小木曽智信・近藤明日子・髙橋雄太・間淵洋子（印刷中）
『昭和・平成書き言葉コーパス』の設計・構築・公開
『情報処理学会誌』65(2)
2023 年 2 月刊行予定

情報処理学会論文誌 Vol.0 No.0 1–14 (Feb. 2024掲載予定)

『昭和・平成書き言葉コーパス』の設計・構築・公開

小木曽 智信[1,a)] 近藤 明日子[2] 髙橋 雄太[3] 間淵 洋子[4]

受付日 2023年5月16日, 採録日 2023年9月12日

概要：日本語の歴史的変化を研究するための基礎資料として昭和・平成期の雑誌・ベストセラー書籍・新聞を収録した『昭和・平成書き言葉コーパス』を構築しオンラインで公開した．このコーパスは 1933 年から 2013 年までの間を 8 年おきに 11 か年分，合計約 3,340 万語を収録した大規模なもので，明治・大正期までの『日本語歴史コーパス』の後を承け，現代に至るまでの日本語の成り立ちを探ることのできる資料として，日本語研究に重要な役割を果たすことが期待される．本稿はこのコーパスの設計と構築方法，公開形態について論じ，さらにこのコーパスを応用した研究例の一端を示す．

キーワード：コーパス，日本語学，形態論情報，著作権，コンコーダンサー

Design, Construction and Publication of the Showa-Heisei Corpus of Written Japanese

Toshinobu Ogiso[1,a)] Asuko Kondo[2] Yuta Takahashi[3] Yoko Mabuchi[4]

Received: May 16, 2023, Accepted: September 12, 2023

Abstract: The "Showa-Heisei Corpus of Written Japanese," which contains magazines, best-selling books, and newspapers from the Showa and Heisei eras, has been constructed and made available online as a basic resource for research on the historical changes in the Japanese language. This is a large-scale corpus that contains a total of approximately 33.4 million words for 11 years, every 8 years from 1933 to 2013, and is expected to play an important role in Japanese language research as a resource for exploring the origins of the Japanese language up to the present day, succeeding the "Corpus of Historical Japanese" that contains materials up to the Meiji and Taisho eras. This paper discusses the design and construction of this corpus, the form in which it is published, and provides some examples of research that has applied this corpus.

Keywords: corpus, study of Japanese language, morphological information, copyright, concordancer

1. はじめに

近年，日本語の実証的研究を支える基礎資料としてコーパスの利用が盛んになり，日本語の歴史的研究においてもコーパスは欠くことのできないものとなっている．こうした研究のために国立国語研究所で構築されたコーパスが広く用いられてきたが，『日本語歴史コーパス』（CHJ）[*1]が前近代から明治・大正期までを対象とし，『現代日本語書き言葉均衡コーパス』（BCCWJ）[*2]が 1976 年以降の主として 2001-2005 年を対象として構築されていることから，両者をつなぐ時期の書き言葉のコーパスが欠けており，明治から現代までを通じた日本語の変化を実証的に研究することが困難な状況にあった．

そこで，昭和・平成期の空白期間を埋め CHJ と BCCWJ をつなぐ書き言葉資料として雑誌・ベストセラー書籍・新聞のコーパスを収録した『昭和・平成書き言葉コーパス』（SHC）を新たに構築・公開することを計画し，科学研究費補助金の援助を得て，これを実現するに至った．

本稿は，このコーパスの設計と構築方法，公開形態について論じるものである．コーパスの設計は先行する他のコー

[1] 人間文化研究機構 国立国語研究所
[2] 東京大学
[3] 明治大学
[4] 和洋女子大学
[a)] togiso@ninjal.ac.jp

[*1] https://clrd.ninjal.ac.jp/chj/

[*2] https://clrd.ninjal.ac.jp/bccwj/

パスとの互換性・一貫性を保証するために，多くは過去のコーパスのものを踏襲したが，対象資料の特性に応じた工夫を行った．また，構築に当たっては，原資料の著作権の扱いについては，改正された著作権法に基づいて著作権処理を行わないで適法に公開するなど，新たな試みを行っている．

以下，SHC の設計，構築，形態論情報の付与，公開の形態について順に説明する．そのうえで，このコーパスを活用した研究事例を紹介し，本コーパスの有用性を示す．

2. 設計

BCCWJ に収録された新聞・雑誌・書籍や，CHJ の雑誌・文学作品・新聞等と比較できるように，同一ジャンルの信頼できるテキストのコーパスが望まれる．

コーパスの設計は既存のコーパスとできる限り共通化し，両者とあわせて利用可能な設計とした．

2.1 既存のコーパスとの関係

近代および現代の大規模な日本語のコーパスとして，BCCWJ と CHJ「明治・大正編」[*3]がある．

BCCWJ は現代日本語書き言葉の全体像を把握するために構築されたコーパスであり，書籍・雑誌・新聞・白書・ブログ・ネット掲示板・教科書・法律などのジャンルにまたがって 1 億語を収録する．BCCWJ は大きく 3 つのサブコーパスに分かれ，収録対象とする期間はサブコーパスにより異なる．主要メディアである書籍・雑誌・新聞の 3 種のレジスターのデータを収録する「出版（生産実態）サブコーパス」は，2001〜2005 年を対象とする．また，図書館に所蔵された書籍のデータを収録する「図書館（流通実態）サブコーパス」は 1986〜2005 年を対象とする．残る「特定目的サブコーパス」は上記 2 つのサブコーパスには収録されないレジスターのデータを収録する．その収録期間はレジスターにより区々であるが，古くは 1976 年以降，新しくは 2009 年までを対象とする．このように，BCCWJ ではメインの収録対象期間は 2001 年〜2005 年の 5 か年になっている．BCCWJ はもとより現代日本語の書き言葉の共時的な調査を念頭に設計されており，一定の年代幅の資料を収録しているとは言え，通時的な研究に用いるには大きな制限がある．また，BCCWJ の主たる収録対象期間から現在までに 15 年以上が経過しており，その期間のデータの拡張も課題であった．

一方，CHJ「明治・大正編」は，明治・大正期の日本語の全体像を見渡せるよう構築されたコーパスである．書き言葉の資料としては雑誌・国定国語教科書・啓蒙書・小説・新聞のデータが収録されており，収録対象期間は 1868〜1925 年（国定国語教科書のみ 1947 年まで）になっている．なかでも大きな割合を占めるのが「明治・大正編 I 雑誌」[*4]であり，ここには 1874〜1925 年の各年代を代表する総合雑誌を 6〜8 年間隔で収録している．これは近代語のコーパスの先駆である『太陽コーパス』[8] を基礎に，それ以前の時期の雑誌を追加し形態論情報を付与して構築されたもので，具体的には 1874・1875 年刊の『明六雑誌』，1881・1882 年刊『東洋学芸雑誌』，1887・1888 年刊の『国民之友』，1895・1901・1909・1917 年・1925 年『太陽』を収録する[*5]．

SHC はこの CHJ と BCCWJ との間の空隙を埋め，さらに BCCWJ 以降の期間を補う書き言葉のコーパスとして設計した．

2.2 収録資料の選定

SHC に収録する資料は，(1) 広く読まれて社会的な影響が大きい，(2) 明治から平成まで継続的に刊行されてきた，(3)CHJ・BCCWJ に収録されておりコーパスを接続可能である，という観点から，雑誌・ベストセラー書籍・新聞の 3 種のレジスターとした．ただし，全ての年にわたってデータを作成することは現実的でないため，長期的な変化の観察に主眼を置き，CHJ「明治・大正編 I 雑誌」で刊行年 6〜8 年間隔で雑誌を収録した方法に倣い，1933 年から 2013 年まで刊行年 8 年間隔に 11 か年分を収録した．これにより昭和・平成期の 80 年間という期間をカバーした．レジスターごとの具体的な収録資料とサンプリング方法を以下に示す．

雑誌

総合雑誌を収録する CHJ「明治・大正編 I 雑誌」との接続を意図して，収録資料として昭和・平成期を代表する月刊総合雑誌『中央公論』『文芸春秋』の 2 誌を選定した．1933・1941・1949・1957 年刊の『中央公論』と 1965・1973・1981・1989・1997・2005・2013 年刊の『文芸春秋』，計 11 か年，各年の通常号 12 冊，全 132 冊の全文テキストを収録した．ただし，(1) 表紙，(2) 目次，(3) 奥付，(4) グラビア等の写真・絵・図表を中心とする記事，(5) 前号の誤植等を謝罪する記事，(6) 広告，(7) 付録，は雑誌の主要なテキストとは見做せないものとしてコーパス収録対象外とした．また，収録対象とした記事中にある写真・絵・図表に付属するテキスト，漢文・外国語・数式からなる段落のテキストは入力の対象外とした．

ベストセラー書籍

1933 年から 1973 年までは国立国語研究所国語辞典編集準備室（1984）[9] に掲載されているコーパスデータ収録対象年の全作品を選定し，1981 年以降は「年間ベストセラーアーカイブ」（トーハン）[2] や『出版年鑑』（出版ニュース

