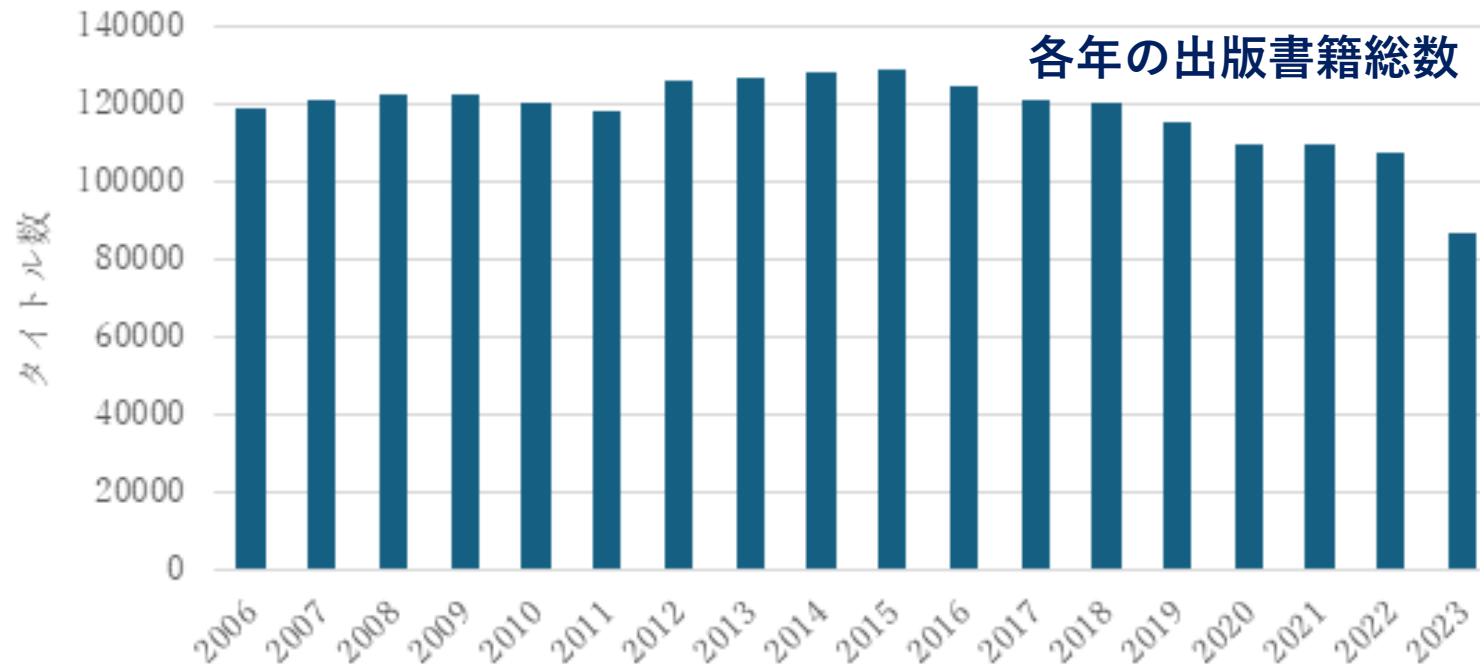


図2-1: BCCWJの構成

BCCWJ2の書籍サンプリング

- BCCWJ1の「出版サブコーパス」を継承・拡大
 - 経年調査には最適（図書館の書籍は経年変化を反映しにくい）
- ⇒ 電子書籍は対象外
- 国立国会図書館の書誌情報データ(JAPAN/MARC)を母集団とみなす

