

《書評》

『コーパスによる日本語史研究：中古・中世編』 青木博史*・岡崎友子**・小木曾智信編***、ひつじ書房、2022 年

小 椋 秀 樹†

I. はじめに

2000 年以降、国立国語研究所（以下、国語研と略す。）が様々なコーパスを構築、公開したことで、日本語研究は大きな転換点を迎えた。国語研が最初に構築、公開したのは、『日本語話し言葉コーパス』（2004 年公開、略称 CSJ）¹である。しかし CSJ は、自発音声を自動認識できる次世代音声認識システムの開発を重要な目的としたことから、収録した音声のほとんどを独話が占めた。また、人文系の研究者が容易に使えるコンコーダンスーも公開されなかったため、日本語研究者が広く利用するには至らなかった。

日本語研究においてコーパス利用が本格化したのは、『現代日本語書き言葉均衡コーパス』（2011 年公開、略称 BCCWJ）²からである。BCCWJ は、2001～2005 年の新聞、雑誌、書籍を中心に多種多様な書き言葉を収録した、延べ語数約 1 億語のコーパスである。BCCWJ の公開形態は複数あり、そのうちの 하나가コーパス検索アプリケーション「中納言」によるオンライン公開である。「中納言」は、人文系の研究者でも容易に、BCCWJ のような大規模コーパスから用例を検索できるように設計されたコンコーダンスーである。この「中納言」版 BCCWJ の公開により、日本語研究におけるコーパス利用が急速に広まった。

その後も、国語研はコーパスの構築、公開を進めていった。その一つが、日本語史研究のための通時コーパス『日本語歴史コーパス』（2012 年から順次公開、略称 CHJ）³である。日本語学（国語学）は、もともと古典語を対象とし、日本語の歴史的変遷を明らかにする学問として発達してきた。そのため、文学作品を中心とする過去の文献を研究資料とし、それら文献資料から効率的に用例を

* 九州大学大学院人文科学研究院教授

** 立命館大学文学部教授

*** 国立国語研究所研究系教授

† 立命館大学文学部教授

h-ogura@fc.ritsumeit.ac.jp

¹ CSJ は、英語名 Corpus of Spontaneous Japanese の略。CSJ の設計及び概要は、国立国語研究所（2006）を参照。

² BCCWJ は、英語名 Balanced Corpus of Contemporary Written Japanese の略。BCCWJ の設計及び概要は、前川監・山崎編（2014）を参照。

³ CHJ は、英語名 Corpus of Historical Japanese の略。CHJ の設計及び概要は、国立国語研究所言語資源研究センターの Web サイトを参照。

収集できるように作られた「総索引」を用いて研究が進められた。しかし「中納言」版 CHJ が公開されたことで、研究の効率化が図られただけでなく、「総索引」といった紙の資料を用いた研究では難しかった研究が可能になった。

CHJ は、国語研の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」（プロジェクトリーダー：小木曾智信、2016～2021 年度）において構築されたコーパスである。このプロジェクトは、国語研内外の多くの研究者が参画して進められ、通時コーパスの構築に加えて、コーパスを活用した新たな日本語史研究の展開も目的として掲げられた。今回取り上げる『コーパスによる日本語史研究 中古・中世編』は、この共同研究プロジェクトの研究成果の一つである。

本書は、〈展望〉〈論文〉〈解説〉の 3 部から成る。〈展望〉では、コーパスを活用した中古語・中世語研究について分野別に概観している。〈論文〉では、CHJ を活用した中古・中世語研究の論文 7 本を収める。〈解説〉では、CHJ を構成するサブコーパスのうち、中古・中世のサブコーパスについて、作業担当者がその仕様と活用方法について説明する。収録された論文、解説は、以下のとおりである。

〈展望〉 岡崎友子「コーパスによる中古・中世語の研究」、近藤泰弘「中古語の資料とコーパス」

〈論文〉 吉田永弘「「思ふ・見る・聞く」の「(ら)る」形—複合動詞と主体敬語—」、大木一夫「テ形補助動詞成立史概略、拾遺」、青木博史「文相当句の名詞化」、北崎勇帆「希望表現の史的変遷—願望を中心に—」、池上尚「中古・中世前期における「ココロヨシ」「ココチヨシ」とその類義表現」、渡辺由貴「短単位 N-gram からみた『天草版平家物語』と『天草版伊曾保物語』の表現の特徴」、片山久留美「コーパスデータから見る大蔵虎明本狂言集の漢字表記語」