[*3] https://clrd.ninjal.ac.jp/chj/meiji_taisho.html

[*4] https://clrd.ninjal.ac.jp/chj/meiji_taisho.html#zasshi

[*5] 総合雑誌の対照資料として 1894〜1895 年刊の『女学雑誌』，1909 年刊の『女学世界』，1925 年刊の『婦人倶楽部』の 3 つの女性誌も収録されている．

社）[11] を参照して上位 20 位までの書籍を選定した．これにより 1933 年から 2013 年までで計 259 組の書籍を収集した．このうち，A. 写真集やコミック，マップなど写真・イラスト・図が中心の書籍 25 組，B. 数式などの特殊な文字列を中心とする書籍 3 組，C. ベストセラーにノミネートされた年より 3 年以前に発表された作品を中心とする書籍（全集や古典作品など）12 組，D. 同一作者による作品が 3 作以上ベストセラーにノミネートされた場合の言語量の偏りを解消するための書籍 9 組，計 49 組を収録対象から除外し，最終的に 210 組の書籍を採用した．ジャンル間比較を可能とするために，ベストセラー書籍は新聞と同程度の規模とすることを目指し，旧字体中心の 1933 年から 1949 年は各年 25 万語前後，1957 年から 2013 年は各年 40 万語前後となるよう設計した．また，ベストセラー書籍ではランダムサンプリングによるテキスト採集を行った．各書籍のコーパステキスト収録対象とする本文のページ（内表紙や目次，前書き，後書き，広告，奥付などを除いたページ）のうち，無作為に抽出したページをランダムサンプリング起点として設定し，その起点のページの前後で，事前に定めた作品あたりの収録規模に合わせて，章や節などのある程度の文章のまとまりを確保した上でコーパステキストを採集した．

新聞

戦前から戦後，さらに現在にかけての公共的な書き言葉の変化を捉えることができる資料として，BCCWJ および CHJ「明治・大正編 V 新聞」[*6]の両コーパスで収録対象になっている『読売新聞』を選定し，コーパステキストを収集した．1933・1941・1949・1957・1965・1973・1981・1989・1997・2005・2013 の各年からの収録対象のサンプリングにおいては，1 か年につき原則として 5 月 2 日・11 月 2 日の 2 日分の全国版朝刊 1 冊のテキストを取得した「明治・大正編 V 新聞」を拡張し，より多くのデータ量を確保することを目的に，原則として奇数月 2 日（2 日が休刊日の場合は 3 日）の計 6 日分の全国版朝刊 1 冊を対象とした．なお，テキストの収録範囲は，「明治・大正編 V 新聞」に倣い，広告記事および固有名詞（人名・地名）や数値の羅列を中心とする記事（叙任・辞令，スポーツの試合結果・株式の取引結果など）と，図表や挿絵の中のテキスト，それらのキャプションなどに相当する文書要素を除外した全文とした．

2.3 データ整備の方針

このコーパスを構築する前提として，限られた時間（科研費の研究計画期間である 4 年間）で大量のテキストを整備する必要があった．また，後述する著作権法上の権利制限規定を踏まえた形で公開を行うために，全文の生データの公開を行うことは出来ず，コーパス検索アプリケーション「中納言」によるオンラインの検索と，統計情報データの形でしか提供できないことが当初より確定していた．

そこで，テキストの整備やアノテーションにあたっては質の面では BCCWJ や CHJ 並みの精度にはこだわらず，確実に昭和・平成期をカバーするしっかりとした量のテキストを確保することを優先することとした．

そのため，文字セットについては，異体字や外字の処理の手間が少なくなるようレジスターごとに適した方法を用いることとし，テキストに対するマークアップについても公開する範囲で必要な最小限のタグセットに留めることとした．また，形態論情報の整備についても，人手による完全な整備を行うことは最初から意図せず，既に整備済みの形態素解析用辞書を活用し，機械解析結果に対して可能な範囲で修正を加えるに留めた（BCCWJ や CHJ における「非コアデータ」に相当）．

このように，BCCWJ や CHJ の構築時のノウハウを活かしつつ，これらと比較して簡素な整備に留めることによってデータ量を確保し，CHJ が扱う前近代から明治・大正期に続く，昭和・平成期の日本語の変化を見渡すことのできるコーパスの完成を目指した．

3. 構築

3.1 テキスト作成

3.1.1 テキスト入力の方針

まず，テキスト作成の方針をレジスターごとに以下に示す．

雑誌・ベストセラー書籍

入力に使用する文字セットは，JIS X 0213: 2004 の文字集合に準拠した[*7]．

文字集合に含まれない非漢字（仮名・記号等）は，形・用途の近い文字集合内の文字によって電子化した．

文字集合に含まれない漢字については，以下の (1)〜(5) の手順で電子化した．

(1) JIS X 0213 の「6.6.3 漢字の字体の包摂規準」に基づき，文字集合内の文字に包摂する．
(2) (1) の包摂規準を適用できない字形差をもつ漢字のうち，微細な字形差を持つ漢字は，近代語用の追加包摂規準（須永・堤・近藤ほか 2013[16]）に基づき文字集合内の文字に包摂する．
(3) (2) の包摂規準を適用できない字形差をもつ漢字は，類似の意味・用法を持ち，同一の漢字部品を有する同音・同訓の文字集合内の文字で代用する．

[*6] https://clrd.ninjal.ac.jp/chj/meiji_taisho.html#shinbun

[*7] ただし，JIS X 0213 附属書 7 2.1 b）に掲載される，戸籍法施行規則付則別表「人名用漢字許容字体表」（昭和 56 年法務省令 51）の漢字，及び常用漢字表（昭和 56 年内閣告示第 1 号）のかっこ書き内の漢字（「いわゆる康熙字典体」）のうち，JIS X 0208 で包摂していた漢字，JIS X 0213: 2004 において UCS との互換のために追加された 10 字，についてはこれを用いない．

(4) (3) による代用が不可能な文字のうち，Unicode に収録されている文字はそれにより電子化する．
(5) (4) による入力が不可能な文字は，「〓」（げた記号，JIS 面区点 1-02-14，U+3013）で表す．

原本の文字のかすれや破損によって判読が困難な文字・記号は，「␣」（空白記号，JIS 面区点 1-07-93，U+2423）によって表した．

そして，入力したテキストに対して，①濁音だが濁点のない仮名が使用されている場合は濁点付き仮名に変換，②踊り字は繰り返す文字列に変換，③誤植と考えられる箇所は訂正後の文字列に変換，等の校訂を行った．これらの校訂においては，校訂前のテキストの情報が参照できる形でマークアップを行った（3.3 参照）．

新聞

文字セットは Uniocde とし，Unicode により表現できる文字はできるだけ原文に忠実に入力した．JIS 規格による包摂等を行わなかった．Unicode による入力が不可能な文字は「〓」（げた記号，U+3013）で代替した．

判読が困難な文字・記号の入力方針や，テキストの校訂方針は，雑誌・ベストセラー書籍と同様である．

3.1.2 テキスト入力の工程

次に，レジスターごとのテキスト入力の具体的な工程を以下に示す．

雑誌

原本またはその複写に基づき，専門業者にテキスト入力を外注した．漢字については，いわゆる新字体・旧字体が入れ替わって入力される等，原本の字体どおりの入力になっていない可能性がある．

ベストセラー書籍

ベストセラー書籍は，雑誌や新聞に比較して段組みなどが簡素であるため，OCR ソフトの自動読み取りによるテキスト作成を中心にテキスト化を行った．原本を裁断・画像化した上で，図表やページ番号などの不要な要素にマスキング処理を施し，読み取りを行った．OCR による読み取りでは，フォントタイプやルビ，傍点や注記号といった諸要因によって精度が揺れるため，作業者と構築者による原本照合のダブルチェックを行ってテキストの精度を高めたほか，漢数字の「二」とカタカナの「ニ」，ひらがなの「へ」とカタカナの「へ」など，読み取り誤りが多い部分については統一的なチェックを行った．ただし，1957 年までにみられる旧字体資料や，紙の劣化や汚れによって読み取り精度が著しく下がる資料については，テキスト入力作業を専門業者に外注してテキストを作成した．

新聞

原本またはその画像，複写等に基づき文字入力して作成した．その際，データ化の対象は本行のみとし，振り仮名等は入力しなかった．

3.2 サンプル単位

以上のようにして作成したテキストは，適当な長さに分割しコーパスに収録した．分割した各テキストを「サンプル」と呼ぶ．サンプルの分割方針と基本統計量をレジスターごとに以下に示す。

雑誌

1 記事を 1 サンプルとしてテキストを分割した．サンプル数は 10003，平均 2739.1 語（最小値 1 語，最大値 58227 語），標準偏差 3782.6 であった。

