第3章　サンプリング

丸山 岳彦　柏野 和佳子　田中 牧郎

3.1　BCCWJ 構築の基本理念

『現代日本語書き言葉均衡コーパス』（以下、BCCWJと略記する）を構築する上での基本理念は、次の4点にまとめられる（第2章参照）。

(1) 現代日本語の縮図となるコーパス
　これまで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るよう設計する。それにより、母集団における言語的諸特性の分布が縮図において過不足なく再現でき、母集団における分布を高い精度で推測できるようになる。
(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、辞書編集や言語政策、日本語教育などでも使えることを意図し、多様な日本語の姿を捉えることができるよう設計する。また、言語変化に対応するためには、同じ設計のコーパスを繰り返し構築するなど定点観測的な工夫も必要である。
(3) 公開可能なコーパス
　収録する著作物の利用許諾を得て、公開を目指す。インターネット上からの簡易検索のほか、共起条件を指定できる検索ツールなどもあわせて提供する。
(4) 既存のコーパスとの調和
　解析単位の仕様を『日本語話し言葉コーパス』に合わせ、短単位、長単位の2種類の解析を行う。

これらの基本理念のうち(1)と(2)は、コーパスの設計、およびサンプリングに関わる問題である。また、(3)は著作権処理、(4)は形態論情報の付与に関わる理念である。サンプリングに関わる問題のうち、(1)については、レジスターごとに母集団を厳密に定義して、層別ランダムサンプリングを実施することにより実現した。(2)については、サンプリングの際、固定長サンプル・可変長サンプルという2種類のサンプルを取得することにより、統計的な研究から文章研究までに対応できるサンプル抽出を実現した。

以下では、BCCWJの設計、およびサンプリング作業の概要について解説する。

3.2　BCCWJ を構成する三つのサブコーパス

まず、BCCWJの構成を、図3-1に示す。

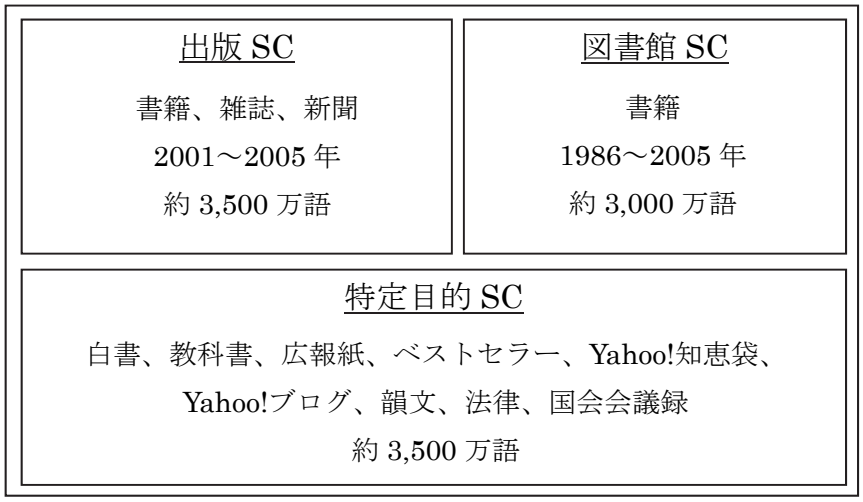

図3-1: BCCWJの構成

各サブコーパス（以下、SCと略記する）の概要を、以下に述べる。

3.2.1 出版（生産実態）SC

出版SCは、書き言葉の出版・生産という側面に着目するSC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・新聞に含まれる文字の総体を母集団として、ランダムサンプリングによって得られる約3,500万語分のデータを収める。書き言葉が実際に出版された結果を、文字数という量的側面からできる限り忠実に反映することで、5年間における書き言葉の出版に関するありさまを捉えることを目的とする。

3.2.2 図書館（流通実態）SC

図書館SCは、書き言葉の流通・流布の実態という側面に着目するSCである。東京都内の公立図書館に所蔵されている書籍（ただし1986年から2005年の20年間に出版されたもの）を対象として、ランダムサンプリングによって得られる約3,000万語分のデータを収める。書き言葉（書籍）が世の中に流通している状態を公立図書館の所蔵状況によって近似的に把握し、世の中に広く行き渡っている書き言葉のありさまを捉えることを目的とする。

3.2.3 特定目的 SC

特定目的SCは、生産・流通という側面からは捉えきれない、あるいは、出版SC・図書館SCの母集団には入らないけれども、書き言葉の研究を遂行する上で必要と思われる種類の書き言葉を収めるSCである。白書、教科書、広報紙、ベストセラー、Yahoo!知恵袋、Yahoo!ブログ、韻文、法律、国会会議録を対象として、約3,500万語分のデータを収める。収録対象期間はレジスターによって異なる。

3.3 BCCWJ を構成する 2 種類のサンプル

三つのSCは、「固定長サンプル」「可変長サンプル」という2種類のサンプルによって構成する。

- 固定長サンプルの設計方針：
  統計的に厳密な言語調査に耐え得る設計にする。
- 可変長サンプルの設計方針：
  文体研究・テキスト研究に耐え得るよう、ある程度の文脈を確保した設計にする。

