複数コーパスの包括的検索系

八木 豊 ピコラボ
中村 壮範 国立国語研究所
浅原 正幸 * 国立国語研究所
前川 喜久雄 国立国語研究所
小木曽 智信 国立国語研究所
小磯 花絵 国立国語研究所
迫田 久美子 広島大学
木部 暢子 国立国語研究所

1. はじめに

国立国語研究所コーパス開発センターでは、コーパス検索環境「中納言」(1)の整備・維持・管理を行っている。現在のところ、次に示す国語研で構築されたコーパスや国語研に移管されたコーパスの検索が可能である：

- 『現代日本語書き言葉均衡コーパス』(BCCWJ) (Maekawa et al. 2014)
- 『日本語歴史コーパス』(CHJ) (小木曽 2016)
- 『日本語話し言葉コーパス』(CSJ) (Maekawa 2003)
- 『多言語母語の日本語学習者横断コーパス』(I-JAS) (迫田ほか 2016)
- 『名大会話コーパス』(NUCC)(藤村ほか 2011)
- 『現日研・職場談話コーパス』(柏野ほか 2018)
- 『日本語日常会話コーパスモニター版』(CEJC) (小磯ほか 2018)

このうち、CSJ については、有償契約者のみ音声配信機能が利用できる。今後、『日本語諸方言コーパス』(COJADS)(木部ほか 2017) や『国語研日本語ウェブコーパス』(NWJC)(Asahara et al. 2014) の検索環境も公開する予定である。「中納言」には、複数のコーパスが格納されているが、各コーパスが異なる設計で構成されており、これらのコーパスを横断的に検索することが困難であった。

本稿では現在開発中である、複数コーパスの包括的検索系について紹介する。包括的検索系は、複数のコーパスを串刺し検索し、その調整頻度情報を可視化する機能を有する。

2. 包括的検索系の諸機能

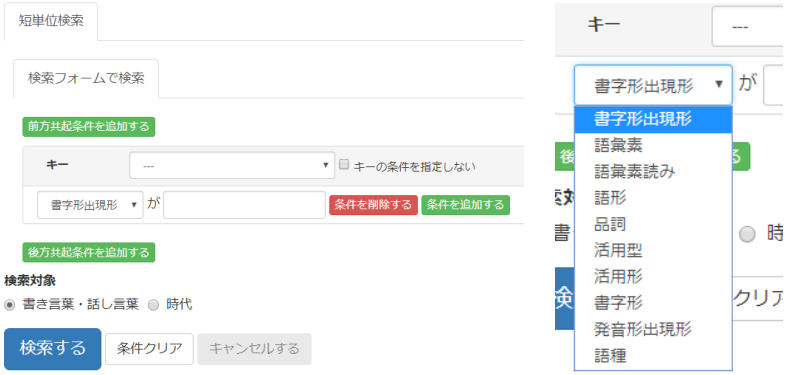


図 1　検索要求画面

検索機能は『中納言』の「短単位検索」機能に準ずる。図 1 に検索要求画面を示す。検索要求は国語

* masayu-a@ninjal.ac.jp
(1) chunagon.ninjal.ac.jp

図 2　相対位置の指定

研短単位に付与された UniDic 品詞体系を含む形態論情報に基づいて構成することができる。1 形態素に対して、「書字形出現形」「語彙素」「語彙素読み」「語形」「品詞（大分類・中分類・小分類）」「活用型（大分類・中分類・小分類）」「活用形（大分類・小分類）」「書字形」「発音形出現形」「語種」に基づいて検索要求を構成することができる。検索キーとなる形態素については、「文頭（節頭・発話頭）」もしくは「文末（節末・発話末）」からの相対位置を指定することができる（図 2）。形態素は前方共起条件と後方共起条件が指定でき、複数の形態素に基づく検索要求を指定することができる。

本検索系では、複数のコーパスを検索することを目的とするが、現在のところ各事例を展開することは行わない。代わりに、調整頻度（相対頻度）分布を可視化する。可視化の軸として「書き言葉・話し言葉」と「時代」の 2 つの軸を設定する。

