

『昭和・平成書き言葉コーパス』  
の公開と研究利用  
—著作権処理をしないコーパスの可能性—

高橋雄太 相田太一 近藤明日子 間淵洋子 小木曾智信

明治大学 東京都立大学 東京大学 和洋女子大学 国立国語研究所

JLR2024 2024年3月15日(金)

# はじめに

- 現代語のコーパスの構築・公開では、著作権処理が構築コストの観点で大きな課題となっていた。
- 平成30年の著作権法の改正に伴い、著作権処理をせずにコーパスを構築・公開することが可能に。
- 本発表では、著作権処理をせずに公開したコーパスのモデルケースとして、『昭和・平成書き言葉コーパス』の設計と公開の方法について述べる。

# 『昭和・平成書き言葉コーパス』(SHC)概要

- 昭和期-平成期の8年おき11カ年分のデータを収録
- 雑誌・ベストセラー書籍・新聞の3レジスタを収録
- 『現代日本語書き言葉均衡コーパス』(BCCWJ)と『日本語歴史コーパス』(CHJ)の間の空白をつなぐコーパス

表1 SHCのレジスタ別の語数

収録年	雑誌	書籍	新聞	計
1933	373万	20万	14万	407万
1941	274万	26万	16万	316万
1949	114万	31万	12万	157万
1957	353万	52万	11万	416万
1965	231万	43万	32万	306万
1973	266万	40万	44万	350万
1981	303万	35万	42万	380万
1989	314万	36万	37万	387万
1997	290万	35万	32万	357万
2005	288万	37万	28万	353万
2013	307万	41万	27万	375万
計	3113万	396万	295万	3804万

# コーパスの構築と著作権処理

- 現代語のコーパスの構築・公開では、収録対象の著作者に許諾を得る著作権処理が課題となっていた。
- 前川(2015)は、BCCWJの構築時の著作権処理について次のように述べている。

BCCWJでは、書籍サンプルだけで約25,000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16,000件について著作権者に連絡をとり、そのうち約10,000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1,000万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。

# 平成30年改正著作権法

- 文化庁著作権課(2019)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」における、日本語研究のためのコーパス利用の事例に対する考え方について、次のように述べている。

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。

# 著作権処理を行わないコーパスの公開

- コーパスの構築は、「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」に該当し、著作者の許諾を得なくとも可能。
- コーパスの公開は、著作物の利用行為が「軽微」であるかどうか問われる。

軽微な利用にとどめるため、SHCは検索アプリケーション「中納言」の原文表示の「文脈長」を短く制限

## 「軽微」な利用

- 著作物の軽微な利用の条件の明確な基準はないが、次のものを参考に、検索キーワードの前後20～30語（約30～50字）であれば条件を満たすと判断。
- 書籍の全文検索サービス「Googleブックス」：  
文脈スニペット表示が約120字
- 日本新聞協会：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
- コーパスの公開方針は、弁護士の確認の上検討した。

# 「中納言」での公開

- コーパス検索アプリケーション「中納言」  
<https://chunagon.ninjal.ac.jp/>
- アカウント登録することで無償で利用可能
- 検索結果をCSV形式でダウンロード可

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

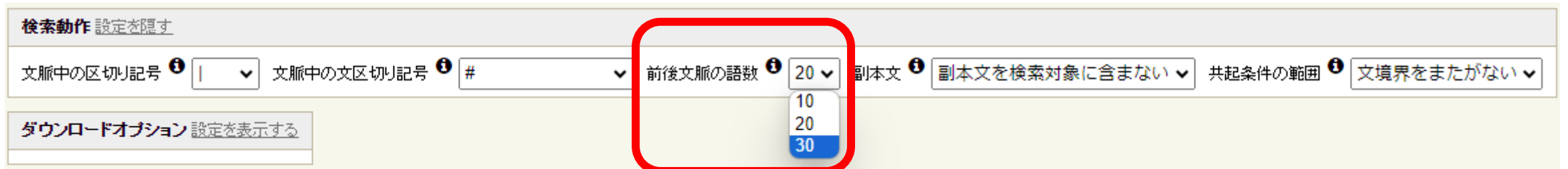
サブコーパス名	サンプルID	開始位置	連番	前文脈	キー	後文脈	語彙素読み	語彙素	語形	品詞	活用型	活用形	原文文字列	振り仮名	本文種別	話者	ジャンル	作品名	成立年	巻名等	作者	生年	底本	ページ番号
昭和・平成-新聞	70P読売 1933_92004	3160	2070	てるんですぜ。い # 猿山川まかう云 乙しながら、(テエブルの上) こころ)とした]	カレー	[を]大[スプーン]に一杯[すく]ひあげ て喰ひまじめた。# 「ね、洗 生、いらいでせう?」]	カレー	カレー	カレー	名詞-普通 名詞一般			カレー				文芸	読売新聞	1933	銀座残暑記	浅原六朗(作)	1895	読売新聞<1933-09-02-第20313号>	4
昭和・平成-雑誌	70M中公 1941_02026	20990	14240	といふふんで、[畫]「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[こ]は私が[女房役]だとい ふふん	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	俳優対談記 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	「もうまく滑り出す事か出来ませ ん。 # みすみす、いらい客がまいつ ても、ライス]	カレー	「一」つ出せぬ[い]んですからね。 # 「一」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 「中納言」SHCでの検索結果画面



