代表性を有するコーパスの設計とサンプリングの実際
—コーパスに基づく言語研究の可能性と限界—

丸山 岳彦

国立国語研究所 言語資源研究系

1　導入

『「言語表現」と「言語」のあいだ』という問題について、現代日本語のコーパスを例に考えてみたい。はじめに、実証的な分析・記述を目指す言語研究にとって、均衡コーパスが必要であることを述べる。その後、『現代日本語書き言葉均衡コーパス』の設計方針と、その中で実施しているサンプリングの過程と問題点を示した上で、コーパスに基づく言語研究の可能性について述べる。

2　実証的な言語研究と均衡コーパスの必要性

20 世紀半ば以降の言語研究の立場は、実際に書かれたり話されたりした言語データ（実データ）を重視するか否かによって、区別することができる (Leech, 1990)。母語話者の言語直観や内省に基づいて言語現象を演繹的に説明しようとする生成文法の立場では、実データは重視されない。一方、言語形式の意味や機能を帰納的に分析しようとする記述言語学や機能言語学、コーパスから観察される現象を定量的に分析するコーパス言語学では、母語話者の内省とともに、実データの参照が不可欠である。このうち後者の立場、すなわち経験主義的・実証的な方法を志向する言語研究では、個別の「言語表現」に関する分析を積み重ね、その結果を一般化して「言語」の性質を記述するという手順が取られる。「言語表現」の分析から「言語」の特質を明らかにするわけである。

コーパス言語学や自然言語処理の研究など、実データを対象とする（対象としなければならない）研究にとって、どのような分析対象を準備するかは、その分析の一般性や妥当性に直結する問題である。極端に分野の限られた言語データ（新聞記事テキストなど）や、出自の分からない言語データ（Web から収集したテキストなど）を分析しても、そこで得られる結果が「現代日本語」の一般的な特徴であるとは必ずしも言い切れない。

得られた分析・記述の結果が、「現代日本語」のどの部分に関することなのかを客観的に位置づけるためには、その基準となる言語資料が必要である。そこで、多様な種類のテキストを含み、ジャンル、メディア、年代、出典情報などが管理された「**均衡コーパス（Balanced Corpus）**」が必要となる。すなわち、個別の「言語表現」の分析結果が、「言語」のどの側面を明らかにしているのかを裏付けるための指標が求められるのである。

英語を対象とした均衡コーパスは、1950 年代の SEU、1960 年代の Brown Corpus、1990 年代の BNC などに代表されるように、すでに長い歴史がある。日本でも、国立国語研究所において 1960 年代から統計的な語彙調査が行なわれてきたが、均衡コーパスの設計やその公開、研究者間での共有という段階までには至らなかった。

2006 年以降、国立国語研究所を中心に『**現代日本語書き言葉均衡コーパス**（Balanced Corpus of Contemporary Written Japanese; 以下 BCCWJ と記す）』の構築が進んでいる (前川, 2008)。この 1 億語規模のコーパスは、日本語では初めての均衡コーパスであり、綿密な調査に基づいた設計と、母集団に対する統計的な代表性を有するという点に、最大の特徴がある。また、書誌情報やジャンル情報などが付与され、そこに含まれる言語表現の客観的な位置づけが可能になっている。さらに、すべてのサンプルについて著作権処理を実施し、研究者間で共有されるコーパスとしての公開を目指している。

本稿では、BCCWJ を構成するサブコーパスの一つである「**出版サブコーパス**（以下 **出版** SC と記す）」について、コーパスの設計とサンプリング作業の実施状況を報告する。出版の実態を捉えるためのコーパスをどのように設計したか、それを実際にサンプリングしようとしたとき、どのような問題が生じたか、それをどのように解決したか、という点について述べる[*1]。

[*1] BCCWJ の設計方針とサンプリングの基準に関する詳細については、丸山・秋元 (2007, 2008)、柏野ほか (2009) を参照。

Copyright © 2010 The Association for Natural Language Processing.
All Rights Reserved.