表 1　「書き言葉・話し言葉」検索時の母集団

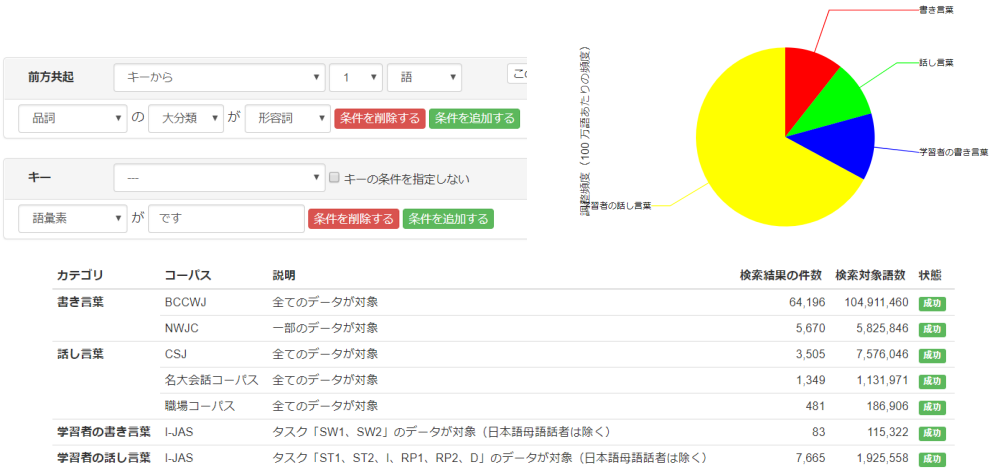
カテゴリ	コーパス	説明	検索対象語数
書き言葉	BCCWJ	全てのデータ	104,911,460
書き言葉	NWJC	一部のデータ	5,825,846
話し言葉	CSJ	全てのデータ	7,576,046
話し言葉	NUCC	全てのデータ	1,131,971
話し言葉	職場	全てのデータ	186,906
学習者の書き言葉	I-JAS	SW1, SW2	115,322
学習者の話し言葉	I-JAS	ST1, ST2, I, RP1, RP2, D	1,925,558

「書き言葉・話し言葉」の機能では、書き言葉・話し言葉の 2 条件だけでなく、日本語母語話者・日本語学習者（非母語話者）の 2 条件も含めた、4 条件で集計を行う。この機能の場合の検索の母集団を表 1 に示す。NWJC は現在 250 億語のうちの 582 万語のみを収録しているが、今後データを拡充する。I-JAS は書き言葉（作文データ）である SW1, SW2 （ストーリーライティング）と、話し言葉（発話データ）である ST1, ST2 （ストーリーテリング）、I （対話）、RP1, RP2 （ロールプレイ）、D （絵描写）の 2 つの群に分割した。なお、学習者のデータからは比較対象として収録している日本語母語話者のデータや実験調査者のデータは除外した。

図 3 に「書き言葉・話し言葉」機能の検索例を示す。例では、語彙素「です」の直前に品詞「形容詞」が出現する例の調整頻度（100 万語あたりの相対頻度）が示されている。表示は帳票形式の結果とともに、棒グラフもしくは円グラフを表示する。図 3 右は円グラフを描画した例である。形容詞＋「です」は、日本語母語話者においては、書き言葉・話し言葉に差がないが、日本語学習者においては、話し言葉で頻出する傾向がみられた。

「時代」の機能では、表 2 に示す母集団に基づき検索が可能である。なお、各時代で収録されているレジスタ差があることに留意して検索する必要がある。

図 4 に「時代」機能の検索例を示す。例では、語彙素「ない」でありかつ、品詞「助動詞」である例の調整頻度が示されている。この例では、時代ごとの調整頻度を棒グラフで示している。助動詞「ない」