3.3.1 固定長（FIXED）サンプル

「固定長サンプル」は、母集団に含まれる全ての文字に対して等確率を与えた上で、ある1文字をランダムに指定し（この1文字を「サンプル抽出基準点」（7.4.5節参照）と呼ぶ）、その文字を始点として1,000文字目までの範囲を抽出するサンプルである。全ての文字に対して等確率を与えるために、母集団に含まれる文字の総数をあらかじめ推計しておく必要がある。母集団（＝推計された総文字数）からの抽出比が明確である点で、基本語彙表や漢字表の作成、語彙・文字調査など、統計的な言語研究に向く。また、母集団の層的かつ量的な構造を忠実に反映する点で、統計的な代表性を備えた均衡コーパスとしての性格を強く持つ。

3.3.2 可変長（VARIABLE）サンプル

「可変長サンプル」は、固定長サンプルと同様、母集団に含まれる全ての文字に対して等確率を与えた上で、ある1文字をランダムに指定し、その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1万字を上限とする）を抽出するサンプルである。文章・談話としてのまとまりを重視したサンプルであるため、テキストの論理構造の把握や文脈の分析、文体の調査などに向く。

なお、可変長サンプルは、三つのSCの全てに対して提供される。一方、固定長サンプルは、統計的な言語調査を行う可能性の高いSC、すなわち、出版SC、図書館SC、および、特定目的SCの一部（白書）に対して提供される。

3.4 BCCWJ に収録するテキストの条件

BCCWJは現代日本語の書き言葉を収録したコーパスであるが、実際にサンプリング作業を実施するにあたり、「現代日本語書き言葉」をどのように定義すればよいか、という問題があった。そこで、「明治初年以降に」「日本語で」「書かれた」言葉を「現代日本語書き言葉」として定義し、これらの条件を満たすことをBCCWJに収録するテキストの条件とした。よって、江戸期以前に書かれた書き言葉は、基本的に（特定目的SC「教科書」レジスターの「国語」の一部を除いて）収録されていない。また、日本語の文章の中に外国

語が混在している場合は可能な限りそのまま収録しているが、例えばひとまとまりの英文が単独のパラグラフを構成している場合、その部分は収録対象から除外した。

3.5 BCCWJ-DVD 版に収録されているサンプルの一覧

BCCWJ-DVD版に収録されているサンプルの一覧を、表3-1に示す。なお、*が付与されているレジスターは、固定長サンプルと可変長サンプルの両方が、表3-1の「サンプル数」分それぞれ収録されている。*が付与されていないレジスターは、可変長サンプルのみが収録されている。

表3-1: 「BCCWJ-DVD版」に収録されているサンプルの一覧

SC	レジスター	対象期間	母集団	サンプル数
出版 SC（生産実態）	書籍 *	2001 年-2005 年	約 485 億文字	10,117
	雑誌 *	2001 年-2005 年	約 105 億文字	1,996
	新聞 *	2001 年-2005 年	約 64 億文字	1,473
図書館 SC（流通実態）	書籍 *	1986 年-2005 年	約 479 億文字	10,551
特定目的 SC	白書 *	1976 年-2005 年	1,006 冊	1,500
	教科書	2005 年-2007 年	145 冊	412
	広報紙	2008 年	100 自治体	354
	ベストセラー	1976 年-2005 年	951 冊	1,390
	Yahoo!知恵袋	2004 年-2005 年	約 312 万質問	91,445
	Yahoo!ブログ	2008 年-2009 年	約 346 万記事	52,680
	韻文	1980 年-2005 年	130 冊	252
	法律	1976 年-2005 年	718 法律	346
	国会会議録	1976 年-2005 年	32,925 会議	159

以下、3.6節では、BCCWJの構築において実施したサンプリング作業の方法について、各SCおよびレジスターごとに、概要を示す。なお、出版SC・図書館SCの設計の詳細については丸山・秋元（2007、2008）を、サンプリングの基準と実施手順の詳細については柏野他（2009）、丸山他（2011）を、それぞれ参照されたい。

3.6 サンプリング方法

以下では、BCCWJの構築において実施したサンプリング作業の方法について、各SC、およびレジスターごとに、その概要を示す。

3.6.1 出版 SC「書籍」

出版SC「書籍」は、2001年から2005年までの5年間に日本国内で出版されたすべての書籍を対象として、ランダムにサンプルを抽出したものである。

母集団の定義

- 国立国会図書館の書誌データ「J-BISC」を用いて、2001年から2005年までの5年間に出版された書籍を同定した。この際、漫画、写真集、電子資料、地図、学習試験図書、一般には流通しない官公庁刊行物、40ページ以下の書籍、ページ数の記録がない書籍などは除外した。その結果、5年間に出版された「書籍」は317,117冊、74,911,520ページという結果を得た。
- これらの書籍に印刷されている総文字数を推計した。「NDC（日本十進分類法）」および判型（本の高さ）の別にランダムに書籍を選び、そこからランダムに選んだページ内の文字数を実測した。合計227冊、1,135ページ分を実測した結果から1ページあたりの平均文字数を算出し、これを74,911,520ページに適用したところ、48,539,925,351文字という結果を得た。この総文字数を、出版SC「書籍」の母集団として定義した。

