国立国語研究所学術情報リポジトリ

『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

メタデータ	言語: Japanese
	出版者:
	公開日: 2020-06-29
	キーワード (Ja):
	キーワード (En):
	作成者: 丸山, 岳彦, 秋元, 祐哉
	メールアドレス:
	所属:
URL	https://doi.org/10.15084/00002840

国立国語研究所内部報告書（LR-CCG-07-01）

『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―

丸山 岳彦・秋元 祐哉

平成20年3月
大規模汎用日本語データベースの構築とその活用に関する調査研究
©2008 独立行政法人国立国語研究所

国立国語研究所内部報告書（LR-CCG-07-01）

『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－

丸山　岳彦
秋元　祐哉

平成20年3月
大規模汎用日本語データベースの構築とその活用に関する調査研究

©2008　独立行政法人国立国語研究所

目 次

第 I 部　本編 **3**

第 1 章　「出版サブコーパス」「図書館サブコーパス」の設計 **5**
1.1　BCCWJ の全体構成（修正版） 5
1.2　出版サブコーパスの設計とサンプリングの手順 7
1.3　図書館サブコーパスの設計とサンプリングの手順 10
1.4　本報告書の目的 10

第 2 章　「図書館サブコーパス」におけるサンプル構成比の算出法 **13**
2.1　図書館サブコーパスの設計 13
2.2　書籍の発行期間に関する修正 15
2.3　東京都内公立図書館の共通蔵書調査 18
2.4　図書館サブコーパスの母集団の定義 20
2.5　サンプル構成比の算出 22

第 3 章　サンプル台帳の作成とサンプルの無作為抽出 **25**
3.1　文字を基準とするサンプリングの方針 25
3.2　サンプル台帳の作成 26
3.3　固定長サンプル・可変長サンプルの抽出 29
3.3.1　書籍の「形態」を構成する要素 29
3.3.2　書籍の「版面」を構成する要素 30
3.3.3　書籍の「本文」を構成する要素 31
3.3.4　書籍の「文字」を構成する要素 31
3.3.5　収録する言語表現の条件 32
3.4　「原サンプル」の作成 33

第 II 部　資料編 **37**

第 4 章　資料集 **39**

表 目 次

1.1 出版サブコーパス全体の推計総文字数と取得サンプル数 8
1.2 出版サブコーパス全体のサンプル構成比 9

2.1 書籍の出版点数の多い出版社（上位 20 社） 17
2.2 都内 52 自治体が所蔵する書籍数 18
2.3 都内 52 自治体における共通蔵書の分布 19
2.4 都内公立図書館 共通蔵書の範囲と冊数 21
2.5 都内公立図書館 共通蔵書の範囲とページ数 21
2.6 都内公立図書館 共通蔵書の範囲と推計総文字数 22
2.7 出版サブコーパスと図書館サブコーパスの母集団 22
2.8 図書館サブコーパス全体のサンプル構成比 23

4.1 書籍の総文字数推計基準表 39
4.2 出版サブコーパスの母集団（発行年・NDC による層別）その 1 40
4.3 出版サブコーパスの母集団（発行年・NDC による層別）その 2 41
4.4 出版サブコーパスの母集団（発行年・NDC による層別）その 3 42
4.5 出版サブコーパスの母集団（発行年・NDC による層別）その 4 43
4.6 出版サブコーパスの母集団（発行年・NDC による層別）その 5 44
4.7 図書館サブコーパスの母集団（発行年・NDC による層別）その 1 45
4.8 図書館サブコーパスの母集団（発行年・NDC による層別）その 2 46
4.9 図書館サブコーパスの母集団（発行年・NDC による層別）その 3 47
4.10 図書館サブコーパスの母集団（発行年・NDC による層別）その 4 48
4.11 図書館サブコーパスの母集団（発行年・NDC による層別）その 5 49
4.12 図書館サブコーパスの母集団（発行年・NDC による層別）その 6 50
4.13 図書館サブコーパスの母集団（発行年・NDC による層別）その 7 51

図 目 次

1.1 BCCWJ の構成 5
1.2 出版サブコーパスの設計とサンプリングの手順 7
1.3 NDC 別のサンプル構成比（出版サブコーパス） 8
1.4 図書館サブコーパスの設計とサンプリングの手順 10

2.1 出版サブコーパスと図書館サブコーパスの設計 15
2.2 ISBN が付与されている冊数と付与されていない冊数 16
2.3 「J-BISC」における ISBN の付与率の推移 16
2.4 主要出版社による ISBN の付与率の推移 17
2.5 東京都内 52 自治体における共通蔵書の分布 19
2.6 NDC 別のサンプル構成比（図書館サブコーパス） 24

3.1 「サンプル抽出基準点」の指定とサンプルの抽出 26
3.2 サンプル台帳の例 27
3.3 サンプル台帳で指定されたページ・座標から 1 文字を特定する例 28
3.4 書籍の形態に関する分類 30
3.5 書籍の版面に関する分類 30
3.6 本文の構成に関する分類 31
3.7 原サンプルの例 33

はじめに

本報告書は,『現代日本語書き言葉均衡コーパス』に含まれる 3 つのサブコーパスのうち,「出版サブコーパス」「図書館サブコーパス」の設計，およびサンプルの無作為抽出法について示すものである。2007 年度に発行した次の研究成果報告書（以下,「前報告書」と呼ぶ）の続編に当たる。

> 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）

『現代日本語書き言葉均衡コーパス』は，1976 年から 2005 年の 30 年間に生産された現代日本語の書き言葉を収録する，1 億語規模のコーパスである。このコーパスの設計にあたり，我々は，いくつかの基礎調査を行った。前報告書では,「出版サブコーパス」の設計に際して実施した「現代日本語書き言葉の文字数調査」について報告した。

本報告書ではまず,『現代日本語書き言葉均衡コーパス』の全体構成，および「出版サブコーパス」「図書館サブコーパス」の設計とサンプリングの手順について確認する。その後,「図書館サブコーパス」の設計に際して実施した「東京都内公立図書館の共通蔵書調査」について報告する。この調査では，1986 年から 2005 年までに国内で発行された書籍のうち，東京都内の公立図書館で所蔵されている書籍の状況を調査した。この分布状況を基礎データとして,「図書館サブコーパス」の母集団を定義し，サンプル構成比を算出した。

本報告書の後半では，一連の基礎調査によって定義した母集団およびサンプル構成比に基づき，サンプリング用の台帳を作成し，母集団に含まれる全ての文字から特定の 1 文字を指定して 2 種類のサンプルを抽出するまでの手続きについて示す。これにより，出版サブコーパス・図書館サブコーパスの母集団を定義し，そこからサンプル台帳に基づいてランダムにサンプルを抽出するまでの，一連の手順が示されることになる。

「東京都内公立図書館の共通蔵書調査」は，2007 年 4 月から 5 月にかけて，丸山岳彦が中心となり実施した。調査結果の集計は，丸山岳彦と秋元祐哉の 2 名が担当した。また，サンプル台帳を作成し，サンプルをランダムに抽出する手順については，2006 年から 2007 年にかけて，サンプリング班（山崎誠，柏野和佳子，丸山岳彦，佐野大樹，秋元祐哉，稲益佐知子，田中弥生）で継続して検討を行った。本報告書の執筆は，丸山岳彦と秋元祐哉の 2 名が担当した。

謝辞：「東京都内公立図書館の共通蔵書調査」の実施にあたり，東京都立中央図書館方より，「東京都公立図書館調査」に関する調査データ（「ISBN 総合目録」）の利用許諾をいただきました。東京都立図書館の御厚情に対し，記して深く感謝の意を表します。

第Ⅰ部

本編

第1章　「出版サブコーパス」「図書館サブコーパス」の設計

本章では，本報告書の前提として,『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）』の構成について確認しておく。その上で，BCCWJ を構成する 3 つのサブコーパスのうち，出版サブコーパス・図書館サブコーパスの設計（コーパスデザイン）とサンプリングの実施手順について述べる。

1.1　BCCWJ の全体構成（修正版）

はじめに，BCCWJ の全体構成を確認しておく。全体構成については前報告書でも示したが，その後いくつかの修正が施されたので，その点も含めて改めて示す。

BCCWJ 全体を構成する 3 つのサブコーパスは，これまで「生産実態サブコーパス」「流通実態サブコーパス」「非母集団サブコーパス」と呼ばれてきたが，それぞれの特性を踏まえて，「出版サブコーパス」「図書館サブコーパス」「特定目的サブコーパス」という別称を設けることにした。BCCWJ の全体構成を図示すると，図 1.1 のようになる。各サブコーパスの概要を，以下に述べる。

<table>
<tr><td>出版サブコーパス　（生産実態）
書籍、雑誌、新聞
約3,500万語　2001年－2005年
固定長サンプル ＋ 可変長サンプル</td><td>図書館サブコーパス　（流通実態）
書籍
約3,000万語　1986年－2005年
固定長サンプル ＋ 可変長サンプル</td></tr>
<tr><td colspan="2">特定目的サブコーパス（非母集団）
白書、国会会議録、Web文書（Yahoo! 知恵袋）、ベストセラー、教科書など
約3,500万語　1976年－2005年
（固定長サンプル ＋ ）可変長サンプル</td></tr>
</table>

図 1.1: BCCWJ の構成

出版サブコーパス　出版サブコーパスは，書き言葉の生産力という側面に着目するサブコーパスである。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生産に関するありさまを捉えることを目的とする。

なお，5 年間に出版されたすべての書籍・雑誌・新聞に含まれる総文字数を推計した「現代日本語書き言葉の文字数調査」の手順と結果については，前報告書で報告した。そこで得られたメディアごとの推計総文字数の比を用いて，各層のサンプル構成比を算出し，層別ランダムサンプリングを実施する。

図書館サブコーパス 図書館サブコーパスは，書き言葉の流通・流布の実態という側面に着目するサブコーパスである。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 年間に発行されたもの）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度合いを公立図書館の所蔵状況によって近似的に把握し，世の中に広く流布している書き言葉のありさまを捉えることを目的とする。

なお，前報告書では書籍の発行期間に関する条件を「1976 年から 2005 年の 30 年間に発行されたもの」としていたが，この条件を「1986 年から 2005 年の 20 年間に発行されたもの」と修正した。この設計変更の経緯については，2.2 節で述べる。

特定目的サブコーパス 特定目的サブコーパスは，生産・流通という側面からは捉えきれない，あるいは，出版サブコーパス・図書館サブコーパス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サブコーパスである。白書，ベストセラー，国会会議録，教科書，WWW 上の文書（Yahoo!知恵袋，ブログ），韻文（俳句・短歌・詩）などを対象として，約 3,500 万語分のデータを収める。収録対象期間はメディアによって異なるが，最長で 1976 年から 2005 年までの 30 年間である。

これら 3 つのサブコーパスに対して,「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。これは，それぞれ以下の 2 つの方針を満たすための設計である。

- 統計的に厳密な言語調査に耐え得るよう，母集団からの抽出比を重視した設計にする。
- 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。

固定長サンプル 「固定長サンプル」は，母集団に含まれるすべ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，その文字を始点として 1,000 文字の範囲を抽出するサンプルである。すべ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バランストコーパスとしての性格を強く持つ。

可変長サンプル 「可変長サンプル」は，固定長サンプルと同様，母集団に含まれるすべての文字に対して等確率を与えた上で，ランダムに指定した 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 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。

可変長サンプルは，3 つのサブコーパスのすべ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サブコーパス，すなわち，出版サブコーパス，図書館サブコーパス，および，特定目的サブコーパスの一部（白書など）に対して，可変長サンプルと同時に提供される。

1.2　出版サブコーパスの設計とサンプリングの手順

以下では，出版サブコーパス・図書館サブコーパスの設計（コーパスデザイン）とサンプリングの手順について述べる。

出版サブコーパスの設計とサンプリングは，以下のような手順で行う。

1. 2001 年から 2005 年までに発行されたすべての書籍・雑誌・新聞を調査し，出版サブコーパスの母集団となる対象を定める。
2. 上記の書籍・雑誌・新聞を，ジャンル（日本十進分類法（NDC），全国紙・ブロック紙・地方紙など）と発行年によって層別し，各層に含まれる総文字数を推計する。推計された文字の総体を母集団として定義する。
3. 推計総文字数の比から，各層のサンプル構成比を算出する。全体で 1,000 万語分の固定長サンプルを取得することを前提として，各層から必要となるサンプル数を比例割当により算出する。
4. 母集団を構成するすべての文字に等確率を与えた上で，ランダムに 1 文字を選び出し，この文字を基準点として，固定長サンプル・可変長サンプルを抽出する。これを各層ごとに必要なサンプル数分くり返す。

図 1.2: 出版サブコーパスの設計とサンプリングの手順

前報告書では，総文字数を推計するために実施した「現代日本語書き言葉の文字数調査」を中心に，上記の 1.〜3. について報告した。調査から得られた書籍・雑誌・新聞の推計総文字数，構成比，取得サンプル数および取得（見込み）語数を，表 1.1 に示す。また，サブコーパス全体のサンプル構成比を図 1.3 に，ジャンルごとの詳細なサンプル構成比を表 1.2 に示す。

参照 出版サブコーパスの母集団全体の冊数・ページ数・推計総文字数を発行年と NDC で層別した一覧については，資料編の表 4.2〜表 4.6 を参照（40〜44 ページ）。

表 1.1: 出版サブコーパス全体の推計総文字数と取得サンプル数

メディア	推計総文字数	構成比	固定長 合計語数	サンプル数	可変長 合計語数
書籍	485.40 億文字	74.14%	741.4 万語	12,604 サンプル	2,891.5 万語
雑誌	105.16 億文字	16.06%	160.6 万語	2,730 サンプル	481.8 万語
新聞	64.16 億文字	9.80%	98.0 万語	1,666 サンプル	98.0 万語
合計	654.72 億文字	100.00%	1,000 万語	17,000 サンプル	3,471.3 万語

推計総文字数の比を 1,000 万語分の固定長サンプルに比例割当することにより，書籍からは 741.4 万語，雑誌からは 160.6 万語，新聞からは 98.0 万語を取得すればよいと算出される。これによって，出版された文字の総体によって定義された母集団の様相を忠実に反映する，統計的な代表性を備えたコーパスを実現する。これに必要なサンプル数は（1 語の平均長を 1.7 文字として）17,000 サンプルであり，書籍は 12,604 サンプル，雑誌は 2,730 サンプル，新聞は 1,666 サンプルとなる。また，固定長サンプルと同時に可変長サンプルも取得することにより，出版サブコーパス全体では約 3,500 万語の可変長サンプルが取得される[1]。

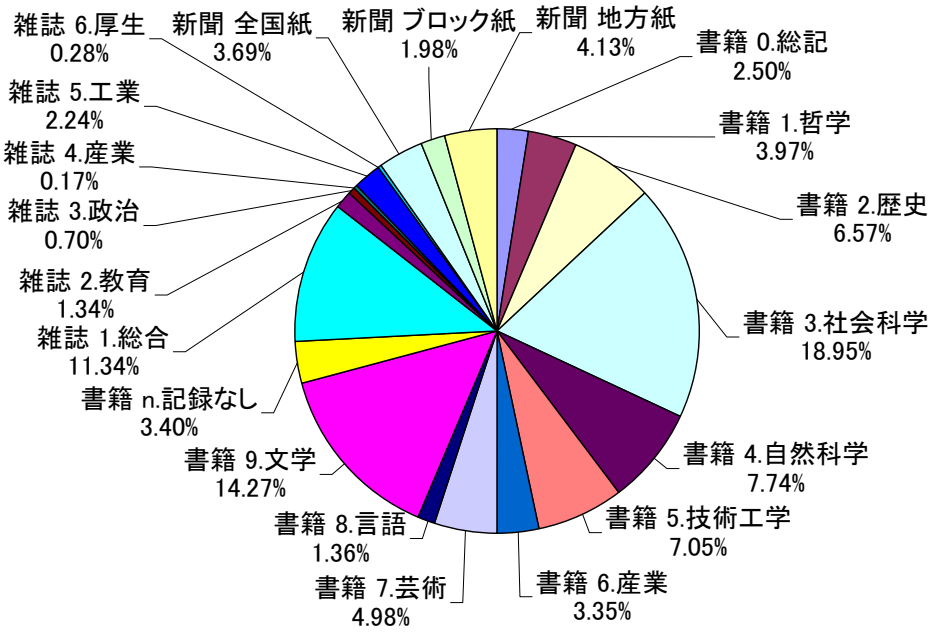


図 1.3: NDC 別のサンプル構成比（出版サブコーパス）

[1] 可変長サンプルの平均長は，書籍が 3,900 文字，雑誌が 3,000 文字，新聞が 1,000 文字と試算している。

表 1.2: 出版サブコーパス全体のサンプル構成比

メディア	ジャンル	推計総文字数	構成比	サンプル数
書籍	0. 総記	1,636,414,548	2.50%	425
	1. 哲学	2,597,610,813	3.97%	674
	2. 歴史	4,301,204,340	6.57%	1,117
	3. 社会科学	12,408,321,943	18.95%	3,222
	4. 自然科学	5,069,594,034	7.74%	1,316
	5. 技術工学	4,615,929,967	7.05%	1,199
	6. 産業	2,196,387,437	3.35%	570
	7. 芸術	3,258,432,447	4.98%	846
	8. 言語	888,800,128	1.36%	231
	9. 文学	9,341,275,486	14.27%	2,426
	n. 記録なし	2,225,954,208	3.40%	578
書籍 小計		48,539,925,351	74.14%	12,604
雑誌	1. 総合	7,421,447,806	11.34%	1,927
	2. 教育	877,875,592	1.34%	228
	3. 政治	456,459,405	0.70%	119
	4. 産業	110,640,958	0.17%	29
	5. 工業	1,468,293,360	2.24%	381
	6. 厚生	180,964,513	0.28%	47
雑誌 小計		10,515,681,634	16.06%	2,730
新聞	全国紙	2,417,622,461	3.69%	628
	ブロック紙	1,296,592,154	1.98%	337
	地方紙	2,701,855,499	4.13%	702
新聞 小計		6,416,070,114	9.80%	1,666
合計		65,471,677,099	100.00%	17,000

