第２章　『現代日本語書き言葉均衡コーパス』の設計

山崎 誠

2.1　はじめに

本章では、『現代日本語書き言葉均衡コーパス』（以下、BCCWJと省略）の設計の概要について説明する。

BCCWJは日本で初めての本格的な書き言葉均衡コーパスである。BCCWJは次のような点で日本語研究の質の向上に貢献する。従来、日本語研究においてコーパスとみなして利用されてきたデータはいくつかあったが、それらは新聞記事データ集や青空文庫などの単一の種類のテキストの集まりであり、書き言葉の一面を捉えているにすぎなかった。それに対して、BCCWJは書籍、新聞、雑誌、白書、ブログ等異なるレジスターのテキストの集まりであり、書き言葉の多様な実態を捉えることができるデータになっている。

また、従来の書き言葉データの多くはプレイン・テキストであり、その使い方は文字列検索が中心であったため正規表現を使っても限界があった。BCCWJは言語単位の情報（形態論情報）や書誌情報などの研究用のアノテーションが施されており、複雑な検索結果をもとに、より深い分析が可能である。

2.2　BCCWJ の設計

2.2.1 基本方針

BCCWJを構築するにあたっては、以下の四つの点を念頭に置いて設計した（前川 2008、山崎 2009）。

（1）現代日本語の縮図となるコーパス

従来、国立国語研究所が行ってきた語彙調査の手法を生かし、コーパスがその母集団の統計的な縮図になり、母集団に対し代表性（representativeness）を持つように設計する。これにより、母集団における言語的諸特性の分布が過不足なく表現できることになり、データの信頼性を高めることが出来る。

（2）汎用的な目的に供するコーパス

言語研究（語彙・文法・文字）以外にも、応用面として日本語教育や国語教育、国語政策、辞書編集、自然言語処理などの分野でも活用することを目的として、多様な日本語の姿を捉えることができるよう設計する。

（3）公開可能なコーパス

収録する著作物について利用許諾を得て公開する。公開形態は、オンラインでの簡易検索のほか、形態論情報を使って共起条件を詳しく指定できるオンライン詳細検索、DVDに

よる全文提供の3種類である。コーパスが学界の共有財産となることによって、研究の追試が可能になったり、日本語を母語としない研究者が研究を行いやすくなるなどのメリットがある。

（4）既存のコーパスとの調和

XMLによる文書構造の記述、2種類の言語単位（短単位、長単位）による形態論情報の付与により、『太陽コーパス』『日本語話し言葉コーパス』との整合性を保つ。

2.2.2 基本概念の定義

BCCWJは、現代日本語の書き言葉を収録するコーパスであるので、「現代」「日本語」「書き言葉」のそれぞれについて、以下のような基準を決めて資料選定にあたった。詳細な取り扱いについては、第3章「サンプリング」及び丸山他（2011a）を参照されたい。

【現代】

明治時代以降を現代とする。したがって、「源氏物語」などの江戸時代より以前の作品は対象外となる。ただし、古典の現代語訳は現代語として扱った。また、短歌、俳句などの韻文で使われる古語は現代語として扱った。

【日本語】

方言を含む日本語が対象である。英語、中国語などの外国語は対象外である。テキストによっては、日本語と外国語が混じっているものがある。そのような場合、段落単位で外国語かどうかの認定を行い、対象範囲を確定した。

【書き言葉】

文字で記録された言葉。インタビューの書き起こしなどを含む。

2.2.3 BCCWJ の基本構成

BCCWJは、出版（生産実態）サブコーパス、図書館（流通実態）サブコーパス、特定目的サブコーパス三つのサブコーパスから構成される（図2-1参照）。

<table>
<tr><td>出版（生産実態）サブコーパス
約3,500万語
書籍、雑誌、新聞
2001年～2005年</td><td>図書館（流通実態）サブコーパス
約3,000万語
書籍
1986年～2005年</td></tr>
<tr><td colspan="2">特定目的サブコーパス
約3,500万語
白書、教科書、広報紙、ベストセラー
Web掲示板、ブログ、韻文、法律、国会会議録
対象期間はさまざま</td></tr>
</table>

図2-1: BCCWJの構成

各サブコーパスは、さらにいくつかのレジスターから構成される。表2-1は各レジスターのサンプル数と短単位で数えた場合の延べ語数を示したものである。語数は品詞欄が空白・補助記号・記号のものは数えていない。また、固定長サンプルと可変長サンプルがあるレジスターについてはそれらを合わせて重複部分を差し引いた範囲を対象としている。

