

国立国語研究所学術情報リポジトリ

はじめに

メタデータ	言語: Japanese 出版者: 研究社 公開日: 2026-07-17 キーワード: 作成者: 石黒, 圭 メールアドレス: 所属: 国立国語研究所
URL	https://doi.org/10.15084/0002000624

This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0
International License.



はじめに

1 本書刊行の背景

本書は、W-CoLeJaという「学習者コーパス」をどのように開発したのか、その開発の過程を紹介する本である。なぜ「学習者コーパス」の開発の過程を紹介する必要があるのか。それは、日本語研究や日本語教育に「学習者コーパス」が欠かせないものになっており、その開発の背景を読者と広く共有することで、第二言語習得研究の深化、教授法や教材開発への応用、さらなる「学習者コーパス」の開発につながると考えるからである。

戦後、とくに1970年代以降の日本語研究は、留学生への日本語教育とともに発展してきたと言える。日本の大学で日本語を学ぶ留学生は、外国語である日本語を苦勞して学び、その過程でたくさんの誤用を生みだすが、その誤用が、日本語という言語のしくみを解明する数々のヒントになってきた。とくに、学習者の誤用の規則性に注目した寺村秀夫という著名な日本語研究者の登場によって、日本語文法の記述が急速に進んだというのが、戦後の日本語研究の歴史だと言える。

一方、世紀の変わり目で研究の進展が落ち着きを見せた日本語そのものの研究に代わり、1990年代ごろから現在に至るまで注目を集めているのが、日本語の習得研究である。つまり、日本語という記号のしくみから、日本語を身に着ける学習者の言語習得のしくみへと研究の関心が移ってきたわけである。とくに、21世紀に入り、コンピュータの性能が急速に向上し、インターネットの接続環境が整った結果、コーパス言語学が言語研究の主流となった。「コーパス」とは、書き言葉・話し言葉といった実際に使われている言語データを大量に収集し、分析・検索・統計処理ができるように整備されたデジタルな言語資料集のことである。「学習者コーパス」とは、とくに学習者が書いたり話したりした言語データを集中的に集めたものを指すが、日本語の学習者コーパスの出現によって、どんな言語を母語とする、どのぐらいの習熟度の学習者が、どのような語彙や文法を使って書いたり話したりするのか、正確かつどころにわかるようになり、研究・教育環境が激変した。第二言語習得の研究に、日本語教師の教授法に、さらに教科書をはじめとする日本語教材の開発に学習者コーパスが役立つことがわかり、21世紀に入ってから多様なコーパスが開発さ

れ、多くの研究者・教育者によって活用されてきた。

しかし、日本語の学習者コーパスの開発は現在も続いている。日本語学習者の母語(何語話者かなど)、年代(子どもか大人かなど)、学習環境(JFL環境かJSL環境かなど)、学習目的(学術かビジネスかなど)、学習媒体(対面かPCかスマホかなど)が多岐にわたるため、それに合った学習者コーパスを整備しようとしても、多様なスロットをすべて埋めるのは気の遠くなるような作業だからである。たとえば、〇〇日本語学校で進学のための日本語を教える先生は、その学校に在籍する学生の卒業後の進路に合った教育を行うための学習者コーパスがほしいだろうし、△△地域の国際交流協会等で会話を教える先生は、そこに来る学習者が地域のコミュニティで日常的に行う会話を集めた学習者コーパスが有用である。さらには、□□大学でアカデミック・ライティングを担当する先生は、その大学の教養課程・専門課程で学術的な文章を教えるのに役立つ論文・レポートに特化した学習者コーパスが必要だろう。

このように、学習者に真に役立つ優れた研究・教育を行うことを考えた場合、それぞれの学習者のニーズに合った学習者コーパスを分析する力、さらには開発する力が不可欠になっている。本書は、こうした学習者コーパスの分析力・開発力をつけるために編まれたものである。

2 本書刊行の目的

国立国語研究所(以下、国語研)は、日本語の多様なコーパスを開発しつづけてきたという点で、国内外を見渡しても最大の研究機関であり、すでに多数の日本語のコーパスを公開してきた多年の実績がある。国語研はそうした過去の経緯を踏まえ、第4期(2022~2027年度)中期目標・中期計画において機関拠点型基幹研究「開かれた言語資源による日本語の実証的・応用的研究」というプロジェクトを立ち上げた。そこで謳われているのは言語資源学という新たな学術分野の創設であり、『シリーズ言語資源学』は、当該プロジェクト期間中に開発された各種の言語資源の概要を紹介することで、国内外の研究機関に所属する研究者による共同利用を推進し、言語資源学創設のための共同研究基盤として活用されることを目指すものである。