1.3 図書館サブコーパスの設計とサンプリングの手順

図書館サブコーパスの設計とサンプリングは，以下のような手順で行う。

1. 出版サブコーパスの「書籍」の母集団（総文字数）と等しい文字数を，母集団のサイズとして設定する。
2. 東京都内の各公立図書館に所蔵されている書籍のリストを入手し，母集団の文字数が含まれる書籍の集合を，より多くの図書館で共通に所蔵されている書籍から構成し，そのリストを作成する。
3. 上記の手順で作られた書籍リストを，NDC と発行年によって層別し，各層に含まれる総文字数を推計する。
4. 推計総文字数の比から，各層のサンプル構成比を算出する。全体で 741.4 万語分の固定長サンプルを取得することを前提として，各層から必要となるサンプル数を比例割当により算出する。
5. 母集団を構成するすべての文字に等確率を与えた上で，ランダムに 1 文字を選び出し，この文字を基準点として，固定長サンプル・可変長サンプルを抽出する。これを各層ごとに必要なサンプル数分くり返す。

図 1.4: 図書館サブコーパスの設計とサンプリングの手順

2 章で詳述するように，図書館サブコーパスの母集団は，出版サブコーパスの書籍部分の母集団と等しいサイズ（文字数）に設定する。すなわち，出版サブコーパスの書籍で母集団として定義した約 485.4 億字を，図書館サブコーパスでも母集団の数値として用いるのである。これは，2 つのサブコーパス（の書籍部分）における固定長サンプル 1,000 万語について，母集団からの抽出比を等しく揃えるためである。次に，母集団である約 485.4 億字を含む書籍の集合を，より多くの公立図書館に共通して所蔵されている書籍から構成し，そこに含まれる文字の総体を母集団として定義する。さらに，書籍の NDC と発行年により母集団を層別し，各層に含まれる総文字数を推計して，サンプル構成比を算出する。その上で，出版サブコーパスと同じ手順によって，12,604 の固定長サンプル，および同数の可変長サンプルを抽出する。この手順により,「出版されたすべての書籍」「広く図書館に収められた書籍」という，異なる性格を持つ 2 つの書籍コーパスが，等しいサイズの母集団から等しい抽出比で作られることになる。

1.4 本報告書の目的

さて，前報告書で示したのは，出版サブコーパスの母集団の定義と，サンプル構成比を算出するために実施した「現代日本語書き言葉の文字数調査」の方法，およびその結果であった。本報告書で報告するのは，次の 2 点である。

- 図書館サブコーパスの母集団の定義と，サンプル構成比の算出法
- 母集団に含まれるすべての文字からランダムに 1 文字を指定し，固定長サンプル・可変長サンプルを抽出する方法

これら 2 点について，次の 2 章・3 章でそれぞれ解説する。

第2章　「図書館サブコーパス」におけるサンプル構成比の算出法

本章では，図書館サブコーパスの母集団をどのように定義し，サンプル構成比をどのように算出したか，その手順を示す。初めに図書館サブコーパスの設計について述べた後,「東京都内公立図書館の共通蔵書調査」を実施した手順について示す。その後，調査の結果を用いて図書館サブコーパスの母集団を定義し，さらにサンプル構成比を算出した手順を示す。

2.1　図書館サブコーパスの設計

図書館サブコーパスは，書き言葉が世の中に流通しているありさまを，公立図書館における書籍の所蔵状況によって近似的に把握しようとするものである。「流通」という言葉からすると，図書館の蔵書というよりもむしろ，書店での販売冊数や売上高のような，物流市場における流通の側面が想定されるかもしれない。しかしながら，市場における書籍の販売冊数や売上高などの実態を網羅的かつ客観的に捉えるためのデータを入手することは，現実的には不可能である[1]。

そこで，書き言葉の流通という観点を，書き言葉が世の中に広く行き渡っている状態として捉え，これを公立図書館における書籍の所蔵状況によって近似的に実現することにした。公立図書館は，さまざまな書き言葉の資料を収集し，地域住民へ提供していくことを任務としている。日本図書館協会図書館政策特別委員会による「公立図書館の任務と目標[2] 」には，次のような一節がある。

> （知る自由の保障）
> 2　住民は，あらゆる表現の記録（資料）に接する権利を有しており，この住民の知る自由を保障することは，公立図書館の重要な責務である。この責務を果たすため，公立図書館は，住民の意思を受けて図書その他の資料を収集し，収集した資料を住民に提供する自由を有する。

ここから，公立図書館における蔵書を対象として母集団を定義し，そこからサンプルを抽出することにより，書き言葉が世の中に流通している実態を把握することができると判断した。特に，より多くの公立図書館で共通に所蔵されている書籍ほど世の中に広く行き渡っている書き言葉であると捉え，それらをサンプリングの対象とすることとした。

---

[1] ただし，各年のベストセラーについては，複数の情報源により情報が入手できる。特定目的サブコーパスには，1976 年から 2005 年までの各年におけるベストセラーを収録する「ベストセラー」という枠が設けられている。

[2] http://www.jla.or.jp/ninmu.htm

次に問題となるのは，公立図書館の蔵書をどのように網羅的に把握するかということである。当初，全国の公立図書館が所蔵する書籍の総体を調査することを計画したが，そのようなデータは入手することができなかった。これに代わるデータとして，東京都内の各自治体における公立図書館の書籍の所蔵状況を東京都立図書館が調査・集計している「東京都公立図書館調査[3] 」に着目した。

そこで，東京都立図書館へ協力を依頼し，BCCWJ でのサンプリングに使用することを目的として,「区市町村立図書館 ISBN 総合目録（2007 年 2 月版）」（以下,「ISBN 総合目録」と記す）のデータを利用する許諾を特別に得た。これは，都内 52 自治体（23 区，29 市町村）ごとに，各自治体の公立図書館で所蔵している書籍を，ISBN（国際標準図書番号）のリストとしてまとめたものである。ISBN は書籍に一意に付与された 10 桁の固有の番号であり[4] ，このリストを集計することにより，東京都内各自治体の公立図書館が所蔵している書籍の分布を知ることができる。図書館サブコーパスの母集団は，この「ISBN 総合目録」を利用して，公立図書館における書籍の所蔵状況を調査し，その結果を用いて定義することにした。ここで行う調査を,「東京都内公立図書館の共通蔵書調査」と呼ぶことにする。

また，図書館サブコーパスの母集団のサイズは，出版サブコーパスの「書籍」の母集団と等しいサイズとし，かつ，そこから等しい量のサンプルを，同じ手順で抽出することにした。これは，2 つのサブコーパス（の書籍部分）のうち，特に文字数が統制されている固定長サンプルに対して，母集団からの抽出比を等しく揃えることが，統計的な言語研究上，有効であると考えたためである。

出版サブコーパスでは，母集団から 12,604 の固定長サンプル（741.4 万語分）を抽出し，同時に，同数の可変長サンプルを抽出する。そこで，図書館サブコーパスでも，約 485.4 億字になるべく近似したサイズの母集団を定義し，そこから 12,604 の固定長サンプル（741.4 万語分），および同数の可変長サンプルを抽出することを計画した。この際の母集団は，より多くの公立図書館で共通に所蔵されている書籍から構成することにした。このような設計により，当該の期間に生産された書き言葉の実態を捉える出版サブコーパスと，多くの図書館で共通に所蔵されており広く行き渡っている書き言葉の実態を捉える図書館サブコーパスという，2 つの異なる性格を持つ（しかし母集団からの抽出比が等しい）書き言葉コーパスを構築することができる[5]。

図書館サブコーパスの設計手順を，出版サブコーパスの設計手順と合わせて，図 2.1 に示す。

以上のような設計に基づき，図書館サブコーパスにおけるサンプリングの計画を立てた。そこで必要となるのは，以下の 3 点である。

1. 「ISBN 総合目録」を集計し，東京都内公立図書館における蔵書の分布を分析すること
2. 上記の調査結果から，約 485.4 億字から構成される母集団を定義すること
3. 取得対象となる 12,604 サンプルの構成比を算出すること

---

[3] 東京都立図書館による「東京都公立図書館調査」は毎年実施され，その主な結果は Web 上でも報告されている。`http://www.library.metro.tokyo.jp/15/15710.html`

[4] 2007 年 1 月 1 日の規格改定により，これ以降に刊行された新刊本・重版本には 13 桁の ISBN が付与されている。ここでは改定前の 10 桁の ISBN を用いている。また，シリーズ本などの場合，1 つの ISBN が書籍を一意に同定しないことがある。

[5] ランダムサンプリングの結果，2 つの書籍コーパスの両方に同じ書籍からのサンプルが採録されることもある（試算したところ，全体の約 0.9%にあたる 227 冊程度）。

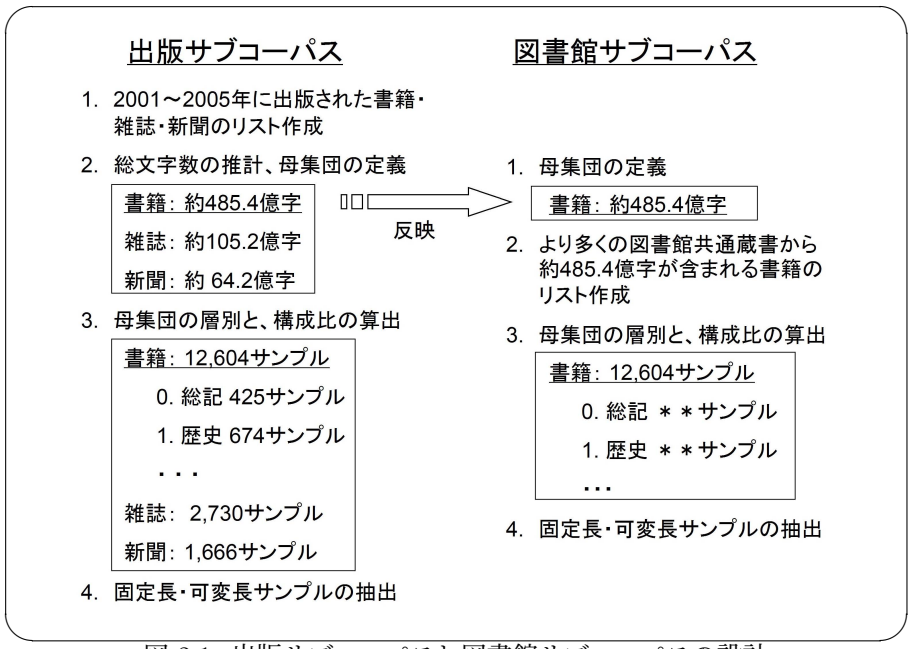


図 2.1: 出版サブコーパスと図書館サブコーパスの設計

以下では，これらの調査・分析を行った結果を示す。

2.2　書籍の発行期間に関する修正

さて，図書館サブコーパスの母集団を定義する際，当初の「1976 年から 2005 年の間に発行された書籍」という条件を修正する必要が生じた。以下では，その経緯について述べる。

「ISBN 総合目録」を集計して母集団を定義する際，問題となったのは，社会における ISBN の普及率である。日本国内で ISBN の付与が始まったのは 1981 年であり，しかも出版社によって導入の時期がまちまちであった。特に 1980 年代前半の時期に出版された書籍を ISBN によって検索すると，ある出版社が出版した書籍はヒットするものの，別の出版社が出版した書籍はヒットしないなど，出版社による対応の違いが反映されてしまい，偏った結果しか得られないという問題が生じる。また，ISBN の導入が 1981 年以降である以上，それ以前に発行された書籍は,「ISBN 総合目録」の集計から定義される母集団にはそもそも入り得ないことになる。

以上のような理由から,「ISBN 総合目録」を用いて母集団を定義するためには，ISBN が普及した時期を見極め，図書館サブコーパスの発行期間に関する条件を適切に修正する必要があると判断した。そこで,「ISBN 総合目録」を集計する前段階の調査として，ISBN の普及率の変遷について調べることにした。調査には，出版サブコーパスの母集団の定義にも用いた，国立国会図書館の蔵書目録である「J-BISC（Japan Biblio disc)」の書誌情報を用いた。

初めに,「J-BISC」に記録されている 1976 年から 2005 年に発行された書籍の総体を対象として，それらに ISBN が付与されているかどうかを調べた。対象となる書籍は，合計 2,314,659 冊である。結果を図 2.2，図 2.3 に示す。

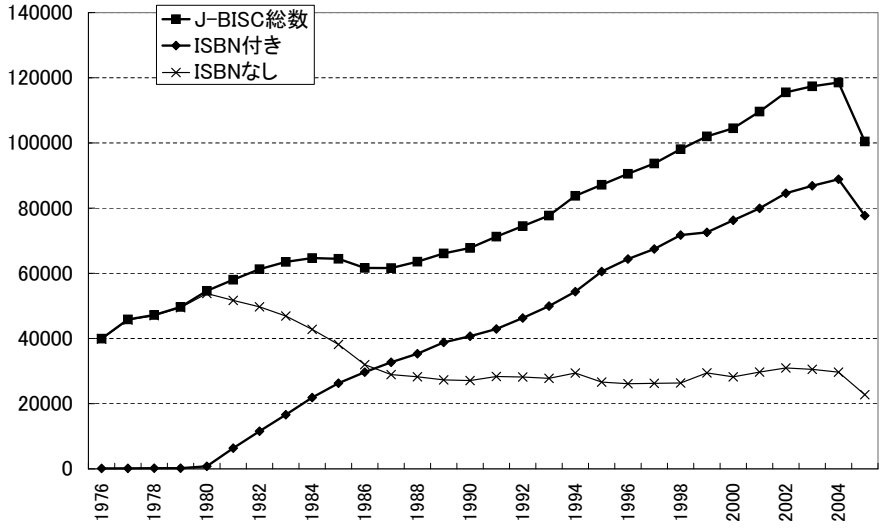


図 2.2: ISBN が付与されている冊数と付与されていない冊数

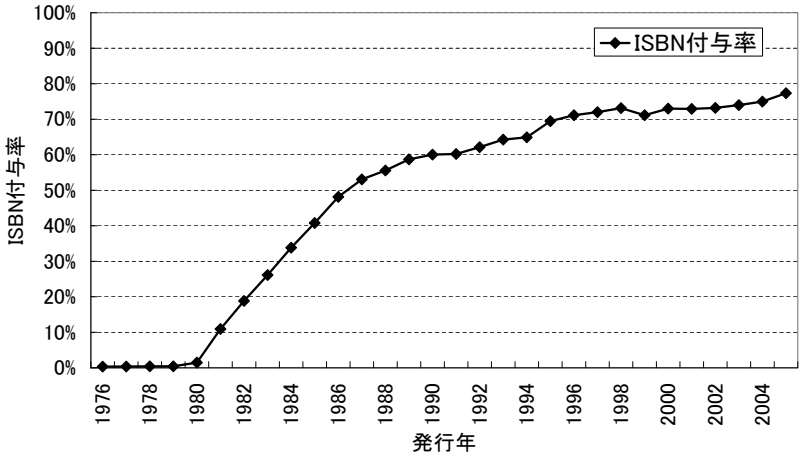


図 2.3: 「J-BISC」における ISBN の付与率の推移

図 2.2 を見ると，ISBN の付いた書籍は，導入の始まった 1981 年から徐々に増え始めていることが分かる。ISBN 付きの書籍の冊数と ISBN のない書籍の冊数とが逆転するのが，1986 年以降である。これは，図 2.3 で ISBN の付与率が 50%を超える時期に相当する。ただし，2005 年に至るまで，全体の約 20%にあたる約 3 万冊の書籍には ISBN が付与されていない。調べてみたところ，これらは，官公庁資料のように一般には流通しない書籍資料が大半であった。すなわち，出版社が ISBN の付与に対応していないために「ISBN 総合目録」の集計に偏りが

生じ得るのは，1981 年から 1985 年までの間であると言える[6]。

次に，主要な出版社における ISBN への対応状況を確認する。まず,「J-BISC」を用いて，1981 年から 2005 年の間に ISBN が付与された書籍を出版した出版社を，出版点数順にリストアップした。出版点数の多い上位 20 社を表 2.1 に示す。

表 2.1: 書籍の出版点数の多い出版社（上位 20 社）

出版社	冊数	出版社	冊数	出版社	冊数
1. 講談社	51,870	8. 学習研究社	11,063	15. 秋田書店	7,169
2. 小学館	31,000	9. PHP 研究所	10,867	16. 明治図書出版	7,012
3. 集英社	22,946	10. 徳間書店	10,727	17. 中央経済社	6,945
4. 角川書店	17,877	11. 文芸社	10,322	18. 筑摩書房	6,838
5. 岩波書店	13,398	12. 新風舎	8,631	19. ポプラ社	6,697
6. ゼンリン	13,200	13. 光文社	8,177	20. 中央公論社	6,678
7. 新潮社	11,910	14. 河出書房新社	7,293	...	