〈解説〉 小木曾智信「『日本語歴史コーパス』中古・中世のデータ構築について」、松崎安子「『日本語歴史コーパス 和歌集編』解説」、池上尚「『日本語歴史コーパス 鎌倉時代編Ⅰ説話・随筆』解説」、鴻野知暁・池上尚「『日本語歴史コーパス 鎌倉時代編Ⅱ日記・紀行』解説」、市村太郎・渡辺由貴「『日本語歴史コーパス 室町時代編Ⅰ狂言』解説」、片山久留美・渡辺由貴「『日本語歴史コーパス 室町時代編Ⅱキリシタン資料』解説」

Ⅱ. 本書の内容

1. 〈展望〉

岡崎論文は、CHJ「平安時代編」が公開された 2012 年以降、2020 年までの、コーパスを活用した中古語・中世語研究全体を広く見渡したものである。近藤論文は、テキストデータベースの時代から CHJ までの中古語資料のコーパス化の歴史をまとめるとともに、通時コーパスの今後の方向性について述べる。

岡崎論文を見ると、2014 年に雑誌『日本語学』で通時コーパスに関する特集が組まれていることが分かる。その後も、『国語語彙史の研究』35 集（2016 年）、『日本語文法史研究』4（2018 年）で特集が組まれている。CHJ は、2012 年から 2022 年にかけて構築、公開されたコーパスであるため、これらの特集が組まれた時期には、CHJ は完成しておらず、使用できるデータも限られていた。そのような時期から特集が組まれたということは、早い時期から日本語史研究者の間で CHJ が注目を集めていたということである。なお、使用できるデータが限られていたことから、岡崎論文で紹介されている論文を見ると、時代や資料を区切ったものが多い。CHJ が完成した今、コーパスを活用

した日本語の“通時”的研究が活発になることが期待される。

2. 〈論文〉

収録された論文を分野で分けると、吉田、青木、大木、北崎論文が文法、池上論文が語彙、渡辺論文が文体、片山論文が表記となる。吉田、青木、大木、北崎、池上論文は「中納言」版 CHJ を、渡辺、片山論文は国語研内のデータベースに格納された CHJ を利用している。また、吉田、大木、北崎、池上、渡辺論文は、CHJ から単語連続を抽出し、分析を行っている。

資料から単語連続を抽出することは、総索引を活用していた時代にも不可能ではなかったが、かなり手間と時間が掛かる作業であった。しかし「中納言」を使えば、CHJ から簡単に単語連続を抽出することができる（最長で、キーの前方、後方を合わせて 10 語の連続まで抽出可能）。単語連続を扱う研究は、コーパスの利点をいかした研究であり、そういう意味で、吉田、大木、北崎、池上、渡辺論文は、コーパスを活用した日本語史研究の好例と言える。

また、北崎、池上論文は、「中納言」を使った検索に当たって、どのような検索条件を用いたか明記している。CHJ の公開によって、日本語史研究者が共通のデータを使って研究し、議論できる環境が整った。そして、共通のデータに基づく議論を真に実現するために必要なことが、検索条件を明示するということである。現在、多くの日本語史研究者が CHJ を活用して研究を行い、成果を公表しているが、「中納言」の検索条件を示さないものが依然として多い。紙の資料を使っていた日本語史研究の時代から、完全に脱しきれていない研究者がまだまだ多いということであろう。実際、本書に収録された吉田、青木、大木論文でも検索条件が示されていない。本書が国語研の共同研究プロジェクトの研究成果であることを考えると、検索条件を示すという作法が守られていないのは残念でならない。一方、北崎、池上論文は、コーパスを活用した研究論文のお手本と言える。

さらに、北崎論文では、用例分類に当たって前接語、後接語の形態論情報を利用するため、「中納言」の「ダウンロードオプション」で「インラインタグを使用」を選択し、kwic データをダウンロードしている。それを研究で利用するだけでなく「本論文集のひとつの目的である「コーパスによる日本語史研究」への導入を期して」（113 頁）、なぜインラインタグが必要なのか、それがどのように利用できるのか、ダウンロードした kwic データをどのように加工すればよいのかということまで、節を一つ立てて解説している。コーパスを活用した日本語研究の歴史はまだ浅いということを見ると、論文中でこのような解説を行うことも必要である。

3. 〈解説〉

コーパスをうまく使いこなすためには、コーパスの検索方法や集計方法を学ぶだけでは十分でない。コーパスがどのように作られているか、どのような情報が付与されているのか、その情報はどのような基準で付与されているのかといったコーパスの仕様について十分に理解する必要がある。そういう意味で、本書の〈解説〉は CHJ を使う上で、必ず読んでおかなければならない重要なものである。本書では、CHJ の和歌集編、鎌倉時代編 I・II、室町時代編 I・II の仕様について、それぞれのデータ作成担当者が執筆を担当している。