層別方法

- 上記で定義した母集団を、以下の二つの基準により、合計55層に層別した。
  - NDC（11層）： 国立国会図書館の蔵書目録「J-BISC」に書籍ごとに付与されているNDCの第1次区分（0～9）に、NDCが付与されていない「記録なし」を加えた、11分類。
  - **出版年（5層）**： 書籍の出版年である2001年から2005年までの、5分類。

サンプリング方法

- 母集団を55層に層別し、全体に対する各層の構成比率を取得サンプル数に比例割当した。
- 各層に含まれる全ページに対してランダムに優先順位を割り振った。優先順位の高い順に、指定された書籍の指定されたページに含まれる文章を一定の手続きにより抽出した。
- 収録した10,117サンプルについて、NDCごとの内訳を、図3-2に示す。

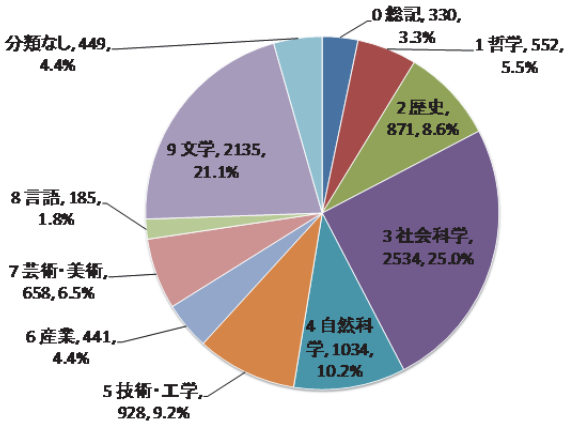

図3-2: サンプルの数と構成比率
（出版SC「書籍」、NDC別）

3.6.2 出版 SC「雑誌」

出版SC「雑誌」は、2001年から2005年までの5年間に日本国内で出版されたすべての雑誌を対象として、ランダムにサンプルを抽出したものである。

母集団の定義

- 『雑誌新聞総かたろぐ』（メディア・リサーチ・センター発行）を用いて、2001年から2005年の間に社団法人日本雑誌協会に加盟していた出版社が出版した定期刊行物を同定した。この際、新聞・通信、コミック、要覧、非日本語による定期刊行物は除外した。その結果、5年間に出版された「雑誌」は、1,259タイトル、55,779冊、10,414,955ページという結果を得た。
- これらの雑誌に印刷されている総文字数を推計した。『雑誌新聞総かたろぐ』のジャンルおよび判型の別にランダムに雑誌を選び、そこからランダムに選んだページ内の文字数を実測した。合計53冊、265ページ分の実測した結果から1ページあたりの平均文字数を算出し、これを10,414,955ページに適用したところ、10,515,681,636文字という結果を得た。この総文字数を、出版SC「雑誌」の母集団として定義した。

層別方法

- 上記で定義した母集団を、以下の二つの基準により、合計30層に層別した。
  - **ジャンル（6層）**：『雑誌新聞総かたろぐ』で分類されているジャンル（1. 総合、2. 教育・学芸、3. 政治・経済・商業、4. 産業、5. 工業、6. 厚生・医療）による6分類。
  - **出版年（5層）**：雑誌の出版年である2001年から2005年までの5分類。

サンプリング方法

- 母集団を30層に層別し、全体に対する各層の構成比率を取得サンプル数に比例割当した。
- 各層に含まれる全ページに対してランダムに優先順位を割り振った。優先順位の高い順に、指定された雑誌の指定されたページに含まれる文章を一定の手続きにより抽出した。なお、著作権処理の観点から、個人情報（一般人の氏名や住所、電話番号など）や出版社から要請のあった箇所に対して伏せ字処理を実施した。

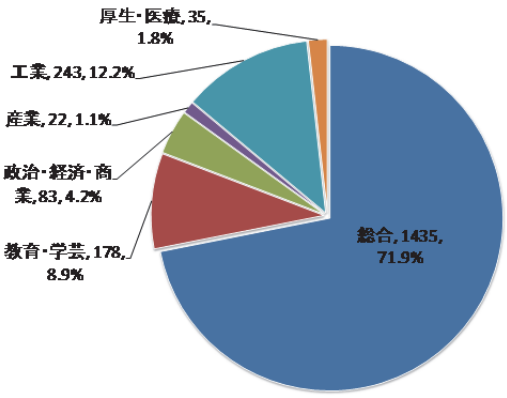

図3-3: サンプルの数と構成比率
（出版SC「雑誌」、ジャンル別）

- 収録した1,996サンプルについて、ジャンルごとの内訳を、図3-3に示す。

3.6.3 出版 SC「新聞」

出版SC「新聞」は、2001年から2005年までの5年間に日本国内で発行されたすべての新聞を対象として、ランダムにサンプルを抽出したものである。

母集団の定義

- 『全国新聞ガイド』（社団法人日本新聞協会発行）を用いて、「全国紙」「ブロック紙」および各地の有力な地方紙をリスト化した。この結果、全国紙（朝日新聞、毎日新聞、読売新聞、日本経済新聞、産経新聞）、ブロック紙（北海道新聞、中日新聞、西日本新聞）、地方紙（河北新報、新潟日報、京都新聞、神戸新聞、中国新聞、高知新聞、愛媛新聞、琉球新報）を同定した。
- 上記の新聞に関するページ数や発行回数などを調査した結果、5年間に発行された「新聞」は、16タイトル、合計49,625冊、1,198,189ページという結果を得た。
- これらの新聞に印刷されている総文字数を推計した。全国紙4紙の朝夕刊を合計8冊を、曜日を考慮してランダムに選び、そこに含まれている211ページに印刷されている全文字数を実測した。この結果から1ページ当たりの平均文字数を面種ごとに算出し、1,198,189ページに適用したところ、6,416,070,114文字という結果を得た。この総文字数を、出版SC「新聞」の母集団として定義した。