『シリーズ言語資源学』で最初に刊行される本書『学習者コーパスの設計と構築』は、国語研が第4期中期目標・中期計画において開発している学習者コーパスのうち、『日本語学習者の縦断作文コーパス』(以下、W-CoLeJaと呼ぶ)を取り上げ、その開発の背景、設計の概要、活用の方法等を紹介する。それにより、

日本語教育研究に関わる研究者の共同利用を推進し、世界で学ぶ日本語学習者の教育に役立つ共同研究基盤として活用されることを目指している。

W-CoLeJaは、「開かれた言語資源による日本語の実証的・応用的研究」のなかの機関型と呼ばれる七つのプロジェクトのうち、日本語教育関係の所員が中心となって運営している「多様な言語資源に基づく日本語非母語話者の言語運用の応用的研究」(プロジェクトリーダー：石黒圭)に位置づけられるものである。当該プロジェクトは五つのサブプロジェクトからなるが、そのなかの一つ「日本語学習者の作文の縦断コーパス研究」プロジェクト(略称「学習者作文」プロジェクト)および科学研究費補助金基盤研究(A)「海外縦断作文コーパスの構築に基づく文章産出能力の発達過程の実証的研究」(研究代表者：石黒圭)の予算を用いて収集しているコーパスである。具体的には、世界の20大学で日本語を専攻する954名(開始当初、現時点での暫定数)の日本語学習者を対象に、学部入学から卒業までの4年間(3年制の一部の大学では3年間)、毎年三つのテーマで書いてもらった作文を収集したものである。若い学部生は、進級、留学、就職・進学準備など、大学生活のなかで多様な変化を経験するものであり、学部4年(3年)間、調査を続けることは困難を極め、半数近くの学生が途中で調査を辞退しており、最終的には500名強の学生が完走の見込みである。しかし、裏を返せば、そうした多感な時期の環境の変化にもかかわらず、過半数の学生が調査を粘り強く継続してくれたわけであり、感謝に堪えない。

W-CoLeJaは2026年3月に全体の9割強の第一次データ公開(作文本文のみの公開)を予定しており、その翌年の2027年3月に同じデータに形態素解析を施した「中納言」によるデータ公開を予定している。データ全体の最終公開は、第4期中期目標・中期計画の終了時期に当たる2028年3月になる予定である。本書は第一次データ公開に合わせ、W-CoLeJaの設計と構築の概要を紹介し、共同研究基盤としての活用を目指すものである。

3 本書の対象読者

本書の対象となる読者として、本書では四つのカテゴリに入る読者を想定している。第一はW-CoLeJaを用いて研究を行いたいと考える研究者であり、本書はW-CoLeJaの研究マニュアルとしての意味を持つ。W-CoLeJaがどのような目的で設計され、どのぐらいの規模のどのような種類のデータを備え、どのような活用が想定されているのかを知ること、W-CoLeJaのより効果的な利活用が期待できる。しかし、本書の対象となる読者はそれにとどまらない。

本書の対象読者の第二のカテゴリは、W-CoLeJaという個別のコーパスに限らず、学習者コーパス全般の理解を深めたいと考える学習者コーパス初心者である。コーパスを対象に研究したいと考える場合、コーパス利用者の立場を離れ、コーパス開発者の立場に立ってみることが大切である。コーパス開発者の目から見て、自分たちの開発したコーパスがもともとの目的や趣旨に反し、無理な使われ方をしていると感じることは少なくない。コーパス開発者が考えた設計の意図、データ収集の規模や範囲、収録されているデータの精度などがわかって初めて十全なコーパス分析ができる。その意味で、学習者コーパス開発の設計や構築の意図や裏話が書かれた本書は、コーパス開発者の身になって考えるとき、参考になる内容が多く盛り込まれているように思われる。

本書の対象読者の第三のカテゴリは、言語教育者である。手前味噌で恐縮であるが、かつて教え子の筒井千絵氏と共著で出版した『留学生のためのここが大切 文章表現のルール』(スリーエーネットワーク)という作文教材がある。2009年に出版されているが、ほぼ毎年増刷され、2025年現在で14刷、40,000部に達している。総合教材でも定番教材でもないこの作文教材が、なぜそこまで市場に受け入れられたのか、正直わからないところもあるが、おそらく筆者が一橋大学留学生センター(当時)に在職中、留学生の作文添削を徹底的に行っていて、そこで収集した誤用例に基づいて作成しているからであろう。いわば自作の作文コーパスを作成していて、そこから多くを学んだのである。言語学習により教材を開発するには、学習者自身に聞いてみるのが一番である。彼らが何に困っているのかに耳を傾け、その困難点の改善につながる教材を開発すれば、現場に受け入れられるものになるはずである。もちろん、学習者のナマの声に耳を傾けることも必要であるが、多くの学習者のナマの声に耳を傾けるには時間的な限界がある。その点で、学習者の困難点を検索を駆使してあぶりだせる学習者コーパスに耳を傾ければ、教室での文法説明も、例文提示も、格段に改善される可能性がある。