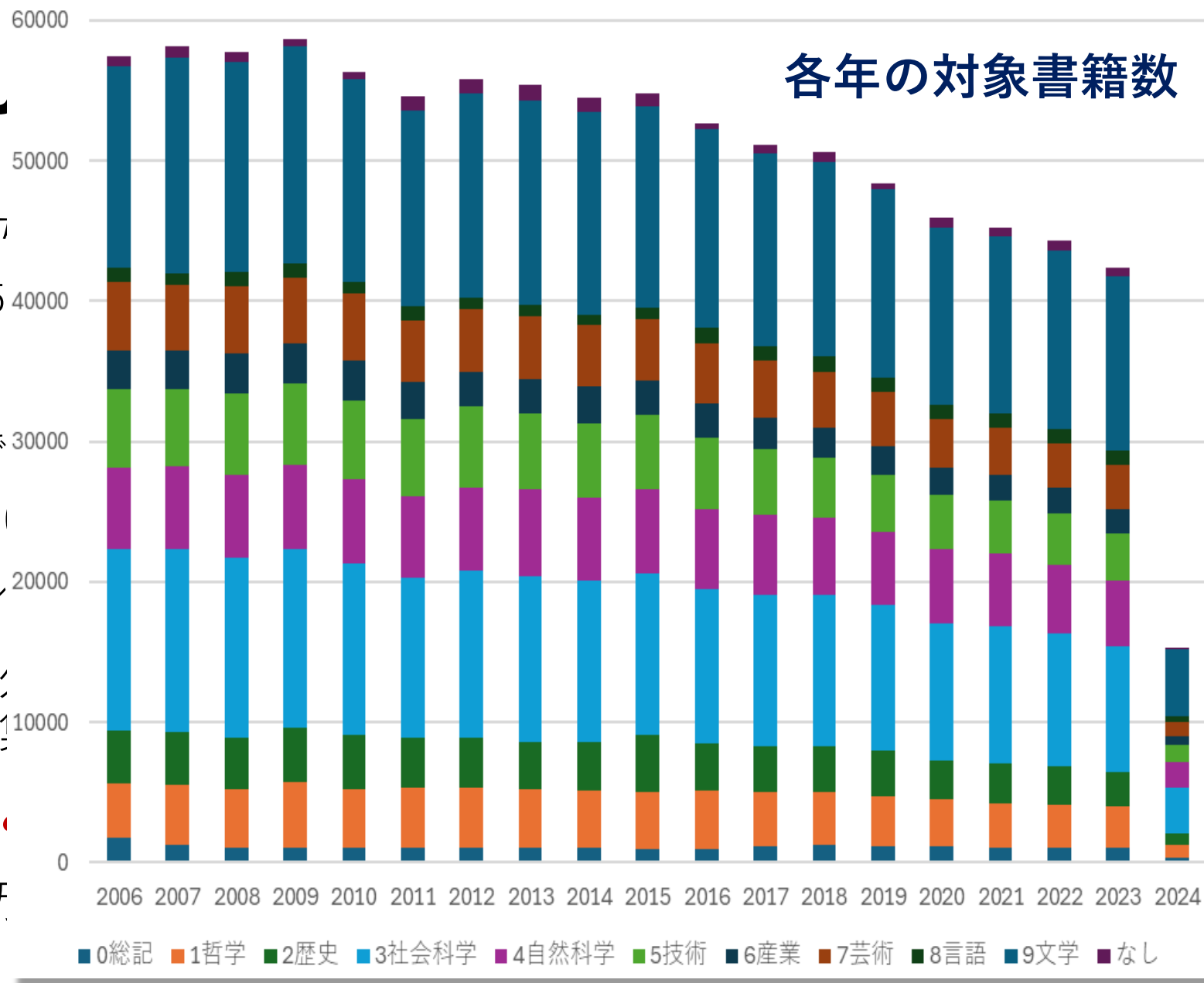


BCCWJ2の書籍サンプリング

- 1年ごとに母集団を想定して各500万語を抽出
 - 2006-2025で1億語
- 書籍を単位として無作為抽出
 - BCCWJ1ではページが単位だったが簡素化
- 固定長（1000字）と可変長（基準ページから前後10ページの範囲）
 - 固定長サンプルは必ず可変長の一部とする（簡素化）
- 対象から除外：俳句・短歌・詩、マンガ、絵本、試験問題集、要覧・名簿、写真集、官庁資料、一般年鑑、地図 etc.
 - ⇒ **BCCWJ1とほぼ同様**
- 全体で400万短単位程度をコアに設定

BCCW.