層別方法

- 上記で定義した母集団を、以下の二つの基準により、合計80層に層別した。
  - **新聞タイトル（16層）**： 新聞タイトルによる16分類。
  - **発行年（5層）**： 新聞の発行年である2001年から2005年までの5分類。

サンプリング方法

- 母集団を80層に層別し、全体に対する各層の構成比率を取得サンプル数に比例割当した。
- 各層に含まれる全ページに対してランダムに優先順位を割り振った。優先順位の高い順に、指定された新聞の指定されたページに含まれる文章を一定の手続きにより抽出した。
- 収録した1,473サンプルについて、新聞タイトルごとの内訳を、図3-4に示す。

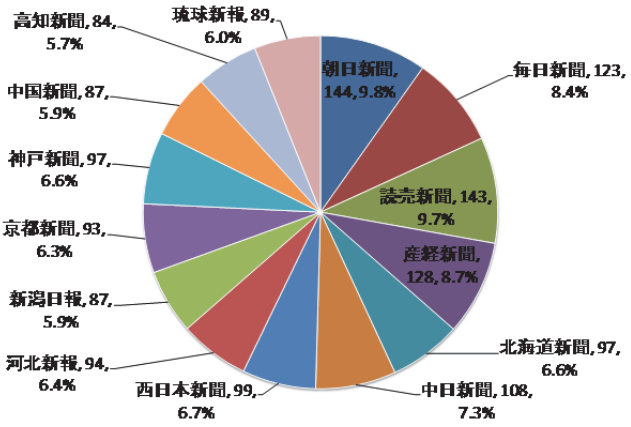

図3-4: サンプルの数と構成比率
（出版SC「新聞」、タイトル別）

3.6.4 図書館 SC「書籍」

図書館SC「書籍」は、1986年から2005年までの20年間に出版された書籍のうち、東京都内の公立図書館に所蔵されている書籍を対象として、ランダムにサンプルを抽出したものである。

母集団の定義

- 東京都立中央図書館作成の「ISBN総合目録」を用いて、東京都内の区市町村立図書館が所蔵する蔵書リストを作成した。
- 集計の結果、東京都内の13自治体以上で共通に所蔵されている335,721冊、85,363,019ページを対象とすると、推計総文字数が47,877,656,072文字となり、出版SC「書籍」の母集団とほぼ等しくなることが判明した。この総文字数を、図書館SC「書籍」の母集団として定義した。

層別方法

- 上記で定義した母集団を、以下の二つの基準により、合計220層に層別した。
  - NDC（11層）： 国立国会図書館の蔵書目録「J-BISC」に書籍ごとに付与されているNDCの第1次区分（0～9）に、NDCが付与されていない「記録なし」を加えた、11分類。
  - **出版年（20層）**： 書籍の出版年である1986年から2005年までの20分類。

サンプリング方法

- 母集団を220層に層別し、全体に対する各層の構成比率を取得サンプル数に比例割当した。
- 各層に含まれる全ページに対してランダムに優先順位を割り振った。優先順位の高い順に、指定された書籍の指定されたページに含まれる文章を一定の手続きにより抽出した。
- 収録した10,551サンプルについて、NDCごとの内訳を、図3-5に示す。

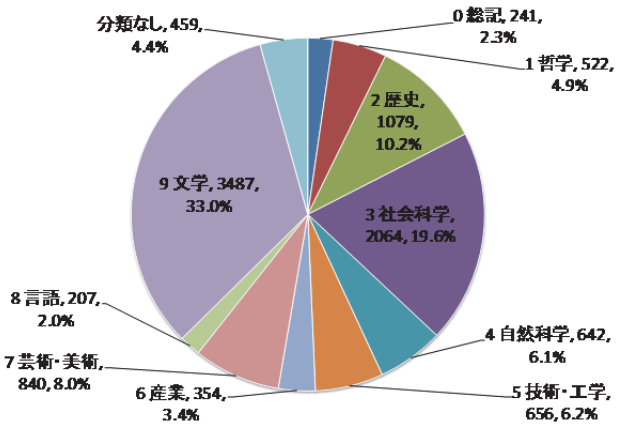

図3-5: サンプルの数と構成比率
（図書館SC「書籍」、NDC別）

3.6.5 特定目的 SC「白書」

特定目的SC「白書」は、1976年から2005年までの30年間に発行された政府系刊行物「白書」を対象として、ランダムにサンプルを抽出したものである。

母集団の定義

- 2001年から2005年までに発行された白書のうち、『官報』に記載のあった白書タイトルを抽出した。これらについて、1976年以降、タイトルの変更や合併などの変遷を調査した。30年間にタイトルの変更や合併などがあったものは、まとめて扱った。例えば『土地白書』は、1989年以前は『国土利用白書』という別タイトルだったが、これは『土地白書（国土利用白書）』という1タイトルにまとめた。この結果、合計で40タイトル、1,006冊の白書が同定された。これらを特定目的SC「白書」の母集団として定義した。