ベストセラー書籍

各書籍の章節項などの階層のうち，最も小さい階層を 1 サンプルとして分割した．章末などに含まれる章まとめの文章やコラム記事なども独立した 1 サンプルとして扱った．ベストセラー書籍のサンプル数は 3223，平均 1069.2 語（最小値 6 語，最大値 17585 語），標準偏差 1322.6 であった。

新聞

1933・1941・1949・1957 年の 4 か年については，原則として 1 ページを単位として，1965・1973・1981・1989・1997・2005・2013 の 7 か年については，1 記事を単位としてサンプルを分割した．1933〜1957 年は整備の都合により記事単位での分割ができなかったため，ページ単位での分割になっているが，ページ内に文芸作品（小説や短歌，俳句，詩など）を含む場合は，その部分のみ記事の単位で切り出した．ページ単位で分割した 1933〜1957 年のサンプル数は 156，平均 3151.1 語（最小値 124 語，最大値 8559 語），標準偏差 2182.6，記事単位で分割した 1965〜2013 年のサンプル数は 4984，平均 456.6 語（最小値 122 語，最大値 6064 語），標準偏差 446.6 であった。

3.3 マークアップ

以上のように作成したテキストに XML を用いて必要な情報をマークアップした．SHC は著作権法上の制限から，全文やソースデータの公開を行うことができないコーパスとして設計した．したがって，当初より豊富なタグ付けを行うことは検討せず，公開形態にそって必要十分な簡素なタグセットでマークアップを行った．たとえば，形態素解析の入力として必要であり，かつ検索条件としても有効な文境界，言語研究に有用な発話・引用等の情報等を，レジスターの特性に合わせて付与した（**表 1**）．

ベストセラー書籍・新聞に比べて雑誌で多くの種類のタグを使用しているのは，雑誌誌面のレイアウトやテキスト構造の複雑さを表現するために必要であったのと同時に，構築が先行した雑誌において当初は CHJ「明治・大正編 I 雑誌」や BCCWJ に倣って詳細なマークアップを行う設計としていたためである．それに対してベストセラー書籍・新聞は，SHC の設計が明確になってから構築を開始したため，必要最小限のマークアップを行った．

表 1 マークアップに用いた XML タグ一覧
Table 1 A list of XML tags.

XML タグ	説明	雑誌	ベストセラー書籍	新聞	「中納言」表示
sample	サンプル	○	○	○	○
p	段落	○	-	-	-
rejectedBlock	入力対象外要素	○	○	○	-
warigaki	割書	○	○	○	-
quotation	会話・引用	○ *	○	-	○
s	文	○	○	○	○
odoriji	踊字	○	○	○	○
fraction	帯分数の中の真分数部分	○	-	-	-
pb	ページ開始位置	○	○	○	○
cb	段開始位置	○	-	-	-
lb	行開始位置	○	-	-	-
br	物理改行	○	○	○	-
ruby	ルビ	○	○	-	○
corr	誤植を訂正した文字	○	○	○	○
g	外字等の特殊文字	○	○	○	○
unclear	不鮮明だが字体推定が可能な文字	○	-	-	-
vMark	濁音だが濁点のない仮名が使用	○	○	○	○
superScript	上付き文字	○	○	-	-
subScript	下付き文字	○	-	-	-
enclosedCharacter	囲み付き文字	○	-	-	-
noteMarker	行中の注参照マーカー	○	-	-	-
noteBodyInline	行の傍らに現れる注記	○	-	-	-

* ジャンルが「文芸」のサンプルのみマークアップ

XML でマークアップした情報は，コーパスの各種公開形態でその一部を公開した．例えば，コーパス検索アプリケーション「中納言」(5.3.1 参照) の検索結果画面において各タグでマークアップした情報が表示されるか否かを表 1 に合わせて示した[*8].

3.4 形態素解析

BCCWJ や CHJ と比較可能なコーパスとするためには，形態素解析によって単語情報を付与することが必須である．さらに，BCCWJ や CHJ の形態素解析で使用された形態素解析用辞書 UniDic を用いることで，単位の互換性を維持することが求められる．

SHC では，形態素解析器 MeCab[1][*9]と下記の形態素解析用の辞書を利用して解析を行った．

- 現代語書き言葉 UniDic[7] [*10]
- 旧仮名口語 UniDic[14] [*11]

少しでも解析精度を向上させるために，公開されている辞書をベースとしつつ，ターゲットとなる資料に合わせて学習用のコーパスを調整し，見出し語の追加を行ったうえで辞書の再学習を行い，カスタマイズした辞書を用いた．

2 種類の UniDic とサンプルの対応関係をレジスター別に見ると，雑誌は 1933〜1957 年の全サンプル，および 1965〜2013 年の旧仮名遣いのサンプルは旧仮名口語 UniDic を，1965〜2013 年の残りのサンプルは現代書き言葉 UniDic を使用した．ベストセラー書籍は 1949 年までと 1957 年に一部に見られる旧字体のサンプルには旧仮名口語 UniDic，1957 年の一部と 1965 年以降の新字体の全サンプルには現代書き言葉 UniDic を使用した．新聞は 1933〜1941 年の全サンプルに旧仮名口語 UniDic を，1949〜2013 年の全サンプルに現代書き言葉 UniDic を使用した．

機械解析結果には多くの誤りが含まれるため，人手による修正を加えることが望ましい．コーパスに付与した形態論情報の詳細と，解析結果の修正についての詳細は 4 で説明する．

3.5 書誌情報

サンプルの書誌情報は XML とは別にリレーショナルデータベースとして作成し，XML を形態素解析したデータを格納する形態論情報データベース [15] と関連付けて管理した．付与する情報の種類は CHJ「明治・大正編」に

[*8] 「中納言」の検索結果画面における各種情報の表示方法の詳細については，SHC のウェブサイトで公開するコーパスの概説書を参照のこと．
[*9] https://taku910.github.io/mecab/
[*10] https://clrd.ninjal.ac.jp/unidic/download.html#unidic_bccwj
[*11] https://clrd.ninjal.ac.jp/unidic/download_all.html#unidic_qkana

倣って,「サンプル ID」「ジャンル」「作品名」「成立年」「巻名等」「作者」「生年」「性別」「底本」「ページ番号」「出版社」とした．これらの情報は「中納言」の検索結果画面に表示されるようにした．このうち,「サンプル ID」はサンプルを一意に定義する ID である．その他の情報についてレジスターごとに以下に述べる．

雑誌

「巻名等」に記事のタイトル,「作者」に記事の著者名・訳者名を原本の表示に基づき一部表記を改めて付与した．また，記事の内容からジャンルを「非文芸」「文芸/小説」「文芸/戯曲」「文芸/詩歌」の 4 種に分類し，情報を付与した．また，記事の掲載された雑誌の奥付に従って,「作品名」に雑誌名（「中央公論」または「文芸春秋」),「成立年」に雑誌の発行年,「底本」に「雑誌名 <YYYY-NN >」の形式で雑誌名・発行年（YYYY）・号数（NN),「出版社」に雑誌の出版社名を付与した．なお，記事の著者・訳者の「生年」「性別」の情報は付与しなかった．

ベストセラー書籍

原本の奥付の情報を基に,「底本」,「成立年」,「出版社」などの書誌情報を付与している．「作品名」には基本的に「底本」と同じタイトルを記載しているが，巻数や副題などの情報を落としているほか，一部の旧字体の作品を現在通用している漢字表記に改めて表示している場合がある．また，ベストセラーの収録時期（前年の 11 月～当年の 11 月）の関係で 1989 年のベストセラーに 1988 年発行の吉本ばなな著の『TSUGUMI』が収録されるといった事例があるが，データ利用のしやすさを重視して「成立年」は一律でベストセラーの収録年を記載している．「ジャンル」には，日本十進分類法（NDC）に準拠した情報を付与しており，900 番台（文学）の書籍には「文芸」の，それ以外には「非文芸」の情報を付与している．著者情報としては「作者」（翻訳作品の場合，訳者もここに含む),「性別」,「生年」の情報が参照できるほか，国立国会図書館典拠データ検索・提供サービス（Web NDL Authorities）[*12]へのリンクも実装している．

新聞

「作品名」は全て共通して「読売新聞」とし，発行日（YYYY-MM-DD）と号数（N）を,「底本」に「読売新聞 <YYYY-MM-DD-N 号 >」の形式で示した．記事を単位として分割したサンプルのうち，1965・1973・1981・1989 年の 4 か年については，記事のタイトルを「巻名等」に示した．残りの 1997・2005・2013 年と，ページを単位として分割した 1933・1941・1949・1957 年のサンプルについては，記事の分割やタイトル情報の整備上の理由から，記事タイトルの代替として面数または紙面記載の面種（国際・経済・社会等）を「巻名等」に示した．また，記事に関する情報の整備が十全なサンプルについては，著者関連情報として「作者」「性別」「生年」と Web NDL Authorities へのリンクを提供している．

[*12] Web NDL Authorities (https://id.ndl.go.jp/auth/ndla)