3 BCCWJ「出版 SC」の設計

3.1 コーパスデザインと母集団の定義

出版 SC は、比較的短い期間に生産された書き言葉の実態を捉えるためのコーパスである。2001 年から 2005 年の間に国内で出版されたすべての書籍・雑誌・新聞を対象に、層別無作為抽出によってサンプルを抽出する。以下ではこのうち、書籍について述べる。

国内における書籍の出版実態を把握するために、国立国会図書館の蔵書目録 J-BISC を資料として用いた。納本制度により、国内で出版される書籍は原則として国立国会図書館に納められることになっており、この目録によって出版された書籍の総体を捉えることができる。2001 年からの 5 年間に発行された書籍のうち、40 ページ以下の書籍、ページ数の記録がない書籍、電子資料、漫画や写真集などを除いたところ、317,337 冊、74,911,520 ページが対象となった。

さらに、これらのページ上に印刷されている文字の数によって、母集団を定義することにした。母集団を量的に捉える上で、文字が最も揺れのない単位であるからである。対象となる書籍を、「発行年」による 5 層、「日本十進分類法（NDC）」による 11 層の、計 55 層に層別した上で、各層に含まれる文字数を推計したところ、48,539,925,351 文字という結果を得た (丸山・秋元, 2007)。この文字数を母集団として定義し、各層に含まれる文字数の比を、抽出するサンプルの構成比として採用した。NDC で層別した場合の総文字数とサンプル構成比を、表 1 に示す。

表 1 **出版** SC 書籍の母集団の構成比

NDC	推計総文字数	構成比
0. 総記	1,636,414,548	3.37%
1. 哲学	2,597,610,813	5.35%
2. 歴史	4,301,204,340	8.86%
3. 社会科学	12,408,321,943	25.56%
4. 自然科学	5,069,594,034	10.44%
5. 技術工学	4,615,929,967	9.51%
6. 産業	2,196,387,437	4.52%
7. 芸術	3,258,432,447	6.71%
8. 言語	888,800,128	1.83%
9. 文学	9,341,275,486	19.24%
n. 記録なし	2,225,954,208	4.59%
合計	48,539,925,351	100%

3.2 抽出単位とサンプルサイズ

抽出単位には、母集団からランダムに抽出した 1 文字を基準として 1,000 文字を取得する「固定長サンプル」と、ランダムに抽出した 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 1 万字を超えない範囲）を取得する「可変長サンプル」の 2 種類を準備した。

出版 SC のサンプルサイズは、「固定長サンプル」を 1,000 万語分集めることとした。1 語を 1.7 文字と仮定し、1,000 文字から成る固定長サンプルを 17,000 個（1,700 万文字分）収集する。このうち、書籍の推計総文字数は**出版** SC 全体での推計総文字数の 74.14% を占めるため、741.4 万語分、12,604 サンプルを収集することになる。このサンプル数を 55 層の構成比に比例割当し、各層から取得すべきサンプル数を算出した。さらに、可変長サンプルの平均長を 3,900 文字と試算し、**出版** SC の書籍全体で約 2,892 万語、**出版** SC 全体で 3,468 万語を取得することを計画した。

3.3 サンプル台帳の作成

母集団である 48,539,925,351 文字の中からランダムに 1 文字を抽出するための近似的な手段として、母集団に含まれるすべてのページからランダムに 1 ページを抽出し、さらにそのページに含まれる 1 文字をランダムに抽出する、2 段抽出法を採ることにした。

そこで、各層に含まれるすべてのページに対してランダムに優先順位を割り振り、さらにページ内の 1 文字を指定するための座標情報をランダムに指定した。指定されたページの、指定された座標に最も近い文字が、母集団からランダムに抽出された 1 文字となる。これらの情報をまとめたものを、サンプル台帳として準備した。

出版 SC の書籍全体で必要なサンプル数は 12,604 であるが、何らかの理由でサンプリングができない場合や著作権処理の過程で問題が生じた場合を考慮し、優先順位で 20,000 位までのサンプル台帳を準備し、この範囲の中でサンプリングを実施することにした。

4 サンプリング作業の実際

4.1 サンプリングの実施に伴う障害

このようなコーパスデザインをもとに、母集団とサンプルサイズが定義され、サンプリングを実施するための台帳の準備が整った。その後は、所定のサンプル数に達成するまでサンプリングを進めていけばよい。

しかしながら、実際にサンプリングを進めていくと、

Copyright © 2010 The Association for Natural Language Processing.
All Rights Reserved.