表2-1: 各レジスターのサンプル数と語数

サブコーパス	レジスター	サンプル（個）	NumTrans 版の語数（万）	非 NumTrans 版の語数（万）
出版サブコーパス	書籍(PB)	10,117	2,855	2,866
	雑誌(PM)	1,996	444	450
	新聞(PN)	1,473	137	138
図書館サブコーパス	書籍(LB)	10,551	3,038	3,044
特定目的サブコーパス	白書(OW)	1,500	488	494
	教科書(OT)	412	93	93
	広報紙(OP)	354	376	383
	ベストセラー(OB)	1,390	374	375
	Yahoo!知恵袋(OC)	91,445	1,026	1,030
	Yahoo!ブログ(OY)	52,680	1,019	1,028
	韻文(OV)	252	23	23
	法律(OL)	346	108	108
	国会会議録(OM)	159	510	510
合計		172,675	10,491	10,542

2.2.4 BCCWJ の規模

表2-1に示すように、BCCWJ全体の規模は短単位で数えて約１億語である。レジスター別では、LB（図書館書籍）が最も大きく約3,000万語、PB（出版書籍）もほぼ同じサイズであり、合わせると、BCCWJ全体の約6割は書籍で占められていることになる。それぞれのレジスターにおける延べ語数が異なるため、レジスター間で出現頻度を比較する場合は、それぞれの語数で割った出現率で比較しなければならない。

2.2.5 各サブコーパスの特徴

以下、各サブコーパスについて概括を述べるが、それぞれのサブコーパスに含まれるレジスターとその選定方法については、第3章「サンプリング」及び丸山他（2011a、b）を参照されたい。

A. 出版サブコーパス

書き言葉を生産する書き手の立場を重視したもので、売れ行きや知名度にかかわらず、出版された書き言葉であれば、どの書籍（雑誌、新聞）も同じ確率で選ばれるようにする。後述の流通実態を捉えたサブコーパスに比べると語彙やコロケーションなど言語的属性の

多様性が確保されることが期待される。

このサブコーパスには成人向けの書籍が一定の割合で含まれている。教育現場で使用する際には注意されたい。

B. 図書館サブコーパス

書き言葉が書き手と読み手との間で、社会的に流通している実態を図書館の所蔵から捉えたサブコーパスである。広い意味で社会の需要を反映している書き言葉とも言える。このサブコーパスは、極端に専門的な書籍や成人向け書籍が排除されることによって、より一般的な用語用字を調べるのに適していると期待される。また、資料年代にある程度の時間的な幅があり、短期間であるが通時的な観察が可能になる。

C. 特定目的サブコーパス

出版サブコーパス、図書館サブコーパスでは十分な分量が集まりにくい資料を中心に収録したサブコーパスである。例えば、政府の白書は上記二つのサブコーパスからでは分析に必要なだけの量が得られないため、白書のみを母集団としたデータからサンプリングを行い、サブコーパスに収録した。同様に、教科書・広報紙・ベストセラー・韻文・法律・国会会議録を収録した。また、ウェブの書き言葉（Yahoo!知恵袋、Yahoo!ブログ）も収録し、紙媒体の言語と比較できるようにした。

2.2.6 コアデータ

BCCWJに付与されている形態論情報などのアノテーションは、ほとんど自動付与であるが、BCCWJ全体の約100分の1の量に相当する約110万語については、人手により解析精度を高めている。この部分を「コアデータ」と呼んでいる。BCCWJ全体の解析精度が約98%であるのに対してコアデータの解析精度は99%以上である。コアデータを構成するレジスターは、出版書籍（PB）、雑誌（PM）、新聞（PN）、白書（OW）、Yahoo!知恵袋（OC）、Yahoo!ブログ（OY）の六つである。

コアデータには、さまざまなアノテーションが施されており、順次、次のURLで公開される予定である。

http://www.ninjal.ac.jp/corpus_center/anno/

2.3 サンプルの長さとタイプ

2.3.1 問題点

コーパスに収録する1サンプルの長さをどのように決めるかはコーパスの設計にとって、コストにも影響する重要な問題である。1サンプルの長さが長くなれば収録するサンプルの数が少なくなり（著作権処理の負担減にも直結する）、労力も少なくて済むが、語彙的なかたよりが生じる。