3.6 著作権処理

現代語のコーパスの構築において，従来は収録対象となる資料の著者の許諾が必要とされ，これが極めて大きな負担となっていた．前川（2009）は BCCWJ の構築時の著作権処理について次のように述べる．[4]

> BCCWJ では，書籍サンプルだけで約 25000 件の著作権処理を行う必要があるのだが，2006 年の 12 月以来，本稿執筆時点までの約 30 月間に約 16000 件について著作権者に連絡をとり，そのうち約 10000 件から利用許諾を得ることができた．この間の経費は研究員の人件費まで含めれば単年度で 1000 万円を大幅に超える水準にある．著作権処理のコストが現代語コーパス構築における最大のあい路といわれる所以である．

SHC のサンプル数を BCCWJ のそれと単純比較することはできないが，多数の雑誌記事を収録していることから，要する費用はこれを大きく下回ることはないと思われる．すなわち，まともな権利処理を行えば本コーパスを構築することは不可能であった．

昭和・平成期の資料をコーパスにすることの必要性が強く意識されつつも，著作権処理コストの点で行うことができない状況にあったが，その状況を変えたのが平成 30 年の著作権法改正である．

3.6.1 平成 30 年改正著作権法

この改正では,「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の整備」が行われた（著作権法第 30 条の 4，第 47 条の 4 及び第 47 条の 5 関係）．文化庁著作権課は，下記のようにコーパスを用いて日本語研究を行う事例を挙げて，これが権利制限の対象となることを示している [17].

> 問 14 日本語の表記の在り方に関する研究の過程においてある単語の送り仮名等の表記の方法の変遷を調査するために，特定の単語の表記の仕方に着目した研究の素材として著作物を複製する行為は，権利制限の対象となるか.
>
> 日本語の表記の在り方に関する研究は，特定の技術の開発や実用化を目的としない基礎研究であるが，当該研究の過程である単語の送り仮名等の表記の方法の変遷を調査するために，特定の単語の表記の仕方に着目した研究の素材として著作物を複製する行為は，あくまで研究の素材として著作物を利用するものであり，当該著作物の視聴等を通じて，視聴者等の知的・精神的欲求を満たすという効用を得ることに向けられた行為ではないものと考えられることから，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行為であると考えられる.

このようにコーパスを意識した内容を含む著作権法の改

正は，著作権処理に大きなコストを払うことなく現代語コーパスを構築する道を拓くものであった．本コーパスの構築は，この著作権法改正の報を受けて平成 30 年中に計画され，翌年度に科研費研究計画が採択されたことで実現したものである．

SHC のようなコーパスを「構築」することについては上述の通り「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行為」として，権利者の許諾を得なくとも問題ないことと考えられる．一方で，こうして作られるコーパスの「公開」については，その著作物の利用行為が「軽微」であるか否かが問われることとなる．国立国語研究所のコーパス検索アプリケーション「中納言」は，例示される所在検索サービスないしは情報解析サービスの一種であると考えられるが，いずれにしても原文を表示するにあたっては，原文の利用が軽微な利用にあたることが必要とされる（著作権法第 47 条の 5）．

3.6.2 文脈長の制限

著作物の軽微な利用としての条件を満たしてコーパスを公開するため，SHC では「中納言」で表示されるテキストの文脈長を制限することとした．認められる文脈の長さとして，明確な基準となるものはないが，

- 書籍の全文検索サービスである Google ブックス[*13]の文脈スニペット表示が約 120 字であること
- 日本新聞協会が 1 記事の 5〜10 %程度であれば軽微利用にあたるとしていること [12]

等を参考に，弁護士と相談のうえで，キーワードの前後それぞれ 20〜30 語（約 30〜50 字）の文脈表示であれば軽微な利用として認められるものと判断した．

そのために，BCCWJ の「中納言」では前後文脈の長さを最大で 500 語まで設定可能としていたところ，SHC では最大で 30 語（標準で 20 語）に制限した．また，この制限は一般的なユーザーインターフェイスを通して利用する場合だけでなく「検索条件式」などの仕組みを利用した場合にも維持されるように設定した．

3.6.3 短い著作物の排除

一般記事については上記の文脈長制限で軽微利用の範囲に絞り込めると考えられるが，これでも問題が生じるケースがありうる．具体的には検索結果の前後 30 語の文脈中に，ごく短い記事や作品，例えば短歌や俳句などの全文が含まれてしまうことによる．これらは各々が独立した著作物であるため，その全文の公開は形式上，違法となる可能性がある．

そこで，このような事例を排除するために，SHC の構築にあたっては，雑誌や新聞の投書欄等，短歌・俳句からなるサンプルは公開対象から外すことで対処した．記事全体がこうした内容からなるものはその全体を除外し，本文の一部としてとられている短歌・俳句については原則として文字数分の「■」（黒四角，JIS 面区点 1-02-03，U+25A0）で入力し伏せ字にして処理した．

レジスター別に見ると，公開までに次のような処理を行って短い著作物を排除することで軽微利用の範囲を守った．雑誌では，俳句・和歌を主な内容とするサンプル，およびジャンルが「文芸/詩歌」のサンプルのうち 121 語以下[*14]のものを非公開とした．ベストセラー書籍では，著作権保護期間内の作者による俳句・短歌のみが該当した．俳句・短歌の作者が不明な場合，あるいは作者の没年が不明な場合は，発表から経過が 50 年未満のもの（SHC では 1973 年以降）に限り伏字処理を行った．創作作品においては基本的に伏字処理の対象外となるが，一部，実在の人物をなぞらえた登場人物（『新・人間革命』における山本伸一＝著者の池田大作など）による作品については，伏字処理を行った．新聞では，短歌・俳句，子どもの詩などを中心とするサンプル，サンプル長が 121 語以下となるサンプルについては，原則として非公開とした．また，ページを単位として分割したサンプル内に含まれる短歌・俳句，子どもの詩等については，その部分の本文を伏せ字化して公開した．

3.7 個人情報保護のための処理

著作権の問題とは別に，コーパスのサンプルとして採用されたテキスト中の個人情報の問題がある．SHC が対象とした雑誌・ベストセラー書籍・新聞については，いずれも広く刊行されたものであり，現在の眼で見たときにも個人情報保護の必要性は必ずしも高くない．しかし，刊行当時の記事における被害者や被疑者・犯人について，今日的には公開することが適当であるとは認めがたい個人名や住所等の情報が含まれている場合がある．そこで，コーパスの公開にあたり，特に雑誌・新聞のテキストについては文字数分の「■」（黒四角，JIS 面区点 1-02-03，U+25A0）で入力し伏せ字にして処理した．ベストセラー書籍は該当する個人情報がなかったため伏せ字化は行わなかった．

以上のコーパス公開の方針に関しては，弁護士の確認[*15]の下で検討し，適法にコーパスの本文の公開を行っている．なお，権利者からの問題指摘については，コーパスの公開ページから意見を受け付け，問題が生じた場合にはすぐに対処可能な体制を整えることとした．

4. 形態論情報の付与

4.1 形態論情報の特徴

3.4 に述べたように，SHC ではレジスター・刊年に合わ

[*13] https://books.google.co.jp/

[*14] 「中納言」の表示可能な最大文脈長 61 語（前後文脈各 30 語＋キー 1 語）の中に，サンプルの半分以上が表示されるものを非公開とするための設定．

[*15] 2023 年 3 月 2 日に高樹町法律事務所の小林利明弁護士に確認．

せてカスタマイズした「現代書き言葉 UniDic」「旧仮名口語 UniDic」を用いた形態素解析結果を形態論情報としている．UniDic の語の単位は「短単位」と呼ばれる比較的短い単位で，例えば「私は国立国語研究所に勤務している.」というテキストであれば「私／は／国立／国語／研究／所／に／勤務／し／て／いる／.」と認定する．短単位は曖昧さや矛盾のない単位認定が可能で，言語研究に必要な単位の斉一性が担保される．また，短単位は語形や表記の変異にかかわらず，同一の語であるかの認定を容易にするため，辞書の見出しに相当する最上位の階層「語彙素」のもとに，語形の変異を区別する「語形」，表記の変異を区別する「書字形」，発音の変異を区別する「発音形」という複数の階層を持つ構造をとる．これにより見出しの同一性が保持され，語形・表記の変異に関する研究が可能となっている（伝・小木曽・小椋ほか 2007[7]）．BCCWJ や CHJ の形態論情報も，UniDic による形態素解析結果を利用した短単位の形態論情報が付与されており，コーパス間の比較研究を可能にするためにも，SHC ではこの短単位の形態論情報の付与を必須とした．なお，BCCWJ では短単位のほかにより長い単位である「長単位」の形態論情報も付与されているが，CHJ「明治・大正編」では付与されていないこともあり，SHC でも付与は行わなかった．