さまざまな原因による問題が発生し、作業の円滑な進行が妨げられた。そもそも、出版されたすべての書籍からランダムにサンプルを抽出するという作業自体、前例のない試みであり、その過程で生じた問題を一つずつ解決していく必要があった。以下では、サンプリングの過程で生じた問題点と、その対処の仕方を示す。

4.1.1 現物の取得可能性に関する問題

まず問題となったのが、指定された書籍をどのように入手・閲覧するかという点である。ランダムに指定された書籍のリストには、著名な作家の単行本もあれば、自費出版の書籍もある。国立国会図書館の蔵書目録で母集団を構成している以上、国立国会図書館ですべての作業ができればよいが、現実的には不可能である。

そこで、(1) 近隣の図書館から該当する蔵書を借りる、(2) 比較的安価な書籍は書店で購入する、という方法を採ることにした。(1) については、立川市図書館を中心に、東京都立図書館、一橋大学附属図書館、八王子市図書館などに個別に依頼し、団体貸出の制度を利用して書籍を借り出し、国語研究所で作業を行なった。この際、蔵書検索のために、各図書館の蔵書データを事前に入手する必要があった。国立国語研究所の研究図書室からも、該当する書籍を借り出した。また、近隣の図書館で所蔵していない書籍について、国立国会図書館へ出向き、特別に許可を得て書庫内での作業を実施した。対象となるすべての書籍が所蔵されているものの、研究所内での作業ほどの効率が得られなかったため、それ以上の作業は断念した。(2) については、近隣の 2 つの書店にリストを渡し、新品または古書を取り寄せた。これによってかなりの冊数を入手することができたが、2001 年から 2005 年に出版された書籍であっても、売場からすぐに消えてしまう書籍や、そもそも市場に流通しない書籍も多数あった。また、書籍を一意に同定するために、発注には ISBN を用いたが、古書の在庫を ISBN で検索できる古書店が非常に限られているという問題もあった。

2 つの書店での探索がほぼ尽きた段階で、Amazon で販売されている古書の購入に切り替えた。先の 2 つの書店では入手できなかった書籍を、安く大量に購入することができたが、1 冊ごとに異なる出品者から購入せざるを得ず、購入にかかる手続きが非常に煩瑣であった。

上記の手段により、最終的には、14,523 冊の書籍（15,162 サンプル分）を入手した。入手先・閲覧先の一覧を、表 2 に示す。

表 2　**出版** SC 書籍の入手先・閲覧先

取得先	冊数	取得先	冊数
オリオン書房	5,538	国語研図書室	222
Amazon	3,122	東京都立図書館	184
立川市図書館	2,633	国立国会図書館	158
高原書店	2,093	八王子市図書館	153
一橋大学図書館	275	その他	145
		合計	14,523

4.1.2 サンプリングの基準に関わる問題

次に問題となるのが、実際に手にした書籍に印刷されている文字列のうち、どの部分をサンプルとして抽出すべきか、という問題である。

抽出対象とするのは、原則として、「現代日本語で書かれた表現」とした。一見、書き言葉の中から「現代日本語で書かれた表現」を取り出すことは簡単な作業のように思われるが、実際には非常に困難である。例えば、日本語と外国語が混じった文章や、数式や化学式などが混じった文章をどう扱うか、表組みのように複雑な構造を持つ部分をどう扱うか、カタログのような様式の印刷紙面上にある文字列のうち、どの部分をどのような順序で取得していけばよいか、などといった問題に直面した。その結果、詳細な規則と判断基準が必要となり、また事例ごとに柔軟な判断が求められる場合が多かった。また、手に取った書籍が「人名録」や「語彙表」のような体裁を持ち、人名や単語の羅列しか記載されていない場合、サンプルとして抽出できる部分はないと判断せざるを得なかった。

4.1.3 著作権処理に関する問題

サンプリングが済んだ個々のサンプルは、コーパスに格納して公開する許諾を著作権（保持）者から得るための作業に回される。現時点（2010 年 1 月）において、著作権者の連絡先が判明しないケースが全体の約 3 割ある。連絡先が判明した約 7 割のサンプルのうち、許諾が得られるのがその約 7 割となり、全体として約 5 割のサンプルで許諾が得られている状況である (前川, 2009)。

著作権者に連絡が取れたものの中で、利用を拒否されるケースは、約 5% あった。さらに、出版社が著作権を保持するサンプルで、出版社側が利用を拒否すると回答してきたケースがあった。これにより、サンプリングしたテキストが利用できなくなったり、その後に入手する予定だった冊数が大幅に減ったりする事態が生じた。

Copyright © 2010 The Association for Natural Language Processing.
All Rights Reserved.