これらの主要な出版社について，ISBN 付与への対応がどのように推移したかを調べた。1980 年から 1992 年までの各年において，上記のうち上位 10 社（ただし，地図の出版が大半であるゼンリンは除く）が発行した書籍に ISBN が付与されていた割合（ISBN 付与率）を集計し，その推移を調べた。結果を，図 2.4 に示す。

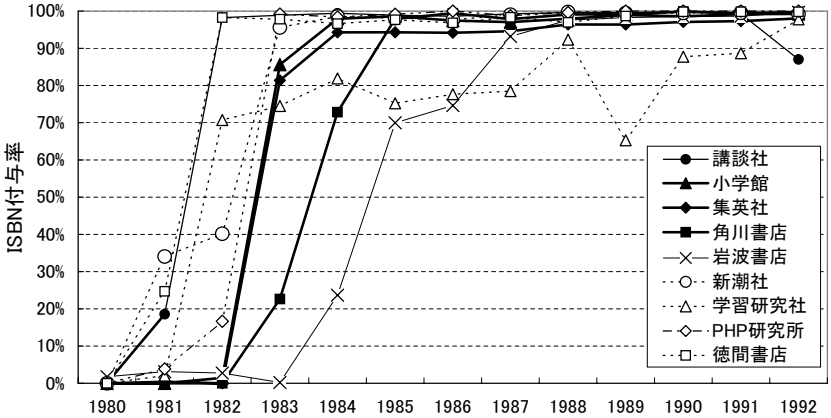


図 2.4: 主要出版社による ISBN の付与率の推移

1981 年以降，出版社によってばらつきがあるものの，徐々に ISBN の普及が進んでいる様子が分かる。1986 年にはすべての出版社で付与率が 75%を超え，かつ大半の出版社では付与率が 95%を超えていることから，ISBN 付与への対応は，この時期にほぼ落ち着いていると見てよい。

[6] なお，図 2.2 で 2005 年の冊数が少なくなっているのは,「J-BISC」のデータを入手したのが 2005 年 10 月であり，2005 年に発行されたすべての書籍が記録されているわけではないことによる。

上記のような検討の結果,「ISBN 総合目録」に基づいて図書館サブコーパスの母集団を定義するためには，ISBN が普及した時期を考慮して，1986 年以降に発行された書籍を対象とすることが適切であると判断した。そこで，当初の図書館サブコーパスにおける発行期間に関する条件「1976 年から 2005 年の 30 年間に発行されたもの」を,「1986 年から 2005 年の 20 年間に発行されたもの」と修正することにした。

2.3 東京都内公立図書館の共通蔵書調査

以下では,「ISBN 総合目録」を集計して東京都内公立図書館の共通蔵書を分析した結果を示す。まず,「ISBN 総合目録」のデータサイズを示す。「ISBN 総合目録」は，都内 52 自治体（23 区，29 市町村）それぞれの公立図書館で所蔵する書籍を，ISBN のリストとして取りまとめたものである。このうち，ISBN が「4」で始まる和書のレコード数は，以下のようになっている。

ISBN 数（延べ）	15,463,769 レコード
ISBN 数（異なり）	1,144,103 レコード

都内 52 自治体の公立図書館には，延べで約 1,546 万冊，異なりで約 114.4 万冊の書籍が所蔵されていることになる[7]。次に，都内 52 自治体が所蔵する書籍数（異なり）を，表 2.2 に示す。

表 2.2: 都内 52 自治体が所蔵する書籍数

自治体	冊数	自治体	冊数	自治体	冊数
1 目黒区	553,131	19 多摩市	341,166	37 小金井市	212,691
2 世田谷区	481,343	20 文京区	340,177	38 稲城市	209,628
3 江東区	477,581	21 港区	339,393	39 青梅市	202,108
4 調布市	464,577	22 品川区	328,361	40 東大和市	199,278
5 足立区	456,979	23 立川市	323,291	41 東久留米市	196,584
6 町田市	453,434	24 荒川区	320,090	42 あきる野市	194,832
7 八王子市	439,889	25 武蔵野市	317,802	43 国立市	185,565
8 大田区	433,329	26 新宿区	306,668	44 昭島市	180,430
9 府中市	424,364	27 渋谷区	303,887	45 羽村市	174,173
10 練馬区	398,892	28 豊島区	294,929	46 福生市	161,931
11 墨田区	395,470	29 日野市	291,198	47 狛江市	153,139
12 板橋区	386,375	30 西東京市	289,129	48 武蔵村山市	144,160
13 江戸川区	385,974	31 三鷹市	278,922	49 千代田区	130,343
14 葛飾区	379,910	32 国分寺市	269,156	50 瑞穂町	117,726
15 中野区	362,211	33 東村山市	264,408	51 日の出町	70,977
16 北区	359,886	34 中央区	255,484	52 奥多摩町	55,424
17 杉並区	347,325	35 台東区	244,123		
18 小平市	341,836	36 清瀬市	224,090		

最も所蔵数が多いのは目黒区で 553,131 冊，次いで世田谷区の 481,343 冊が続いている。52 自治体全体の平均は，297,380 冊となる。

[7] ただし，各自治体における ISBN は異なりで集計されているため，各自治体内にある複本を含めた数は，延べの冊数には入っていない。また以下では，便宜上，1 つの ISBN が 1 冊の書籍に対応するものと見なす。

さて,「ISBN 総合目録」に記録された約 114.4 万冊の書籍の中には，52 自治体すべてで所蔵されている書籍もあれば，1 自治体でしか所蔵されていない書籍もある。ここで必要となるのは，これら約 114.4 万冊の書籍の中から，より多くの自治体で共通に所蔵されている書籍を集めて，約 485.4 億字を含む集合を定め，図書館サブコーパスの母集団として定義することである。このためには，約 114.4 万冊の書籍が各自治体にどれだけ共通して所蔵されているか，その分布を知る必要がある。

そこで，約 114.4 万冊の書籍が各自治体間でどれだけ共通に所蔵されているかを，各自治体間の ISBN リストがどれだけ重複しているかによって調べた。この結果を，図 2.5 および表 2.3 に示す。共通自治体数が増えるにしたがって，共通蔵書の冊数が減少していく様子が分かる。

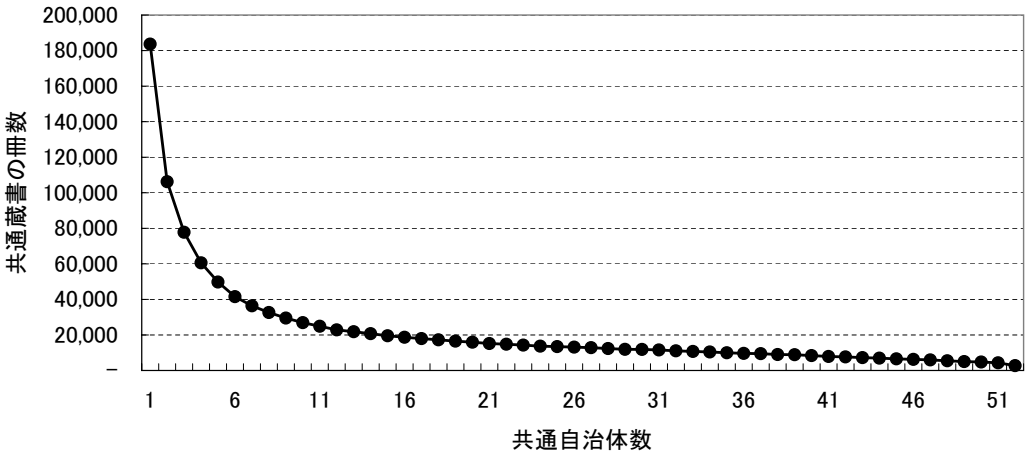


図 2.5: 東京都内 52 自治体における共通蔵書の分布

表 2.3: 都内 52 自治体における共通蔵書の分布

自治体数	冊数	累積	自治体数	冊数	累積	自治体数	冊数	累積
1	183,663	183,663	19	16,498	825,236	37	9,481	1,047,212
2	106,206	289,869	20	15,897	841,133	38	8,969	1,056,181
3	77,792	367,661	21	15,179	856,312	39	8,862	1,065,043
4	60,584	428,245	22	14,753	871,065	40	8,383	1,073,426
5	49,755	478,000	23	14,252	885,317	41	7,870	1,081,296
6	41,539	519,539	24	13,705	899,022	42	7,654	1,088,950
7	36,404	555,943	25	13,409	912,431	43	7,245	1,096,195
8	32,630	588,573	26	13,119	925,550	44	6,919	1,103,114
9	29,509	618,082	27	12,813	938,363	45	6,561	1,109,675
10	26,937	645,019	28	12,346	950,709	46	6,252	1,115,927
11	24,886	669,905	29	11,891	962,600	47	5,943	1,121,870
12	22,845	692,750	30	11,854	974,454	48	5,448	1,127,318
13	21,796	714,546	31	11,556	986,010	49	4,996	1,132,314
14	20,742	735,288	32	11,041	997,051	50	4,756	1,137,070
15	19,528	754,816	33	10,698	1,007,749	51	4,303	1,141,373
16	18,706	773,522	34	10,401	1,018,150	52	2,730	1,144,103
17	17,957	791,479	35	9,952	1,028,102			
18	17,259	808,738	36	9,629	1,037,731			

52 自治体のすべてで所蔵されている書籍の数は，2,730 冊であった。これらは，東京都内の広い範囲に渡って存在している書き言葉であると見なすことができる。一方，1 自治体でしか所蔵されていない書籍の数は，183,663 冊であった。これらは，図書館に収められてはいるものの，極めて狭い地域にしか行き渡っていない書き言葉であると見なすことができる。

以上,「ISBN 総合目録」の集計により，共通自治体数と共通蔵書数の分布が明らかになった。これによって，どの範囲の書籍を図書館サブコーパスの母集団とすべきかを検討する材料が整った。

2.4 図書館サブコーパスの母集団の定義

以下では,「東京都内公立図書館の共通蔵書調査」の結果に基づいて，より多くの図書館で共通に所蔵している書籍から，図書館サブコーパスの母集団を定義するまでの手順を示す。2.1 節で述べたように，図書館サブコーパスの母集団は，出版サブコーパスの書籍の母集団である約 485.4 億字になるべく近似させ，そこから 741.4 万語分に相当する 12,604 の固定長サンプル，および，同数の可変長サンプルを抽出する。そこで，共通自治体数の範囲を変えながら，そこに含まれる全書籍の推計総文字数の変動を調査し，約 485.4 億字を含む書籍の集合をどのように定めればよいかを検討することにした。総文字数の推計は，丸山・秋元 (2007) で示した「現代日本語書き言葉の文字数調査」と同じ手順で行う。

ここで，総文字数を推計するために必要な「ページ数」「判型」「NDC」に関する情報を得るために,「ISBN 総合目録」と「J-BISC」の書誌情報を，ISBN をキーにして結合した。この結果に対して，1986 年から 2005 年の間に発行された書籍に対象を絞り込み，さらに前報告書で示した「適切性条件[8]」をクリアする書籍のみを抽出した。その結果,「ISBN 総合目録」に含まれる 1,144,103 レコードの ISBN のうち，上記の条件を満たす書籍として残ったのは 695,950 レコードであった。そこで，この冊数を対象として，母集団の範囲を検討することとした。

まず，いくつ以上の自治体で共通に所蔵されている書籍の冊数が，出版サブコーパスの母集団に含まれる書籍の冊数と最も近似するかを調べた。ここでは，52 自治体のうち「16 以上の自治体」から「12 以上の自治体」までの範囲における共通蔵書の調査結果を示す。この範囲における共通蔵書の冊数と，出版サブコーパスの母集団 317,117 冊との差分について，表 2.4 に示す。

[8] コーパスへの収録対象としては不適切な書籍を母集団から除外するための条件。「40 ページ以下の書籍」「ページ数の記録がない書籍」「官公庁刊行物のうち非流通物」「学習試験図書」「電子資料，地図資料」「漫画」「写真集・図画集」「複製，覆刻」などの書籍を,「J-BISC」上の書誌情報を用いて除外する。

表 2.4: 都内公立図書館 共通蔵書の範囲と冊数

範囲	共通蔵書 冊数	差分	
16 以上の自治体	292,659	–24,458	(92.29%)
15 以上の自治体	306,485	–10,632	(96.65%)
14 以上の自治体	320,895	3,778	(101.19%)
13 以上の自治体	335,721	18,604	(105.87%)
12 以上の自治体	351,086	33,969	(110.71%)

冊数という観点からは，14 以上の自治体で共通に所蔵されている 320,895 冊が，出版サブコーパスの母集団の冊数と最も近似する。

次に，出版サブコーパスの母集団に含まれる書籍のページ数と最も近似する共通蔵書の範囲を調べた。共通自治体の範囲と，その共通蔵書に含まれるページ数，および出版サブコーパスの母集団 74,911,520 ページとの差分について，結果を表 2.5 に示す。

表 2.5: 都内公立図書館 共通蔵書の範囲とページ数

範囲	共通蔵書 ページ数	差分	
16 以上の自治体	74,453,550	–457,970	(99.39%)
15 以上の自治体	78,002,542	3,091,022	(104.13%)
14 以上の自治体	81,635,598	6,724,078	(108.98%)
13 以上の自治体	85,363,019	10,451,499	(113.95%)
12 以上の自治体	89,202,899	14,291,379	(119.08%)

共通蔵書のページ数という観点からは，16 以上の自治体で所蔵されている 74,453,550 ページが，出版サブコーパスの母集団のページ数と最も近似する。

最後に，出版サブコーパスの母集団に含まれる書籍の推計総文字数と最も近似する共通蔵書の範囲を調べた。総文字数の推計には，丸山・秋元 (2007) で示した「総文字数推計基準表」を用いる。これは，書籍の NDC と判型の別によって，1 ページあたりに含まれる平均文字数を推計したものであり，NDC・判型ごとのページ数に掛け合わせることにより，総文字数を推計することができる。

参照 「総文字数推計基準表」については，資料編の表 4.1 を参照（39 ページ）。

共通自治体の範囲と，その共通蔵書に含まれる推計総文字数，および出版サブコーパスの母集団 48,539,925,351 文字との差分について，結果を表 2.6 に示す。

表 2.6: 都内公立図書館 共通蔵書の範囲と推計総文字数

範囲	推計総文字数	差分	
16 以上の自治体	41,542,127,686	–6,997,797,665	(85.58%)
15 以上の自治体	43,595,361,035	–4,944,564,316	(89.81%)
14 以上の自治体	45,694,295,469	–2,845,629,881	(94.14%)
13 以上の自治体	47,877,656,072	–662,269,279	(98.64%)
12 以上の自治体	50,143,594,390	1,603,669,039	(103.30%)

共通蔵書の推計総文字数という観点からは，13 以上の自治体で所蔵されている 47,877,656,072 文字が，出版サブコーパスの母集団の文字数と最も近似する。