4.2 形態論情報の修正

4.2.1 修正方法

UniDic による形態素解析結果に基づいて形態論情報を付与するとはいえ，解析結果には誤りが含まれるため，できる限り人手による修正を行い，形態論情報の精度を高めることが望ましい．一方で，コーパスの大量のデータ全てに対して人手修正を行うことは現実的でないため，BCCWJ や CHJ では，ほぼ完全な修正を行った高精度なコアデータと，機械解析結果に一定程度の修正を行っただけの非コアデータに分けて整備することが行われてきた．短期間で構築した SHC においては，形態論情報の修正にかけられる労力や時間が限られていたため，コアデータの作成は行わず全体を非コアデータ相当とし，特に形態論情報の誤りが生じやすい部分を重点的に修正した．サブコーパス共通で取り組んだ修正としては，例えば「未知語」処理が挙げられる．「未知語」とは，当該の文字列が BCCWJ や CHJ をはじめとする過去のコーパスの構築過程で出現しなかったために UniDic に登録されてこなかった語や語形，書字形に対して，形態素解析で付与される特殊品詞である．「未知語」は外来語や外国の固有名詞のようなカタカナ文字列や，現代では使われない漢字列に付与されることが多く，語彙素の同定ができる範囲で UniDic への登録および形態論情報の付与を行った．修正には，BCCWJ の構築以来用いられているコーパス修正ツール「大納言」を利用した [15]．「未知語」の処理のほか，各レジスターで行った修正は以下の通りである．

雑誌

形態素解析後に修正を行う方法ではなく，登録語の整備による UniDic のカスタマイズの作業に注力し，形態素解析の精度自体を高める方法をとった．テキスト入力が先行していた 1933〜1997 年のテキストを既存の UniDic で仮に形態素解析し，その結果から，(1) 品詞が「未知語」の語，(2) 漢字 1 字を書字形とする語の連続，(3) カタカナを書字形とする語の連続，(4) 文語活用動詞，を抽出し，その中から既存の UniDic に未登録の語を探し追加登録した．また，既存の旧仮名口語 UniDic は近代のテキストに対応するため，近代に使用される語も登録されていた．その中には，昭和期には使用されなくなった語が含まれており，それが昭和期の旧仮名遣いのテキストの解析精度を下げる一因となっていた．そこで，UniDic の登録語に付与されている時代区分の情報 ([13]) を利用して，旧仮名口語 UniDic の時代区分が「近代」の登録語のうち，1933〜1997 年の仮解析結果では低頻度な語や高頻度であっても誤解析で用いられている語は登録から外した．

以上のようにカスタマイズした新たな UniDic を用いて，整備の完了した 1933〜2013 年のテキストを解析し，公開する形態論情報のベースとした．そこに含まれる (1) 品詞が「未知語」の語，(2) 漢字 1 字を書字形とする語の連続，を主な対象とし，形態論情報の修正を行った．

ベストセラー書籍

「未知語」の処理に加えて，精度向上のための修正作業を 3 点行った．1 点目に，テキストのルビと読みの照合による形態論情報の修正，2 点目に，複数の短単位にわたってカタカナが連続するレコードや動詞が連続するレコードなど，結合候補となるレコードの修正がある．3 点目に，固有名詞（人名・地名）の統一的チェックがある．1 記事，1 著者あたりの言語量の少ない雑誌や新聞に比較して，書籍は 1 作品当たりの言語量が多いため，同じ人物や場所，テーマ語が多数出現しやすい特徴がある．各作品の一部分をチェックし，主要な人物や地名，テーマ語をリストアップし，統一的なチェックを行った．

新聞

サンプルに出現する品詞「未知語」，JIS 外字に相当する Unicode 漢字に起因する誤解析，1 字漢字の連続として実現された誤解析を中心に形態論情報整備を行った．

4.2.2 修正箇所数

公開用に整備したデータのうち，人手による修正が行われた記録を持つ箇所は**表 2** に示すとおりである[*16]．自動解析結果に対する修正では，語の分割や結合を伴う修正によって修正箇所の数も全体の語数も変化するため，厳密に

[*16] ここでの総語数には記号等を含む．

は修正を行った数をそのまま反映するものではないが，作業量の大きさを示すものである．

修正の割合はベストセラー書籍が 0.47804%と最も高くなっているが，これは当該テキストに特に誤りが多かったからではなく，他のジャンルと比較して形態論情報の修正を丁寧に行う時間をとることができたためである．新聞の修正の割合は 0.42213%とベストセラー書籍に次いで高いが，JIS 外字に相当する Unicode 漢字に起因する誤解析が多く，それを集中的に修正したためである．雑誌の修正の割合は 0.10486%と他のレジスターより低いが，総語数が格段に多いためで，修正箇所は他のレジスターより多い．雑誌を 1933〜1997 年と 2005〜2013 年に分けて見ると，修正の割合はそれぞれ 0.10221%・0.19058%で 1933〜1997 年のほうが低い．これは，解析に用いた UniDic が 1933〜1997 年に出現する語に基づいてカスタマイズしたものであり，誤解析の割合が 1933〜1997 年のほうが 2005〜2013 年より低かったことによるものと考えられる．

表 2 修正箇所数
Table 2 Numbers of corrections.

	雑誌	ベストセラー書籍	新聞
修正箇所	32602	18990	12398
総語数	31120433	3972489	2937041
%	0.10486	0.47804	0.42213

5. データと公開形態

5.1 コーパスの語数

上述したような設計・整備を行った結果，公開用の SHC の収録語数は**表 3** に示すとおりとなった．総語数約 3340 万語（記号等を除く）のうち，雑誌が約 82.0%（約 2740 万語）を占め，ベストセラー書籍が約 10.3%（約 345 万語）[*17]，新聞が 7.7%（約 256 万語）となっている．

5.2 形態論情報の精度

公開用のデータについて，形態論情報の精度評価を行った．精度評価は形態素解析以前のテキスト整備の精度評価と，形態論情報の精度評価を UniDic の階層構造に対応した 4 つのレベルに分けて行った．まず，テキスト（書字形）が正しく入力されているかを見る「Lv. 0 テキスト」を設定した．これは直接的には形態論情報の精度を評価するものではないが，その前提となるテキストが十分に実用的な精度で入力できているかを確認するためものである．形態論情報の精度評価としては，まず語の単位境界の認定が正しく行われているかを見る「Lv. 1 単位境界」を設定した．

[*17] 延べ語数での文芸ジャンルと非文芸ジャンルの比率はそれぞれ 52.9%，47.1%であった．時代別には，昭和期の文芸比率が 57.6%，平成期が 45.1%であった．

表 3 収録語数
Table 3 Numbers of words

年	雑誌	ベストセラー書籍	新聞	合計
1933	3,291,739	178,895	128,016	3,598,650
1941	2,460,554	229,811	146,149	2,836,514
1949	1,016,658	272,594	116,868	1,406,120
1957	3,134,703	457,663	100,545	3,692,911
1965	2,025,871	373,622	279,162	2,678,655
1973	2,323,584	343,542	379,015	3,046,141
1981	2,658,012	297,230	355,731	3,310,973
1989	2,744,385	315,787	316,743	3,376,915
1997	2,541,480	306,927	271,339	3,119,746
2005	2,523,450	315,032	236,960	3,075,442
2013	2,679,038	355,067	228,549	3,262,654
合計	27,399,474	3,446,170	2,559,077	33,404,721

その次に，Lv. 1 に加えて UniDic の語形の階層に相当する品詞・活用型・活用形の認定が正しく行われているかを見る「Lv. 2 品詞」を設定した．その次に，Lv. 1〜2 に加えて UniDic の語彙素の階層に相当する語彙素読み・語彙素・語種の認定が正しく行われているかを見る「Lv. 3 語彙素」を設定した．Lv. 3 は例えば「主」を見出し語「主（シュ）」ではなく「主（ヌシ）」，「かわす」を見出し語「躱す」ではなく「交わす」と正しく認定できているかを評価するものである．最後に，Lv. 1〜3 に加えて UniDic の発音形の階層に相当する発音形の認定が正しく行われているかを見る「Lv. 4 発音形」を設定した．Lv. 4 は例えば見出し語「何（ナニ）」を文脈に沿って読みを「ナニ」ではなく「ナン」，見出し語「通り（トオリ）」を「トオリ」ではなく「ドオリ」と正しく認定できているかを評価するものである．

その結果を**表 4** に示す．評価対象はレジスターごとに空白・補助記号を除きランダムに抽出した 4000 語である．自動解析結果と正しく整備した結果とで語数が変化するため，評価値として，適合率（Precision），再現率（Recall），F 値を示した．適合率は調査対象語数そのまま（4000）を分母，正解であった語数を分子とした値，再現率は調査対象語を正しく整備したときの総語数を分母，正解であった語数を分子とした値である．F 値は適合率と再現率の調和平均である．

まず Lv. 0 を見ると，ベストセラー書籍のレジスターに 1 件の誤入力が見られたが，雑誌や新聞のレジスターでは今回評価の対象とした 4000 語のテキストの誤入力は見られなかった．4000 語の書字形の総文字数は，雑誌・ベストセラー書籍・新聞それぞれで 6378 文字・6567 文字・6575 文字である．その文字数の中の誤入力が 1 件にとどまることから，実用に堪えるテキスト入力の精度であることが確認できた．