- 1年ごとに
- 2006-2025
- 書籍を単位
- BCCWJ1で
- 固定長（10）
- 固定長サン
- 対象から除
名簿、写真
- ⇒ BCCWJ1
- 全体で400万



BCCWJ2の教科書データ

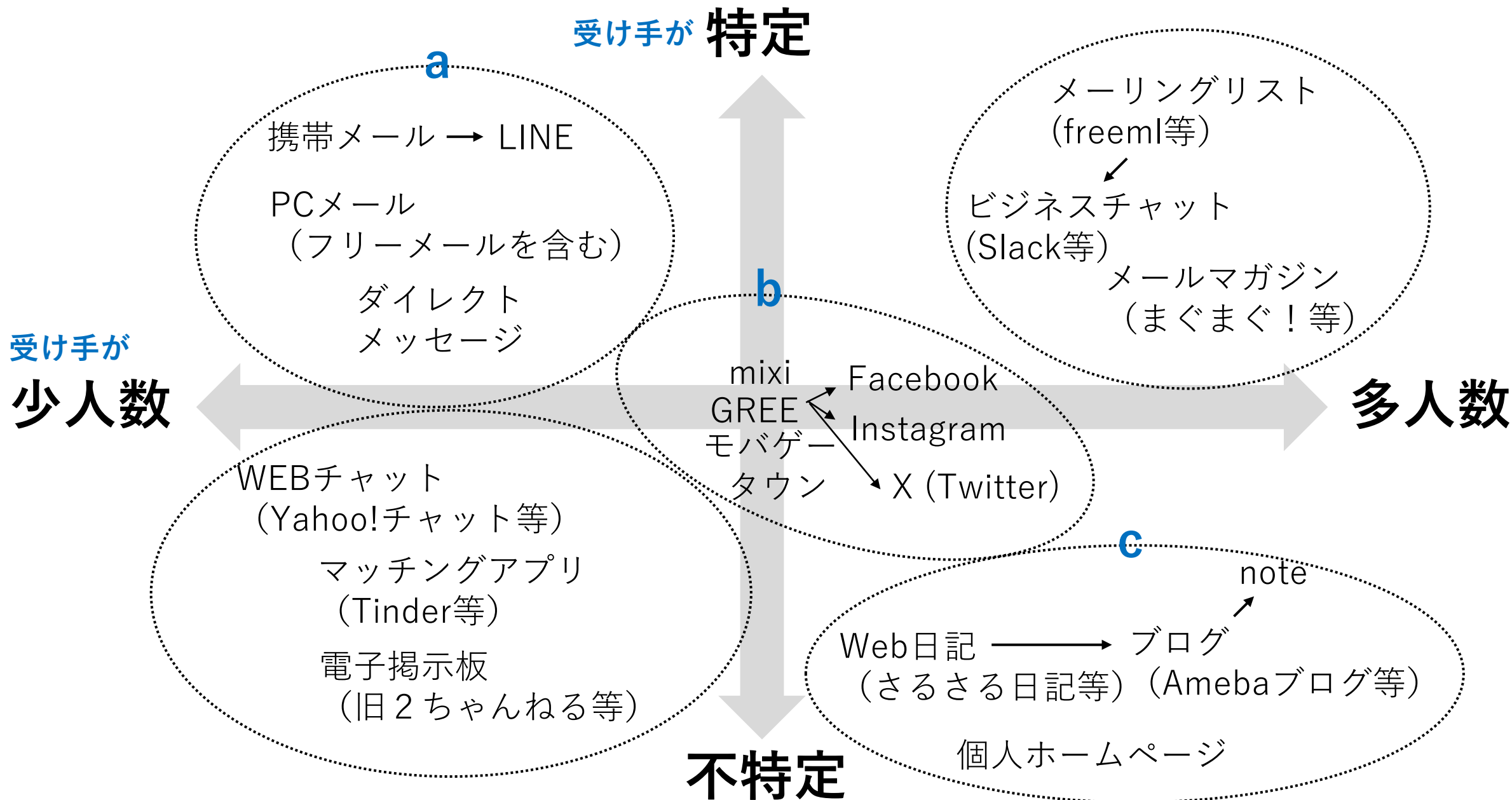
- BCCWJ1ではサンプリング
- BCCWJ2では全文（小・中・高の全学年ほぼ全科目、全国占有率の最も高いもの各1冊）
- 過去3回の学習指導要領改訂にそれぞれ対応した使用年度3か年分