層別方法

- 上記で定義した母集団を、以下の二つの基準により、合計54層に層別した。
  - **ジャンル（9層）**： 白書の内容に基づいて設定した、「安全」「外交」「科学技術」「環境」「教育」「経済」「国土交通」「農林水産」「福祉」という9分類。
  - **発行年（6層）**： 白書の発行年（1976年～2005年）の30年間を5年刻みにした、6分類。

    第1期：1976～1980年、第2期：1981～1985年、第3期：1986～1990年、
    第4期：1991～1995年、第5期：1996～2000年、第6期：2001～2005年

サンプリング方法

- 第1期から第6期のそれぞれから250サンプルずつ、全体で1,500サンプル（約500万語）の取得を計画した。40タイトルごとに総ページ数を集計し、1,500サンプルに比例割当して、各期・各タイトルから取得するサンプル数を算出した。
- 各層に含まれる全ページに対してランダムに優先順位を割り振った。優先順位の高い順に、指定された白書の指定されたページに含まれる文章を一定の手続きにより抽出した。
- 収録した1,500サンプルについて、ジャンルごとの内訳を、図3-6に示す。

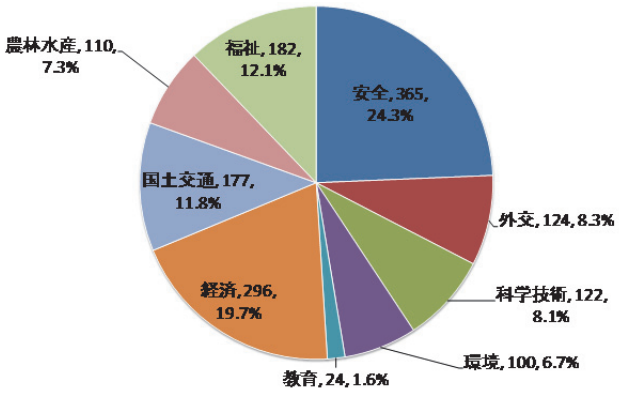

図3-6: サンプルの数と構成比率
（特定目的SC「白書」、ジャンル別）

3.6.6 特定目的 SC「教科書」

特定目的SC「教科書」は、小学校・中学校・高等学校で採用された各教科の教科書から、ランダムにサンプルを抽出したものである。

母集団の定義

- 小学校・中学校・高等学校の各学習指導要領（平成10～11年文部省告示、平成15年一部改正）に基づき、2005年度から2007年度に実際に使用された検定教科書を対象とした。ただし、専門に分化した高等学校の一部の科目（「農業」「商業」など）は除外した。
- 各校種・各学年・各教科から1種ずつの教科書を選出した。その際、できるだけ発行部数の多い教科書から順に選出した。この結果、145冊の教科書（推計総文字数7,859,456文字）が同定された。これらを、特定目的SC「教科書」の母集団として定義した。

層別方法

- 以下の二つの基準により、母集団を合計25層に層別した。
  - **教科（10層）**：「国語」「数学」「理科」「社会」「外国語」「技術家庭」「芸術」「保健体育」「情報」「生活」の10分類。
  - **校種（3層）**：「小学校」「中学校」「高等学校」の3分類。

サンプリング方法

- 母集団を25層に層別し、全体に対する各層の構成比率を取得サンプル数に比例割当した。
- 各層に含まれる全ページに対してランダムに優先順位を割り振った。優先順位の高い順に、指定された教科書の指定されたページに含まれる文章を一定の手続きにより抽出した（ただし、教科書であることを考慮し、書籍等の基準とは一部異なっているところがある）。
- 収録した412サンプルについて、教科ごとの内訳を、図3-7に示す。

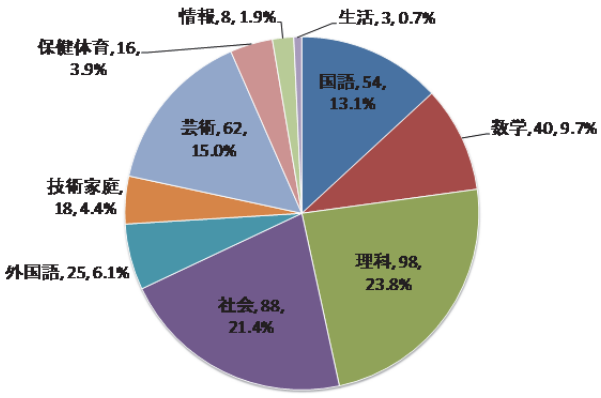

図3-7: サンプルの数と構成比率
（特定目的SC「教科書」、教科別）

3.6.7 特定目的 SC「広報紙」

特定目的SC「広報紙」は、日本の地方自治体において発行されている「広報紙」から、ランダムにサンプルを抽出したものである。

母集団の定義

- 全国各地から地域や人口構成比などを考慮して100の自治体（区市町村）を抽出し、その100自治体で2008年度に発行された広報紙を母集団として定義した。

層別方法

- 広報紙が発行している自治体の地域に応じて、母集団を合計8層に層別した。
  - **地域（8層）**： 北海道地方、東北地方、関東地方、中部地方、近畿地方、中国地方、四国地方、九州・沖縄地方

