

「抄物コーパス」の構築をめざして

青木, 博史
九州大学大学院 : 教授

<https://doi.org/10.15017/7325723>

出版情報 : 文獻探究. 62, pp. 57-56, 2024-03-31. 文献探究の会
バージョン :
権利関係 :



「抄物コーパス」の構築をめざして

青木博史

現在、日本語の歴史的研究は、コーパス（コンピュータによる処理を前提とした大規模な電子テキスト）の使用が欠かせないものとなっている。中でも、国立国語研究所によって構築された『日本語歴史コーパス』（以下、CHJ）は、奈良時代から現代までの様々な資料を収めた「通時コーパス」の資格を備えたものとなっており、検索用の Web アプリケーション「中納言」を用いることによって、文字列検索では不可能だった様々な検索を可能にしている。こうした方法論に基づき、近藤泰弘ほか編『コーパスと日本語史研究』（ひつじ書房、2015 年）をはじめ、小木曾智信ほか編『コーパスによる日本語史研究 中古・中世編、近世編、近代編』（ひつじ書房、2022 年、2023 年、2021 年）など、重要な成果が次々と発表されている。

日本語史上の大きな転換期を示す中世後期（室町時代）は、抄物資料・キリシタン資料・狂言資料という、いわゆる「三大口語資料」をバランスよく用いることで深化を遂げてきた。中でも、抄物資料は、①15 世紀から 16 世紀にかけて多く作成されたこと、②言語量が十分に含まれること、③資料的性格が明らかなこと、そして何より、④当時の口語を多分に反映していること、などから、中世後期の言語資料としては最も重要なものと言える。しかし、逆にその膨大さ、あるいは本文の難解さが障壁となり、コーパス化が進められることはなかった。『CHJ』から漏れた抄物資料は、日本語史研究の表舞台から 1 歩下がったところに位置するものとなっている。

この問題の発展的な解決を図るために、JSPS 科研費「「抄物コーパス」の構築とコーパスを応用した日本語史研究」（基盤研究 A、2021～2024 年度）を獲得し、研究分担者 10 名、研究協力者 2 名とともに、「抄物コーパス」の構築に着手することとなった。本事業において、コーパスの構築・公開を行うことを目指す抄物は、以下の 5 点である。

書名	成立年	作成者		分量
史記抄	1477 年	桃源瑞仙	五山	全 20 巻，約 100 万字
中華若木詩抄	16C 中頃	如月寿印	五山	全 4 巻，約 15 万字
毛詩抄	1539 年	清原宣賢講 林宗二聞書	博士家	全 20 巻，約 100 万字
中興禅林風月集抄	1550 年頃	惟高妙安	五山	全 1 巻，約 7 万字
玉塵抄	1563 年～	惟高妙安	五山	全 55 巻，約 350 万字

選出した5点は、抄物資料群全体を見渡し、桃源瑞仙を中心とした「前期抄物」と、惟高妙安を中心とした「後期抄物」に分けられることをふまえ、成立年が偏らないよう配慮している。また、五山僧だけでなく、博士家が作成したものも含め、また、抄（注釈）の原典として、経書・史書・詩文集・韻書など多様なものとなることにも配慮している。これらの抄物資料は、影印や索引の類は公刊されており、断片的な利用は従来からもなされてきたが、コーパス化によって新たな研究の展開が期待される。

コーパスの構築・公開に際しては、①本文制定、②テキスト入力、③校訂、④XML化、⑤形態素解析、⑥データ修正、⑦データ公開、といったプロセスに沿い、Webアプリケーション「中納言」を通して一般に公開することになる。2024年2月現在、『中華若木詩抄』については、「XML化」まで終え、国立国語研究所が開発した検索システム『ひまわり』での利用に供する形で公開している。『ひまわり』は、言語研究用に設計された全文検索システムで、XML文書から特定の文字列を高速に検索する機能と、検索結果をKWIC (KeyWord In Context) 形式で表示する機能などを持ち合わせている。『ひまわり』版での試用を経ながら、適宜データの修正を行っていく。形態素解析には、狂言台本をコーパス化した際の解析用辞書「中世口語（狂言）UniDic」を試行的に用いて専用の解析用辞書を開発し、誤解析によるデータ修正を繰り返しながら精度を担保する。ここから2年以内に『CHJ』に組み込むことを目指して取り組んでいくものである。

『CHJ』に収められる多くの文献資料は、すでに活字化されたものを底本にしているため、テキストデータ化は比較的しやすい。しかし、我々の抄物データは、写本や版本など一次資料を底本とし、他本と校合しながら本文を定めていくため、本文の確定までにかかなりの手間と時間を要する。たとえば『中華若木詩抄』では、最も善本とされる古活字版十八行本を底本とし、写本6種、古活字版2種、整版3種と校合し、校訂本文を作成した。他の抄物についても、同様の入念な検討を経て、独自の校訂本文を作成していく。歴史コーパスの利活用に際しては、当該文献の資料的性格に十分に留意することが大前提であり、文献学的研究の成果を正しく反映させることを心がけていく。

幸い、2024年度から新たにJSPS科研費「抄物コーパス」の構築・公開と通時コーパスを利用した総合的な日本語史研究」（基盤研究A、2024～2028年度）が採択されることが決定した。現在、抄物本文のテキストデータ化は、『中興禅林風月集抄』の全文と、『毛詩抄』の約半分まで終えることができた。大部の『史記抄』、『玉塵抄』に取り組むには困難も予想されるが、その先には、初期に成立した『応永本論語抄』、あるいは洞門抄物の『巨海代抄』なども手掛けていきたいと考えている。「抄物コーパス」の夢は広がっていくが、まずは数年後、またここで進捗状況を報告できたらと思う。

（あおき ひろふみ・九州大学大学院教授）