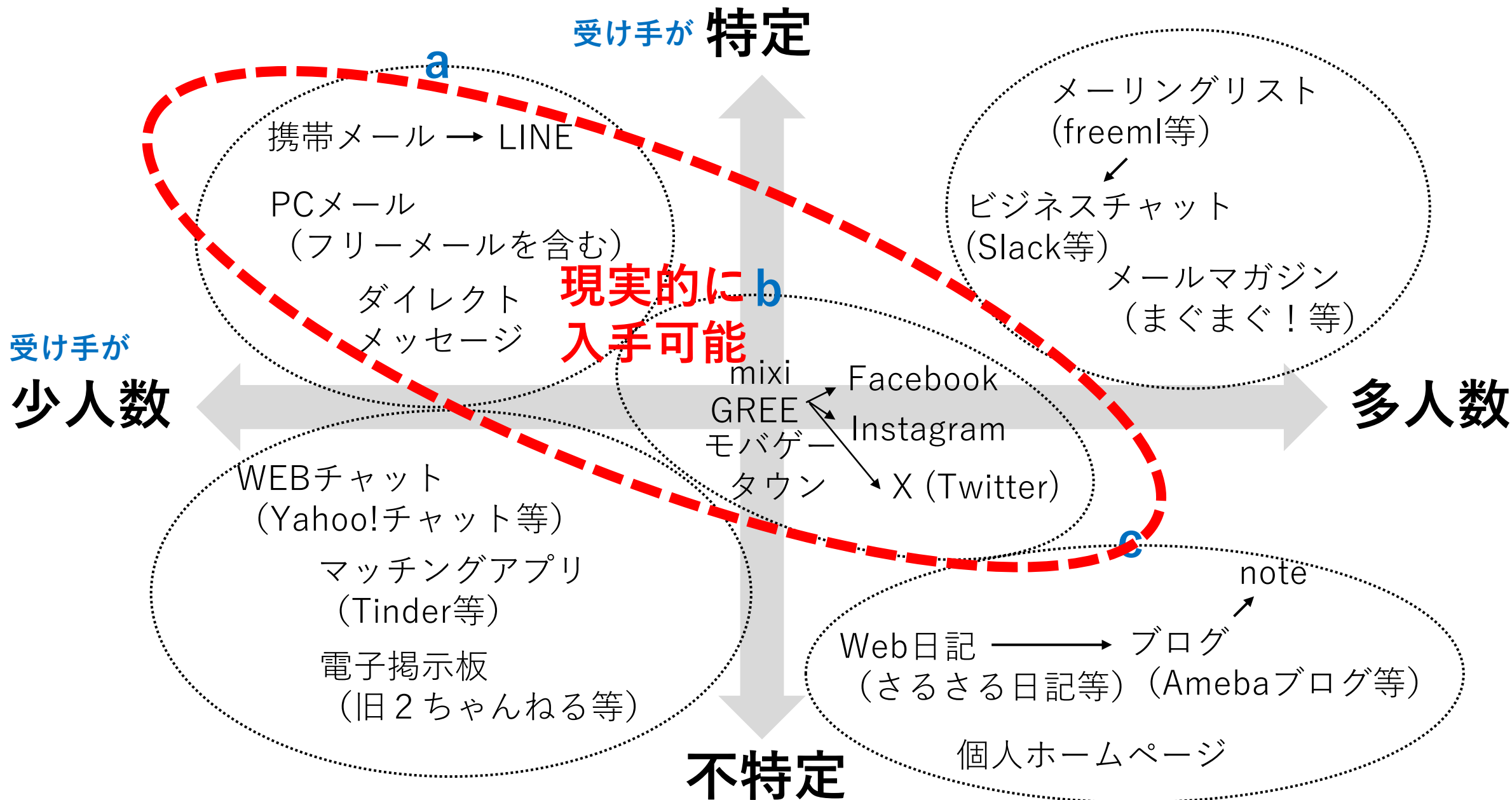
学習指導要領	小学校	中学校	高等学校
① 平成10-11（1998-1999）年改訂	平成18（2006） 60万語	平成17（2005） 91万語	平成18（2006） 349万語
② 平成20-21（2008-2009）年改訂	平成26（2014）	平成26（2014） 133万語	平成27（2015）
③ 平成29-30（2017-2018）年改訂	令和7（2025）	令和7（2025）	令和7（2025）