サンプリング方法

- 1自治体から6万字程度を取得することにした。入手した各自治体の広報紙からランダムに1冊（1号）を選び、そこに含まれる全文をサンプルとして取得した。
- また、著作権処理の観点から、外部著者による「寄稿」や、個人情報（一般人の氏名や住所、電話番号など）に相当する部分は伏せ字処理を実施した。
- 各自治体で6万字程度が取得できるまで、冊の取得を繰り返した結果、354サンプルを取得した。地域ごとの内訳を、図3-8に示す。

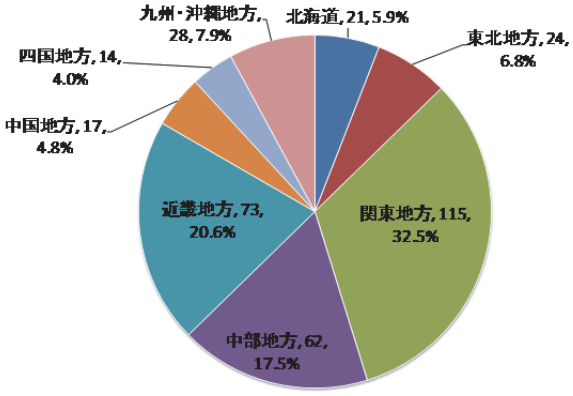

図3-8: サンプルの数と構成比率
（特定目的SC「広報紙」、地域別）

3.6.8 特定目的 SC「ベストセラー」

特定目的SC「ベストセラー」は、1976年から2005年までの30年間にベストセラーとなった書籍を対象として、ランダムにサンプルを抽出したものである。

母集団の定義

- 1976年から2005年までの30年間において、『出版年鑑』（出版ニュース社）および『出版指標年報』（全国出版協会出版科学研究所）のどちらかに、各年のベストセラーとして上位20位までに挙げられた書籍を調査した。その結果、951冊が同定された。これ

らを特定目的SC「ベストセラー」の母集団として定義した。

- なお、1971年に出版された本が1976年のベストセラーになったなど、出版年とベストセラーになった年との間にずれがあるものがある。

層別方法

- 「ベストセラー」という性格上、層別は実施しなかった。

サンプリング方法

- 1冊からランダムに2サンプルずつを取得することにした。
- 各冊に含まれる全ページに対して、ランダムに優先順位を割り振った。優先順位の高い順に、指定された書籍の指定されたページを開け、そこに印刷されている文章を一定の手続きにより抽出した。
- 951冊からは、合計1,902サンプルが取得できることになるが、作業上の理由（サンプリングできる箇所がない、当該の書籍が入手できないなど）により、すべてが取得できたわけではない。
- 収録した1,390サンプルについて、NDCごとの内訳を図3-9に示す。

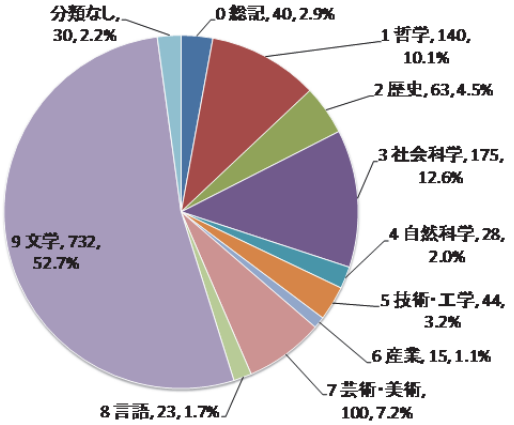

図3-9: サンプルの数と構成比率
（特定目的SC「ベストセラー」、NDC別）

3.6.9 特定目的 SC「Yahoo!知恵袋」

特定目的SC「Yahoo!知恵袋」は、Q&A形式のナレッジコミュニティサービス「Yahoo!知恵袋」の投稿データからランダムにサンプルを抽出したものである。

母集団の定義

- 「Yahoo!知恵袋」の元データには、2004年10月から2005年10月にかけて投稿された3,120,839の質問と、それに対する複数の回答が含まれていた。これらを、特定目的SC「Yahoo!知恵袋」の母集団として定義した。

層別方法

- 「Yahoo!知恵袋」の質問は、その質問内容に応じて、ある「カテゴリ」に分類されている。カテゴリは、15個の大カテゴリ・82個の中カテゴリ・279個の小カテゴリという

