『日本語歴史コーパス』活用入門

小木曽智信，服部紀子，松崎安子

国立国語研究所で設計と構築が進められてきた『日本語歴史コーパス』(Corpus of Historical Japanese: 以下 CHJ) は，現在までに下の表に☑で示した多くのサブコーパス（CHJ を構成する時代・ジャンル別の資料群）が利用可能になっている。2020 年 3 月には，下表に■で示したサブコーパスを新たに公開した。本コーパスを検索するウェブ上のアプリケーション「中納言」も，さまざまな機能の拡張が行われてきた。その結果として CHJ の活用の幅が広がった反面，使い方については複雑さが増し，特に初心者にとってとっつきにくいものとなってきたことは否めない。

表　『日本語歴史コーパス』 ver.2020.3 収録資料

奈良時代	☑万葉集　■宣命	
平安時代	☑仮名文学	☑和歌集
鎌倉時代	☑説話・随筆　☑日記・紀行　☐軍記	
室町時代	☑狂言　☑キリシタン資料	
江戸時代	☑洒落本　☑人情本　■近松浄瑠璃	
明治・大正	☑雑誌　☑教科書　☑明治初期口語資料　☐文学作品　☐新聞	

そこで本ワークショップでは，CHJ をこれから研究に活用しようと考える研究者や学生に向けて，CHJ の基本的な使い方や，使うために知っておかなければならないことについて解説する。下記の発表順にしたがって説明を行い，質疑応答を行ったのち，ラウンドセッションを設ける。

〇発表 1：小木曽智信『日本語歴史コーパス』ver.2020.3 の概要

CHJ は，上代から近代までの日本語の用例を通時的に検索することのできる日本語研究の基礎資料として国立国語研究所で構築・公開中のコーパスである。ここでは，このコーパスの特長と設計方針，コーパス全体にわたる利用上の注意点，最新状況と今後の拡張予定等について解説する。

〇発表 2：服部紀子『日本語歴史コーパス』の形態論情報 ―「中納言」検索における注意点―

本発表では CHJ が採用する言語単位（短単位・長単位）と，各語が持つ階層化された形態論情報について解説する。付与された形態論情報は語彙素・語形・書字形といった階層構造を持っているため，表記や語形の揺れに関わらず一括して検索することが可能となっている。コーパスに付与された形態論情報は，通時的な一貫性を重視しつつも，歴史的な変化に対応して各時代特有の処理を行った部分があるため，時代別の検索や通時的検索に際して特に注意すべき箇所がある場合には適宜補足しながら解説を行う。

〇発表 3：松崎安子『日本語歴史コーパス』を「中納言」で検索する方法 ―和歌集編における掛詞の検出―

CHJ を「中納言」で利用する方法についてデモンストレーションを交え説明する。特に「中納言」ver.2.5.0 以降で利用可能になった多重化された形態論情報について解説する。日本語資料には，本行の文字に対して単純な読み仮名以外のふりがなが付されたり，一つの語句に対し共通の音を持つ語句が重ねられたりした複層的・重層的な本文を持つものがある。CHJ ではこうした本文について，本文の同一箇所に対し複数の形態論情報を持たせることでどちらの語としても検索できるようにしている。本発表では，主として八代集に付与された多重化情報としての「掛詞」の検出を行うとともに，その検索結果画面表示の読み取りについて解説する。

〇ラウンドセッション

上記の発表内容に関する質疑応答を行ったのち，参加者が CHJ の利用方法について発表者と議論し，理解を深めるためのラウンドセッションを設ける。具体的なテーマに即して実際の検索・集計方法を紹介するなど，発表者とのやり取りを通して，参加者が自身の研究に CHJ を活用する方法について考えられる場とする。