①～③期で2000万語を予定

SNS/CMCデータについて

- BCCWJ1のサンプリング対象期間後に生まれたメディア
 - **Twitter** 2006年3月サービス開始
 - **Facebook** 2006年9月に一般向けサービス開始
 - **LINE** 2011年6月サービス開始
 - Cf. **iPhone** 2007年6月販売開始
- 言語生活の実態に近い
 - **国民の43%がTwitter (X)を, 91%が LINEを利用 (総務省調査)**
- マスコミ等で様々な調査に利用
- データの入手可能性・価格設定はさまざま





入手可能なSNSデータ: a

- LINEのデータを収集
- 自動収集は不可。協力者からのデータ収集
 - 研究利用・公開についての許諾を得る
 - 3年計画(2024年～2026年)でLINEのテキストチャットのデータを収集
 - 協力者は東京出身の20代～30代延べ500名程度
 - 対照群として関西も予定
- 3年間で70万語程度を収集予定

入手可能なSNSデータ: **b**

- X (Twitter)は本文の再配布不可かつAPI有償
- 代替としてBlueskyとMisskeyのデータを収集中
 - **Bluesky** 国内登録ユーザー 150万人程度 (2025年5月時点)
 - **Misskey** アカウント数 90万程度 (2024年時点)
- 想定入手データサイズ (当初1年間で)
 - Bluesky 50億語程度
 - Misskey 15億語程度

BCCWJ2のアノテーション・メタデータ

- 形態論情報

- 短単位と長単位で二重解析
- 短単位は従来通り MeCab + UniDic（辞書の充実により自動解析精度は向上）
- 長単位用に解析器を新規開発
- 「コア」を設定して手修正で精度向上

- タグとメタデータ

- BCCWJ1では多数の研究用タグを考案し付与したがあまり使われていないので簡素化
- 『中納言』での検索に用いるデータを優先して付与

BCCWJ2の権利処理

- BCCWJの時代

日本の著作権法は権利者保護を重視しており、これがサイバー社会の進展の障壁となっていた。また同じく2005年に個人情報保護法が完全実施されたことやGoogle社によるGoogle books計画が始まっていたことも逆風の要因となった

⇒ BCCWJ1では当時の著作権法を遵守して約24,000件を処理。高い権利処理コストを負担

- BCCWJ後の著作権法改正

- 情報処理研究・産業化を念頭において権利制限規定を拡大

新たに導入された権利制限規定

- 30 条の 4（2012年改正）：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 ⇒ コーパスの利用全般はこれに該当する
 - 47 条の 4（2018年改正）：電子計算機における著作物の利用に付随する利用等 ⇒ コーパス全般の構築・配信はこれに該当する
 - 47 条の 5（2018年改正）：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 ⇒ 『中納言』などのコーパス配信サービスはこれに該当する
- ⇒ **権利制限規定の範囲内で構築・配信・利用可能なコーパスとすることで開発コストを軽減し、コーパスの価値も高める**

権利制限規定の範囲内でのコーパス利用方策

- 電子媒体（DVD等）での有償配布
 - ⇒ 契約によって30条の4に該当する範囲での利用に制約する
- 『中納言』等による無償オンライン公開
 - ⇒ 表示される文脈をBCCWJ1よりも制限することで軽微利用の範囲に収める
 - ⇒ 『昭和平成書き言葉コーパス』（SHC）の公開方式（小木曾ほか2024）