3階層に分かれている。このうち、小カテゴリによって、母集団を合計279の層に層別した。

サンプリング方法

- 母集団から、ひとつの質問とそれに対するひとつの回答の組を抽出して1サンプルとすることにした。複数の回答がある場合、「ベストアンサー」と呼ばれる回答を利用した。
- 全体で約1,000万語分のサンプルを取得することとし、1サンプルの平均長を試算して、対象データ全体から91,450サンプルを取得することを計画した。
- 279の各層に含まれる質問数を集計し、91,450サンプルに比例割当して、各小カテゴリから取得するサンプル数を算出した。この結果、取得対象となるのは14個の大カテゴリ、59個の中カテゴリ、130個の小カテゴリとなった。
- 各小カテゴリに含まれる質問から必要数をランダムに取得し、その質問に対する回答も同時に取得して、全体で91,445サンプルを取得した。大カテゴリごとの内訳を、図3-10に示す。

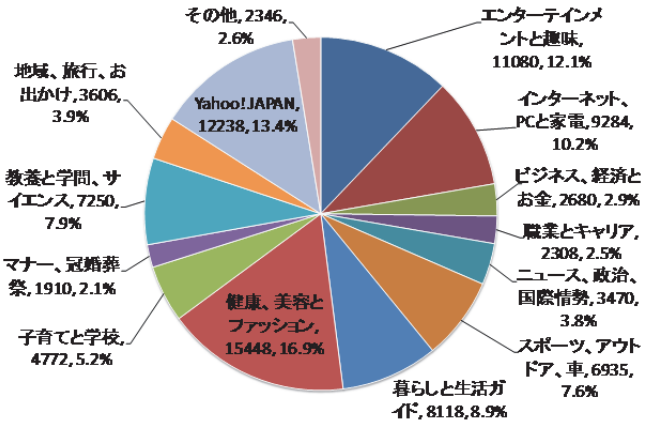

図3-10: サンプルの数と構成比率

(特定目的SC「Yahoo!知恵袋」、大カテゴリ別)

3.6.10 特定目的 SC「Yahoo!ブログ」

特定目的SC「Yahoo!ブログ」は、「Yahoo!ブログ」の記事データからランダムにサンプルを抽出したものである。

母集団の定義

- 「Yahoo!ブログ」の元データには、合計3,463,413の記事(ただし、以下の条件を満たすもの)が含まれていた。これらを、特定目的SC「Yahoo!ブログ」の母集団として定義した。
  1. 2008年4月26日から2009年4月25日までに投稿された記事。
  2. 抽出時点で1,000記事以上あるブログからの記事。
  3. 抽出時点で1か月以上掲載されており、かつ「公開」モードである記事。
  4. 転載(Yahoo!ブログ内のほかの記事の内容をコピーして、自分のブログに掲載す

ること）による記事は除外する。

5. ひとつの記事が全角20文字以下のものは除外する。

層別方法

- 「Yahoo!ブログ」の記事は、その内容に応じて、ある「カテゴリ」に分類される。カテゴリは、15個の大カテゴリ・54個の中カテゴリ・316個の小カテゴリという3階層に分かれているが、事前の層別には用いなかった。

サンプリング方法

- 全体で約1,000万語分のサンプルを取得することとした。サンプルは、記事タイトルやトラックバックを含まない、記事本文として記述されたテキストのみで構成するものとした。
- 対象データ全体を、投稿日時によって記事ごとに並び替え、等間隔サンプリングによって全体の1.8%を抽出した。
- ここから広告のみからなる記事などを除外した。結果、「ブログ」として、52,680サンプルを取得した。大カテゴリごとの内訳を、図3-11に示す。

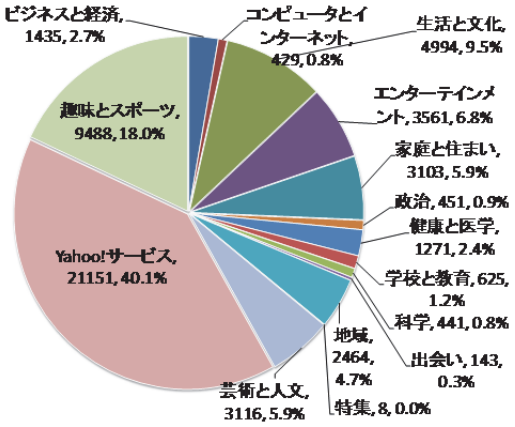

図3-11: サンプルの数と構成比率

（特定目的SC「Yahoo!ブログ」、大カテゴリ別）

3.6.11 特定目的 SC「韻文」

特定目的SC「韻文」は、短歌・俳句・詩の3種類について、代表的な作品からサンプルを抽出したものである。

母集団の定義

- 以下の作品を母集団として定義した。
  - 短歌: 『現代短歌全集』（筑摩書房、2002年刊） 第14巻～第17巻
  - 俳句: 『増補現代俳句大系』（角川書店、1980年～1982年刊） 第8巻～第15巻
  - 詩: 「現代詩文庫」シリーズ（思潮社、1986年～2005年刊） 118冊

層別方法

- 「短歌」「俳句」「詩」という3種類によって層別した。

サンプリング方法

- 短歌・俳句・詩からそれぞれ約5万語ずつを取得することとし、各歌集・句集・詩集からほぼ等量ずつのサンプルを抽出した。
- 収録した252サンプルについて、内訳を図3-12に示す。

