『日本語歴史コーパス』の今とこれから

小木曽智信，松崎安子，村山実和子，近藤明日子，南雲千香子，高田智和，片山久留美

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』(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）が利用可能になり，広く研究に用いられるようになってきた。2019 年 3 月には，下表に■で示したサブコーパスが新たに公開されるほか，下線を付したサブコーパスのデータ更新を行った。Web 上のコーパス検索アプリケーション「中納言」も，原文文字列の表示や外部サイトの原本画像の表示などにも対応するなどの機能拡張を行ってきたが，今回は新たに掛詞や洒落などの多重の形態論情報の利用を可能にしている。

『日本語歴史コーパス』構築の進捗状況（2019 年 3 月）

<table>
<tr><td>奈良時代</td><td colspan="2">☑万葉集　☐宣命</td></tr>
<tr><td>平安時代</td><td>☑仮名文学</td><td rowspan="2">■和歌集</td></tr>
<tr><td>鎌倉時代</td><td>☑説話・随筆　☑日記・紀行　☐軍記</td></tr>
<tr><td>室町時代</td><td colspan="2">☑狂言　<u>☑キリシタン資料</u></td></tr>
<tr><td>江戸時代</td><td colspan="2"><u>☑洒落本</u>　■人情本　☐近松</td></tr>
<tr><td>明治・大正</td><td colspan="2"><u>☑雑誌</u>　☑教科書　■明治初期口語資料　☐文学作品　☐新聞</td></tr>
</table>

本ワークショップの前半では，下記の通り，CHJ の現状について司会が概要を説明したのち，各サブコーパスの新規公開や更新の詳細についてそれぞれのサブコーパス構築担当者が発表を行う。

司会者発表：『日本語歴史コーパス』の今：小木曽智信

CHJ の現状と構築計画について紹介し，今後コーパスに追加すべき資料についてのアンケートの説明を行う。

発表１：CHJ「和歌集編」（八代集）の構築と公開：松崎安子

新たに公開した CHJ「和歌集編」について報告する。国文学研究資料館が公開している正保版本『二十一代集』のうちの八代集に短単位情報を付与し，原本画像や「新大系」本文へのリンクを可能にした。

発表２：CHJ「江戸時代編Ⅰ洒落本」の拡張と「江戸時代編Ⅱ人情本」の公開：村山実和子

CHJ「江戸時代編Ⅰ洒落本」のアップデートと，新たに公開した「江戸時代編Ⅱ人情本」について報告する。臨時的なふりがなや洒落などに対応するため，新たに多重に形態論情報を付与する試みを行った。

発表３：CHJ「明治・大正編Ⅲ明治初期口語資料」の構築と公開：近藤明日子

明治 0 年代から 10 年代にかけて刊行された口語体資料 10 作品をコーパス化した CHJ「明治・大正編Ⅲ明治初期口語資料」について報告する。『安愚楽鍋』と啓蒙書『交易問答』『百一新論』等の資料を収録している。

発表４：CHJ「明治・大正編Ⅰ雑誌」の拡充 ―『東洋学芸雑誌』のコーパス化と公開―：南雲千香子・近藤明日子

CHJ「明治・大正編Ⅰ雑誌」は概ね 8 年おきに大正末までカバーしていたが，明治 14，15 年の資料が不足していた。これを補うために新たに『東洋学芸雑誌』(1～15 号) を収録し，アップデートしたことを報告する。

発表５：大英図書館所蔵 天草版『平家物語』『伊曽保物語』『金句集』の画像公開とコーパス連携：高田智和・片山久留美

イギリス大英図書館の協力の下，天草版のカラー画像を国立国語研究所ウェブサイトから公開することが実現した。この画像公開サイトと CHJ「室町時代編Ⅱキリシタン資料」から画像へのリンクについて報告する。

ワークショップの後半では，会場で配布・回収したアンケートにもとづいて，前半の発表に対する質疑応答を行ったのち，「**『日本語歴史コーパス』のこれから**」と題して討論を行う。質と量の両面で充実してきた CHJ だが，まだまだ不足する点は多い。たとえば，収録資料が主要文学作品に偏っていることから，ジャンルの幅を広げていく必要がある。特に近世・近代では残されている資料が膨大であるため，追加すべき資料の選定が重要である。また，文節境界の情報付与，地の文・会話文・和歌等を区別する「本文種別」や話者情報の充実，などアノテーションの充実も課題である。長期的な展望を描くためにも，アンケートをもとに「通時コーパス」を今後どのように拡張していくべきか，議論を進めたい。