次に，形態論情報の精度を Lv. 3 の F 値で見ると，雑誌

表 4 形態論情報の精度
Table 4 Accuracy of morphological information

評価レベル	評価項目	雑誌	ベストセラー書籍	新聞
Lv. 0 テキスト	Precision	1.0000	0.9998	1.0000
	Recall	1.0000	0.9999	1.0000
	F 値	1.0000	0.9999	1.0000
Lv. 1 境界	Precision	0.9915	0.9928	0.9871
	Recall	0.9962	0.9979	0.9873
	F 値	0.9939	0.9953	0.9872
Lv. 2 品詞	Precision	0.9778	0.9925	0.9823
	Recall	0.9827	0.9977	0.9825
	F 値	0.9801	0.9951	0.9824
Lv. 3 語彙素	Precision	0.9735	0.9835	0.9798
	Recall	0.9781	0.9889	0.9800
	F 値	0.9758	0.9861	0.9799
Lv. 4 発音形	Precision	0.9705	0.9808	0.9773
	Recall	0.9751	0.9859	0.9775
	F 値	0.9728	0.9833	0.9774

が 0.9758，ベストセラー書籍が 0.9861，新聞が 0.9799 となっている．BCCWJ 全体の精度約 98%（前川 2015[5]）と比較して，SHC の精度は同等あるいは若干下回る程度の値を示しており，実用に問題ない精度であることが確認できた．また 4.2.2 で示したように，形態論情報の修正の割合が十分に高いとは言えない中で実用に問題ない精度のデータが作成できたということは，SHC の形態素解析に用いた UniDic がカスタマイズによって解析精度の高い辞書となっていたことを示している．

レジスター間を比較すると，Lv.1〜4 の F 値すべてでベストセラー書籍の精度が雑誌・新聞を上回っている．これは 4.2.2 で述べたように，ベストセラー書籍では形態素解析結果の修正作業を丁寧に行ったことによるものと考えられる．そして，雑誌・新聞はベストセラー書籍より精度が低いとはいえ，大きな差は見られない．形態論情報の修正作業にあてられる労力と時間が限られる中，新聞では修正対象をピンポイントで選別できたこと，雑誌では UniDic のカスタマイズによって誤解析そのものを減らす方針が有効であったことによるものと考えられる．

5.3 データ公開形式

5.3.1 コーパス検索アプリケーション「中納言」

SHC はコーパス検索アプリケーション「中納言」[*18]を通して無償で一般公開した（要ユーザー登録）．「中納言」はコーパスの文字列やコーパスに付与されている形態論情報をもとに複数の条件を指定して検索できるアプリケーションであり，アカウント登録を行えば無償で利用できる．3.3 のマークアップした情報や 3.4 の形態素解析による形態論情報，3.5 の書誌情報といったデータは，**図 1** のような検索結果の形で参照できる．また，検索結果をエクセルデータでダウンロードすることも可能であり，特に CHJ とは表示される列を共通にしているため，CHJ と SHC でそれぞれにダウンロードした検索結果の Excel データを結合して利用することができる．

[*18] https://chunagon.ninjal.ac.jp

表 5 書字形 n-gram データの例
Table 5 An example of orth-token n-gram data.

レジスター	刊行年	n	n-gram	頻度
雑誌	1965	4gram	て い た .	2551
雑誌	1965	4gram	の で ある .	2491
雑誌	1965	4gram	で あっ た .	2331
雑誌	1965	4gram	し て い た	1155

表 6 語彙素 IDn-gram データの例
Table 6 An example of lemma ID n-gram data.

レジスター	刊行年	n	n-gram	頻度
雑誌	1965	4gram	24874 2585 21642 25	2707
雑誌	1965	4gram	28990 22916 1216 25	2607
雑誌	1965	4gram	22916 1216 21642 25	2344
雑誌	1965	4gram	21642 28990 22916 1216	1477

5.3.2 語彙表

SHC は検索サービスのほかに統計データを公開している．語彙表は SHC に出現する各語の出現頻度を集計してタブ区切りテキストデータである．語の形態論情報およびレジスター・作品名・成立年・本文種別等の情報を列項目とし，その項目すべてが一致する語の頻度を集計し最終列に出力した．Excel 等に取り込んで利用することを想定しており，フィルター機能・ピボットテーブル機能により，特定の範囲の総語数を集計することができる．形式は CHJ の語彙表 [10] と統一しており，両語彙表を結合して利用することもできる．

5.3.3 *n*-gram 頻度データ

単語の n-gram とその頻度のタブ区切りデータである．近藤・相田・小木曽（2022）[18] において新聞レジスターで出力した形式と同様のものを他のレジスターでも出力するものである．レジスター・刊行年ごとに「書字形 n-gram」と「語彙素 IDn-gram」の 2 種類を作成し，それぞれ 1-gram から 5-gram まで集計した．

「書字形 n-gram」は単語の書字形出現形（表層形）の n-gram である（**表 5**）．

それに対し，「語彙素 IDn-gram」は，各単語を UniDic に収録されている語彙素 ID に変換した n-gram である（**表 6**）．

語彙素 ID は UniDic の見出し語に相当する階層・語彙素を一意に示す ID で，UniDic の辞書アーカイブ[*19]中の語彙表（lex.csv）に含まれる語彙素 ID 列と対応付けることで，

[*19] https://clrd.ninjal.ac.jp/unidic/download.html


228 件の検索結果が見つかりました。
検索対象語数: 38,030,044　記号・補助記号・空白を除いた検索対象語数: 33,404,844　検索対象サンプル数: 17,910

サブコーパス名	サンプルID	開始位置	連番	前文脈	キー	後文脈	語彙素読み	語彙素	語形	品詞	活用型	活用形	原文文字列	振り仮名	本文種別	話者	ジャンル	作品名	成立年	巻名等	作者	生年	底本	ページ番号
昭和・平成-新聞	70P読売 1933_92004	3160	2070	てる|ん|です|ぜ。|」＃狼山|は|かう|云ひながら、|テエブル|の|上|ニ　ニ　ニ　ニ|リ|とした|	カレー	|を|大|スプーン|にー杯|すくひあげ|て|瞰ひ|はじめた。＃「|ね|、先生|、|いい|でせう|？|」	カレー	カレー	カレー	名詞-普通名詞一般			カレー				文芸	読売新聞	1933	銀座殘暑記	淺原六朗(作)	1895	読売新聞 <1933-09-02-第20313号>	4
昭和・平成-雑誌	70M中公 1941_02026	20990	14240	と|い|ふ|ん|で、|晝食|は|「牛肉」|『ゆ|であづき』|或は、|「ライス|	カレー	」|と|いつた|獻立|に|致します。＃それ|に|は|私|が|女房|役だ|と|い|ふ|ん	カレー	カレー	カレー	名詞-普通名詞一般			カレー				非文芸	中央公論	1941	俳優對談記 2 長十郎・國太郎の巻	三宅周太郎(作)/河原崎國太郎(作)/河原崎長十郎(作)		中央公論 <1941-02>	本欄230
昭和・平成-ベストセラー書籍	70B浮雲 1949_00028	20260	14080	は|ひまく|滑り|出す|事|が|出來ません。＃みすみす、|いい|客|が|はいつて|も、|ライス|	カレー	ー|つ|出せ|ないン|ですから|ね|。＃ーーー|何しろ、|密告|が|やかましく|て、|あぶなくて	カレー	カレー	カレー	名詞-普通名詞一般			カレー		会話	亭主	文芸	浮雲	1949	二十八	林芙美子(作)	1903	浮雲	195

図 1　SHC「中納言」検索結果画面
Fig. 1　Search results of Chunagon SHC

UniDic の語彙素（代表表記）・語彙素読み・語種等の情報を取り出すことができる．これにより，書字形 n-gram とは異なる，異表記や異語形をまとめあげた n-gram が利用できる．

5.3.4　SVMlight 形式データ

語の共起情報を SVMlight 形式で出力したデータである．近藤・相田・小木曽（2022）[18] において新聞レジスターで出力した形式と同様のものを他のレジスターでも出力するもので，主として自然言語処理技術を活用した言語変化の研究で利用されることを意図している．以下にデータ例をあげる．

```
0 0:58 1:17 10030:1 100370:1 10079:1 10129:1 1018:1 10183:2
1019:1 10196:2 10237:1 10282:3 1031:2 10310:1 103238:1
103421:1 10357:3 10388:1 104550:1 10479:1 10518:5 1054:2
10567:1 1061:8 10634:1 10636:2 10637:1 10652:1 10654:2
10673:1 10701:1 10829:5 10841:4 10847:1 10862:1 10865:2
10884:1 10889:1 10906:1 10909:1 10912:1 10945:3（下略）
```