# 「中納言」での文脈長の制限

- 前後の文脈長の表示を制限（最大で30語）
  - BCCWJ, CHJなど他のコーパスでは最大で300語まで表示可
- ダウンロードした検索結果のデータでも同様に制限



検索動作 設定を隠す

文脈中の区切り記号 ⓘ | ▼ 文脈中の文区切り記号 ⓘ # ▼ 前後文脈の語数 ⓘ 20 ▼ 副本文 ⓘ 副本文を検索対象に含まない ▼ 共起条件の範囲 ⓘ 文境界をまたがない ▼

ダウンロードオプション 設定を表示する

図2 SHCの検索動作



# 語彙統計情報の公開

- 著作権の関係で生テキストは公開できない代わりに、データを集計した各種語彙統計情報の公開はできる
  - 統合語彙表
  - 語数表
  - n-gram 頻度形式データ
  - SVMlight形式データ

# SHC 統合語彙表

- <https://repository.ninjal.ac.jp/records/2000043>
- 各語彙素の年あたりの出現回数を集計した語彙表
- 雑誌・書籍・新聞の3つに分割して公開
- CHJの統合語彙表などを結合して利用可

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	語彙素読み	語彙素	語種	語彙素ID	品詞	語形	書字形	時代	サブコー	作品名	部	成立年	コアフラク	本文種別	文体	freq
471114	トウケイ	統計	漢	25948	名詞-普通	トウケイ	統計	8昭和	昭和・平成	中央公論		1933	0			133
471115	トウケイ	統計	漢	25948	名詞-普通	トウケイ	統計	8昭和	昭和・平成	中央公論		1941	0			101
471116	トウケイ	統計	漢	25948	名詞-普通	トウケイ	統計	8昭和	昭和・平成	中央公論		1949	0			39
471117	トウケイ	統計	漢	25948	名詞-普通	トウケイ	統計	8昭和	昭和・平成	中央公論		1957	0			170
471118	トウケイ	統計	漢	25948	名詞-普通	トウケイ	統計	8昭和	昭和・平成	文芸春秋		1965	0			75
471119	トウケイ	統計	漢	25948	名詞-普通	トウケイ	統計	8昭和	昭和・平成	文芸春秋		1973	0			48
471120	トウケイ	統計	漢	25948	名詞-普通	トウケイ	統計	8昭和	昭和・平成	文芸春秋		1981	0			111
471121	トウケイ	統計	漢	25948	名詞-普通	トウケイ	統計	9平成	昭和・平成	文芸春秋		1989	0			77
471122	トウケイ	統計	漢	25948	名詞-普通	トウケイ	統計	9平成	昭和・平成	文芸春秋		1997	0			42
471123	トウケイ	統計	漢	25948	名詞-普通	トウケイ	統計	9平成	昭和・平成	文芸春秋		2005	0			36
471124	トウケイ	統計	漢	25948	名詞-普通	トウケイ	統計	9平成	昭和・平成	文芸春秋		2013	0			72

図4 SHC 統合語彙表 (図は雑誌の例)

# SHC 語数表

- <https://clrd.ninjal.ac.jp/shc/stats.html>
- サンプル（雑誌・新聞の記事、書籍の章節項）ごとの語数をまとめた表
- CHJの語数表などを結合して利用可