上記の検討の結果，出版サブコーパスの書籍部分と図書館サブコーパスの母集団サイズを総文字数によって近似させる場合，13 以上の自治体で共通に所蔵している書籍の集合が最適であることが分かった。そこで,「都内公立図書館のうち，13 以上の自治体で共通して所蔵している書籍」を，図書館サブコーパスの母集団として定義することにした。検討結果をまとめて，表 2.7 に示す。

表 2.7: 出版サブコーパスと図書館サブコーパスの母集団

	出版サブコーパス	図書館サブコーパス
冊数	317,117	335,721
ページ数	74,911,520	85,363,019
文字数	48,539,925,351	47,877,656,072

2.5 サンプル構成比の算出

最後に，上記の手順で得られた母集団をどのように層別し，サンプル構成比をどのように算出したかについて示す。

まず，この母集団を層別する基準について示す。ここでは，出版サブコーパスと同様,「NDC」および「発行年」を用いて層別を行った。「NDC」は書籍の内容により，大きく「0. 総記」「1. 哲学」「2. 歴史」「3. 社会科学」「4. 自然科学」「5. 技術・工学」「6. 産業」「7. 芸術・美術」「8. 言語」「9. 文学」という 10 カテゴリに分類される。「J-BISC」に付与されている NDC（1 桁目）の 10 分類に加え，NDC が付与されていないレコードを「n（null；記録なし)」として，合計 11 の層に分類した。これに「発行年」として 1986 年から 2005 年までの各年を 20 の層として重ね合わせ，母集団全体を合計 220 の層に分割した。

母集団から 12,604 サンプルを取得する際，その構成比は，220 の各層に含まれる推計総文字数の比を比例割当することによって算出する。推計総文字数の比をサンプル構成比として採用することにより，図書館に収められている書籍を文字数という絶対量によって層別し，その分布のありさまを反映するサブコーパスを構築するわけである。

以下，図書館サブコーパスの母集団に含まれる 47,877,656,072 文字について，NDC ごとの推計総文字数とサンプル構成比を表 2.8 に示す。また，サンプル構成比を比例割当して得られる各層からの取得サンプル数を，図 2.6 に示す。

参照 実際には，220 の各層に対して推計総文字数を比例割当することにより，図書館サブコーパス全体のサンプル構成比が算出されることになる。図書館サブコーパスの母集団全体の冊数・ページ数・推計総文字数を発行年と NDC で層別した一覧については，資料編の表 4.7〜表 4.11 を参照（45〜49 ページ）。

表 2.8: 図書館サブコーパス全体のサンプル構成比

NDC	総文字数	構成比	サンプル数
0. 総記	1,003,528,880	2.096%	264
1. 哲学	2,343,849,711	4.895%	617
2. 歴史	5,010,749,621	10.466%	1,319
3. 社会科学	8,946,058,392	18.685%	2,355
4. 自然科学	3,028,276,363	6.325%	797
5. 技術工学	3,149,144,051	6.577%	829
6. 産業	1,690,150,481	3.530%	445
7. 芸術	4,057,291,256	8.474%	1,068
8. 言語	956,625,910	1.998%	252
9. 文学	15,485,091,056	32.343%	4,077
n. 記録なし	2,206,890,351	4.609%	581
合計	47,877,656,072	100.00%	12,604

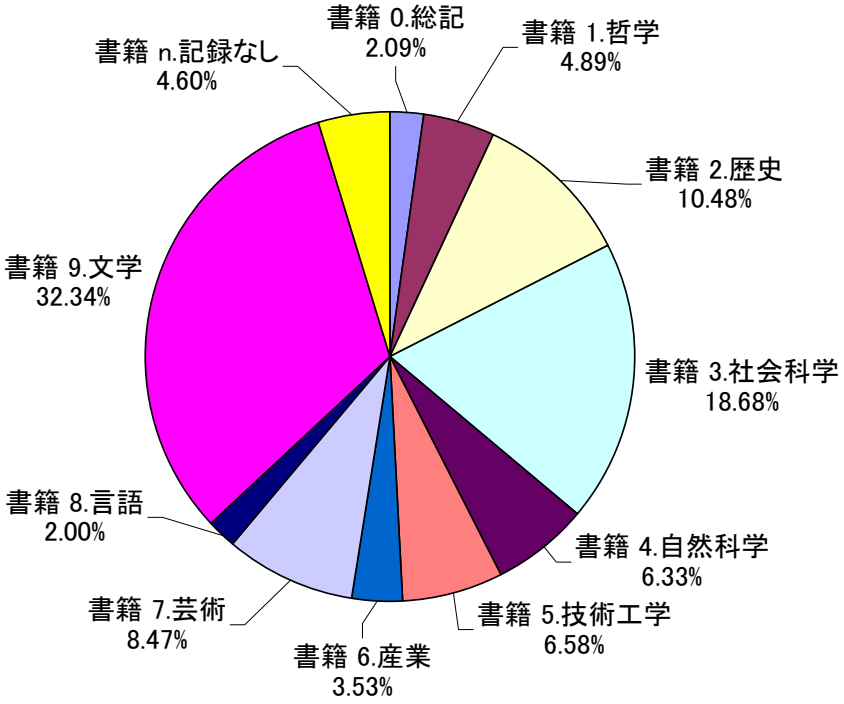


図 2.6: NDC 別のサンプル構成比（図書館サブコーパス）

これで，出版サブコーパスのサンプル構成比（9 ページ）と図書館サブコーパスのサンプル構成比が出揃った。

以上,「東京都内公立図書館の共通蔵書調査」の結果にもとづいて図書館サブコーパスの母集団を定義し，サンプル構成比を算出した手順について示した。

第3章　サンプル台帳の作成とサンプルの無作為抽出

本章では，前章までで示した出版サブコーパス・図書館サブコーパスの設計に基づいて，実際にどのようにサンプルを取得していくか，その手順について述べる。母集団を構成するすべての文字に対して等確率を与え，その中の1文字をランダムに指定するためのサンプル台帳を作成し，固定長サンプル・可変長サンプルを抽出するまでの手続きを示す。

3.1　文字を基準とするサンプリングの方針

従来，国語研究所が行ってきた統計的な漢字調査・語彙調査の中では，主として「エリアサンプリング」が採用されてきた。例えば「現代新聞の漢字調査」では，新聞1ページを30のエリア（縦15段×横2段）に分割し，乱数表によってあるエリアを選び，結果として母集団全体の1/60の面積に含まれる言語表現が選ばれるようなサンプリング方法を取っている。

BCCWJで採用するサンプリングは，エリアサンプリングではなく，いわば「文字サンプリング」とでも呼ぶべき方法である。これは，エリアサンプリングのように言語を載せる媒体（面積）の物理的な計測に基づいてサンプルを抽出するのではなく，あくまでも言語そのものの絶対量，すなわち文字量に基づいてサンプルを抽出することを意図するものである。

このためにまず，各層に含まれる全ての文字に対して等確率を与え，ランダムに1文字を指定して,「サンプル抽出基準点」を決定する。そして，ランダムに指定された1文字を基準として，2種類のサンプルを抽出するのである。これは，母集団を構成する1文字目から最後の文字（出版サブコーパスでは65,471,677,099文字目，図書館サブコーパスでは47,877,656,072文字目）までを1列に配置した上で，ランダムに指定された1文字を「サンプル抽出基準点」として，そこから1,000文字という範囲[1]，およびその文字を含む「章」「節」などの言語的なまとまりを持つ範囲を，それぞれ固定長サンプル・可変長サンプルとして同時に抽出することを意図している。このことを，出版サブコーパスを例に図示すると，図3.1のようになる。

母集団は文字数によって定義されているため，そのすべての文字に等確率を与えて1文字をランダムに選び出すことは理論上は可能であるが，しかしながら現実的には非常に困難である。そこで我々は，母集団に含まれる全てのページに対して等確率を与えた上で，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文字をランダムに指定する，という2段階の方法を取ることにした。

[1] 厳密には，サンプル抽出基準点を含む文の文頭，およびサンプル抽出基準点から1,000文字目を含むの文の文末までが取得される。

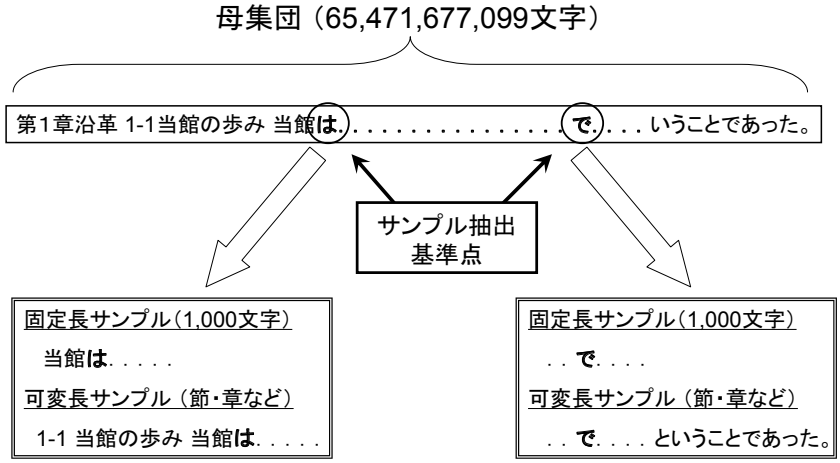


図 3.1: 「サンプル抽出基準点」の指定とサンプルの抽出

3.2　サンプル台帳の作成

以下では，書籍の場合を例として，サンプル台帳の作成と 1 文字の指定の仕方について示す。

表 2.7 に示したように，出版サブコーパスの「書籍」の母集団には 74,911,520 ページが，図書館サブコーパスの母集団には 85,363,019 ページが，それぞれ含まれている。この全ページを，各サブコーパスの各層 —出版サブコーパスでは 55 層，図書館サブコーパスでは 220 層— に分類し，書誌情報つきのテーブルとしてリレーショナル・データベース[2]上に展開した。その上で，各層に含まれる全てのページをシャッフルして，各ページに対してランダムに優先順位を割り振った。

さらに，各ページに対して，ページ内の 1 点を指定する座標情報をランダムに指定した。これは，ページに 10×10 の座標枠を割り当て，指定された座標の交点に最も近い文字を「サンプル抽出基準点」として指定するためのものである。座標情報は，横軸を 0〜9，縦軸を A〜J として,「0A」から「9J」まで 100 通りの交点を指定した。

ただし，サンプル台帳で指定されたあるページが白紙であった場合,「サンプル抽出基準点」となる文字を指定することができないため，次に優先順位の高いページに移らなければならない。これはほとんどの場合，その場で手に取った書籍を放棄し，次の該当ページを含む書籍を新たに探し出す必要がある。しかしながら，ランダムに指定された特定の書籍を探し出すことは実際には非常に手間のかかる作業であり，同一の書籍内から次候補を探し出す方がはるかに効率的である。そこで，作業進行上の効率を考慮して，ある書籍のうち最も優先順位が高いページから文字が抽出できなかった場合は，同じ書籍の中で次に優先順位の高いページに移ってよいこととした。これを上位 20 位まで繰り返してよいこととし，20 位までのページから 1 文字も指定できなかった場合は，その書籍をサンプリングの対象から除外することとした[3]。

[2] Microsoft SQL Server 2005 を用いた。
[3] 上位 20 位までのページから 1 文字も指定できない書籍には，図鑑や図説，デザイン集やカット集，統計資料集のように，図や表が主体となって書籍全体が構成されているものや，古文や外国語など非現代日本語のみで書籍全体が構成されているものなどが多い。

また，交点の直近が図・写真であったり，交点の直近に文字がなかったりする場合,「サンプル抽出基準点」となる文字を指定することができない。そこで，1 ページあたり 10 通りの交点を準備し，それぞれに優先順位を付した。第 1 位の交点で文字が指定できなかった場合，第 2 位の交点に移ってよいこととした。第 10 位までの交点でも文字が指定できなかった場合，次に優先順位の高いページに移ることとした。

以上のような手順によって，出版サブコーパスでは 74,911,520 ページ分の，図書館サブコーパスでは 85,363,019 ページ分のサンプル台帳を，それぞれ作成した。このサンプル台帳により，母集団に含まれる全てのページから特定の 1 ページを，さらにそのページに含まれる特定の 1 文字を，ランダムに指定することができる。実際のサンプリング作業では，サンプル台帳にしたがって，優先順位のより高いページの，優先順位のより高い座標に近い文字を「サンプル抽出基準点」として指定していくことになる。

そこで，サンプル台帳には，同一の書籍から優先順位の高い上位 20 位までのページ番号と，それに付随する 10 位までの座標情報を含めることにした。さらに，書籍のタイトルや，サンプル管理用の ID などを表示して，1 枚のサンプル台帳としてまとめた。サンプル台帳の例を，図 3.2 に示す。

サンプル取得数　　借り出し日
1　SampleID: PB58_00482　BibID: 20770388　タイトル: 伝え合いの言葉　2006/12/28

NDLsize　　配架情報
21cm　　国語研図書館
分類せず

優先順位	対象頁	有効	乱数1	乱数2	乱数3	乱数4	乱数5	乱数6	乱数7	乱数8	乱数9	乱数X	備考
一位	49		4E	2H	2D	0D	7D	4J	2I	7A	5I	1A	
二位	225		8D	8A	8J	6H	1C	9I	5H	4E	9D	0F	
三位	78		4J	0H	0I	7J	7C	8D	7B	2E	6E	6J	
四位	20		9J	2B	0J	0G	4I	2E	1E	7H	5J	3B	
五位	115		4D	5F	9F	8I	9E	4A	9J	2I	6H	3B	
六位	5		0H	8A	3H	1G	5I	2A	2D	6D	9E	4D	
七位	108		0G	5D	1E	0I	2I	7C	5A	2A	4H	1I	
八位	12		0C	6D	2E	3D	3F	6E	9A	0G	8J	2H	
九位	201		2J	3A	8B	0F	5E	0H	7D	4B	2B	7C	
十位	152		1E	4E	7C	1C	0F	9G	7G	4J	9D	5D	
十一位	232		2G	1D	7J	7C	0F	2H	8E	7F	5D	4C	
十二位	242		3H	2F	9J	4G	4J	6D	1G	3J	4C	2B	
十三位	51		8D	7I	4H	3E	6J	1J	1D	0J	0B	9C	
十四位	44		1I	2I	6I	7F	4H	0E	0I	0F	2E	6J	
十五位	69		5A	2E	6I	9A	7F	1J	4G	7I	4H	5C	
十六位	233		3F	1J	4I	3D	8D	5F	5C	0H	1E	1A	
十七位	153		4I	1C	2C	3G	0H	4A	4B	4G	8E	2G	
十八位	159		5E	3J	0H	3I	0G	1D	9F	8I	7H	6E	
十九位	158		7A	4B	0H	0I	2C	0C	6A	2J	8D	3I	
二十位	193		6E	6B	0F	8I	5B	8J	6G	7D	3B	2H	

図 3.2: サンプル台帳の例

実際のサンプリング作業では，サンプル台帳で指定されたページおよび座標情報にしたがって，実際の書籍を手に取り，指定された座標に最も近い 1 文字を見つけて「サンプル抽出基準点」として指定する。上記の台帳の例では，この書籍の中で優先順位が「一位」である「49 ページ」の交点「4E」に最も近い文字を探し出すことになる。この過程を，図 3.3 に示す[4]。

[4] 実際の作業では，座標の枠を印刷した透明のシート（「サンプル抽出基準点」指定シート）を判型ごとに用意し，印刷紙面に当てることにより，1 文字を特定している。