権利制限規定の範囲内でのコーパス利用

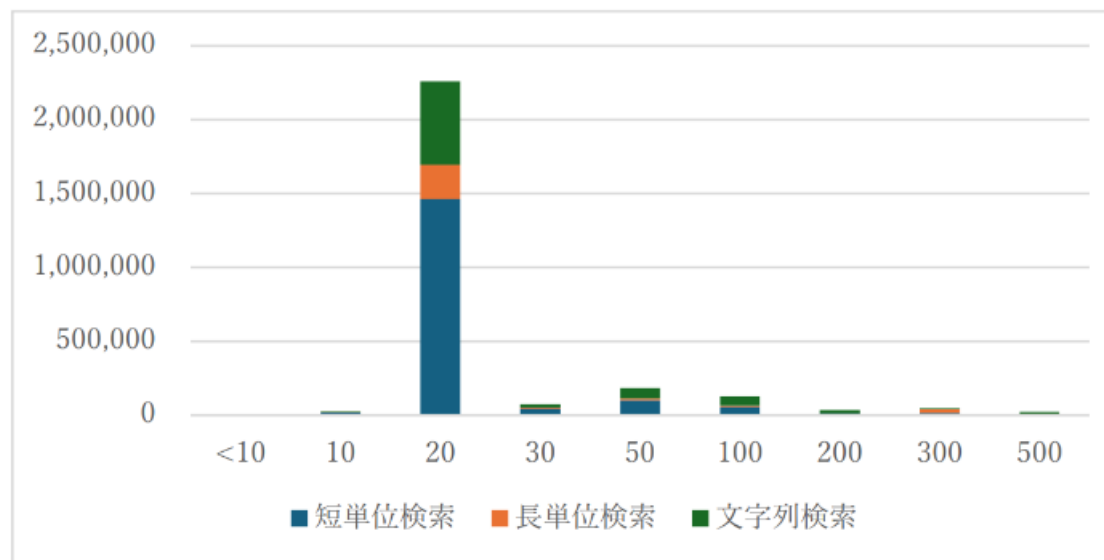
現在は最長500短単位まで表示可能

- BCCWJ1のユーザーはどのくらいの文脈長を必要としているか？

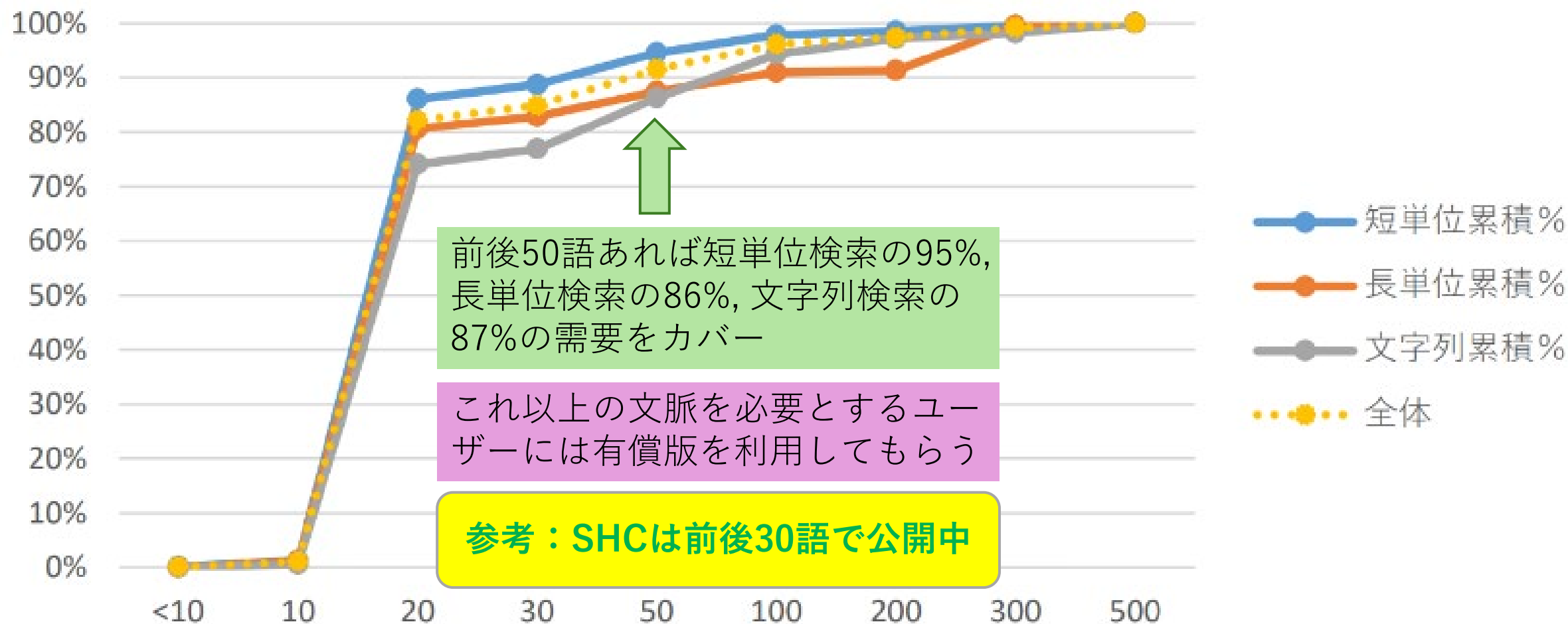
⇒ 『中納言』 検索ログ(278万)を分析

⇒ ほとんどがデフォルト設定（前後各20語）での利用

文脈長	<10	10	20	30	50	100	200	300	500
短単位検索	3,143	18,554	1,463,731	46,075	100,345	56,482	12,369	16,763	8,163
長単位検索	0	3,317	230,558	6,532	13,290	10,191	1,022	23,843	1,354
文字列検索	2	4,283	565,016	21,993	71,711	61,760	22,264	7,680	13,791
全体	3,145	26,154	2,259,305	74,600	185,346	128,433	35,655	48,286	23,308



権利制限規定の範囲内でのコーパス利用



BCCWJ2ロードマップ案

年 度	R6(2024)	R7(2025)	R8(2026)	R9(2027)	R10(2028)
システム	クラウド化検討 ・コーパス構築システム	クラウド化検討 ・コーパス構築システム /コーパス公開システム インターフェイス改修 (中納言・少納言)			
構築	母集団の決定(2006-2023) データ整備(2006-2010分) ・PDF化/入力/タグ付け ・形態素解析 ・DBへの格納 PDF化のみ(2016-2019)	母集団の決定(2024年) データ整備(2011-2015分) ・PDF化/入力/タグ付け ・形態素解析 ・DBへの格納	母集団の決定(2025年) データ整備(2016-2020分) ・PDF化(2020)/入力/タグ 付け ・形態素解析 ・DBへの格納	データ整備(2021-2025) ・PDF化/入力/タグ付け ・形態素解析 ・DBへの格納	データ整備 ・PDF化/入力/タグ付け ・形態素解析 ・DBへの格納
