『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

著者	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
ページ	1-122
発行年	2011-02-25
シリーズ	国立国語研究所内部報告書 ; LR-CCG-10-02
URL	http://doi.org/10.15084/00002852

国立国語研究所内部報告書（LR-CCG-10-02）

『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

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

平成23年2月

国立国語研究所内部報告書（LR-CCG-10-02）

『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

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

平成23年2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

目 次

はじめに

2006 年度に『現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese; 以下 BCCWJ）』の構築が開始されてから，5 年が経過した。コーパス本体の構築を担う「データ班」では,「サンプリング」「著作権処理」「電子化」「形態論情報」という 4 つのサブグループに分かれて，BCCWJ の構築を分担して進めてきた。本報告書は，このうちサンプリングを担当した我々のグループ（SSG; サンプリングサブグループ）の最終報告書である。

2006 年度から活動を開始したサンプリングサブグループでは，BCCWJ を構成する 3 つのサブコーパス「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」の設計，およびサンプリングの実作業を担当してきた。特に「出版サブコーパス」「図書館サブコーパス」の作業過程では，無作為抽出によって選ばれた 3 万冊以上にもおよぶ書籍・雑誌・新聞を入手し，そこに現れた「現代日本語」をサンプリングするという作業を継続してきた。前例のないこのような作業の実施は時に困難を極めたが，原本の入手方法を模索したり，サンプリングの基準と手順を探索的に規定したりしながら，着実に結果を積み重ねてきた。2011 年 1 月現在，当初の設計方針に基づいて継続してきたサンプリング作業は，すべて完了している。

また，サンプリングの実作業と並行して，コーパスに格納されたサンプルの出自を表わすデータベース「書誌情報データ」の設計と実装を進めてきた。あるサンプルに関する書誌情報—例えば，書籍のタイトル，編著者，発行年，出版社，ジャンル，といったような情報— をデータベース化しておくことにより，コーパスをより柔軟に検索したり，コーパスの検索結果を書誌情報と関連づけて解釈したりすることができる。コーパス本体のデータと書誌情報データを関連付けて利用することにより，均衡コーパスの持つ真価が発揮されると言えるだろう。

さらに，サンプリングの設計方法や抽出基準，その過程で生じた問題点，または完成したデータを用いた分析の結果などについて，論文を執筆したり，学会や研究会，ワークショップなどで発表したりすることによって，その成果を対外的に発信してきた。これらは，サンプリングサブグループによる研究成果である。

以下，本報告書の構成を示す。第 I 部では BCCWJ に含まれるサンプルの全体像について示す。第 II 部では「書誌情報データ」の設計と実装方法について示す。第 III 部は，サンプリングサブグループから発表された研究成果の一覧とその一部の再掲である。これら 3 点について報告することで，サンプリングサブグループの活動の最終報告書とする。

なお，2006 年以降，サンプリングサブグループにスタッフとして参加したのは，秋元祐哉，稲益佐知子，大矢内夢子，柏野和佳子，佐野大樹，田中弥生，丸山岳彦，山崎誠，吉田谷幸宏の 9 名であった。安部達雄，市原乃奈，井上陽子，遠藤直子，久古直，佐藤真奈美，志賀里美，田口久美子，田中美恵子，立花幸子，趙恩英，長門美帆子，服部紀子，三浦智子，保田祥，吉田奈央らが，アルバイトとして，これを助けた。

謝辞

BCCWJ のサンプリング作業を実施するにあたり，以下の各機関・各社より多大なご協力をいただきました。記して感謝申し上げます。

大阪市立中央図書館，オリオン書房，学習研究社，国立国会図書館，
埼玉県立浦和図書館，埼玉県立久喜図書館，埼玉県立熊谷図書館，
自治大学校図書室，　小学館，　湘北短期大学図書館，　高原書店，
立川市図書館，　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，　八王子市図書館，
一橋大学附属図書館，　ヤフー株式会社，　横浜市中央図書館

（五十音順）

第I部

BCCWJに含まれるサンプル

第1章　BCCWJの基本構成

1.1　BCCWJを構成する3つのサブコーパス

BCCWJは,「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3つのサブコーパス（以下，SCと略記する）から構成される。BCCWJの内部構成を，図1.1に示す。

図 1.1: BCCWJの内部構成

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。

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。

特定目的SCは，出版・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

期間はメディアによって異なる。

1.2　BCCWJ を構成する 2 種類のサンプル

BCCWJ に収録されるサンプルには,「**固定長サンプル**」「**可変長サンプル**」という 2 種類がある。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。

- **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ような設計にする。
- **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。

「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 文字目まで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均衡コーパスとしての性格を強く持つ。

「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，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，3 つの SC 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なう可能性の高い 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。

第2章　3つのSCの設計とサンプリングの結果

2.1　「出版SC」「図書館SC」の設計とサンプリングの結果

2.1.1　「出版SC」「図書館SC」の設計方針

BCCWJの設計時において，出版SC・図書館SCの設計方針を，以下のように定めた。

- 「出版SC」は，2001年から2005年までに国内で発行された書籍・雑誌・新聞を対象とし，そこに含まれる総文字数（推計65,471,677,099文字）によって母集団を定義する。
- 「図書館SC」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13自治体以上の公立図書館で共通に所蔵されている書籍を対象とし，そこに含まれる総文字数（推計47,877,656,072文字）によって母集団を定義する。
- 母集団を「ジャンル」「発行年」によって層別し，層別ランダムサンプリングを実施する。
- 母集団の中から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，そこから1,000文字の範囲を「固定長サンプル」として取得する。また,「サンプル抽出基準点」を含む章や節のまとまりを「可変長サンプル」として取得する。
- 「出版SC」の固定長サンプルを1,000万語取得することを基準として，各層に含まれる文字数の比例割当により，各層から取得するサンプル数を定める。