そして、本書の対象読者の第四のカテゴリは、コーパス開発者である。コーパスは研究目的に合わせて設計・構築されるものであり、自分の知りたいことがわかる既存のコーパスがなければ、自分で設計・構築する必要がある。コーパスと呼べる規模でなくても、自作のデータベースを作って研究する大学院生は少なくない。そのさい、どのように作成するかは見よう見まねであり、多くの調査協力者の手を煩わせてデータベースを作っても、不十分な設計の結果、せっかくのデータベースが使い物にならないこともある。しかし、こうした悲

惨な結果を招かないようにコーパス開発の手引きを参考にしたいくても、公開されることがまれで、参考となる情報を入手することは困難な現状がある。その意味で、本書はW-CoLeJaの研究マニュアルであるとともに、コーパス開発マニュアルでもある。

第二言語コーパスの構築はとくに英語教育を中心に進んでいるが、日本語・英語にかぎらず、参考になる開発の手引きは少ない。本コーパスは、日本語学習者コーパスでありながら、中国語、韓国語、ベトナム語、タイ語、フランス語、英語などを含んでおり、本コーパス構築のノウハウは第二言語コーパス構築全般に役に立つことが期待される。さらに、第一言語である国語教育の作文コーパスの出現も望まれているところであり、本書刊行が各言語のコーパス研究者の協力する一つのきっかけになれば、これほど嬉しいことはない。

4 本書の構成

本書は、W-CoLeJaを使って研究したい方、W-CoLeJaにかぎらず学習者コーパス全般を学びたい方、学習者コーパスを教育現場に応用したい方、自作の学習者コーパスを新たに開発したい方を対象読者とする旨を述べたが、本書の構成は、そうした対象読者を意識したものとなっている。

第1章「学習者コーパスの意義と役割」では、12の母語、世界20地域の1,000名の日本語学習者を対象にした巨大な横断コーパスである『多言語母語の日本語学習者横断コーパス (I-JAS)』の開発者である迫田久美子氏 (広島大学) から、学習者コーパスの意義と役割について語っていただく。

それに続く第2章「日本語学習者コーパス開発の歴史」では、石黒圭・呉楚琦氏 (国立国語研究所) が、日本語の学習者コーパスとしてこれまでどんなコーパスが開発されてきたか、学習者の技能、学習者の母語、学習者の習熟度という三つの観点から学習者コーパス開発の歴史を振り返り、その発展の過程と現状の課題を確認する。冒頭のこの二つの章を読めば、日本語教育における学習者コーパスの基本情報が頭のなかで整理され、学習者コーパスの利活用、さらには自作コーパス開発の必要性が理解できるはずである。

第3章「学習者コーパスの設計」は、第2章「日本語学習者コーパス開発の歴史」を踏まえ、W-CoLeJaはどのような理念に基づいて設計され、コーパス構築が行われてきたのか、石黒圭 (国立国語研究所) が、W-CoLeJa開発の概要を紹介する。この第3章は、W-CoLeJa開発の詳細を記した第4章から第15章までの内容を整理し予告する章なので、第3章を熟読したうえで第4章から第15章

までをお読みいただけると、理解が一層深まるはずである。

第4章から第15章については、内容を簡単に示すと、次のとおりである。この計12章の内容を読めば、学習者コーパスの開発に必要な知識や技術、さらにその背後にある考え方や心構えが一通り押さえられるようになっている。

- ・第4章：調査実施のサポート体制について
- ・第5章：内製の作文入力システムの仕様について
- ・第6章：学習者作文の母語対訳データの活用法について
- ・第7章：協力者の日本語能力測定テストの作成と妥当性について
- ・第8章：協力者の背景情報を収集する質問紙の作成法について
- ・第9章：協力者の学習環境と学習リソースの把握について
- ・第10章：学習者作文の形態素解析の課題と対処法について
- ・第11/12章：学習者作文の誤用分析法（人手/生成AI利用）について
- ・第13章：学習者作文を読んだ読み手の印象評価情報の付与について
- ・第14章：全文検索システム『ひまわり』によるデータ処理について
- ・第15章：Web上でのデータの公開法と公開後のメンテナンスについて

上述のような本書の構成をとoshi、学習者コーパスを用いて分析する勘どころ、さらには学習者コーパスを新たに構築する勘どころについて精通し、ご自身の研究に役立てていただくことを、執筆者一同、心から願っている。

編者代表
石黒 圭