いずれの担当者もページを割いて解説しているのが、本文の作成についてである。例えば、鎌倉時代編 I 説話・随筆に収録されている『今昔物語集』は、『新編日本古典文学全集』（小学館）を底本としている。既に活字化されたものを底本としているため、コーパスに収録するテキストを作成

する際に、何も問題は起きないように思うかもしれない。しかし、実際はそうではない。『新編日本古典文学全集』の『今昔物語集』は、以下に示すような漢字片仮名交じり文である。

征^そ箭^やシ、コソ射^{いさぶら}候へ。ヒキメハ更^{さら}ニ否^えヤ不^{いつけさぶらはざ}射付候ラム。箭^やノ道^{みち}ニ落^{おち}テ候ハムハ、射^い□シテ候^{さぶら}
ハムヨリモ嗚呼^{をここにさぶらふべ}奇^き候シ

形態素解析システムは、もともと現代語を対象として開発されたものであるため、このような漢字片仮名交じり文には対応していない。さらに上の例では。漢文と同様の返読（「不射付」）、欠損（「射□シテ」）、踊り字（「シ、」）がある。これらにも形態素解析システムは対応していない。

このような本文に対して形態素解析を施すには、形態素解析用の校訂本文を作る必要がある。校訂本文では、漢字片仮名交じり文を漢字平仮名交じり文に変換した上で、返読、踊り字等の処理について方針を立て、形態素解析が可能な本文としている。〈解説〉では、形態素解析用の校訂本文がどのような仕様で作成されているかについて詳しく説明されている。

形態論情報については、BCCWJと同様に短単位、長単位をCHJでも採用している⁴。ただし、基本的な設計方針は現代語のBCCWJと共通するものの、CHJに収録された資料は古典語であり、文法体系や一語化しているか否かも時代によって異なるため、各サブコーパスにおいて必要な拡張が行われている。例えば、BCCWJでは「書き付ける」のような「動詞連用形＋動詞」という複合動詞は、1短単位としているが、鎌倉時代編では1短単位とせず、「書き／付く」のように2短単位に分割している。

CHJを利用するに当たって、校訂本文の仕様を理解していなければ、『今昔物語集』の本文は、漢字平仮名交じりだなどと誤った認識をしてしまうことになる。また形態論情報の仕様について理解していなければ、「中納言」での用例検索に支障が生じるほか、集計結果の解釈を誤ることにともなかねない。〈解説〉は、本書の肝と言ってもよいであろう。

Ⅲ. 終わりに

本書の刊行前後に、『コーパスによる日本語史研究』の近代編、近世編が刊行された。いずれも本書と同様、〈展望〉〈論文〉〈解説〉の3部構成となっている。CHJを使いこなし、質の高い研究を展開するためには、この一連のシリーズをしっかりと読む込むことが必要である。

通時コーパスに関する国語研の新しい共同プロジェクトとして、2022年度から「開かれた共同構築環境による通時コーパスの拡張」（プロジェクトリーダー：小木曾智信）が開始された。CHJには上代から近代までの様々な日本語史資料が収録されているが、これまでの日本語史研究で使われた資料を網羅しているわけではない。今後は、国語研の外部の研究者が日本語史資料をコーパス化し、CHJを拡張していくことが期待される。その際、『コーパスによる日本語史研究』に収録された〈解説〉が大きな役割を果たすことは間違いない。

本書には15本の論文、解説が収められているが、紙幅の関係、また何より評者の力不足により、その一部について評者の感想を述べるにとどまった。また、誤読の箇所もあろうかと思う。これら

⁴ BCCWJの短単位、長単位の設計及び認定規程については、前川監・山崎編（2014）、小椋・小磯・富士池ほか（2011）を参照。

のことについては、編者、著者の寛容を乞う次第である。

参考文献

小椋秀樹・小磯花絵・富士池優美・宮内佐夜香・小西光・原裕（2011）『『現代日本語書き言葉均衡コーパス』形態論情報規程集 第4版』上・下，国立国語研究所内部報告書。

<https://doi.org/10.15084/00002855>, <https://doi.org/10.15084/00002856>（2024年2月26日閲覧）

国立国語研究所（2006）『日本語話し言葉コーパスの構築法』

<https://doi.org/10.15084/00001357>（2024年2月26日閲覧）

前川喜久雄監・山崎誠編（2014）『講座日本語コーパス2 書き言葉コーパス——設計と構築』，朝倉書店。

関連 URL

国立国語研究所言語資源開発センター：<https://clrd.ninjal.ac.jp/index.html>