明治・大正期の文学作品コーパスの設計とその課題

髙橋雄太，服部紀子，小木曽智信

本発表では，明治・大正期における文学作品コーパスの設計と課題について報告する。本コーパスは国立国語研究所共同研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」の成果，『日本語歴史コーパス　明治・大正編』のサブコーパスの一つとして，検索アプリケーション「中納言」を通して 2020 年度末に公開される予定である。

構築対象となる文学作品は，国立国語研究所に設置された国語辞典編集準備室で 1980 年に作成された「用例採集のための主要文学作品目録」（以下「目録」）を基に選定した。コーパス構築の観点から文学作品を調査すると，①作品毎の規模の差，②作家毎の作品数の差，③74 作品で 500 万語超の大規模，といった問題点が浮き彫りとなった。そこで，（1）「目録」の中で重要度の高いとされる作品を優先，（2）一作家につき一作品，（3）各年代・各時代から均等に採用，（4）「目録」に多くの作品が選定される作家を優先，の 4 方針を立て，21 作品を選定した。その際，10 年刻みの各年代に 5 作品ずつ，明治と大正にそれぞれ 11 作品と 10 作品を採用し，偏りを極力減らすよう設計した。また，5 万語を超えるような大規模な作品につき，日本語研究に全文が必要なのかを検討するため，実験を行った。『吾輩は猫である』と『或る女』を例に，1 万語規模ごとの段階的なデータセットをそれぞれ 8 段階準備し，各段階で共通して出現する語と一方に独自に出現する語の特徴を調査し，また雑誌コーパスとの比較も行った。その結果，機能語（助詞/助動詞/形式名詞/文法的機能を持つ動詞・形容詞）やコソアドの類などの，基礎的な語彙は両作品と雑誌に共通して出現することや，登場人物や舞台，ストーリー性にまつわる語が各作品に独自に出現することを確認した。

なお，発表ではコアデータと非コアデータの設計についても述べる。本コーパスでは全ての作品にコアデータを設定し，作品に独自に登場する固有名詞や表記を解析用の辞書に登録することにより，非コアデータの解析精度の向上を目指すことを述べる。