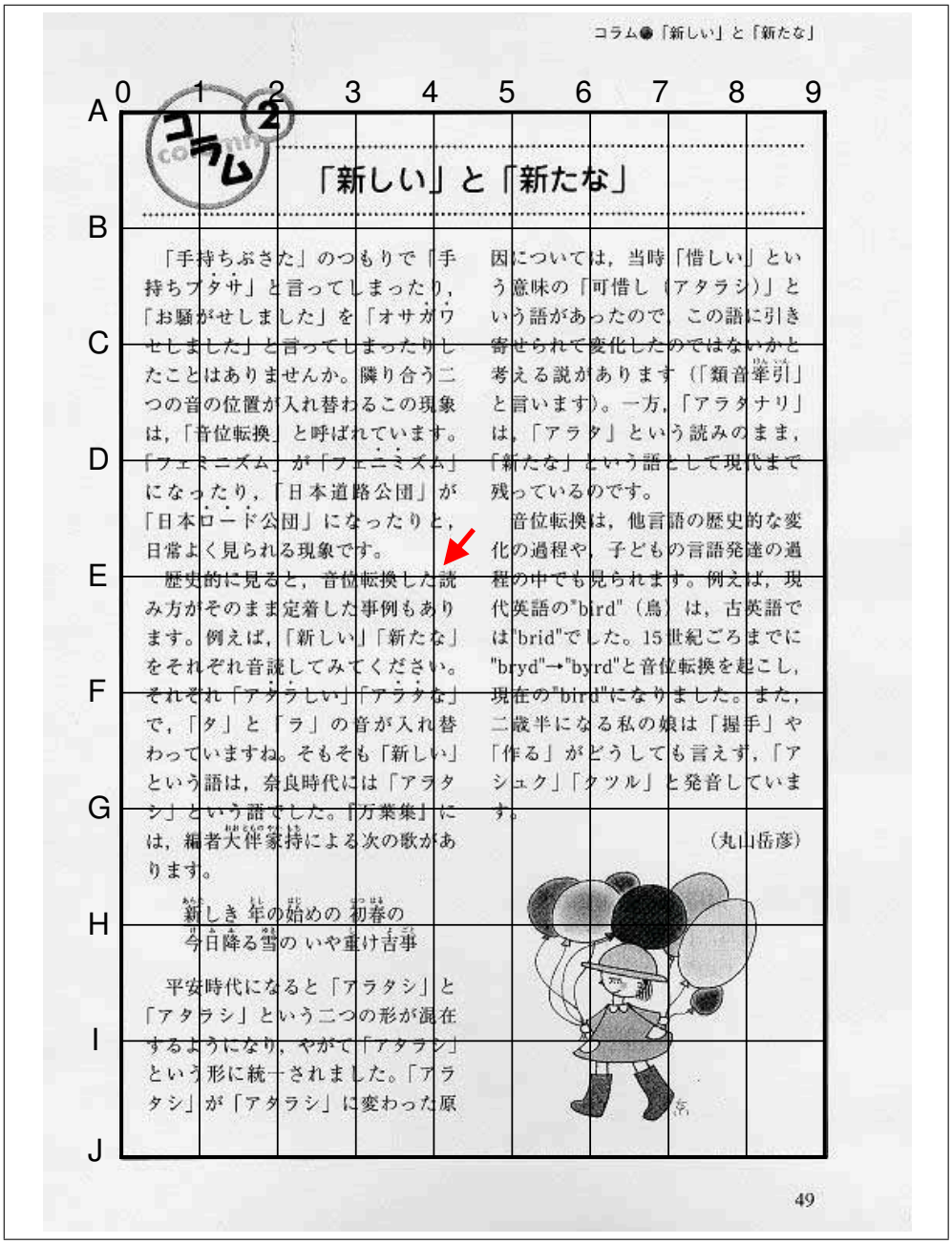


図 3.3: サンプル台帳で指定されたページ・座標から 1 文字を特定する例

図 3.3 を見ると，サンプル台帳で指定された「49 ページ」上の交点「4E」に最も近い文字は,「た」である。そこで，この文字を「サンプル抽出基準点」として指定する。その後，この文字を基準として，固定長サンプル・可変長サンプルの範囲を抽出していくことになる。

3.3 固定長サンプル・可変長サンプルの抽出

以上の手続きにより，サンプル台帳を作成し，母集団を構成する文字の集合からランダムに 1 文字を指定することができた。次に必要になるのは，この「サンプル抽出基準点」を基準として，固定長サンプル・可変長サンプルを抽出していくための方法論を検討することである。

具体的な書き言葉の版面（印刷紙面）からコーパスに格納する言語表現をサンプルとして抽出するためには，さまざまな事例を検討し，詳細な基準を作成しなければならない。以下ではその基本的な方針のみ示すことにする。詳細なサンプリングの基準については，稿を改めることとする[5]。

出版サブコーパス・図書館サブコーパスに格納されるのは，書籍・雑誌・新聞の書き言葉である。これらの書き言葉の実体は，紙面上に印字された文字列である。ここからサンプルを抽出する際，書き言葉の多様な構造をどのように一元的に見なすか，という問題が生じる。言い換えれば，さまざまな物理的・論理的な構造を含む版面上の書き言葉から，図 3.1 に示したような 1 次元の文字列をどのように抽出するか，という問題である。

書き言葉は，それが実現されている文書中において,「本文」「見出し」「注」「ルビ」「目次」「表」など，さまざまな論理的要素から構成されている。書き言葉の版面からサンプルを抽出するためには，版面を構成する要素のうち，どの要素を抽出し，どの要素を抽出しないのかを前もって決めておかなければならない。このためには，書き言葉が持つ構造をあらかじめ体系的に把握しておいた上で，個別の事例について対処していく必要がある。

また，版面上に現れる「本文」「見出し」「注」などの要素をどのように区別するか（区別できるか）ということも問題となる。これらの要素の区別は一見自明的であるように思われるが，しかしながら，論理的な構造に関する情報が文書中に明示的に表示されているわけではない。むしろ，版面上のある言語表現が「見出し」であり，別の言語表現が「本文」であることは，意識的であれ無意識的であれ，読み手が能動的に読み取っている情報である。

BCCWJ では，サンプルは最終的に XML 形式で電子化され,「本文」「見出し」「キャプション」などの要素に対して文書構造情報を表すタグが付与されることになる（山口ほか, 2008）。しかしながら，サンプリングの段階において，特に 1,000 文字の固定長サンプルを抽出する場合には，あらゆる実現様式・あらゆる論理構造に関わる書き言葉を一元的に見立て，1 次元の文字の連鎖として把握しなければならないのである。

以下では，版面上の構成・構造がもっとも多様である書籍を例として，書き言葉の物理的・論理的な構造を把握し，どの部分がサンプリングの対象となるのかについて概略を述べる。書籍の構造を,「形態」「版面」「本文」「文字」という 4 つの側面によって段階的に捉え，その中からコーパスに収録するサンプルとして抽出される部分と，それが満たすべき基準について示す。

3.3.1 書籍の「形態」を構成する要素

初めに，書籍の「形態」という側面について示す。1 冊の書籍の形態は，おおむね，図 3.4 のように分類できる。

[5] 書籍から得られるサンプルの多様性については，柏野ほか (2008) を参照。

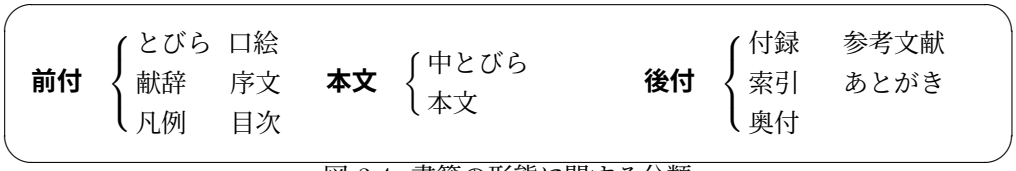


図 3.4: 書籍の形態に関する分類

このうち，主として文章表現によって実現されるのは,「序文」「本文」「あとがき」である。そこで，これらのカテゴリに相当する部分はサンプリングの対象とする。「中とびら」は章立てを表す要素の一つと考え，やはりサンプリングの対象とする。「とびら」「凡例」「目次」「参考文献」「索引」「奥付」には現代日本語が現れるものの，箇条書きであったり図的な扱いであったりする場合が多く，書き言葉コーパスに収録する対象としてはふさわしくないため，サンプリングの対象外とする。「口絵」は言語表現ではないため，対象外とする。「献辞」「付録」は，文章表現によって構成される場合とそうでない場合があるため，収録対象とするか否かは個別に判断する。

3.3.2　書籍の「版面」を構成する要素

次に，書籍の「版面」という側面について示す。書籍の版面は，おおむね，図 3.5 のような構成要素から成り立っている。

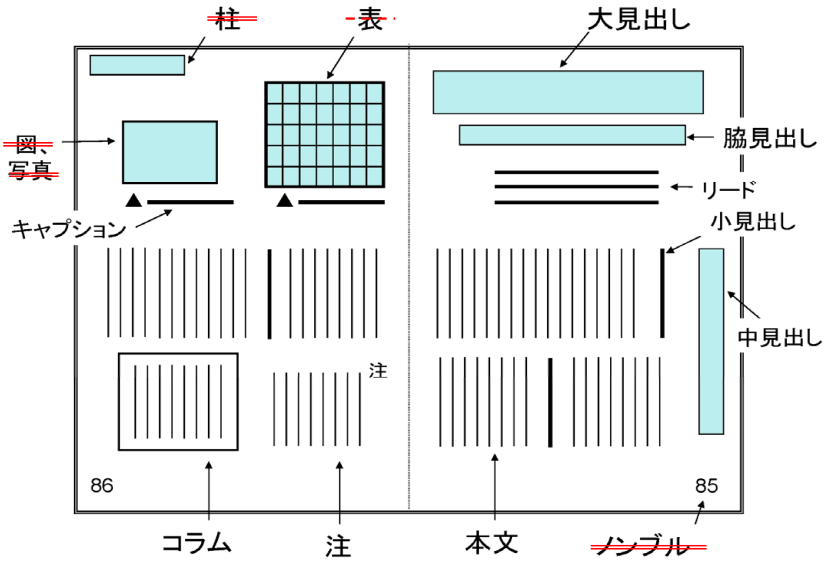


図 3.5: 書籍の版面に関する分類

このうち，主として文章表現によって実現されるのは,「大見出し」「脇見出し」「リード」「中見出し」「小見出し」「本文」「コラム」である。そこで，これらのカテゴリに相当する部分はサンプリングの対象とする。「キャプション」「注」は文章表現によって実現される場合とそう

でない場合（キャプションが「← 19,800 円」のみである場合，注が「山崎 (2007) 参照。」のみである場合など）があるが，これらは一括してサンプリングの対象とする。「図」「写真」は言語表現ではないため，サンプリングの対象から外す。仮に図・写真の中に言語表現が含まれていても，それが図・写真の範囲内にあるものであれば，一括してサンプリングの対象から外す。「柱」「ノンブル」は書籍のメタ的な構造に関わる部分であるため，サンプリングの対象から外す。「表」は，基本的にはサンプリングの対象外とする。ただし，その内部に文章表現を含み，かつそのページ全体が大きな表組みによって成立しているような場合は，表とは見なさず，サンプリングの対象とする。

3.3.3 書籍の「本文」を構成する要素

次に，書籍の版面を構成する要素のうち,「本文（ほんぶん）」部分そのものの構成について示す。本文部分は，おおむね，図 3.6 のような構成要素から成り立っている。

- 主たる文
- 箇条書き
- ルビ，グロス
- 注番号，添え字

図 3.6: 本文の構成に関する分類

「主たる文」は，本文の中でも特に主になっている文を指す。発言が引用される部分は「引用文」と呼ばれることもあるが，ここでは一括して扱う。「箇条書き」は，行頭に番号や記号などが付されてリスト状の体裁になっている部分を指す。これらの要素は，基本的にすべてサンプリングの対象とする。

3.3.4 書籍の「文字」を構成する要素

最後に，書籍の「文字」という側面について述べる。サンプリングの対象となった部分に含まれる文字は，JIS X 0213:2004 に依拠してすべて電子化されることになる[6]。このうち,「サンプル抽出基準点」を起点とした 1,000 文字の範囲を固定長サンプルとして抽出することになるが，句読点や記号などを含むすべての文字を 1,000 文字としてカウントするわけではない。文字種によって，固定長サンプルを構成する文字としてカウントするか否かを定めている。

固定長サンプル 1,000 文字のカウント対象とする文字種は，以下のようなものである。

1. 仮名文字（平仮名・片仮名・変体仮名）
2. 漢字
3. 準仮名・漢字（「ー」「々」「ゝ」等）
4. 数字（アラビア数字・ローマ数字）
5. アルファベット（ローマ字・ギリシャ文字）

[6] 再現できない漢字や記号などは，タグによって記述される。山口ほか (2008) 参照。

逆に，固定長サンプル 1,000 文字のカウント対象としない文字種は，以下のようなものである。

1. 句読点類（「,」「。」「,」「.」「…」「・」「：」「；」等）
2. 疑問符，感嘆符（「？」「！」等）
3. 括弧類（「(」「「」「{」「＜」「《」「【」「〔」等）
4. 線記号類（「-」「—」「～」等）
5. 矢印類（「→」「↓」「⇒」「⇔」等）
6. 算術記号類（「＋」「－」「×」「÷」「＝」「±」等）
7. 通貨・単位記号類（「£」「$」「¥」「%」「‰」等）
8. 音符類（「♪」等）
9. 絵文字（携帯電話の絵文字など）
10. その他記号類（「○」「▲」「※」「#」「&」「☆」等）

このような区別は，純粋な言語表現を構成する文字種に限定して 1,000 文字を取得することにより，より精密な文字調査や語彙調査を実現しようという，研究用途上の要請によるものである。

3.3.5　収録する言語表現の条件

上記までの諸基準に加えて，本文部分に含まれる言語表現そのものに関する条件が設けてある。それは，BCCWJ が「現代日本語書き言葉」のコーパスである以上，サンプルとして収録する言語表現は現代日本語として書かれたものでなければならないという条件である。したがって，以下のような表現が出現した場合，その部分はサンプリングの対象から外す。

1. 非日本語（英語，フランス語，中国語等）
2. 非現代日本語（明治元年より前に書かれた日本語）
3. 非言語（数式，化学式等）

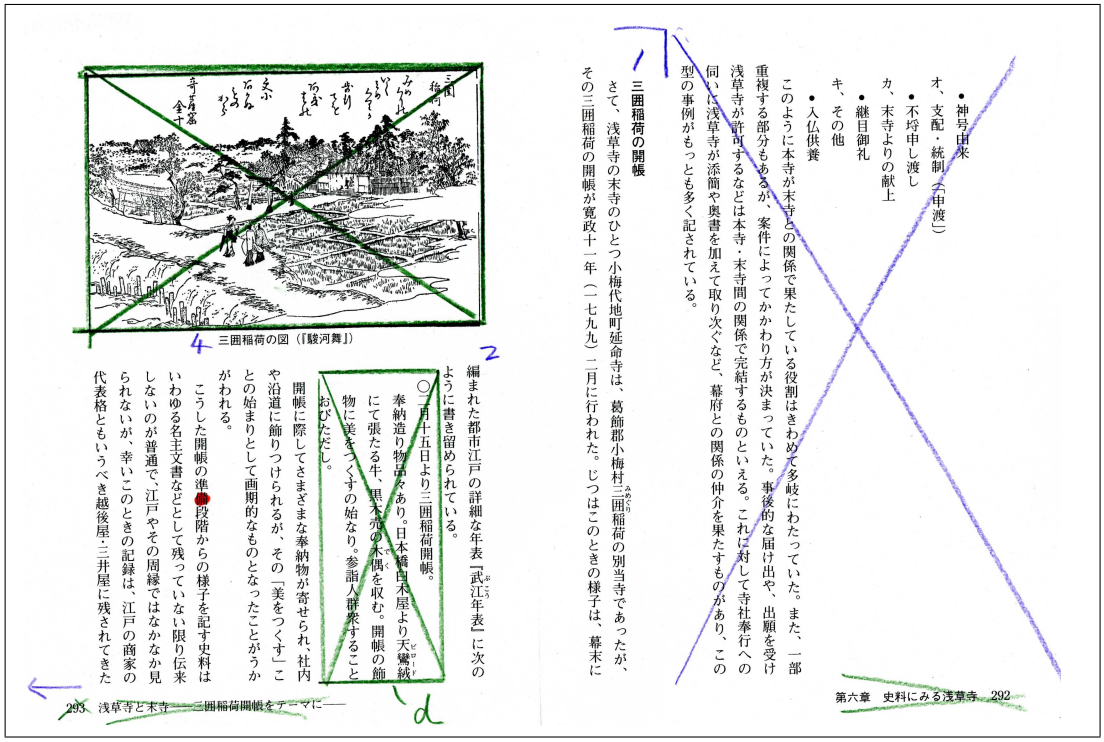
ただし，主たる文の中に非日本語・非現代日本語が混じっている場合がある。この部分までをサンプリングの対象外とすると，言わばそこだけ「穴の開いた」サンプルが多く抽出されることになり，研究用途上，好ましくない。そこで，これらの表現が主たるの文の中に現れている場合には，除外することはしないことにした。上記の各表現がサンプリングの対象外となるのは，典型的には，前後に改行を伴い，主たる文からインデントされてブロック状に引用されているような場合である。

以上，書き言葉の構造を段階的に捉え，それぞれの段階に応じて構成要素ごとに基準を立てることによって，サンプリングの対象を絞り込んでいく手続きを示した。さまざまな様式・体裁を持つ書き言葉からサンプルを抽出するためには，書き言葉の多様性に即した，分析的なサンプリング手順が必要となる。