語は UniDic の語彙素 ID に変換した．頻度 5 回以上の語彙素 ID を集計対象とし，前後 5 つの語彙素 ID の共起情報を示した．上記例では，対象の語彙素 ID「0」に続けて，共起する語彙素 ID とその共起回数を「共起する語彙素 ID: 共起回数」の形式で示している．

6.　応用・活用例

ここでは，SHC の価値を示す活用例として日本語学での利用例について述べる．

6.1　接続詞の通時的変化・レジスター差

現代日本語において，接続詞がレジスターによって使用頻度に違いがあることが指摘されている（石黒 2007[6]，髙野・上村 2017[3] 等）．ただし，これまで近代語と現代語をつなぐコーパスが存在しなかったため，現代語におけるレジスター差がどのような過程を経て形作られたのかといった通時的視点からの実態の把握は困難であった．それが SHC を使用することで可能になる．

例として，SHC の雑誌・新聞の 2 レジスターのジャンルが「非文芸」のサンプルから品詞「接続詞」の語を抽出し，年代別の粗頻度上位 10 語をリストアップしたものを**表 7**，**表 8** として示す．

ここにあがっている接続詞は昭和・平成期の雑誌・新聞で広く使用されたものと言える．これらの中には順位に通時的変化が見られるものがある．例えば，雑誌・新聞ともに順位が次第に上がっていくものとして「いっぽう」「さらに」「ただ」（表中でセルに赤系統の色を付けた）が，順位が次第に下がっているものとして「および」「すなわち」「ないし」（表中でセルに青系統の色を付けた）があげられる．これらの語の変化の程度にはレジスター差が見られる．順位が上昇する「いっぽう」「さらに」「ただ」の語群を見てみる．雑誌では，「ただ」と「さらに」が 1933・1941 年に 10 位に位置し，その後「ただ」＞「さらに」の関係を保ちつつ，2005・2013 年には 4〜6 位まで上昇する．それに対し，「いっぽう」は 1989 年に 10 位に登場し，以降 2013 年まで「さらに」の下位の 8〜9 位に位置する．一方の新聞では，1941 年は「さらに」のみが 10 位に位置していたのが，1949 年になると 6 位に「いっぽう」，8 位に「ただ」が登場し 10 位の「さらに」の上位に位置するようになる．以後，この 3 語は順位を入れ替えながら順位が上昇し，2005・2013 年には 2 位〜5 位の位置を占めるようになる．雑誌と新聞を比較してもっとも差異があるのが「いっぽう」で，雑誌では 1989 年に初めて 10 位に登場し 2013 年も 8 位にとどまるのに対し，新聞では 1957 年に 6 位に登場し，2013 年には 2 位を占めるまでになり，今日の新聞で高頻度な接続詞となったことが分かる．

表 7　雑誌レジスターの接続詞順位の通時的変化
Table 7　Diachronic change in conjunctions ranking of magazine register.

順位	1933	1941	1949	1957	1965	1973	1981	1989	1997	2005	2013
1	しかし	しかし	しかし	しかし	しかし	しかし	しかし	しかし	しかし	しかし	しかし
2	また	また	そして	そして	そして	そして	そして	そして	そして	そして	そして
3	そして	および	また	また	また	また	また	また	また	また	また
4	すなわち	そして	および	あるいは	しかも	しかも	しかも	ただ	しかも	ただ	ただ
5	および	すなわち	あるいは	しかも	あるいは	あるいは	あるいは	しかも	あるいは	しかも	さらに
6	あるいは	しかも	しかも	および	ただ	ただ	ただ	あるいは	ただ	さらに	しかも
7	しかも	あるいは	すなわち	すなわち	さらに	さらに	さらに	さらに	さらに	あるいは	あるいは
8	かつ	ないし	ただ	ただ	および	すなわち	さて	もっとも	いっぽう	ただし	いっぽう
9	が	かつ	ないし	が	すなわち	さて	もっとも	が	で	いっぽう	ただし
10	ただ	さらに	かつ	もっとも	さて	が	が	いっぽう	が	および	および

表 8　新聞レジスターの接続詞順位の通時的変化
Table 8　Diachronic change in conjunctions ranking of newspaper register.

順位	1933	1941	1949	1957	1965	1973	1981	1989	1997	2005	2013
1	および	および	また	また	しかし	また	また	しかし	しかし	また	しかし
2	しかし	また	および	しかし	また	しかし	しかし	また	また	しかし	いっぽう
3	また	しかし	しかし	および	そして	そして	そして	そして	さらに	ただ	ただ
4	すなわち	すなわち	かつ	なお	および	さらに	さらに	いっぽう	いっぽう	いっぽう	また
5	そして	しかも	なお	しかも	しかも	および	いっぽう	さらに	ただ	さらに	さらに
6	あるいは	そして	しかも	いっぽう	いっぽう	いっぽう	ただ	ただ	そして	そして	そして
7	しかも	しかして	あるいは	そして	なお	あるいは	しかも	が	しかも	しかも	あるいは
8	しかして	ないし	さらに	ただ	ただ	なお	および	しかも	が	かつ	ただし
9	ないし	あるいは	そして	あるいは	あるいは	ただ	あるいは	あるいは	および	なお	かつ
10	かつ	さらに	すなわち	さらに	さらに	しかも	なお	および	あるいは	が	および

SHC の利用により，以上のような簡単な用例抽出だけからでも接続詞の通時的変化やレジスター差に関する興味深い事象を把握することができる．ただし，接続詞は「したがっ／て」「だ／が」「ところ／で」等，複数短単位からなるものも多く，より本格的な研究を行うためにはこの研究例のように品詞が「接続詞」の短単位のみに注目するのでは不十分である．その場合，5.3 であげた n-gram 頻度データを利用して接続詞を抽出する手法が有効である．

6.2　日本語表記の通時的変化と国語政策の影響

昭和・平成期には，日本語史上の大きな出来事として「当用漢字表」（1946 年）に始まる一連の国語政策があるが，CHJ と BCCWJ では国語政策による日本語の変化を実証・検証することは困難であった．髙橋（2022）では，CHJ「明治・大正編 I 雑誌」と開発中途段階にあった SHC の雑誌を用いて，副詞の高頻度語 100 語を対象に近現代における日本語の表記の変化の一つである仮名表記化の過程を調査した．また，凡例にて副詞をなるべく仮名書きにすることを記載した「当用漢字表」をはじめとする一連の国語政策の与える仮名表記化への影響について考察した [19]．

図 2 は，髙橋（2022）[19] 掲載の各副詞の各年における仮名表記率 x の分布を示した図を一部簡略化して示したものである．

明治期から大正前期までは漢字表記中心の語が多いが，1917 年頃から昭和中期にかけて仮名表記化が進行し，1965 年以降は安定して仮名表記が用いられるようになる変化がみてとれる．仮名表記率が 80 %以上の語が最も増えたのが「当用漢字表」の施行前後の 1941 年と 1949 年の間であることから，同表の出版物に対する強制力は強かったと考えられる．その一方で，同表の施行以前から仮名表記化の進行が見られることから，大正期から昭和前期にかけて副詞を仮名表記する機運が高まり，「当用漢字表」によって仮名表記化が加速したと見てとれるだろう．一方で，1973 年の「当用漢字改定音訓表」や 1981 年の「常用漢字表」の施行による大きな動きは見られないことから，副詞の表記は 1957 年ごろから安定しつつあったといえる．

7.　おわりに

以上，CHJ を承けて昭和・平成時代までをつなぐ SHC の構築について述べた．このコーパスの公開によってはじめて，コーパスを利用して，真に通時的に，今日の日本語の成り立ちを探ることが可能になったと言える．

