ID	KK300703
種別	書評
タイトル	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
Title	YAMAZAKI Makoto (ed.) (2014). *Corpus of Written Japanese: its Design and Structure.* Series of the Japanese Corpus, Vol.2. Tokyo: Asakura Shoten.
著者	鯨井　綾希
Author	KUJIRAI Ayaki
掲載号	30巻7号
発行日	2016年12月20日
開始ページ	427
終了ページ	433
著作権者	計量国語学会

計量国語学 30 巻 7 号（2016 年 12 月）pp.427-433.

書評

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

鯨井　綾希（東北大学）

1．はじめに

本書は，国立国語研究所を中心にして行われてきた研究プロジェクトをもとに企画された『講座日本語コーパス』というシリーズの一冊である．『講座日本語コーパス』には，電子化されたコーパスの概論から，コーパスと日本語学，あるいはコーパスと日本語研究に関連する諸領域との関わりに至るまでがシリーズの各書を通じて幅広く収められている．

その中で本書は，国立国語研究所がこれまで構築してきた言語研究用の書き言葉コーパスを例として，その設計方針や構築方法を具体的に示した書籍である．コーパスを用いた研究の普及に伴い，コーパスをテーマとした入門書や叢書も増えつつあるが，コーパスの基本設計・標準設計となりうる情報を示せるのは，大規模な日本語コーパスの構築を主導してきた国立国語研究所のみである．その点で本書は，次巻に位置づけられる『話し言葉コーパス―設計と構築―』と共に，『講座日本語コーパス』が他の書籍やシリーズと一線を画するための，最も重要な書籍であると考えられる．

本書の基本的な構成と各章の担当執筆者を以下に示す．

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）

本書は，書き言葉コーパスの一つである『現代日本語書き言葉均衡コーパス』（以下 BCCWJ と表記）を例としたコーパスの設計の概略（第 1 章）から始まり，コーパスを構成するデータの内実を決定付ける対象資料の選定方法およびそのサンプリングと電子化の手法（第 2 章・第 3 章），データに付与する言語分析用のアノテーションとしての形態論情報（第 4 章・第 5 章），それらの全てを踏まえて進められている『日本語歴史コーパス』構築の現在（第 6 章）へと続く．基礎的・横断的なコーパスの設計像から，各設計箇所の解説を経て総合的な構築例へと論が展開されていく点で，本書には構成上の明解さがある．

このうち，特に第 2 章・第 3 章・第 4 章は，将来的に日本語研究者の各々が自らコーパ

スを構築していく上で求められる，コーパスのイメージとコンセンサスの形成のために必要な部分であり，かつ本書を特徴づける重要な章であるように感じられる．以下では，紙幅の都合上，この三つの章に絞って，その内容について検討していきたい．

２．論評

2.1　第２章　サンプリング

コーパスを構築するにあたっては，どのような資料をどのように収集するかという点を予め明確にしておくことが大前提となる．第２章では，BCCWJ において定められた資料収集上の方針と方法について解説されている．以下では，本書で示される BCCWJ の収集方針と収集方法の一部について，概略的な紹介を行う．

2.1.1　テキストの収集方針とサブコーパスの設計（本書 2.2.1 項）

コーパス構築にあたっては，何よりもまず，どのような資料を収集するかという点を考えなければならない．本書によれば，BCCWJ における資料収集上の方針は，以下の三点を重視したとされる（23 頁）．

①「生産」「流通」「受容」という書き言葉の伝達過程の諸側面をできるだけ反映する設計であること
②母集団を数量的に定義できること
③母集団を適切に代表するサンプリングを実施できること

①については，合わせて「現代日本語書き言葉の研究にとって主要な対象となると考えられるタイプの書き言葉を優先的に収集する」とされる（23 頁）．また，BCCWJ は現代日本語を代表する，均衡の取れたコーパスを目指しているため，②や③の方針に従い統計的な処理が施された，母集団とサンプリングに関する設計が考慮されている．

