

書評

前川喜久雄 (編) (2013)

『講座日本語コーパス 第 1 巻 コーパス入門』

鯨岡 さつき (学習院大学大学院生)

本書は、「講座日本語コーパス」シリーズの第 1 巻として 2013 年に出版された。「講座日本語コーパス」は、国立国語研究所教授の前川喜久雄が監修を手掛けた、全 8 巻からなるシリーズである。第 1 巻は『コーパス入門』(前川喜久雄編)、第 2 巻は『書き言葉コーパス—設計と構築』(山崎誠編)、第 3 巻は『話し言葉コーパス—設計と構築』(小磯花絵編)、第 4 巻は『コーパスと国語教育』(田中牧郎編)、第 5 巻は『コーパスと日本語教育』(砂川有里子編)、第 6 巻は『コーパスと日本語学』(田野村忠温編)、第 7 巻は『コーパスと辞書』(伝康晴・荻野綱男編)、そして第 8 巻は『コーパスと自然言語処理』(松本裕治・奥村学編)ということになっている。このシリーズの出版元である朝倉書店のホームページ¹によると、2016 年 1 月現在、第 1, 2, 3, 4, 6 巻が発行されており、2016 年 2 月には第 5 巻が発行される予定となっている。第 1 巻の巻頭にある「本講座の刊行にあたって」によると、本書は文部科学省科学研究費特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築: 21 世紀の日本語研究の基盤整備 (略称: 日本語コーパス)」の「成果を一般に広く還元することを目標として企画立案されたものであり、各巻の編者には特定領域研究計画研究班の班長があたっている」²とのことである。また本書の監修者である国立国語研究所教授の前川喜久雄は、この特定領域研究の代表者である³。

さて、今回紹介する第 1 巻は、続く他の 7 つの巻の導入として位置づけられた本で、まさにタイトル通り「コーパス入門」を使命としている。本書の構成は、全 6 章と 2 つの付録からなる。第 1 章がコーパスの存在意義 (前川喜久雄)、第 2 章がコーパスと計算言語学 (辻井潤一)、第 3 章が言語教育とコーパス (投野由紀夫)、第 4 章が言語処理とコーパス (徳永健伸)、第 5 章が日本語コーパスの発展 (丸山岳彦)、第 6 章が語彙調査の系譜とコーパス (山崎誠) である。そして 2 つの付録では、短単位検索サイト『中納言』の使い方 (小木曾智信・中村壮範)、そして全文検索システム『ひまわり』の使い方 (山口昌也) が実際の操作画面の画像付きで紹介されている。

まず第 1 章は、第 1 巻および「講座日本語コーパス」全体の導入である。コーパス研究の基本的な概念について説明した上で、コーパスの利用が言語研究にとって有効な方法の一つであることを確認している。第 2 章は Montague と Chomsky の 2 つの流れからなる合理主義の言語学の変遷を追った上で、科学として言語学が発展する際にコーパスと計算機が果たす役割の大きさについて説明している。第 3 章はコーパスから得た情報を言語教

1 http://www.asakura.co.jp/G_11.php?sreiesname=307

2 前川 (2013: i) 参照。

3 文部科学省の平成 23 年度科学研究費補助金・事後評価のページを参照した。
(http://www.mext.go.jp/a_menu/shinkou/hojyo/chukan-jigohyouka/1316703.htm)

育に活かそうとする取り組みについて、コーパスを間接利用することで作られた辞書、文法書、会話教材や、コーパスを直接利用した学習プログラムなどの英語教育の分野からの例を交えて紹介している。第4章では、「我々が日常で使っている自然言語をコンピュータによって理解することを目的とする」⁴言語処理の研究においてコーパスがどのように関わってきたかを論じている。第5章は日本語コーパスの歴史を概観するもので、「黎明期（1950年代～）」「萌芽期（1980年代後半～）」「進展期（1990年代後半～）」「林立期（2000年代前半～）」「普及期（2000年代後半～）」の5つに時代区分した上で、日本語コーパスの発展を詳細に説明している。そして第6章では、今度は国立国語研究所による語彙調査について、「準備期」「発展期」「展開期」「停滞期」「探索期」「回帰期」の6つに時代区分して解説を行い、この語彙調査における流れが与えたコーパス日本語学に対する影響を概観することを試みている。