3.4 「原サンプル」の作成

最後に，我々がどのような形で実際のサンプルを抽出・作成しているかについて触れておく。

サンプリングを実施する作業者は，サンプル台帳で指定された書籍の現物を手に取り，あらかじめ定められた手続きと基準に従って，固定長サンプル・可変長サンプルを抽出する。その結果は，書籍の該当範囲をコピーした紙面の上に転記される。この紙媒体を,「原サンプル」と呼ぶ。原サンプルの一部を，図 3.7 に示す。


●神号由来
オ、支配・統制（申渡）
●不埒申し渡し
カ、末寺よりの献上
●継目御礼
キ、その他
●入仏供養
このように本寺が末寺との関係で果たしている役割はきわめて多岐にわたっていた。また、一部重複する部分もあるが、案件によってかかわり方が決まっていた。事後的な届け出や、出願を受け浅草寺が許可するなどは本寺・末寺間の関係で完結するものといえる。これに対して寺社奉行への伺いに浅草寺が添簡や奥書を加えて取り次ぐなど、幕府との関係の仲介を果たすものがあり、この型の事例がもっとも多く記されている。

三囲稲荷の開帳

さて、浅草寺の末寺のひとつ小梅代地町延命寺は、葛飾郡小梅村三囲稲荷の別当寺であったが、その三囲稲荷の開帳が寛政十一年（一七九九）二月に行われた。じつはこのときの様子は、幕末に編まれた都市江戸の詳細な年表『武江年表』に次のように書き留められている。

○二月十五日より三囲稲荷開帳。
奉納造り物品々あり。日本橋白木屋より天鵞絨にて張たる牛、黒木綿の木偶を収む。開帳の飾物に美をつくすの始なり。参詣人群衆すること おびただし。

開帳に際してさまざまな奉納物が寄せられ、社内や沿道に飾りつけられるが、その「美をつくす」ことの始まりとして画期的なものとなったことがうかがわれる。

こうした開帳の準備段階からの様子を記す史料はいわゆる名主文書などとして残っていない限り伝来しないのが普通で、江戸やその周縁ではなかなか見られないが、幸いこのときの記録は、江戸の商家の代表格ともいうべき越後屋・三井屋に残されてきた

三囲稲荷の図（『駿河舞』）

293　浅草寺と末寺――三囲稲荷開帳をテーマに――

第六章　史料にみる浅草寺　292

図 3.7: 原サンプルの例

図 3.7 は見開き 2 ページ分の原サンプルの例である。このうち，左ページに転記された「●」の記号は,「サンプル抽出基準点」として指定された 1 文字を示す[7]。作業者は,「サンプル抽出基準点」から，3.3.4 節で示した基準に即して 1,000 文字目までを数え，固定長サンプルの範囲を確定する。この例では，1,000 文字の範囲がページ内に収まらないので，固定長サンプルの終端が次ページ以降に存在することを表す「←」の記号がページの最後に付与されている。

[7] 図 3.7 では塗りつぶされているが，実際には色鉛筆で赤く塗られている。

また，サンプル抽出基準点を含む言語的まとまりのうち，10,000 文字を上限とする最大の範囲を見定め，可変長サンプルの範囲を確定する。図 3.7 では，見開き右ページの左側に，可変長サンプルの開始点を示す記号「⎤」が付与されている。

これらの範囲指定に加えて，サンプルを一次元の文字列として把握するために，どのような順序で読み進めていくかを指示する連番が付与される。図 3.7 では図のキャプションに「4」という数字が付与されているが，これは次ページ以降にある「3」の後にキャプション部分を読み込むことを指示するものである。さらに，柱やノンブル，ブロック引用として引用されている非現代日本語の部分は，サンプリングの対象外となるので,「×」の記号が付けられる。可変長サンプルに含まれない本文部分（右ページの大半）もまた,「×」記号によって削除される。

上記のような手続きにより，出版サブコーパス・図書館サブコーパス合わせて 25,208 の原サンプルを確保すべく，サンプリング作業を続けている。

以上，本報告書では,『現代日本語書き言葉均衡コーパス』の全体構成，および「出版サブコーパス」「図書館サブコーパス」の設計とサンプリングの手順について確認した。また,「東京都内公立図書館の共通蔵書調査」について報告し,「図書館サブコーパス」の母集団の定義,およびサンプル構成比の算出法について示した。さらに，サンプリング用の台帳を作成し，母集団に含まれる全ての文字から特定の 1 文字を指定して 2 種類のサンプルを抽出するまでの手続きについて示した。これにより，出版サブコーパス・図書館サブコーパスの母集団を定義し，そこからサンプル台帳（抽出台帳）に基づいてランダムにサンプルを抽出するまでの，一連の手順が示された。

なお，具体的な紙面からコーパスに格納する言語表現をどのような手続き・基準で抽出するかについては，さまざまな事例をもとに詳細な基準を立てる必要がある。これについては別稿に委ねる。また，書籍の版面上に現れた書き言葉の多様な論理構造・体裁の分類については,柏野ほか (2008) を参照されたい。

参考文献

柏野和佳子・丸山岳彦・秋元祐哉・稲益佐知子・佐野大樹・田中弥生・山崎誠 (2008). 『『現代日本語書き言葉均衡コーパス』における書籍サンプルの多様性』, 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-07-02), 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 (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・柏野和佳子・稲益佐知子・秋元祐哉・吉田谷幸宏・山崎誠 (2007) 書き言葉の構造を捉える —書き言葉の多様な構造とサンプリング手法—. 『言語処理学会 第 13 回年次大会 発表論文集』. 言語処理学会.

丸山岳彦・柏野和佳子・山崎誠・佐野大樹・秋元祐哉・稲益佐知子・吉田谷幸宏 (2007)「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概要. 『特定領域「日本語コーパス」平成 18 年度公開ワークショップ（研究成果報告会）予稿集』. 国立国語研究所.

丸山岳彦・柏野和佳子・山崎誠・佐野大樹・秋元祐哉・稲益佐知子・田中弥生 (2008) 「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概要 (2) —流通実態サブコーパスの設計—. 『特定領域「日本語コーパス」平成 19 年度公開ワークショップ（研究成果報告会）予稿集』. 国立国語研究所.

山口昌也・高田智和・北村雅則・間淵洋子・小林正行・西部みちる (2008) 『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.2.0. 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-07-03), 特定領域研究「日本語コーパス」データ班.

山崎誠 (2007) 『現代日本語書き言葉均衡コーパス』の基本設計について. 『特定領域「日本語コーパス」平成 18 年度公開ワークショップ（研究成果報告会）予稿集』. 国立国語研究所

第II部

資料編

第4章　資料集

表 4.1: 書籍の総文字数推計基準表

⟹ 参照元 2.4 節

NDC	15cm	16cm	17cm	18cm	19cm	20cm	21cm	22cm	23cm
0. 総記	*410.8*	*442.0*	*473.3*	530.8	557.8	389.8	660.9	502.2	591.2
1. 哲学	544.0	418.5	*473.3*	542.2	497.0	529.6	743.9	655.8	*660.8*
2. 歴史	516.0	434.0	*473.3*	466.6	471.9	529.8	652.9	747.1	*660.8*
3. 社会科学	413.8	*442.0*	*473.3*	456.0	602.5	603.0	618.6	885.9	*660.8*
4. 自然科学	*410.8*	*442.0*	290.8	457.0	469.1	631.6	680.8	801.1	*660.8*
5. 技術工学	430.4	396.0	*473.3*	340.8	409.9	539.0	672.8	815.5	*660.8*
6. 産業	*410.8*	*442.0*	*473.3*	558.8	513.9	511.6	448.3	1192.7	*660.8*
7. 芸術	503.6	492.8	*473.3*	451.2	573.4	654.5	667.1	638.8	*660.8*
8. 言語	600.6	345.0	*473.3*	522.2	632.0	561.6	910.4	1052.0	*660.8*
9. 文学	435.2	487.2	*473.3*	482.8	447.8	501.4	753.3	585.3	*660.8*
n. null	*410.8*	*442.0*	*473.3*	*504.5*	*535.8*	*567.0*	*598.3*	*629.5*	*660.8*

NDC	24cm	25cm	26cm	27cm	28cm	29cm	30cm	31cm	null
0. 総記	467.6	*723.3*	1026.4	*785.8*	*817.1*	*848.3*	569.4	*910.8*	*629.9*
1. 哲学	*692.1*	*723.3*	*754.6*	*785.8*	*817.1*	*848.3*	*879.6*	*910.8*	*675.1*
2. 歴史	*692.1*	*723.3*	1234.6	851.6	*817.1*	*848.3*	*879.6*	*910.8*	*700.6*
3. 社会科学	*692.1*	*723.3*	1272.6	*785.8*	*817.1*	*848.3*	1674.1	*910.8*	*757.6*
4. 自然科学	*692.1*	*723.3*	1159.7	473.6	*817.1*	*848.3*	580.4	*910.8*	*650.0*
5. 技術工学	860.8	*723.3*	702.4	544.8	*817.1*	636.2	962.5	*910.8*	*641.0*
6. 産業	*692.1*	*723.3*	1254.6	1123.2	*817.1*	*848.3*	441.5	*910.8*	*707.2*
7. 芸術	*692.1*	*723.3*	318.5	*785.8*	*817.1*	955.8	1510.8	778.8	*688.1*
8. 言語	*692.1*	*723.3*	506.6	*785.8*	*817.1*	*848.3*	*879.6*	*910.8*	*701.3*
9. 文学	*692.1*	*723.3*	*754.6*	*785.8*	*817.1*	*848.3*	*879.6*	*910.8*	*661.1*
null	*692.1*	*723.3*	*754.6*	*785.8*	*817.1*	*848.3*	*879.6*	*910.8*	*660.8*

※ 表中の斜体は推計値，その他は実測値を表す。前報告書の第 3 章を参照。

表 4.2: 出版サブコーパスの母集団（発行年・NDC による層別）その 1

⟹ 参照元 1.2 節

		2001 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	2,503	0.59%	667,715	0.77%	382,248,739	0.58%
	1. 哲学	3,474	0.82%	900,008	1.04%	515,756,270	0.79%
	2. 歴史	5,206	1.23%	1,391,806	1.61%	938,250,607	1.43%
	3. 社会科学	12,813	3.03%	3,279,637	3.79%	2,536,580,019	3.87%
	4. 自然科学	5,469	1.29%	1,295,428	1.50%	959,112,273	1.46%
	5. 技術	7,211	1.71%	1,539,416	1.78%	1,079,958,087	1.65%
	6. 産業	3,278	0.78%	709,456	0.82%	483,579,299	0.74%
	7. 芸術	5,298	1.25%	1,083,312	1.25%	682,099,095	1.04%
	8. 言語	1,317	0.31%	304,599	0.35%	224,414,946	0.34%
	9. 文学	13,488	3.19%	3,558,890	4.11%	1,772,246,415	2.71%
	n. 記録なし	2,914	0.69%	400,720	0.46%	258,618,003	0.40%
書籍 小計		62,971	14.90%	15,130,987	17.4%	9,832,863,752	15.02%
雑誌	1. 総合	7,585	1.80%	1,371,908	1.59%	1,431,244,354	2.19%
	2. 教育・学芸	1,100	0.26%	200,884	0.23%	180,149,335	0.28%
	3. 政治・経済・商業	656	0.16%	92,560	0.11%	89,172,756	0.14%
	4. 産業	125	0.03%	25,116	0.03%	24,668,378	0.04%
	5. 工業	1,701	0.40%	357,048	0.41%	352,242,172	0.54%
	6. 厚生・医療	203	0.05%	34,272	0.04%	34,621,619	0.05%
雑誌 小計		11,370	2.69%	2,081,788	2.41%	2,112,098,615	3.23%
新聞	朝日新聞	638	0.15%	18,266	0.02%	92,389,283	0.14%
	毎日新聞	638	0.15%	13,864	0.02%	79,815,633	0.12%
	読売新聞	638	0.15%	18,657	0.02%	91,447,895	0.14%
	日本経済新聞	638	0.15%	18,902	0.02%	137,621,605	0.21%
	産経新聞	638	0.15%	15,605	0.02%	82,250,076	0.13%
	北海道新聞	638	0.15%	16,176	0.02%	84,146,071	0.13%
	中日新聞	638	0.15%	17,308	0.02%	91,026,288	0.14%
	西日本新聞	638	0.15%	16,176	0.02%	84,146,071	0.13%
	河北新報	638	0.15%	14,334	0.02%	73,737,775	0.11%
	新潟日報	638	0.15%	13,624	0.02%	70,209,696	0.11%
	京都新聞	638	0.15%	13,668	0.02%	71,551,663	0.11%
	神戸新聞	638	0.15%	14,375	0.02%	75,434,008	0.12%
	中国新聞	638	0.15%	13,624	0.02%	70,209,696	0.11%
	高知新聞	638	0.15%	11,494	0.01%	59,625,460	0.09%
	愛媛新聞	355	0.08%	9,940	0.01%	49,393,104	0.08%
	琉球新報	638	0.15%	13,624	0.02%	70,209,696	0.11%
新聞 小計		9,925	2.35%	239,638	0.28%	1,283,214,023	1.96%

表 4.3: 出版サブコーパスの母集団（発行年・NDC による層別）その 2

⟹ 参照元 1.2 節

		2002 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	2,346	0.56%	633,884	0.73%	360,518,330	0.55%
	1. 哲学	3,635	0.86%	929,325	1.07%	537,184,714	0.82%
	2. 歴史	4,950	1.17%	1,287,397	1.49%	858,931,868	1.31%
	3. 社会科学	12,938	3.06%	3,313,592	3.83%	2,548,447,478	3.89%
	4. 自然科学	5,733	1.36%	1,357,459	1.57%	1,013,964,132	1.55%
	5. 技術	6,692	1.58%	1,424,416	1.65%	996,560,860	1.52%
	6. 産業	3,083	0.73%	659,593	0.76%	432,238,970	0.66%
	7. 芸術	5,199	1.23%	1,069,813	1.24%	678,634,912	1.04%
	8. 言語	1,143	0.27%	260,842	0.30%	191,010,162	0.29%
	9. 文学	14,217	3.36%	3,710,845	4.29%	1,838,128,351	2.81%
	n. 記録なし	4,417	1.05%	666,743	0.77%	470,198,491	0.72%
書籍 小計		64,353	15.23%	15,313,909	17.70%	9,925,818,268	15.16%
雑誌	1. 総合	7,651	1.81%	1,422,765	1.64%	1,475,024,010	2.25%
	2. 教育・学芸	1,148	0.27%	197,584	0.23%	176,544,881	0.27%
	3. 政治・経済。商業	692	0.16%	98,850	0.11%	97,193,945	0.15%
	4. 産業	117	0.03%	24,636	0.03%	23,580,545	0.04%
	5. 工業	1,514	0.36%	316,861	0.37%	312,234,063	0.48%
	6. 厚生・医療	239	0.06%	42,960	0.05%	41,224,514	0.06%
雑誌 小計		11,361	2.69%	2,103,656	2.43%	2,125,801,958	3.25%
新聞	朝日新聞	638	0.15%	18,266	0.02%	92,389,283	0.14%
	毎日新聞	638	0.15%	13,864	0.02%	79,815,633	0.12%
	読売新聞	638	0.15%	18,657	0.02%	91,447,895	0.14%
	日本経済新聞	638	0.15%	18,902	0.02%	137,621,605	0.21%
	産経新聞	638	0.15%	15,605	0.02%	82,250,076	0.13%
	北海道新聞	638	0.15%	16,176	0.02%	84,146,071	0.13%
	中日新聞	638	0.15%	17,308	0.02%	91,026,288	0.14%
	西日本新聞	638	0.15%	16,176	0.02%	84,146,071	0.13%
	河北新報	638	0.15%	14,334	0.02%	73,737,775	0.11%
	新潟日報	638	0.15%	13,624	0.02%	70,209,696	0.11%
	京都新聞	638	0.15%	13,668	0.02%	71,551,663	0.11%
	神戸新聞	638	0.15%	14,375	0.02%	75,434,008	0.12%
	中国新聞	638	0.15%	13,624	0.02%	70,209,696	0.11%
	高知新聞	638	0.15%	11,494	0.01%	59,625,460	0.09%
	愛媛新聞	355	0.08%	9,940	0.01%	49,393,104	0.08%
	琉球新報	638	0.15%	13,624	0.02%	70,209,696	0.11%
新聞 小計		9925	2.35%	239,638	0.28%	1283214023	1.96%