また、1サンプルの長さについて、それが一定かどうかという、サンプルのタイプも重要な問題である。一定の長さのサンプルは計量的な分析に向いているが、多くの場合文が途

中で切れてしまうことになり、文脈を把握するような分析には向いていない。意味的なまとまりを重視するとサンプルの長さがまちまちになる。

BCCWJでは、サンプルの長さをとタイプについて、それぞれの長所を生かす以下のような設計を行った。

2.3.2 サンプルのタイプ

A. 固定長（FIXED）サンプル

固定長サンプルは、ひとつのサンプルの長さを1,000字とする（句読点などの補助記号は含めない）。固定長サンプルは、母集団からの抽出比率に基づいた統計的な処理、語彙表や漢字表の作成に適している。ちなみに、1サンプル1,000字は短単位で約590語であり、文庫本でいうと見開きより少し多いくらいの言語量である。

固定長サンプルのデータは、係り受けの関係が理解できるよう、サンプルの開始点を含む文の文頭からサンプルの終了点を含む文の文末までが収録されている。そのため、実際のひとつのサンプルの文字数は1,000字より多いが、サンプルの開始点と終了点がマークアップされており、その間がちょうど1,000文字となる。

B. 可変長（VARIABLE）サンプル

可変長サンプルは、文章のまとまりをもとに長さを決める。そのためひとつのサンプルの長さは一定ではない。多くの書籍では、節、章などのまとまりが1サンプルとなる。ただし、無制限に長いサンプルができるとそのサンプルの影響が強く出てしまうので、長さの上限を1万字としている。可変長サンプルは文章の論理構造を対象とした分析に適している。

2.3.3 サンプルの重なり

コーパス構築に当たって固定長サンプルと可変長のサンプルを別々に取得するのは作業コストがかかりすぎるため、BCCWJでは1回のサンプリングで当たった同一箇所から固定長と可変長の二つのサンプルを取得している。そのため、固定長サンプルと可変長サンプルとの間には包含関係を基本とする3種類のパターンが生じる。いちばん多いパターンは、固定長サンプルが可変長サンプルの中に完全に含まれる場合である。次に多いのが、固定長サンプルが可変長サンプルの末尾からはみ出す場合である。また、数は少ないが、固定長サンプルと可変長サンプルが重なり合わないパターンもある。

2.3.4 レジスターとサンプルのタイプ

表2-2にレジスターとサンプルのタイプの関係を示した。可変長サンプルは全てのレジスターにあるが、固定長サンプルは、出版サブコーパス全体、図書館サブコーパス全体と特定目的サブコーパスの白書だけに存在する。

表2-2: レジスターとサンプルのタイプ

サブコーパス	レジスター	サンプルのタイプ
出版サブコーパス	書籍(PB)	固定長、可変長
	雑誌(PM)	固定長、可変長
	新聞(PN)	固定長、可変長
図書館サブコーパス	書籍(LB)	固定長、可変長
特定目的サブコーパス	白書(OW)	固定長、可変長
	教科書(OT)	可変長
	広報紙(OP)	可変長
	ベストセラー(OB)	可変長
	Yahoo!知恵袋(OC)	可変長
	Yahoo!ブログ(OY)	可変長
	韻文(OV)	可変長
	法律(OL)	可変長
	国会会議録(OM)	可変長

2.4 電子化

2.4.1 文字入力

出版サブコーパスおよび図書館サブコーパスのように原文が紙媒体（原資料の媒体についての詳細は表4-1を参照）である場合には、電子化するための基準が必要である。文字入力については、以下の方針を立てた。

(1) JIS X 0213:2004 規格に基づき字形を詳細に区別する

この文字セットの採用により、ほとんどの文字を入力し分けることができる。詳細は、高田他（2009）を参照されたい。

(2) 記号・改行の意味による統制、統一的な表記

例えば、「コーパス」という語を表記する際の2文字目の中央位置横線は、通常「ー（長音符号）」が用いられるが、資料によっては「－（マイナス）」や「―（ダッシュ）」が用いられているものや、形状からはどの文字かを判別できない場合がある。また、「ー（マイナス）」を用いた「コーパス」という表記を、そのままコーパス本文に採用すると、語の検索や形態素解析を困難にする。そのため、原文における見え方ではなく、その意味によって入力し分ける。ダッシュ、ハイフン、長音、漢数字の「一」、丸記号、漢数字の「〇」、ローマ字の「０」などが対象となる。また、改行やスペースは、レイアウトではなく、論理的に意味をもつもののみを再現する。例えば、語や文を句切る空白、段落冒頭の1字字下げは入力するが、レイアウトのための空白は入力しない。