本書によれば，BCCWJ は上記の①を鑑み，書き言葉の生産性に関わる側面を重視した出版サブコーパスと，書き言葉の流通実態に関わる側面を重視した図書館サブコーパス，およびそれらに収まらないけれども日本語研究上重要な資料となると考えられるものを集めた特定目的サブコーパスという三つのサブコーパスによる資料収集を行ったという（23 頁）．また，23 頁から 24 頁には，各コーパスの概略も記されている．出版サブコーパスは，2001 年から 2005 年までの短期間を対象とし，該当期間に出版された書籍・雑誌・新聞から②と③に沿う形でテキストを抽出している．図書館サブコーパスは，1986 年から 2005 年までの長期間を対象とし，東京都内の公立図書館にある書籍から，②と③に沿う形でテキストを抽出している．特定目的サブコーパスは，全体としては 1976 年から 2009 年までに出版・Web 配信された，白書・教科書・広報誌・ベストセラー・Yahoo!知恵袋・Yahoo!ブログ・韻文・法律・国会会議録という九種類の書き言葉を収録している．

以上からも分かる通り，コーパスは一定の収集方針の下で複数の下位資料を収めることが可能である．したがって，コーパスの利用者は，自らの目的に適したサブコーパスを主体的に選択することで，分析結果の確実性を高めることができる．

2.1.2　サンプルの抽出単位（本書 2.2.3 項）

資料中のどの部分をどの程度集めるのかという点もコーパスを構築する上で重要であ

る．第 2 章で取り上げられている BCCWJ の場合は，個々の資料中に含まれる文字列を「固定長サンプル」「可変長サンプル」という二つのサンプルとして取り出し，それをサンプルの収集単位として設定しているという．固定長サンプルは「統計的に厳密な言語調査を想定した，母集団からの抽出比を重視したサンプル」であり，「ある 1 文字を無作為に指定し，その文字を始点として 1000 文字目までの範囲を抽出」したものである（26 頁）．また，可変長サンプルは「文体研究・テキスト研究を想定した，ある程度の文脈を確保したサンプル」であり，「無作為に指定した 1 文字を含む言語的な構造のまとまり（「章」や「節」など，ただし概ね 1 万字を上限とする）を抽出するサンプルである」（26 頁）．

日本語を含む言語の研究にあたっては，言語の量的側面だけでなく，文や文脈上の意味を始めとした質的側面も重要である．BCCWJ において，統計学的厳密性を保持した「固定長サンプル」のみならず，言語研究で重要な文脈的問題にも対処しうる「可変長サンプル」を導入したことは，言語研究のためのコーパスの構築という目的に適うものであると言える．本書においても，「代表性と汎用性という 2 つの側面を可能な限り両立する方策として」，上記の二つのサンプルを設計したとあり（26 頁），苦慮のほどが窺える．

なお，以上のサンプリング方法は，現代日本語を代表する，均衡性の高いコーパスを構築することを目的とした BCCWJ におけるものである．当然ながら，コーパスの設計方針が異なれば，サンプリング方法もそれに応じて変わってくる．本書評では紙幅の都合で詳しくは述べないが，この点について本書は，さらに第 6 章 6.2 項で，『日本語歴史コーパス』の構築に際したサンプリングのあり方についての検討内容を紹介している．第 2 章と第 6 章を合わせて考えることで，サンプリング方法の手本を知るだけでなく，設計方針の多様性と本質について，より深く理解できるようになるだろう．

2.2　第 3 章 文書構造の電子化

日本語研究の中にコーパスを用いた研究を定着させるためには，日本語研究において共通基盤となりうる設計方針の下に，コーパスという分析資料を構築していく必要がある．そして，今後のコーパス構築手法における共通基盤の一つとなりうるものが，第 3 章で取り上げられる XML（Extensible Markup Language）による文書の構造化と電子化である．以下では，第 3 章の中でも特に文字媒体を電子化するときの文字入力の仕様（3.2 項）と，電子化された文書の構造を示す文書構造タグの仕様（3.3 項）の紹介を行う．