表 4.4: 出版サブコーパスの母集団（発行年・NDC による層別）その 3

⟹ 参照元 1.2 節

		2003 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	2,248	0.53%	584,306	0.68%	333,607,966	0.51%
	1. 哲学	3,616	0.86%	887,380	1.03%	506,856,721	0.77%
	2. 歴史	5,121	1.21%	1,310,844	1.51%	873,088,165	1.33%
	3. 社会科学	13,292	3.15%	3,395,390	3.92%	2,620,601,954	4.00%
	4. 自然科学	6,041	1.43%	1,446,005	1.67%	1,087,798,636	1.66%
	5. 技術	6,424	1.52%	1,406,853	1.63%	973,775,399	1.49%
	6. 産業	3,152	0.75%	674,493	0.78%	444,077,805	0.68%
	7. 芸術	5,262	1.25%	1,067,470	1.23%	675,617,526	1.03%
	8. 言語	942	0.22%	214,818	0.25%	156,449,050	0.24%
	9. 文学	15,192	3.60%	3,905,502	4.51%	1,937,829,829	2.96%
	n. 記録なし	4,428	1.05%	651,185	0.75%	498,946,836	0.76%
書籍 小計		65,718	15.55%	15,544,246	17.97%	10,108,649,887	15.44%
雑誌	1. 総合	7,668	1.81%	1,441,592	1.67%	1,486,735,314	2.27%
	2. 教育・学芸	1,179	0.28%	212,468	0.25%	189,340,521	0.29%
	3. 政治・経済。商業	630	0.15%	95,508	0.11%	92,843,411	0.14%
	4. 産業	131	0.03%	24,180	0.03%	23,082,064	0.04%
	5. 工業	1,349	0.32%	284,317	0.33%	276,547,213	0.42%
	6. 厚生・医療	210	0.05%	37,152	0.04%	35,586,644	0.05%
雑誌 小計		11,167	2.64%	2,095,217	2.42%	2,104,135,166	3.21%
新聞	朝日新聞	638	0.15%	18,266	0.02%	92,389,283	0.14%
	毎日新聞	638	0.15%	13,864	0.02%	79,815,633	0.12%
	読売新聞	638	0.15%	18,657	0.02%	91,447,895	0.14%
	日本経済新聞	638	0.15%	18,902	0.02%	137,621,605	0.21%
	産経新聞	638	0.15%	15,605	0.02%	82,250,076	0.13%
	北海道新聞	638	0.15%	16,176	0.02%	84,146,071	0.13%
	中日新聞	638	0.15%	17,308	0.02%	91,026,288	0.14%
	西日本新聞	638	0.15%	16,176	0.02%	84,146,071	0.13%
	河北新報	638	0.15%	14,334	0.02%	73,737,775	0.11%
	新潟日報	638	0.15%	13,624	0.02%	70,209,696	0.11%
	京都新聞	638	0.15%	13,668	0.02%	71,551,663	0.11%
	神戸新聞	638	0.15%	14,375	0.02%	75,434,008	0.12%
	中国新聞	638	0.15%	13,624	0.02%	70,209,696	0.11%
	高知新聞	638	0.15%	11,494	0.01%	59,625,460	0.09%
	愛媛新聞	355	0.08%	9,940	0.01%	49,393,104	0.08%
	琉球新報	638	0.15%	13,624	0.02%	70,209,696	0.11%
新聞 小計		9925	2.35%	239,638	0.28%	1283214023	1.96%

表 4.5: 出版サブコーパスの母集団（発行年・NDC による層別）その 4

⟹ 参照元 1.2 節

		2004 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	2,185	0.52%	540,869	0.63%	311,214,584	0.48%
	1. 哲学	3,993	0.95%	1,007,377	1.16%	579,725,751	0.89%
	2. 歴史	5,137	1.22%	1,338,088	1.55%	893,356,324	1.36%
	3. 社会科学	12,924	3.06%	3,275,960	3.79%	2,559,493,082	3.91%
	4. 自然科学	6,212	1.47%	1,451,958	1.68%	1,082,972,822	1.65%
	5. 技術	6,012	1.42%	1,263,110	1.46%	861,521,544	1.32%
	6. 産業	3,181	0.75%	690,668	0.80%	463,080,360	0.71%
	7. 芸術	5,200	1.23%	1,045,430	1.21%	662,141,479	1.01%
	8. 言語	974	0.23%	225,671	0.26%	172,823,399	0.26%
	9. 文学	16,011	3.79%	4,034,730	4.66%	1,991,206,040	3.04%
	n. 記録なし	4,897	1.16%	735,215	0.85%	562,058,122	0.86%
書籍 小計		66,726	15.79%	15,609,076	18.04%	10,139,593,508	15.49%
雑誌	1. 総合	7,692	1.82%	1,452,792	1.68%	1,505,343,555	2.30%
	2. 教育・学芸	1,005	0.24%	186,446	0.22%	166,023,480	0.25%
	3. 政治・経済。商業	593	0.14%	88,260	0.10%	85,924,961	0.13%
	4. 産業	119	0.03%	21,580	0.02%	20,591,796	0.03%
	5. 工業	1,324	0.31%	279,794	0.32%	274,676,145	0.42%
	6. 厚生・医療	214	0.05%	37,520	0.04%	35,806,010	0.05%
雑誌 小計		10,947	2.59%	2,066,392	2.39%	2,088,365,948	3.19%
新聞	朝日新聞	638	0.15%	18,266	0.02%	92,389,283	0.14%
	毎日新聞	638	0.15%	13,864	0.02%	79,815,633	0.12%
	読売新聞	638	0.15%	18,657	0.02%	91,447,895	0.14%
	日本経済新聞	638	0.15%	18,902	0.02%	137,621,605	0.21%
	産経新聞	638	0.15%	15,605	0.02%	82,250,076	0.13%
	北海道新聞	638	0.15%	16,176	0.02%	84,146,071	0.13%
	中日新聞	638	0.15%	17,308	0.02%	91,026,288	0.14%
	西日本新聞	638	0.15%	16,176	0.02%	84,146,071	0.13%
	河北新報	638	0.15%	14,334	0.02%	73,737,775	0.11%
	新潟日報	638	0.15%	13,624	0.02%	70,209,696	0.11%
	京都新聞	638	0.15%	13,668	0.02%	71,551,663	0.11%
	神戸新聞	638	0.15%	14,375	0.02%	75,434,008	0.12%
	中国新聞	638	0.15%	13,624	0.02%	70,209,696	0.11%
	高知新聞	638	0.15%	11,494	0.01%	59,625,460	0.09%
	愛媛新聞	355	0.08%	9,940	0.01%	49,393,104	0.08%
	琉球新報	638	0.15%	13,624	0.02%	70,209,696	0.11%
新聞 小計		9925	2.35%	239,638	0.28%	1283214023	1.96%

表 4.6: 出版サブコーパスの母集団（発行年・NDC による層別）その 5

⟹ 参照元 1.2 節

		2005 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	1,850	0.44%	433,019	0.50%	248,824,929	0.38%
	1. 哲学	3,349	0.79%	805,239	0.93%	458,087,357	0.70%
	2. 歴史	4,210	1.00%	1,121,037	1.30%	737,577,377	1.13%
	3. 社会科学	11,019	2.61%	2,794,537	3.23%	2,143,199,410	3.27%
	4. 自然科学	5,290	1.25%	1,221,108	1.41%	925,746,173	1.41%
	5. 技術	5,038	1.19%	1,047,540	1.21%	704,114,077	1.08%
	6. 産業	2,638	0.62%	564,103	0.65%	373,411,003	0.57%
	7. 芸術	4,428	1.05%	887,506	1.03%	559,939,435	0.86%
	8. 言語	835	0.20%	190,910	0.22%	144,102,572	0.22%
	9. 文学	14,808	3.50%	3,678,311	4.25%	1,801,864,850	2.75%
	n. 記録なし	3,884	0.92%	569,992	0.66%	436,132,755	0.67%
書籍 小計		57,349	13.57%	13,313,302	15.39%	8,532,999,935	13.03%
雑誌	1. 総合	7,789	1.84%	1,474,932	1.70%	1,523,100,574	2.33%
	2. 教育・学芸	1,024	0.24%	185,842	0.21%	165,817,376	0.25%
	3. 政治・経済。商業	597	0.14%	94,104	0.11%	91,324,332	0.14%
	4. 産業	107	0.03%	19,660	0.02%	18,718,175	0.03%
	5. 工業	1,213	0.29%	255,780	0.30%	252,593,766	0.39%
	6. 厚生・医療	206	0.05%	37,584	0.04%	33,725,726	0.05%
雑誌 小計		10,936	2.59%	2,067,902	2.39%	2,085,279,949	3.19%
新聞	朝日新聞	638	0.15%	18,266	0.02%	92,389,283	0.14%
	毎日新聞	638	0.15%	13,864	0.02%	79,815,633	0.12%
	読売新聞	638	0.15%	18,657	0.02%	91,447,895	0.14%
	日本経済新聞	638	0.15%	18,902	0.02%	137,621,605	0.21%
	産経新聞	638	0.15%	15,605	0.02%	82,250,076	0.13%
	北海道新聞	638	0.15%	16,176	0.02%	84,146,071	0.13%
	中日新聞	638	0.15%	17,308	0.02%	91,026,288	0.14%
	西日本新聞	638	0.15%	16,176	0.02%	84,146,071	0.13%
	河北新報	638	0.15%	14,334	0.02%	73,737,775	0.11%
	新潟日報	638	0.15%	13,624	0.02%	70,209,696	0.11%
	京都新聞	638	0.15%	13,668	0.02%	71,551,663	0.11%
	神戸新聞	638	0.15%	14,375	0.02%	75,434,008	0.12%
	中国新聞	638	0.15%	13,624	0.02%	70,209,696	0.11%
	高知新聞	638	0.15%	11,494	0.01%	59,625,460	0.09%
	愛媛新聞	355	0.08%	9,940	0.01%	49,393,104	0.08%
	琉球新報	638	0.15%	13,624	0.02%	70,209,696	0.11%
新聞 小計		9925	2.35%	239,638	0.28%	1283214023	1.96%

表 4.7: 図書館サブコーパスの母集団（発行年・NDC による層別）その 1

⟹ 参照元 2.5 節

		1986 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	167	0.05%	44,135	0.05%	21,575,113	0.05%
	1. 哲学	329	0.10%	92,831	0.11%	51,241,453	0.11%
	2. 歴史	707	0.21%	198,815	0.23%	112,112,029	0.23%
	3. 社会科学	878	0.26%	244,367	0.29%	153,371,935	0.32%
	4. 自然科学	436	0.13%	99,403	0.12%	59,079,266	0.12%
	5. 技術	397	0.12%	86,877	0.10%	47,893,669	0.10%
	6. 産業	224	0.07%	51,013	0.06%	31,319,389	0.07%
	7. 芸術	669	0.20%	152,773	0.18%	95,472,812	0.20%
	8. 言語	111	0.03%	30,620	0.04%	19,312,305	0.04%
	9. 文学	2,646	0.79%	791,133	0.93%	386,782,855	0.81%
	n. 記録なし	1,075	0.32%	158,199	0.19%	96,087,163	0.20%
書籍 小計		7,639	2.28%	1,950,166	2.28%	1,074,247,989	2.24%

		1987 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	141	0.04%	35,898	0.04%	17,566,258	0.04%
	1. 哲学	388	0.12%	112,109	0.13%	62,198,363	0.13%
	2. 歴史	789	0.24%	224,097	0.26%	129,083,031	0.27%
	3. 社会科学	1,178	0.35%	323,347	0.38%	207,032,821	0.43%
	4. 自然科学	527	0.16%	120,365	0.14%	70,255,393	0.15%
	5. 技術	499	0.15%	106,977	0.13%	60,014,558	0.13%
	6. 産業	328	0.10%	79,899	0.09%	52,449,453	0.11%
	7. 芸術	824	0.25%	186,502	0.22%	115,479,963	0.24%
	8. 言語	138	0.04%	36,791	0.04%	24,550,687	0.05%
	9. 文学	3,312	0.99%	1,025,368	1.20%	505,156,663	1.06%
	n. 記録なし	1,125	0.34%	170,364	0.20%	103,292,913	0.22%
書籍 小計		9,249	2.75%	2,421,717	2.84%	1,347,080,103	2.81%

		1988 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	207	0.06%	54,606	0.06%	27,445,573	0.06%
	1. 哲学	403	0.12%	114,485	0.13%	64,665,539	0.14%
	2. 歴史	984	0.29%	271,852	0.32%	158,102,407	0.33%
	3. 社会科学	1,294	0.39%	350,626	0.41%	222,895,187	0.47%
	4. 自然科学	550	0.16%	127,754	0.15%	76,691,202	0.16%
	5. 技術	567	0.17%	117,446	0.14%	67,379,325	0.14%
	6. 産業	284	0.08%	66,492	0.08%	41,378,839	0.09%
	7. 芸術	842	0.25%	195,634	0.23%	123,508,234	0.26%
	8. 言語	186	0.06%	47,657	0.06%	33,011,371	0.07%
	9. 文学	3,724	1.11%	1,141,821	1.34%	562,731,930	1.18%
	n. 記録なし	1,083	0.32%	160,826	0.19%	97,217,974	0.20%
書籍 小計		10,124	3.02%	2,649,199	3.10%	1,475,027,582	3.08%

表 4.8: 図書館サブコーパスの母集団（発行年・NDC による層別）その 2

⟹ 参照元 2.5 節

		1989 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	237	0.07%	67,685	0.08%	33,789,142	0.07%
	1. 哲学	471	0.14%	136,293	0.16%	77,943,044	0.16%
	2. 歴史	1,084	0.32%	306,846	0.36%	179,526,333	0.37%
	3. 社会科学	1,481	0.44%	407,872	0.48%	259,470,559	0.54%
	4. 自然科学	661	0.20%	157,197	0.18%	94,969,678	0.20%
	5. 技術	689	0.21%	146,878	0.17%	85,362,847	0.18%
	6. 産業	386	0.11%	91,712	0.11%	59,401,422	0.12%
	7. 芸術	978	0.29%	228,216	0.27%	144,342,671	0.30%
	8. 言語	197	0.06%	51,524	0.06%	35,283,287	0.07%
	9. 文学	4,083	1.22%	1,220,790	1.43%	602,945,850	1.26%
	n. 記録なし	1,221	0.36%	173,350	0.20%	105,481,738	0.22%
書籍 小計		11,488	3.42%	2,988,363	3.50%	1,678,516,570	3.51%

		1990 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	214	0.06%	57,811	0.07%	27,925,443	0.06%
	1. 哲学	569	0.17%	168,537	0.20%	94,948,215	0.20%
	2. 歴史	1,098	0.33%	312,071	0.37%	177,325,594	0.37%
	3. 社会科学	1,782	0.53%	494,614	0.58%	312,887,151	0.65%
	4. 自然科学	719	0.21%	168,125	0.20%	100,737,994	0.21%
	5. 技術	719	0.21%	154,320	0.18%	88,091,456	0.18%
	6. 産業	367	0.11%	82,615	0.10%	49,164,514	0.10%
	7. 芸術	1,025	0.31%	246,823	0.29%	154,534,324	0.32%
	8. 言語	200	0.06%	53,298	0.06%	36,259,930	0.08%
	9. 文学	4,782	1.42%	1,429,051	1.67%	697,840,189	1.46%
	n. 記録なし	1,357	0.40%	192,870	0.23%	118,874,316	0.25%
書籍 小計		12,832	3.82%	3,360,135	3.94%	1,858,589,127	3.88%

		1991 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	236	0.07%	65,937	0.08%	32,230,243	0.07%
	1. 哲学	621	0.18%	177,940	0.21%	99,235,670	0.21%
	2. 歴史	1,240	0.37%	349,143	0.41%	201,047,974	0.42%
	3. 社会科学	2,047	0.61%	548,049	0.64%	349,595,114	0.73%
	4. 自然科学	840	0.25%	197,266	0.23%	119,452,768	0.25%
	5. 技術	804	0.24%	160,930	0.19%	91,907,950	0.19%
	6. 産業	394	0.12%	92,563	0.11%	58,455,617	0.12%
	7. 芸術	1,113	0.33%	258,804	0.30%	163,316,060	0.34%
	8. 言語	202	0.06%	52,123	0.06%	34,897,203	0.07%
	9. 文学	5,028	1.50%	1,497,995	1.75%	730,716,012	1.53%
	n. 記録なし	1,288	0.38%	185,455	0.22%	112,552,106	0.24%
書籍 小計		13,813	4.11%	3,586,205	4.20%	1,993,406,718	4.16%