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
書き言葉	BCCWJ	全てのデータが対象	64,196	104,911,460	成功
	NWJC	一部のデータが対象	5,670	5,825,846	成功
話し言葉	CSJ	全てのデータが対象	3,505	7,576,046	成功
	名大会話コーパス	全てのデータが対象	1,349	1,131,971	成功
	職場コーパス	全てのデータが対象	481	186,906	成功
学習者の書き言葉	I-JAS	タスク「SW1、SW2」のデータが対象（日本語母語話者は除く）	83	115,322	成功
学習者の話し言葉	I-JAS	タスク「ST1、ST2、I、RP1、RP2、D」のデータが対象（日本語母語話者は除く）	7,665	1,925,558	成功

図 3　「書き言葉・話し言葉」機能の検索例：形容詞＋「です」

表 2　「時代」検索時の母集団

カテゴリ	コーパス	説明	検索対象語数
奈良	CHJ	サブコーパス「奈良」	98,499
平安	CHJ	サブコーパス「平安」	856,827
鎌倉	CHJ	サブコーパス「鎌倉」	822,905
室町	CHJ	サブコーパス「室町」	358,419
江戸	CHJ	サブコーパス「江戸」	204,519
明治・大正	CHJ	サブコーパス「明治・大正」	13,259,330
昭和・平成	BCCWJ	全てのデータ	104,911,460

は、江戸時代以降出現することが確認できる。

3.　おわりに

本稿では現在開発中である複数コーパスの包括的検索系について紹介した。検索系『中納言』のアカウント所持者は、登録申請ののち、`chunagon.ninjal.ac.jp/integrated/` から利用可能である。

謝　辞

検索系『中納言』の維持・管理に従事している国立国語研究所コーパス開発センター諸氏に感謝の意を表します。本研究は国立国語研究所コーパス開発センター共同研究プロジェクトによるものです。

文　献

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). “Balanced Corpus of Contemporary Written Japanese.” *Language Resources and Evaluation*, 48, pp. 345–371.

小木曽智信 (2016).「『日本語歴史コーパス』の現状と展望」　國語と國文學, 93:5, pp. 72–85.

Kikuo Maekawa (2003). “Corpus of Spontaneous Japanese: its design and evaluation.” *Proceedings of The ISCA IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*,

graph.pdf

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
奈良	CHJ	サブコーパス「奈良」のデータが対象	0	98,499	成功
平安	CHJ	サブコーパス「平安」のデータが対象	0	856,827	成功
鎌倉	CHJ	サブコーパス「鎌倉」のデータが対象	0	822,905	成功
室町	CHJ	サブコーパス「室町」のデータが対象	0	358,419	成功
江戸	CHJ	サブコーパス「江戸」のデータが対象	307	204,519	成功
明治・大正	CHJ	サブコーパス「明治・大正」のデータが対象	34,995	13,259,330	成功
昭和・平成	BCCWJ	全てのデータが対象	637,068	104,911,460	成功

図 4　「時代」機能の検索例：助動詞「ない」

pp. 7–12.

迫田久美子・小西円・佐々木藍子・須賀和香子・細井陽子 (2016).「多言語母語の日本語学習者横断コーパス International Corpus of Japanese as a Second Language」　国語研プロジェクトレビュー　6 巻, pp. 93–110.

藤村逸子・大曽美恵子・大島 ディヴィッド義和 (2011).「会話コーパスの構築によるコミュニケーション研究」　藤村逸子・滝沢直宏 (編)『言語研究の技法：データの収集と分析』　ひつじ書房 pp. 43–72.

柏野和佳子・大村舞・西川賢哉・小磯花絵 (2018).「『現日研・職場談話コーパス』中納言版公開データの作成」　言語資源活用ワークショップ 2018 発表論文集, pp. 494–509.

小磯花絵・天谷晴香・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・西川賢哉・伝康晴 (2018).「『日本語日常会話コーパス』モニター公開版の概要」　言語資源活用ワークショップ 2018 発表論文集, pp. 484–493.

木部暢子・佐藤久美子・中西太郎・中澤光平 (2017).「『日本語諸方言コーパス』の構築について」　言語資源活用ワークショップ 2016 発表論文集, pp. 57–68.

Masayuki Asahara, Kikuo Maekawa, Mizuho Imada anda Sachi Kato, and Hikari Konishi (2014). “Archiving and Analysing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan.” *Alexandria*, 26:1–2, pp. 129–148.