(3) 組み文字・半角文字を使わない

㈱、㌢のようないわゆる組み文字は「（株）」、「センチ」のようにすべて１字ずつ切

り離して入力する。また、半角文字は使用せずすべて全角で入力する。

文字入力の具体的な記述は西部他（2011）を参照されたい。

2.4.2 タグの仕様

BCCWJのタグの特徴は、形態論情報のタグだけでなく、『太陽コーパス』で行ったタグ付けの経験を生かし、文書構造が的確に再現されるようにしている点である。以下に主なタグの種別と特徴を挙げる。タグの詳細は、第4章および山口他（2011）を参照されたい。

(1) 文書構造情報

記事、見出し、段落、引用、文などのタグを付与し、文章を構造化・階層化して表現する。

(2) 文字情報

文字の読みに関するルビ、誤植などの校正注、文字集合に含まない文字や記号（外字）などの情報を付与する。

(3) 形態論情報

短単位、長単位についての形態論情報（語彙素、出現形、品詞、語彙素読み、語種など）を付与する。

(4) サンプリング情報

サンプリング時に決定するサンプル抽出基準点（乱数による縦横交叉点から決まる文字。7.4.5節参照）の情報を付与する。

2.5 解析単位（短単位、長単位）

BCCWJでは柔軟な検索・分析に対応するために「短単位」「長単位」という2種類の言語単位を用いている。短単位はコーパスからの用例収集に適した単位であり、長単位はBCCWJに格納したレジスターの言語的特徴の解明に適した単位である。

解析単位は、大量のデータをコンピュータで処理するのに向いているという性質が必須である。BCCWJの構築にあたってもその趣旨に則って、解析単位を揺れの少ない規則の集合として定義した。その詳細は、第5章および小椋他（2011a）を参照されたい。

BCCWJはすべてのサンプルが短単位と長単位の二つの単位で解析されている。解析精度は品詞も含めた見出し語の認定のレベルで98%以上である（レジスターによって解析精度に若干差がある）。

短単位、長単位は、元々は国立国語研究所の語彙調査で開発された調査単位であり『日本語話し言葉コーパス』の構築においても使用された。前者は最小単位（形態素）の一次結合までを最大とする言語単位であり、後者はほぼ文節に近い長さの言語単位である。例えば、「国立国語研究所は人間文化研究機構に移管される。」という文は、短単位で「/国立/国語/研究/所/は/人間/文化/研究/機構/に/移管/さ/れる/。/」と14単位に分割されるが、長単位では、「/国立国語研究所/は/人間文化研究機構/に/移管さ/れる/。/」と7単位になる。

参考文献

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011a）「『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）」国立国語研究所内部報告書 LR-CCG-10-05-01

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011b）「『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 （下）」国立国語研究所内部報告書 LR-CCG-10-05-02.

高田智和・小林正行・間淵洋子・大島一・西部みちる・山口昌也（2009）「JIS X 0213:2004 運用の検証」特定領域研究「日本語コーパス」平成21年度研究成果報告書JC-D-09-01.

西部みちる・大島一・間淵洋子・小林正行・田島孝治・高田智和・山口昌也（2011）「『現代日本語書き言葉均衡コーパス』における電子化テキストの構築」国立国語研究所内部報告書LR-CCG-10-04.

前川喜久雄（2008）「KOTONOHA『現代日本語書き言葉均衡コーパス』の開発」日本語の研究, 4 (1), 82-95.

丸山岳彦・山崎誠・柏野和佳子・佐野大樹・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子（2011a）「『現代日本語書き言葉均衡コーパス』におけるサンプリングの原理と運用」国立国語研究所内部報告書LR-CCG-10-01.

丸山岳彦・山崎誠・柏野和佳子・佐野大樹・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子（2011b）「『現代日本語書き言葉均衡コーパス』に含まれるサンプルおよび書誌情報の設計と実装」国立国語研究所内部報告書LR-CCG-10-02.

山口昌也・高田智和・北村雅則・間淵洋子・大島一・小林正行・西部みちる（2011）「『現代日本語書き言葉均衡コーパス』における電子化フォーマット ver.2.2」国立国語研究所内部報告書LR-CCG-10-04.

山崎誠（2009）「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), 623-631.