※形態素解析には、コアデータの整備を含む					
公開		2006-2010分 (約2500万語)	2011-2015分 (約2500万語追加)	2016-2020分 (約2500万語追加)	2021-2025分 (約2500万語追加) 公開完了
※データ整備作業は5年間でPDF化・入力・タグ付け等を行う。作業が終了したものから順次データを精査し、公開の準備を進めるため、公開時期とは異なってくる。					
その他	日本語学会発表 言語処理学会発表	日本語学会発表 言語処理学会発表	イベント(仮)	イベント(仮)	シンポジウム

まとめ

- BCCWJ1の出版サブコーパスを拡張する形で2006-2025の20年間に出版された書籍を母集団として1億語のBCCWJ2を構築中
 - 雑誌・新聞データは『昭和平成書き言葉コーパス』が部分的に提供
 - 書籍と別に教科書とSNS/CMCデータのコーパスを構築中 ⇒ BCCWJ2に統合予定
- 著作権法の権利制限規定の範囲内での利用を前提として権利処理が不要な形で公開
- 2025年度末から段階的に公開
- 成果発表会的なイベントも開催

参考文献

小木曾智信・近藤明日子・高橋雄太・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).

落合哉人「「打ちことば」の特質：分類と実態の解明」日本語学, 44(2), 14-25, 明治書院, 2025.

前川喜久雄「コーパス構築と著作権保護」人工知能学会誌25(5), 2010:09の初版原稿
(<https://clrd.ninjal.ac.jp/bccwj/doc.html> よりダウンロード可能)

前川喜久雄監修・山崎誠編『書き言葉コーパス：設計と構築』朝倉書店, 2014.

山崎誠・小木曾智信・高田智和「『現代日本語書き言葉均衡コーパス (BCCWJ)』の拡張計画について」日本語学会2024年春季大会予稿集, 2024.

山崎誠・高橋雄太・小木曾智信「『現代日本語書き言葉均衡コーパス』の拡張とその設計」日本語学会2025年春季大会予稿集, 2025.

BCCWJ2ホームページ <https://www2.ninjal.ac.jp/BCCWJ2/>