SHC でカバーされるレジスターは限られており，収録された年代幅は 8 年と大きく，用例の初出年や流行語を

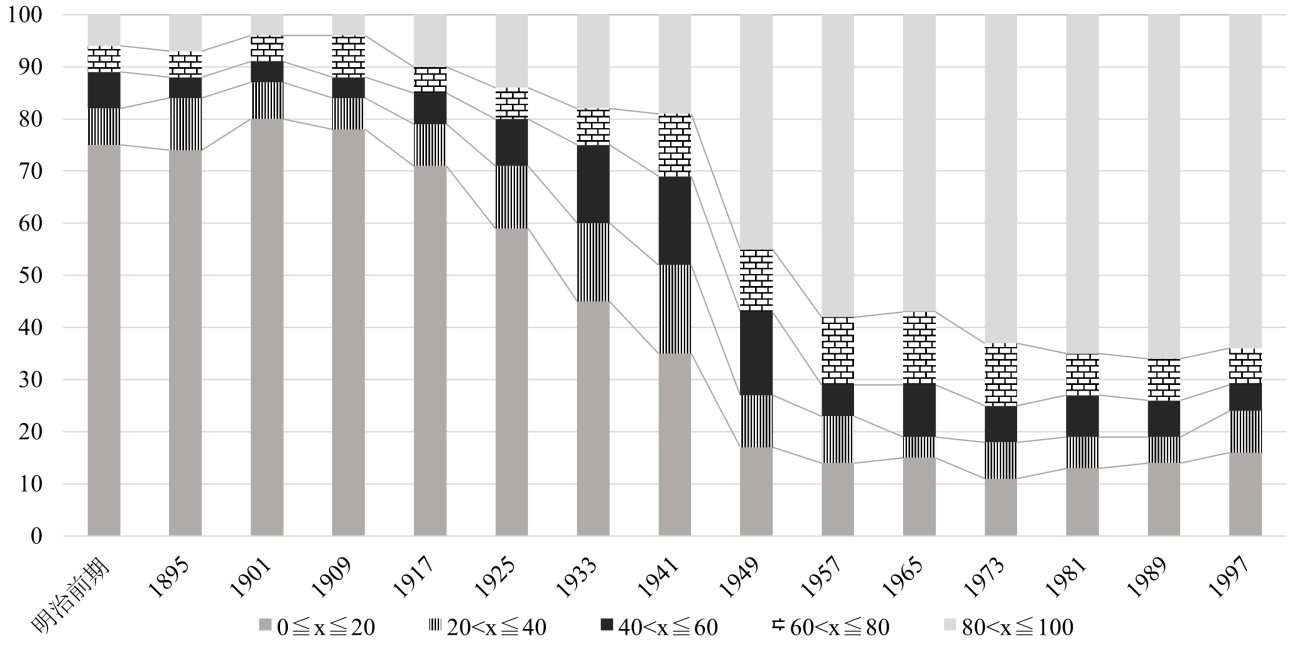


図 2 副詞 100 語の仮名表記率 x の通時的変化
Fig. 2 Diachronic changes in the kana notation rate of adverb

調査するのには不十分な点もある．それでも，本コーパスの構築中に国立国会図書館より公開された NDL Ngram viewer[*20]などと合わせ用いることで，日本語の歴史を精密に記述する環境が整ったと言えるだろう．

また，こうした大規模データと比較したときに，SHC は検索対象となる分母となる語数が明確に分かるという利点が際立ったものとして見えてくる．単に用例を探すだけならば大規模テキストデータを検索すれば良いが，その用例がどの程度の頻度で出現し，全体の中でどれだけの割合を占めていたかを探るためには，SHC のように全語に対して単語情報が付されていることが必要だからである．

SHC のように大きな規模の現代語コーパスを，改正された著作権法に基づいて，著作権者の許諾を得ないで（著作権法上の権利制限規定の下で）公開する試みは，これが初めてのものである．一定の制限の下とは言え，このようにコーパスの公開が行えるようになったことで，言語研究の面はもちろん，自然言語処理など各方面での活用が期待される．SHC の公開を契機に，今後さまざまな言語コーパスが構築・公開され，広く活用されることを願うものである．

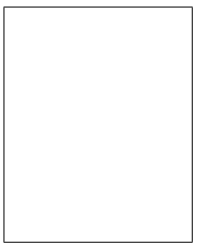
謝辞 本研究は，JSPS 科研費 19H00531「昭和・平成書き言葉コーパスによる近現代日本語の実証的研究」および国立国語研究所共同研究プロジェクト「開かれた共同構築環境による通時コーパスの拡張」による成果の一部である．

[*20] https://lab.ndl.go.jp/service/ngramviewer/

参考文献

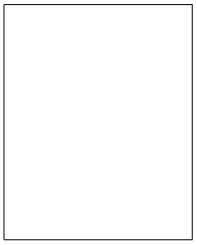
[1] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.
[2] トーハン. 年間ベストセラーアーカイブ, 1997-2013. https://www.tohan.jp/bestsellers/past.html.
[3] 髙野愛子, 上村圭介. レジスター別出現頻度に基づく順接接続詞の文体差の評価：現代日本語書き言葉均衡コーパス (BCCWJ) の用例分析から. 語学教育研究論叢, No. 34, pp. 273–293, 2017.
[4] 前川喜久雄. 代表性を有する大規模日本語書き言葉コーパスの構築（特集 日本語コーパス）. 人工知能, Vol. 24, No. 5, pp. 616–622, 2009.
[5] 前川喜久雄. 『現代日本語書き言葉均衡コーパス』入門. 『現代日本語書き言葉均衡コーパス』利用の手引き 第 1.1 版, pp. 1–18, 2015.
[6] 石黒圭, 阿保きみ枝, 佐川祥予, 中村紗弥子, 劉洋. 接続表現のジャンル別出現頻度について. 一橋大学留学生センター紀要, No. 12, pp. 73–85, 2009.
[7] 伝康晴, 小木曽智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵. コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用. 日本語科学, No. 22, pp. 101–123, 2007.
[8] 国立国語研究所. 太陽コーパス ―雑誌『太陽』日本語データベース―. 博文館新社, 2005. 国立国語研究所資料集 15.
[9] 国立国語研究所国語辞典編集準備室. 用例採集のためのベストセラー目録. 国立国語研究所, 1984. 国語辞典編集準備資料 4.
[10] 国立国語研究所通時コーパスプロジェクト, 小木曽智信. 『日本語歴史コーパス』統合語彙表（バージョン 2022.03）, 2022. http://doi.org/10.15084/00003541.
[11] 出版ニュース社出版年鑑編集部. 出版年鑑. 出版ニュース社, 1981-2013.
[12] 日本新聞協会新聞著作権小委員会. 著作権法第 47 条の 5 と新聞記事の利用について Q&A, 2021.

https://www.pressnet.or.jp/statement/20220215.pdf.
[13] 鴻野知暁, 小木曽智信. 見出し語の時代情報を付与した電子化辞書の構築. 言語処理学会第 20 回年次大会発表論文集, pp. 3–17. 言語処理学会, 2014.
[14] 小木曽智信. 旧仮名遣いの口語文を対象とした形態素解析辞書. じんもんこん 2012 論文集, pp. 25–32. 情報処理学会, 2012.
[15] 小木曽智信, 中村壮範. 『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用. 自然言語処理, Vol. 21, No. 2, pp. 301–332, 2014.
[16] 須永哲矢, 堤智昭, 近藤明日子, 木川あづさ, 服部紀子. 明治中期雑誌の異体漢字と JIS 漢字: 『国民之友』を事例として. じんもんこん 2013 論文集, pp. 201–208. 情報処理学会, 2013.
[17] 文化庁著作権課. 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), October 2019.
[18] 近藤明日子, 相田太一, 小木曽智信. 近現代雑誌通時コーパスの語彙統計情報の公開. 言語処理学会第 28 回年次大会発表論文集, pp. 1695–1698. 言語処理学会, 2022.
[19] 髙橋雄太. 近現代における副詞の仮名表記化. 論究日本近代語, Vol. 2, pp. 221–234, 2022.



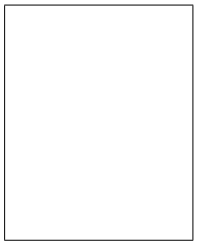
小木曽 智信 (正会員)

1971 年生．1995 年東京大学文学部日本語日本文学 (国語学) 専修課程卒業．1997 年東京大学大学院人文社会系研究科日本文化研究専攻修士課程修了．2001 年同博士課程中途退学．2014 年奈良先端科学技術大学院大学情報科学研究科博士課程修了．博士 (工学)．2001 年明海大学講師，2006 年独立行政法人国立国語研究所研究員を経て，2009 年人間文化研究機構国立国語研究所准教授，2016 年より教授．専門は日本語学，自然言語処理．情報処理学会，言語処理学会，日本語学会各会員．



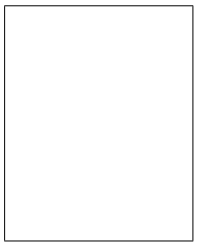
近藤 明日子

1972 年生．1994 年学習院大学文学部卒業．1996 年東京大学大学院人文社会系研究科修士課程修了．2001 年同博士課程単位取得退学．2018 年同博士課程再入学．2019 年同博士課程修了．博士 (文学)．2021 年大学共同利用機関法人人間文化研究機構センター研究員 (特任助教)．2022 年東京大学助教．専門は日本語学．日本語学会，言語処理学会，計量国語学会各会員．



髙橋 雄太

1991 年生．2014 年明治大学国際日本学部国際日本学科卒業．2016 年同大学大学院博士前期課程修了．2020 年同博士後期課程修了．博士 (国際日本学)．2020 年明治大学国際日本学部助教．専門は日本語学．日本語学会，社会言語科学会会員．



間淵 洋子

1972 年生．1997 年東京都立大学大学院修士課程修了．2005 年同博士課程単位取得満期退学．2018 年明治大学大学院博士後期課程修了．博士 (国際日本学)．2018 年大学共同利用機関法人人間文化研究機構センター研究員 (特任助教)．2021 年和洋女子大学准教授．専門は日本語学，計量言語学．計量国語学会理事．日本語学会，社会言語科学会，言語処理学会，全国大学国語国文学会各会員．