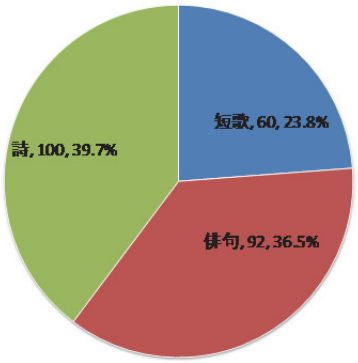

図3-12: サンプルの数と構成比率
（特定目的SC「韻文」）

3.6.12 特定目的 SC「法律」

特定目的SC「法律」は、1976年から2005年までの30年間に公布され、2009年時点でも施行されているすべての法律を対象として、そこからランダムにサンプルを抽出したものである。

母集団の定義

- Web上の「法令データ提供システム」（http://law.e-gov.go.jp/）から、1976年から2005年までの間に公布され、2009年9月の時点でも施行されている法律を検索したところ、718法律を得た。これらを特定目的SC「法律」の母集団として定義した。

層別方法

- 公布年により、母集団を合計6層に層別した。
  - **公布年（6層）**： 1976年から2005年までの30年間を5年刻みにした6分類。
    第1期：1976～1980年、第2期：1981～1985年、第3期：1986～1990年、
    第4期：1991～1995年、第5期：1996～2000年、第6期：2001～2005年

サンプリング方法

- 第1期から第6期のそれぞれから30万文字ずつを取得し、約100万語分のサンプルを取得した。
- 各層に含まれる全法律に対して、それぞれ200箇所の文字を優先順位付きでランダムに選び、その文字を基準にして1万字を超えない一定範囲（条、節など）を取得した。その際、公布時以降に付け加えられた「附則」は取得の対象外とした。
- 収録した346サンプルについて、ジャンルごとの内訳を表3-2に示す。

表3-2: 取得したサンプルの数

（特定目的SC「法律」、ジャンル別）

憲法	2	国土開発	5	文化	2	航空	1
国会	3	土地	1	産業通則	18	貨物運送	3
行政組織	22	都市計画	7	農業	11	郵務	4
国家公務員	3	道路	1	林業	5	電気通信	13
行政手続	1	災害対策	6	水産業	3	労働	9
地方自治	4	建築・住宅	8	鉱業	2	環境保全	12
地方財政	1	財務通則	4	工業	10	厚生	17
司法	5	国税	18	商業	13	社会福祉	15
民事	36	専売・事業	4	金融・保険	40	防衛	1
刑事	7	国債	3	陸運	11	外事	6
警察	4	教育	3	海運	4	合計	346

3.6.13 特定目的 SC「国会会議録」

特定目的SC「国会会議録」は、1976年から2005年までの30年間における「国会会議録」からランダムにサンプルを抽出したものである。

母集団の定義

- Web上の「国会会議録検索システム」（http://kokkai.ndl.go.jp/）で公開されているデータのうち、第77回国会から第163回国会までに開かれた32,986会議の会議録データを特定目的SC「国会会議録」の母集団とした。
- このうち、「両院協議会」で開かれた61会議、発言部分の文字数が1,000文字以下の6,401会議、第77回国会のうち1975年に開催された33会議は除外した。

層別方法

- 以下の三つの基準により、母集団を合計48層に層別した。
  - **開催院（2層）**： 「衆議院」「参議院」による、2分類。
  - **開催時期（6層）**： 1976年から2005年までを5年刻みにした6分類。
    第1期：1976～1980年、第2期：1981～1985年、第3期：1986～1990年
    第4期：1991～1995年、第5期：1996～2000年、第6期：2001～2005年
  - **会議種別（4層）**： 「常任委員会」「特別委員会」「本会議」「その他」による4分類。

サンプリング方法

- 全体で約500万語分のサンプルを取得することを計画した。1サンプルは、1会議に含まれる発言部分のみで構成することにした。
- 48の各層に含まれる発言文字数を集計し、各層から取得するサンプル数を比例割当により算出した。各層に含まれる会議から必要数をランダムに取得し、全体で159サンプルを取得した。
- 収録した159サンプルについて、開催院・会議種別ごとのサンプル数と構成比率を図3-13に示す。

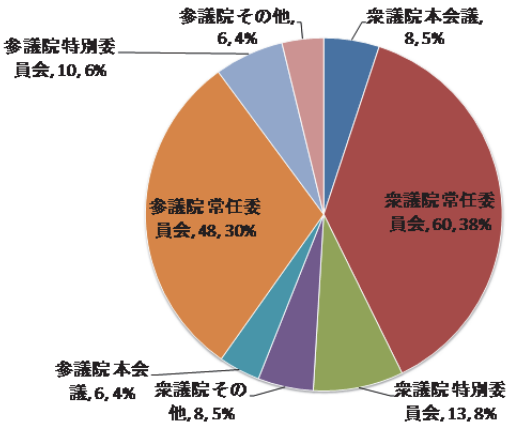

図3-13:サンプルの数と構成比率

（特定目的SC「国会会議録」、開催院・会議種別）

参考文献

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」国立国語研究所内部報告書 LR-CCG-08-01.

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－」国立国語研究所内部報告書 LR-CCG-06-02.

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－」国立国語研究所内部報告書 LR-CCG-07-01.

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　(2011)「『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.