以上の本文に加え、さらに2つの付録では、本書の様々な箇所で見ることになる2つのコーパス検索ツール、短単位検索サイト『中納言』（付録A）と全文検索システム『ひまわり』（付録B）の紹介がされている。『中納言』の紹介では、まず『現代日本語書き言葉均衡コーパス』（BCCWJ）のデータ（約1億短単位、約8500万長単位）を検索できる『中納言』の概要から始まる。BCCWJは全文検索サイト『中納言』でも検索することができるが、『中納言』では「短単位検索」「長単位検索」「文字列検索」の3つの方法でBCCWJのデータを検索することができる。また3つの検索方法のどちらにおいても、検索結果の表示方法など細かい設定ができる。さらには検索履歴や検索条件式などについても丁寧な説明がされている。いずれの項目においても実際の操作画面の画像が示されているので、理解がしやすい。続く付録では、全文検索システム『ひまわり』の説明がされている。『ひまわり』は、「コーパスをはじめとした様々な言語資料から、指定した文字列を検索し、その結果を見やすい形で表示する」⁵ことのできるシステムである。『ひまわり』が検索対象としているのはXMLで記述された言語資料で、それらを全文検索し、その検索された文字列に伴っている前後の分脈や、文字列をマークアップしているタグの属性を取得することができる。その他にも『ひまわり』用コーパスの作成支援や、コーパスの配布、追加的なアノテーションといった、『ひまわり』のコーパス研究を支援する機能を紹介している。またこの付録でも、『中納言』の紹介と同じように、『ひまわり』の操作画面の画像やXML文書の例が示されているので、事柄の把握のしやすいものとなっている。

前述の「本講座の刊行にあたって」には、このシリーズの想定する読者として「日本語学の研究を志す学部・大学院の学生、これからコーパスを利用した研究に挑戦しようとしている日本語研究者、そして関連領域としてのコーパス日本語学に興味を持つ情報系研究者」が挙げられている。ただし、書評者はコーパスを利用した研究に興味を持つ大学院生であるものの、実は日本語研究者ではなくドイツ語研究者である。そのため書評者は当初、日本語研究者でない者にとって本書およびこのシリーズがはたしてどのような意味があるものかについて、不確かに思っていた。しかし、実際に読み始めてみると、本書にはドイ

4 前川（2013：79）参照。

5 前川（2013：170）参照。

ツ語学研究者にも多岐にわたり役に立つ情報が詰まっていた。例えば、代表性、均衡性、規模などのコーパス構築上の問題は、どの言語でも考えなければならない問題である。ドイツ語の大規模コーパスではドイツ語研究所 (IDS, Institut für deutsche Sprache) の DeReKo (das deutsche Referenzkorpus) が有名で、収録語数は 2000 年には約 6 億語だったのが、2015 年には約 270 億語にまで拡大している。⁶しかし小説などいくつかのジャンルでは、他のジャンルと比べてまだ用例数が不十分であるという IDS 関係者による報告もある。⁷これは一見規模の面では十分にも思えるコーパスでも、均衡性の面から見ると欠点があることの一例であろう。また、国立国語研究所の語彙調査に関する技術・ノウハウの発展過程とといった、日本語コーパスの歴史をとってみても、同語異語判別 (レンマ化) は他言語コーパス構築の際にも考慮に入れなければならない。例えばドイツ語の場合、動詞だけでなく、名詞・形容詞なども性・数・格に応じて語形変化が生じる。この点について西欧語コーパス処理の研究支援ソフトである Tecely (岐阜経済大学山田善久研究室作成)⁸では、様々な変化形を記載した「ドイツ語レマ辞書」⁹を採用することで、Tecely 内でレンマ処理をすることを可能にしている。

このように、シリーズ「講座日本語コーパス」の第 1 巻である本書『コーパス入門』は、日本語研究者に限らず、コーパスに興味を持つ他言語研究者にとっても、日本語のコーパス研究を知るためだけでなく、自身のコーパス研究の手引き書としても最適な書である。第 2 巻以降においても、様々な分野の人に、コーパスに関する様々なきっかけや情報を与えてくれるものになっているだろう。

文献

Kupietz, Marc/ Lungen, Harald (2014): Recent Developments in DeReKo. In: Calzolari, Nicoletta et al. (eds.): Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik: ELRA, 2378-2385.

< http://www.lrec-conf.org/proceedings/lrec2014/pdf/842_Paper.pdf >

(2016.1.27. 確認)

恒川元行・大藪正彦 (編) (2014) 『コーパス利用に基づくドイツ語研究—幅広いデータ収集と頻度から見直す—』日本独文学会研究叢書 98 号。

(2016 年 1 月 16 日受付・2016 年 1 月 27 日再受付)

6 <http://www1.ids-mannheim.de/kl/projekte/korpora/archiv.html#Umfang>

7 Kupietz/Lungen (2014: 2383) 参照.

8 Tecely および「ドイツ語レマ辞書」については、<http://www.gifu-keizai.ac.jp/~yamada/> を参照した.

9 阿部一哉 (跡見学園女子大学) 作成, 黒田廉 (富山大学) 増補.