〈著書紹介〉 前川喜久雄 監修/山崎誠 編 山崎誠,前川喜久雄,丸山岳彦,柏野和佳子,山口昌也,小椋秀樹,小木曽智信,田中牧郎 著『書き言葉コーパス-設計と構築-』

著者	山崎 誠
雑誌名	国語研プロジェクトレビュー
巻	6
号	1
ページ	27-28
発行年	2015-06
URL	http://doi.org/10.15084/00000801

〈著書紹介〉

前川喜久雄 監修／山崎誠 編
山崎誠，前川喜久雄，丸山岳彦，柏野和佳子，山口昌也，小椋秀樹，小木曽智信，田中牧郎 著
『書き言葉コーパス―設計と構築―』
講座　日本語コーパス 2
2014 年 12 月　朝倉書店　A5 判　149 ページ　3,000 円＋税

山崎 誠

1. 本書について

本書は『講座 日本語コーパス』の中の 1 巻である。この講座は，2006～2010 年度にかけて国立国語研究所を中心に行われた特定領域研究「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」（研究代表者：前川喜久雄）の成果を全 8 巻にまとめるものである。既に第 1 巻『コーパス入門』が 2013 年 7 月に刊行され，本プロジェクトレビューの 4 巻 2 号にもその紹介がある。本書は講座の第 2 冊目にあたり，「現代日本語書き言葉均衡コーパス」（BCCWJ）の構築過程を中心として記述されている。執筆に当たったのは実際に構築に携わったメンバーである。

2. 本書の構成

本書は，以下に示す 6 章と付録とから構成されている。

第 1 章　コーパスの設計（山崎誠・前川喜久雄）
第 2 章　サンプリング（丸山岳彦・柏野和佳子）
第 3 章　文書構造の電子化（山口昌也）
第 4 章　形態論情報（小椋秀樹）
第 5 章　形態素解析（小木曽智信）
第 6 章　歴史コーパス（田中牧郎）
付　録　形態素解析ツール（小木曽智信）

3. 本書の内容

本書は 2011 年に公開を開始した BCCWJ 及び近代語を中心とした歴史コーパスについてそれらの構築に当たってどのような方法や技術が必要かを具体的に記述したものである。以下，章を追って簡単に各章のポイントを記す。

第 1 章は，一つのプロジェクトとして書き言葉コーパスを構築する際の基本概念及び構築作業上で注意する点が紹介されている。基本概念として重要なのは，「代表性」と「均衡」である。この概念はかつて国語研究所が実施してきた語彙調査においても重要視されてきた考え方であり，後述の言語単位の設計と合わせて，国語研究所で培われた伝統的な手法がコーパスにも継承されていると言える。

第 2 章は，代表性を確保するための手段であるサンプリングの過程を説明する。BCCWJ を構成する 13 のレジスターについてそれぞれどのように母集団を決めてサンプルの抽出を行ったかが記述されている。併せてメタデータとしての書誌情報データベースの必要性についても触れている。

第 3 章は，文字入力の仕様と従来のコーパスでは重視されてこなかった文書構造の電子化について説明する。文書構造の電子化は，既に『太陽コーパス』で実現された技術だが，BCCWJ ではさらにそれを拡張して用いている。

第 4 章は，日本語の計量調査に際して避けて通れない「語」の認定に関する問題を扱う。日本語は通常分かち書きをしないので，語の境界が決めにくいことがある。客観的な調査のためには人為的な言語単位を設計し，それに基づいてコーパスに情報を付与する必要がある。本章では BCCWJ で用いた二つの言語単位（短単位と長単位）が詳細かつゆれの少ない規則の集合として規定されていることが紹介されている。

第 5 章は，言語単位への分割を行う形態素解析システムとそこで利用される形態素解析用辞書 UniDic の解説である。UniDic は従来の工学系の言語処理システムが持っていた語の長さや見出し語の同一性の問題を解決した，言語学的な語を見出し語とする電子辞書である。

第 6 章は，『太陽コーパス』などの近代語のコーパスを例として，歴史コーパスの設計・構築について述べる。過去の言語は時代を遡れば遡るほど資料が少なくなるため，資料選定の問題が重要になる。また，異体字の取り扱いとして「包摂」「代用」などを逐一決める必要がある。本章ではこれらの問題を具体例を挙げながら説明している。

付録では，第 5 章で採り上げた UniDic を利用して形態素解析を行う「茶まめ」の利用法を紹介している。「茶まめ」は入門的なツールであり，これを足掛かりとして「ChaKi」などの解析結果を活用するツールへのステップアップが期待される。

4. データの更新について

本書で扱われている BCCWJ は，Web 上の検索インターフェース（「少納言」「中納言」）及び DVD により公開されている。2015 年 3 月，文境界の情報を更新した BCCWJ Version 1.1 がリリースされたが，本書の内容にはそれが反映されていない。更新情報についての詳細は国立国語研究所コーパス開発センターのホームページに掲載される予定なので，それを参照されたい。

山崎　誠（やまざき・まこと）

国立国語研究所言語資源研究系准教授。博士（学術）（東京学芸大学）。国立国語研究所研究員，同室長，同領域長等を経て，2009 年 10 月より現職。

主な著書・論文：『複合辞研究の現在』（共編著，和泉書院，2006），「代表性を有する現代日本語書籍コーパスの構築」（『人工知能学会誌』24(5)，2009），『言語研究のための統計入門』（共著，くろしお出版，2010），*A frequency dictionary of Japanese*（共編著，Routledge，2013）．

社会活動：計量国語学会理事，言語処理学会理事．