	A	B	C	D	E	F	G	H	I
1	時代	サブコー	サンプルID	ジャンル	作品名	成立年	巻名等	記号抜き語	語数
2	8昭和	昭和・平成	70B女の1933_20101	文芸	女の一生	1933	私生児 一	615	733
3	8昭和	昭和・平成	70B女の1933_20107	文芸	女の一生	1933	私生児 七	814	940
4	8昭和	昭和・平成	70B女の1933_20103	文芸	女の一生	1933	私生児 三	626	735
5	8昭和	昭和・平成	70B女の1933_20102	文芸	女の一生	1933	私生児 二	609	720
6	8昭和	昭和・平成	70B女の1933_20105	文芸	女の一生	1933	私生児 五	640	728
7	8昭和	昭和・平成	70B女の1933_20106	文芸	女の一生	1933	私生児 六	699	781
8	8昭和	昭和・平成	70B女の1933_20104	文芸	女の一生	1933	私生児 四	656	742
9	8昭和	昭和・平成	70B女の1933_10907	文芸	女の一生	1933	第一の出産	533	660
10	8昭和	昭和・平成	70B女の1933_10909	文芸	女の一生	1933	第一の出産	680	828
11	8昭和	昭和・平成	70B女の1933_10905	文芸	女の一生	1933	第一の出産	644	784

図5 SHC語数表

# SHC n-gram頻度データ

- <https://github.com/alda4/shc-data>
- 単語（出現形）および語彙素IDの1~5-gramとその頻度を出力したデータ

表2 データ例 (magazine\_2013-2013\_5gram.txt)

ていました。	957	24874 2585 35697 21642 25	1003
しています。	549	28990 22916 29321 27442 5569	564
のではないか	547	19537 24874 2585 35697 25	551
していた。	439	24874 2585 21642 28990 22916	482
たのである。	429	21642 28990 22916 1216 25	463

# SHC SVMlight形式データ

- <https://github.com/alda4/shc-data>
- 語彙素IDに関する共起情報をSVMlightの形式で出力したデータ

表3 データ例 (newspaper\_2013-2013\_threshold-20\_svmlight.txt)

```
0 0:440 15:1 16:3 17:1 23:16 24:151 25:70 33:24 34:22 37:1 38:1 44:45 45:1 46:243 49:172
55:5 180:1 383:6 1216:5 1569:1 1571:10 1802:1 2050:25 2267:1 2414:1 2585:6 2868:2
2891:1 2957:1 3089:1 3186:3 3819:1 3965:1 4149:1 4453:2 4579:4 4855:1 4875:1 4890:1
5043:3 5360:1 5577:1 5580:6 5582:1 5606:2 5614:2 5679:1 5738:15 5823:1 5830:1 5985:2
6004:1 6121:1 6217:1 6689:1 6885:3 7219:17 7457:1 7577:1 7581:4 7743:1 7792:1 7836:1
7855:6 7888:1 7889:47 7918:5 7932:3 7998:1 8098:1 8121:3 8220:1
```

- n-gram頻度データ・SVMlightデータの使用用途と使用例については相田ほか(2024)を参照

## おわりに

- SHCのような信頼できるコーパスの公開により、自然言語処理，日本語学などでの研究の発展が期待される。
- SHCの構築・公開は、今後言語研究のために構築される現代語のコーパスの設計・公開方針の指標となる。



# 参考文献

- 相田太一, 近藤明日子, 小木曾智信(2024)「「昭和・平成書き言葉コーパス」の語彙統計情報の公開」言語処理学会第30回年次大会
- 小木曾智信, 近藤明日子, 高橋雄太, 間淵洋子(2024)「『昭和・平成書き言葉コーパス』の設計・構築・公開」『情報処理学会誌』第65巻2号, pp.278-291.
- 日本新聞協会新聞著作権小委員会(2021)「著作権法第47条の5と新聞記事の利用について Q&A 」<https://www.pressnet.or.jp/statement/20220215.pdf>.
- 文化庁著作権課(2019)「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)」[https://www.bunka.go.jp/seisaku/chosakuken/hokaisei/h30\\_hokaisei/pdf/r1406693\\_17.pdf](https://www.bunka.go.jp/seisaku/chosakuken/hokaisei/h30_hokaisei/pdf/r1406693_17.pdf)
- 前川喜久雄(2015)「『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.

# 関連URL

- コーパス検索アプリケーション「中納言」<https://chunagon.ninjal.ac.jp/>
- 「昭和・平成書き言葉コーパス」(SHC) <https://clrd.ninjal.ac.jp/shc/index.html>
- SHC「統合語彙表」<https://repository.ninjal.ac.jp/records/2000043>
- SHC「語数表」<https://clrd.ninjal.ac.jp/shc/index.html>
- SHC・n-gram頻度形式／SVMlight形式データ<https://github.com/alda4/shc-data>

# 付記

- 本発表は、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。