また、翻訳書の著作権処理において、翻訳エージェンシーによっては利用料金を請求してくるところがあった。対象となるサンプル数を考えると非常に高額になることから、その翻訳エージェンシーが仲介している翻訳書のサンプリングは見送らざるを得なかった。

4.1.4 当初の設計の修正

さらに、作業の進行に伴い、設計の修正が必要となることが分かった。当初は可変長サンプルを平均 3,900 文字と試算していたが、電子化作業が進むことにより、平均で約 4,300 文字程度になることが判明した。このため、設計通りに 12,604 サンプルを取得すると、可変長サンプルのサイズが大幅に増大する見込みとなった。

試算の結果、当初の設計のうち、80% のサンプル数を確保することで、可変長サンプルの合計がほぼ予定通りの語数に達することが判明した。そこで、全体で取得する固定長サンプルの合計を 800 万語とし、必要な書籍のサンプル数の合計を 10,083 サンプルと修正した。

4.2 サンプリングの結果

出版 SC 書籍のサンプリングは、上記のような手続き上の問題、著作権処理をめぐる問題、設計上の問題などに対処しながら進めた。2006 年 11 月より作業を開始し、現時点までで 11,140 サンプルの作業が完了している。当初の設計（12,604 サンプル）に対して、サンプリングが完了した比（達成率）を、図 1 に示す。

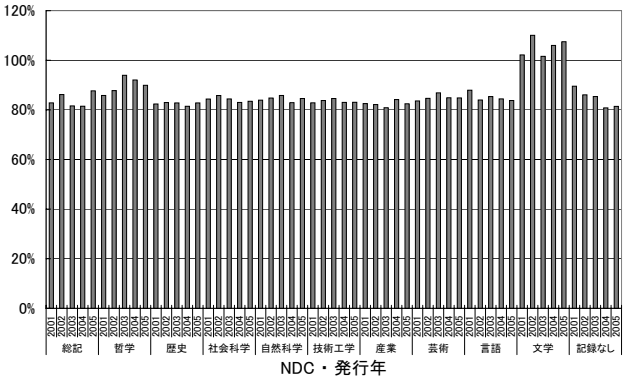

図 1　**出版** SC 書籍： 55 層の達成率

5 コーパスに基づく言語研究の可能性と限界

BCCWJ に含まれる言語現象の分析は、今後の日本語研究にとって大きな課題である。コーパスに基づく語彙調査、文字調査、文法研究、変異研究、文体研究など、さまざまな実証的研究へ応用するための方法論自体が、今後模索される必要がある。

また、Biber (1993) が述べるように、コーパスが備えるべき代表性をさまざまな側面から検討し、母集団の定義やコーパスサイズなど、その設計の妥当性を評価することも必要である。サンプリングの手続き上の問題や、どのような仕様のコーパスを作ればどのような結果が得られるのかなどを、経験的に探っていく必要がある。

さらに、コーパスを使った言語研究の限界についても認識しておく必要がある。どれだけ質のよいコーパスができても、稀な言語現象は含まれていないこともある。また、「日本語表現」の具体的な事例を収集・分析できたとして、それが母集団が備える特質に過ぎないのか、広く「現代日本語」一般に適用できる言語事実なのかを見極め、位置づける作業は、最終的には分析者に任される。内省に基づく主観的判断と、コーパスの客観的な分析結果とを相補的に組み合わせながら、言語現象の記述とその検証を進めていく姿勢が求められると思われる。

謝辞　本研究は、文部科学省科学研究費補助金特定領域研究「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」（平成 18〜22 年度、領域代表者：前川喜久雄）による補助を得ています。

参考文献

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257.

柏野和佳子・丸山岳彦・稲益佐知子・田中弥生・秋元祐哉・佐野大樹・大矢内夢子・山崎誠. (2009). 『現代日本語書き言葉均衡コーパス』における収録テキストの抽出手順と事例. 特定領域研究「日本語コーパス」平成 20 年度研究成果報告書 (JC-D-08-01) 特定領域研究「日本語コーパス」データ班.

Leech, G. (1990). The value of a corpus in english language research: A reappraisal. ことばの饗宴 (pp. 115–126). くろしお出版.

丸山岳彦・秋元祐哉. (2007).『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法 –現代日本語書き言葉の文字数調査–. 特定領域研究「日本語コーパス」平成 18 年度研究成果報告書 (JC-D-06-02) 特定領域研究「日本語コーパス」データ班.

丸山岳彦・秋元祐哉. (2008). 『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法 (2) –コーパスの設計とサンプルの無作為抽出法–. 特定領域研究「日本語コーパス」平成 19 年度研究成果報告書 (JC-D-07-01) 特定領域研究「日本語コーパス」データ班.

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.

前川喜久雄. (2009). KOTONOHA『現代日本語書き言葉均衡コーパス』における著作権処理. 漢字文献情報処理研究, *10*, 30–35.

Copyright © 2010 The Association for Natural Language Processing.
All Rights Reserved.