2.2.1　文字入力の仕様（本書 3.2 項）

文字入力の仕様は，コーパスのもととなる原資料をどのような形で電子化するかという点を考える上で，非常に重要である．第 3 章では，BCCWJ を例として，文字入力に関わる仕様が解説されている．文字入力仕様の設計方針は次の通りである（47 頁）．

①装飾，レイアウトなどの図形的情報を除いて，文字を入力する．
②すべての文字種の入力に，いわゆる全角文字を用いる．
③文字合成は行わない．
④上記条件に抵触しない範囲で，原則として，原文を忠実に転記する．

また，上記の基本的な方針に合わせて，BCCWJ の構築にあたって入力に用いる文字集

合および文字符号化方式も定められている．本書によれば，BCCWJ では，文字集合として日本工業標準調査会による JIS X 0213：2004 に準拠したものが設定され，文字符号化方式は UTF-8 が設定されている．これらの規格は，当該コーパスが扱える文字の範囲がどのくらいであり，またその文字をコンピュータに表示するためにどのような表示方式を選択しなければいけないかを明示するものである．

そのようにして文字入力上の仕様が明確に定められる一方，それに伴う注意事項も存在する．例えば，3.2.4 項では，データ入力に際した文字集合の設定による字体の包摂基準が定められており，一定基準に従って文字の置き換えが行われていることが述べられている．これは，原資料の文字が忠実に電子化されているわけではないということを意味し，コーパスにおいて文字に関わる研究を行う場合の注意点を示唆している．

上記のように，コンピュータ上で表示されている文字には，当然ながら電子的な処理が施されている．そうである以上，コーパスを設計・構築しようとする研究者は，原資料の文字を電子化する際の文字入力の仕様についても強く意識しておく必要がある．

2.2.2　文書構造タグの仕様（本書 3.3 項）

書き言葉コーパスの収集対象となるテキストの多くは，単なる文の集積ではなく，基本的に文よりも大きい文章という構造体として存在している．そうした文章上の構造に関わる情報を書き込むために用いられるのが文書構造タグであり，その仕様は BCCWJ の場合，以下の設計方針のもとで定められているという（54 頁）．

> ①多様な文書形式に対応できるようにする．
> ②文書記述言語として，XML を用いる．
> ③情報付与の対象は，文書中の論理的な役割が明確であり，かつ，紙面上の物理的な構造が明確な文書要素とする．

さて，ここで注目したいのは，①や③の設計方針を具体化する際に用いるツールとして，XML を利用することにしたという②の記述である．本書中では，XML を採用した理由が以下のように述べられている（54 頁）．

> XML は，拡張性に優れた文書記述言語であり，利用目的に応じて文書型を定義できる．また，XML はコーパスの記述だけでなく，データ一般の記述に広く用いられており，データ形式の検証，変換，検索などを行う際に，既存のツールを利用できるという利点もある．

上記引用部分のうち，特に最初の一文は，今後日本語研究においてコーパスが様々に構築されていくために重要である．「利用目的に応じて文書型を定義できる」とあるが，実際には，XML には「文書」の型のような文章相当の単位での情報のみならず，形態論情報を始めとした様々な情報を付与していくことができる．本書中では，第 6 章 6.3.3 項で取り上げられている『太陽コーパス』の構造化テキストの例が最も分かりやすい．すなわち，XML によってデータ構築を行えば，研究者の各々が自らの研究視座に応じて，コーパスに付与する情報タグを自由に考えることができるということである．さらに，XML によ

るタグ付与が日本語研究者のコーパス構築における共通基盤となれば，研究視座に合わせて既存のコーパスに情報を追加し，多様な視点を盛り込んだ共用コーパスを構築できるようになる．その意味で，本書で示される BCCWJ の設計を例とした XML によるデータ記述は，コーパス構築の将来において非常に重要な役目を果たすと考えられる．

