「昭和・平成書き言葉コーパス」の設計

近藤明日子，小木曽智信，髙橋雄太，田中牧郎，間淵洋子

近年，近現代の日本語の研究にとってコーパスは欠かせない資料になりつつある。国立国語研究所構築の『日本語歴史コーパス』(CHJ) の「明治・大正編」と『現代日本語書き言葉均衡コーパス』(BCCWJ) がその代表的なものである。しかし，CHJ「明治・大正編」が主に明治・大正期の資料を収録対象とし，BCCWJ が主に 2001～2005 年の資料を収録対象としているため，その間の期間や BCCWJ 以降の期間については，コーパスを用いた実証的な研究が困難な状況にある。そこで発表者らは，CHJ と BCCWJ の空隙を埋め，さらに BCCWJ 以降の期間をも補う新たな書き言葉コーパスとして，「昭和・平成書き言葉コーパス」の構築を計画し，2022 年度中に公開を目指して設計と試作を開始した。本コーパスの構築により，CHJ・BCCWJ とあわせて明治期から平成期までの約 150 年間の日本語の変化を追う実証的な研究が可能となる。

コーパスに収録する資料は，新聞（『読売新聞』）・雑誌（『中央公論』『文芸春秋』）・書籍（ベストセラー）の 3 種とし，CHJ・BCCWJ に接続するように，1933～2013 年の期間から刊行年 8 年おきに収録する予定である。そして，CHJ・BCCWJ にならい，コーパスデータは XML 形式で構築し，公開はコーパス検索アプリケーション「中納言」を通じて行う。本文テキストには短単位による形態論情報をはじめとする各種アノテーションを付与し，「中納言」の検索結果に表示する。また，コーパスの総語数等の統計情報もデータとして公開する予定である。なお，本コーパスでは収録資料の著作権処理は行わない。これは，平成 30 年法律第 30 号「著作権法の一部を改正する法律」によって，本コーパスで予定している公開方式であれば，著作権者の許諾なくデータ提供を行うことが可能になったためである。

発表では，本コーパスの構築目的・設計について述べた上で，多くの意見を頂戴し，近現代語研究に有用なコーパス完成に向け議論を深めたい。