表 4.9: 図書館サブコーパスの母集団（発行年・NDC による層別）その 3

⟹ 参照元 2.5 節

		1992 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	288	0.09%	83,810	0.10%	41,738,599	0.09%
	1. 哲学	641	0.19%	180,100	0.21%	99,227,778	0.21%
	2. 歴史	1,489	0.44%	421,182	0.49%	239,601,023	0.50%
	3. 社会科学	2,244	0.67%	596,137	0.70%	378,574,068	0.79%
	4. 自然科学	975	0.29%	224,019	0.26%	134,183,944	0.28%
	5. 技術	967	0.29%	195,487	0.23%	115,141,842	0.24%
	6. 産業	436	0.13%	99,401	0.12%	60,150,468	0.13%
	7. 芸術	1,284	0.38%	310,080	0.36%	192,740,441	0.40%
	8. 言語	248	0.07%	67,127	0.08%	45,169,865	0.09%
	9. 文学	5,255	1.57%	1,550,108	1.82%	757,439,244	1.58%
	n. 記録なし	1,225	0.36%	166,786	0.20%	102,793,668	0.21%
書籍 小計		15,052	4.48%	3,894,237	4.56%	2,166,760,941	4.53%

		1993 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	352	0.10%	104,577	0.12%	50,758,548	0.11%
	1. 哲学	734	0.22%	212,817	0.25%	116,414,379	0.24%
	2. 歴史	1,656	0.49%	462,132	0.54%	261,120,682	0.55%
	3. 社会科学	2,581	0.77%	684,017	0.80%	435,958,129	0.91%
	4. 自然科学	1,096	0.33%	247,025	0.29%	145,847,588	0.30%
	5. 技術	1,104	0.33%	222,018	0.26%	130,354,549	0.27%
	6. 産業	464	0.14%	103,131	0.12%	59,334,921	0.12%
	7. 芸術	1,462	0.44%	350,045	0.41%	219,479,239	0.46%
	8. 言語	287	0.09%	75,417	0.09%	49,647,395	0.10%
	9. 文学	5,489	1.63%	1,640,827	1.92%	803,223,543	1.68%
	n. 記録なし	1,274	0.38%	173,883	0.20%	106,562,402	0.22%
書籍 小計		16,499	4.91%	4,275,889	5.01%	2,378,701,374	4.97%

		1994 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	373	0.11%	99,293	0.12%	48,894,099	0.10%
	1. 哲学	783	0.23%	222,643	0.26%	121,962,593	0.25%
	2. 歴史	1,594	0.47%	435,451	0.51%	245,381,108	0.51%
	3. 社会科学	2,658	0.79%	714,738	0.84%	455,343,330	0.95%
	4. 自然科学	1,116	0.33%	254,639	0.30%	149,445,341	0.31%
	5. 技術	1,273	0.38%	240,597	0.28%	140,424,266	0.29%
	6. 産業	565	0.17%	130,130	0.15%	78,780,223	0.16%
	7. 芸術	1,465	0.44%	347,924	0.41%	215,227,980	0.45%
	8. 言語	249	0.07%	65,000	0.08%	41,566,114	0.09%
	9. 文学	5,693	1.70%	1,708,738	2.00%	836,725,747	1.75%
	n. 記録なし	1,280	0.38%	181,163	0.21%	108,886,467	0.23%
書籍 小計		17,049	5.08%	4,400,316	5.15%	2,442,637,267	5.10%

表 4.10: 図書館サブコーパスの母集団（発行年・NDC による層別）その 4

⟹ 参照元 2.5 節

		1995 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	367	0.11%	99,597	0.12%	49,122,544	0.10%
	1. 哲学	889	0.26%	235,296	0.28%	128,397,689	0.27%
	2. 歴史	1,730	0.52%	488,617	0.57%	274,381,725	0.57%
	3. 社会科学	3,050	0.91%	805,087	0.94%	511,962,406	1.07%
	4. 自然科学	1,220	0.36%	277,476	0.33%	161,532,490	0.34%
	5. 技術	1,373	0.41%	269,982	0.32%	157,551,447	0.33%
	6. 産業	577	0.17%	133,141	0.16%	82,063,628	0.17%
	7. 芸術	1,663	0.50%	382,586	0.45%	236,859,113	0.49%
	8. 言語	314	0.09%	80,863	0.09%	52,190,195	0.11%
	9. 文学	5,941	1.77%	1,811,288	2.12%	886,726,510	1.85%
	n. 記録なし	1,409	0.42%	212,156	0.25%	128,746,856	0.27%
書籍 小計		18,533	5.52%	4,796,089	5.62%	2,669,534,603	5.58%

		1996 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	424	0.13%	116,925	0.14%	57,531,443	0.12%
	1. 哲学	988	0.29%	267,354	0.31%	145,878,255	0.30%
	2. 歴史	1,606	0.48%	439,107	0.51%	250,630,577	0.52%
	3. 社会科学	2,992	0.89%	779,427	0.91%	496,641,232	1.04%
	4. 自然科学	1,343	0.40%	309,138	0.36%	181,225,486	0.38%
	5. 技術	1,443	0.43%	276,535	0.32%	162,261,484	0.34%
	6. 産業	651	0.19%	147,145	0.17%	87,272,226	0.18%
	7. 芸術	1,757	0.52%	404,628	0.47%	249,586,740	0.52%
	8. 言語	324	0.10%	81,742	0.10%	53,344,928	0.11%
	9. 文学	5,892	1.76%	1,823,234	2.14%	893,244,191	1.87%
	n. 記録なし	1,253	0.37%	177,469	0.21%	108,409,960	0.23%
書籍 小計		18,673	5.56%	4,822,704	5.65%	2,686,026,522	5.61%

		1997 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	438	0.13%	122,780	0.14%	61,364,916	0.13%
	1. 哲学	1,024	0.31%	279,441	0.33%	152,302,984	0.32%
	2. 歴史	1,866	0.56%	509,319	0.60%	294,572,638	0.62%
	3. 社会科学	3,217	0.96%	843,914	0.99%	533,502,529	1.11%
	4. 自然科学	1,392	0.41%	324,584	0.38%	190,754,125	0.40%
	5. 技術	1,520	0.45%	287,677	0.34%	165,721,431	0.35%
	6. 産業	753	0.22%	165,814	0.19%	99,563,757	0.21%
	7. 芸術	1,773	0.53%	402,189	0.47%	246,894,232	0.52%
	8. 言語	340	0.10%	82,959	0.10%	56,102,490	0.12%
	9. 文学	5,855	1.74%	1,807,296	2.12%	884,819,865	1.85%
	n. 記録なし	1,425	0.42%	198,173	0.23%	121,317,092	0.25%
書籍 小計		19,603	5.84%	5,024,146	5.89%	2,806,916,060	5.86%

表 4.11: 図書館サブコーパスの母集団（発行年・NDC による層別）その 5

⟹ 参照元 2.5 節

		1998 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	481	0.14%	128,376	0.15%	64,665,981	0.14%
	1. 哲学	1,003	0.30%	269,552	0.32%	146,334,787	0.31%
	2. 歴史	1,702	0.51%	485,455	0.57%	282,690,454	0.59%
	3. 社会科学	3,293	0.98%	861,443	1.01%	546,884,270	1.14%
	4. 自然科学	1,301	0.39%	306,290	0.36%	183,530,389	0.38%
	5. 技術	1,747	0.52%	322,291	0.38%	191,116,210	0.40%
	6. 産業	822	0.24%	175,389	0.21%	104,399,798	0.22%
	7. 芸術	1,771	0.53%	397,280	0.47%	243,211,545	0.51%
	8. 言語	348	0.10%	88,180	0.10%	59,085,312	0.12%
	9. 文学	5,634	1.68%	1,769,735	2.07%	870,812,055	1.82%
	n. 記録なし	1,389	0.41%	188,427	0.22%	116,174,911	0.24%
書籍 小計		19,491	5.81%	4,992,418	5.85%	2,808,905,712	5.87%

		1999 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	422	0.13%	113,403	0.13%	58,041,746	0.12%
	1. 哲学	987	0.29%	265,427	0.31%	143,979,847	0.30%
	2. 歴史	1,745	0.52%	479,774	0.56%	273,328,731	0.57%
	3. 社会科学	3,430	1.02%	895,801	1.05%	565,744,176	1.18%
	4. 自然科学	1,378	0.41%	318,631	0.37%	189,478,913	0.40%
	5. 技術	1,957	0.58%	366,020	0.43%	219,630,451	0.46%
	6. 産業	817	0.24%	178,093	0.21%	107,323,803	0.22%
	7. 芸術	1,805	0.54%	406,777	0.48%	253,619,278	0.53%
	8. 言語	372	0.11%	90,423	0.11%	58,214,700	0.12%
	9. 文学	5,438	1.62%	1,730,751	2.03%	849,998,290	1.78%
	n. 記録なし	1,317	0.39%	175,020	0.21%	108,025,652	0.23%
書籍 小計		19,668	5.86%	5,020,120	5.88%	2,827,385,587	5.91%

		2000 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	526	0.16%	133,549	0.16%	68,451,907	0.14%
	1. 哲学	1,019	0.30%	269,026	0.32%	147,320,402	0.31%
	2. 歴史	1,902	0.57%	526,748	0.62%	308,262,725	0.64%
	3. 社会科学	3,366	1.00%	864,254	1.01%	536,278,117	1.12%
	4. 自然科学	1,411	0.42%	324,272	0.38%	192,085,614	0.40%
	5. 技術	2,085	0.62%	387,821	0.45%	234,585,591	0.49%
	6. 産業	932	0.28%	201,023	0.24%	118,507,241	0.25%
	7. 芸術	1,739	0.52%	386,746	0.45%	234,957,872	0.49%
	8. 言語	375	0.11%	91,888	0.11%	61,200,106	0.13%
	9. 文学	5,419	1.61%	1,720,810	2.02%	844,082,340	1.76%
	n. 記録なし	1,438	0.43%	209,526	0.25%	125,518,570	0.26%
書籍 小計		20,212	6.02%	5,115,663	5.99%	2,871,250,485	6.00%

表 4.12: 図書館サブコーパスの母集団（発行年・NDC による層別）その 6

⟹ 参照元 2.5 節

		2001 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	511	0.15%	135,359	0.16%	68,597,494	0.14%
	1. 哲学	922	0.27%	243,510	0.29%	131,882,289	0.28%
	2. 歴史	1,881	0.56%	497,578	0.58%	296,697,252	0.62%
	3. 社会科学	3,492	1.04%	901,740	1.06%	556,773,448	1.16%
	4. 自然科学	1,355	0.40%	299,792	0.35%	175,437,774	0.37%
	5. 技術	2,239	0.67%	411,877	0.48%	252,181,396	0.53%
	6. 産業	918	0.27%	198,263	0.23%	119,071,780	0.25%
	7. 芸術	1,728	0.51%	383,343	0.45%	231,717,728	0.48%
	8. 言語	403	0.12%	92,454	0.11%	61,581,160	0.13%
	9. 文学	5,141	1.53%	1,648,390	1.93%	801,210,908	1.67%
	n. 記録なし	1,555	0.46%	214,431	0.25%	132,868,489	0.28%
書籍 小計		20,145	6.00%	5,026,737	5.89%	2,828,019,718	5.91%

		2002 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	518	0.15%	132,749	0.16%	68,279,040	0.14%
	1. 哲学	995	0.30%	251,326	0.29%	135,949,102	0.28%
	2. 歴史	1,964	0.59%	498,820	0.58%	311,306,283	0.65%
	3. 社会科学	3,729	1.11%	953,133	1.12%	594,552,965	1.24%
	4. 自然科学	1,505	0.45%	331,414	0.39%	196,692,400	0.41%
	5. 技術	2,269	0.68%	393,628	0.46%	238,532,625	0.50%
	6. 産業	929	0.28%	199,035	0.23%	115,340,037	0.24%
	7. 芸術	1,782	0.53%	397,977	0.47%	241,830,486	0.51%
	8. 言語	457	0.14%	105,902	0.12%	69,260,392	0.14%
	9. 文学	5,613	1.67%	1,771,613	2.08%	860,948,799	1.80%
	n. 記録なし	1,677	0.50%	250,589	0.29%	162,218,388	0.34%
書籍 小計		21,438	6.39%	5,286,186	6.19%	2,994,910,518	6.26%

		2003 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	502	0.15%	129,411	0.15%	68,611,340	0.14%
	1. 哲学	1,044	0.31%	252,782	0.30%	137,079,726	0.29%
	2. 歴史	2,307	0.69%	565,949	0.66%	360,385,891	0.75%
	3. 社会科学	4,012	1.20%	996,487	1.17%	622,929,005	1.30%
	4. 自然科学	1,581	0.47%	335,364	0.39%	199,467,121	0.42%
	5. 技術	2,278	0.68%	389,979	0.46%	236,590,016	0.49%
	6. 産業	965	0.29%	208,264	0.24%	122,132,308	0.26%
	7. 芸術	1,831	0.55%	393,719	0.46%	239,401,851	0.50%
	8. 言語	390	0.12%	87,643	0.10%	57,602,309	0.12%
	9. 文学	5,999	1.79%	1,824,470	2.14%	888,429,633	1.86%
	n. 記録なし	630	0.19%	106,819	0.13%	78,705,041	0.16%
書籍 小計		21,539	6.42%	5,290,887	6.20%	3,011,334,241	6.29%

表 4.13: 図書館サブコーパスの母集団（発行年・NDC による層別）その 7

⟹ 参照元 2.5 節

		2004 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	541	0.16%	132,325	0.16%	72,726,904	0.15%
	1. 哲学	1,152	0.34%	281,456	0.33%	153,571,072	0.32%
	2. 歴史	2,196	0.65%	540,511	0.63%	344,478,292	0.72%
	3. 社会科学	4,022	1.20%	988,797	1.16%	620,436,942	1.30%
	4. 自然科学	1,709	0.51%	363,680	0.43%	219,887,279	0.46%
	5. 技術	2,331	0.69%	397,777	0.47%	243,910,419	0.51%
	6. 産業	1,049	0.31%	220,303	0.26%	130,064,277	0.27%
	7. 芸術	1,950	0.58%	413,269	0.48%	248,859,888	0.52%
	8. 言語	377	0.11%	84,646	0.10%	59,976,285	0.13%
	9. 文学	6,417	1.91%	1,952,396	2.29%	952,286,705	1.99%
	n. 記録なし	737	0.22%	123,727	0.14%	91,334,277	0.19%
書籍 小計		22,481	6.70%	5,498,887	6.44%	3,137,532,339	6.55%

		2005 年					
	ジャンル	冊数		ページ数		文字数	
書籍	0. 総記	493	0.15%	116,755	0.14%	64,212,546	0.13%
	1. 哲学	1,007	0.30%	245,264	0.29%	133,316,526	0.28%
	2. 歴史	1,896	0.56%	490,006	0.57%	310,714,872	0.65%
	3. 社会科学	3,704	1.10%	924,146	1.08%	585,225,007	1.22%
	4. 自然科学	1,559	0.46%	314,032	0.37%	187,521,599	0.39%
	5. 技術	2,064	0.61%	359,098	0.42%	220,492,519	0.46%
	6. 産業	920	0.27%	192,045	0.22%	113,976,780	0.24%
	7. 芸術	1,643	0.49%	341,817	0.40%	206,250,798	0.43%
	8. 言語	345	0.10%	73,017	0.09%	48,369,874	0.10%
	9. 文学	5,918	1.76%	1,795,870	2.10%	868,969,728	1.81%
	n. 記録なし	644	0.19%	110,905	0.13%	81,822,366	0.17%
書籍 小計		20,193	6.01%	4,962,955	5.81%	2,820,872,616	5.89%

付記

『現代日本語書き言葉均衡コーパス』の「出版サブコーパス」「図書館サブコーパス」を対象としたサンプリング作業では，以下の各機関より,「J-BISC」,「ISBN 総合目録」等のデータの提供，書籍の貸し出し等のご協力をいただいています。
記して感謝申し上げます。

- 国立国会図書館
- 立川市中央図書館
- 東京都立多摩図書館
- 東京都立中央図書館
- 東京都立日比谷図書館
- 日本図書館協会
- 一橋大学附属図書館　　（五十音順）

研究開発部門言語資源グループ（サンプリングサブグループ）

山崎 誠　（研究開発部門グループ長（副））
柏野 和佳子　（研究開発部門研究員）
丸山 岳彦 *　（研究開発部門研究員）
佐野 大樹　（研究開発部門特別奨励研究員）
秋元 祐哉 *　（研究開発部門研究補佐員）
稲益 佐知子　（研究開発部門研究補佐員）
田中 弥生　（研究開発部門研究補佐員）

（* は執筆者）

国立国語研究所内部報告書（LR-CCG-07-01）

『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—

平成 20 年 3 月 21 日
執 筆 者　丸山 岳彦　秋元 祐哉
発 行 者　独立行政法人国立国語研究所
〒 190–8561 東京都立川市緑町 10 番地の 2
電 話　042（540）4300（代表）

©2008 独立行政法人国立国語研究所　（平 19–9）

国立国語研究所