2.3　第 4 章 形態論情報

第 4 章では，コーパスを構築する上で必要な言語単位の考え方や，言語単位の設計方針，コーパスへの言語単位の反映の仕方が述べられている．本書評では，コーパスにおける言語単位の詳細が説明される 4.2 項と，BCCWJ を例とした具体的な言語単位の設計が説明される 4.3 項を取り上げて論評を行う．

2.3.1　コーパスの言語単位（本書 4.2 項）

4.2 項では，コーパスの言語単位の考え方が述べられている．その中でも重要なのが，言語単位が不統一であることが引き起こす問題である．本書では，コーパスを用いる上で語の区切りの統一・不統一がもたらす問題について，以下の二つを論じている．

一つ目は用例の検索に関わる問題である．BCCWJ を始めとした言語研究用のコーパスは，単に文字資料を電子化しただけでなく，種々のアノテーションを施している．そのうち，語という単位のアノテーションは，用例の検索を始めとして，言語研究上の利便性を高める上で重要なものである．しかし，言語単位が不統一では，検索の際に本来拾わなければいけない用例を取り逃してしまうことになる．よって，コーパスを用例検索のツールとして確実なものとするためには，言語単位の統一が必要である．

二つ目は計量分析に関わる問題である．本書では宮島達夫（1969）「総索引への注文」（『国語学』76，110-120）に基づき，複合動詞をまるごと捉えるか分割して捉えるかによって計量結果が変わることを例に挙げつつ説明している．コーパスは量的な分析と相性が良いと言えるが，定量的分析を行うためには集計上の言語単位を統一的に扱わねばならない．

そうした問題を考慮し，BCCWJ の構築に際しては，語の斉一性という考え方が重視されている．本書によれば，語の斉一性とは「同じ語が常に同じに区切られて」おり，かつ「同じ構造の語が常に同じに区切られていることも求められる」（70 頁）．斉一性というキーワードの下に，言語分析に適いつつも，揺れの生じにくい，客観的な語の区切りを検討した結果が，4.3 項で紹介される「短単位」「長単位」という言語単位である．

2.3.2　BCCWJ における言語単位（本書 4.3 項）

BCCWJ では，語のレベルでの言語単位が考慮されており，そこに語の見出しとその表記，品詞，活用型，活用形，語種などの情報を加え，これを「形態論情報」と呼んでコーパス内のテキストに付与していく作業が行われている（68 頁）．4.3 項では，語のレベルの言語単位を定める上での方針として以下の三つが挙げられている（73 頁）．

方針 1：コーパスに基づく用例収集，各ジャンルの言語的特徴の解明に適した単位を設計する．
方針 2：『日本語話し言葉コーパス』と互換性のある形態論情報を設計する．
方針 3：国立国語研究所の語彙調査における知見を活用する．

BCCWJ の構築に際しては，方針 2 にあるように先行して構築された『日本語話し言葉コーパス』や，方針 3 にあるように国立国語研究所が長年取り組んできた語彙調査を踏まえつつ，方針 1 を最も重要なものとみなして言語単位の設計が検討されたという．そして，方針 1 にある「コーパスに基づく用例収集」と「各ジャンルの言語的特徴の解明」のそれぞれに適した単位を別々に採用することとし，前者に対応させた単位として「短単位」を，後者に対応させた単位として「長単位」を設けたとされる．また，それぞれの言語単位は具体的な認定規程に基づいて決定されていることが，本書で示される．

「短単位」と「長単位」という二つの言語単位をコーパスにおける形態論情報として組み込むことで，BCCWJ は用例収集用の検索にも，言語的特徴の分析にも堪えうるコーパスとなっている．目的に合わせて二つの言語単位を設計するという試みは，原資料のサンプリングの際の「固定長サンプル」「可変長サンプル」と同様に，コーパス構築における形式的な統一性と意味的な理解可能性の間に生じる問題への対処法として学ぶべき点がある．

2.4　コーパスの設計・構築における共通基盤形成という観点から見た本書の問題点

ここまで，電子化コーパスの構築において本書の中でも特に重要だと思われる三つの章に言及してきた．以上の三章を概観しただけでも，本書で提示されるコーパスの設計方針と構築方法が，日本語研究のための将来的なコーパスの構築と利用において，非常に重要な基盤的考え方を含んでいることが分かる．ただ，本書をコーパス設計・構築の共通基盤形成のための教科書とみなした場合に，個人的に気になった点も若干ながら存在した．

その一つとして，第 3 章で説明される XML による文書構造の電子化と，第 4 章で説明される形態論情報とに，コーパス設計上の隔たりがあるように見える点を取り上げたい．これは，XML を用いてマークアップされた文書構造ファイルと，形態素解析器を用いて形態論情報が付与された形態素解析済みファイルを，各研究者が将来的に行うであろうコーパスの設計・構築過程でどのように連結させるかという点が，少なくともコンピュータ言語に慣れていない者から見れば，本書から明確には読み取れないという問題である．

両者の連結は，例えば BCCWJ であれば，M-XML（形態論情報付き統合形式 XML）を見れば具体的なイメージを掴むことができる．本書においても，第 6 章 6.4.1 項で『明六雑誌コーパス』による形態論情報の XML での記述例が示されている．しかし，第 6 章で第 3 章と第 4 章をつなぐものが提示されていたとしても，各章の独立性の高さゆえに，それに気づける読者は多くないように思える．

もちろん，各章の内容を簡潔かつ十全に説明するためには，章ごとの結びつきに紙幅を割くわけにはいかない．ただ，そうは言っても，電子的な処理に詳しくない研究者にも広くコーパスの設計を知ってもらい，将来本書に即したコーパスを構築してもらうために，そうした周辺的な部分への配慮があれば，なお良かったように思われる．

2.5　その他の章について

その他の章も，本書評では紙幅の都合で論評することが適わなかったが，コーパスの設計と構築を学ぶ上での貴重な知識を提供してくれる．第 1 章はコーパスの設計・構築の基本を考える上で押さえておくべきことを簡略かつ明解に示している点で，今後コーパスを

構築しようとしている研究者にとって有益である．第 5 章はコーパスのアノテーションにおいて機械化された部分の実態を把握するのに貢献し，コーパス構築作業のブラックボックス化を防ぐ手だてとなる．第 6 章は第 5 章までで説明されてきた設計が構築段階においてどのように生かされるのかを実感するのに役立つだけでなく，現代語を資料とした BCCWJ とは異なる歴史資料の電子化に際した問題をあぶり出している．付録における形態素解析ツールの紹介は日本語研究者がコーパスを構築する際の手段と可能性を広げる契機となりうる．これらの章もまた，本書全体の構成において重要かつ必須な部分である．

3．おわりに

本書評では，日本語の電子化コーパスが今後幅広く構築されていくために必要な共通基盤の形成に対して本書が果たす役割に重点を置き，論評を行った．その結果，章ごとに検討すべき部分が多くなり，章を絞った内容紹介となってしまったが，本書が持つ重要性については一定程度示せたように思う．

日本語のコーパスを構築するための環境は，ここ数年で急速に整ってきていると言える．それに伴い，今後は研究者ごとに様々なコーパスが作られるようになるであろう．しかし，コーパスの設計・構築上の理念や具体的構造については，現状では日本語研究者の間で十分なコンセンサスが得られていないように感じられる．本書の刊行により，日本語研究用のコーパスを構築する際の共通基盤が出来上がり，それを踏まえて各人が各人らしいコーパスを作ったり，相互に利用したりできるようになることを期待したい．

（2016 年 5 月 31 日受付）