上記の方針に基づき，取得するサンプル数とそこから得られる固定長サンプル・可変長サンプルの語数を，表2.1のように試算した。この際，可変長サンプルの平均文字数を，書籍で3,900文字，雑誌で3,000文字，新聞で1,000文字と仮定した。また，1語は1.7文字で構成されると仮定した。

この試算により，出版SCでは約3,500万語，図書館SCでは約3,000万語が取得できることになり，特定目的SCの約3,500万語と合計して，BCCWJ全体を構成する語数である「1億語」を達成することができると見積もった。

表 2.1: 出版 SC・図書館 SC の設計

SC	メディア	サンプル数	固定長サンプル語数	可変長サンプル語数
出版 SC	書籍	12,604	7,414,118	28,915,059
	雑誌	2,730	1,605,882	4,817,647
	新聞	1,666	980,000	980,000
	合計	17,000	10,000,000	34,712,706
図書館 SC	書籍	12,604	7,414,118	28,915,059

2.1.2　作業の進捗に伴う設計の見直し

2006 年度からサンプリングの設計を開始し，以降 5 年間，ランダムに選ばれた書籍・雑誌・新聞を入手してサンプルを抽出する作業を継続した。この結果が電子テキスト化され，サンプルの数が蓄積されることにより，可変長サンプルの平均文字数について正確な見積もりが得られるようになった。これによると，可変長サンプルの平均文字数は，書籍で平均 4,534 文字，雑誌で平均 3,873 文字，新聞で平均 980 文字となり，新聞を除いて当初の見積もりを上回る結果となった。このため，設計通りに出版 SC で 17,000 サンプル，図書館 SC で 12,604 サンプルを取得すると，可変長サンプル全体の語数が大幅に増大してしまう見込みとなった。そこで，当初の設計の 80%が達成されていることを最低条件として，当初に見積もった取得サンプル数を下方修正した。

2.1.3　サンプリングの最終結果

サンプリング作業の完了が近づくにつれて，最終的に公開されるサンプルが当初に設計した構成比になるべく近似するように，各層から取得するサンプル数を細かく調整した。例えば，当初から予想されたことであるが，著作権処理の過程において著作権者から利用を拒否する旨の回答が来たため，公開することができなくなったサンプルが多数生じた。そこで，サンプリング作業の進捗にあわせて各層の「許諾率」を計算し，許諾率の低い層からは当初の計画より多めにサンプルを取得するよう調整しながら作業を進めた。書籍・雑誌・新聞のメディア別，ジャンル別，発行年別に層を分けた上で，各層の構成比，および許諾率を計算し，当初の設計から不足している層には必要な数のサンプルを補填した。全体の構成比を見極めながら微調整を進め，2010 年 5 月をもって，当初に設計した構成比に可能な限り近似させた形で公開可能な候補を絞り込み，サンプリング作業を完了することができた。

サンプリングの最終結果から，表 2.1 に相当する部分のみを示すと，表 2.2 のようになる。

最終的に取得したサンプル数は，当初の設計に対して，出版 SC の書籍で 89.0%，雑誌で 91.0%，新聞で 89.4%，という結果になった。全体の取得サンプル数を算出する基準とした，「出版 SC」の固定長サンプルを 1,000 万語取得するという点については，最終的には 89.3%の

表 2.2: 出版 SC・図書館 SC のサンプリング結果

SC	メディア	サンプル数	固定長サンプル語数	可変長サンプル語数
出版 SC	書籍	11,212 (89.0%)	6,595,294 (89.0%)	29,541,361 (102.2%)
	雑誌	2,483 (91.0%)	1,460,588 (90.9%)	5,687,485 (118.0%)
	新聞	1,490 (89.4%)	876,471 (89.4%)	864,364 (88.1%)
	合計	15,185 (89.3%)	8,932,353 (89.3%)	36,093,211 (104.0%)
図書館 SC	書籍	11,242 (89.2%)	6,612,941 (89.2%)	30,053,412 (103.9%)

※ 下段は当初の設計に対する達成率

約 893 万語となった。図書館 SC においても，89.2%というほぼ同等の結果となった。一方，可変長サンプルの語数は，当初の設計に対して，出版 SC の書籍で 102.2%，雑誌で 118.0%，新聞で 88.1%，図書館 SC の書籍で 103.9%という結果になり，新聞のみ設計を下回ったものの，全体的には当初の設計を上回る語数が得られた。

サンプリングの設計時におけるサンプル数と語数の試算，およびその最終結果について，出版 SC・図書館 SC のジャンル別に，表 2.3，2.4 に示す。列名にある「S」は「サンプル」を表わす。さらに，出版年（出版 SC は 2001 年から 2005 年までの 5 期，図書館 SC は 1986 年から 2005 年の 20 年間を 5 年刻みで分けた 4 期）およびジャンルごとのサンプル数と語数の試算，およびその最終結果としての達成率について，表 2.5 から表 2.13 に示す。

2.1.4 著作権処理と公開サンプル数

先述のとおり，取得した全サンプルのうち，公開対象となるのは著作権処理を経て公開可能と判断されたもののみであり，表 2.2 に示したすべてのサンプルが公開されるわけではない。したがって，公開サンプル数は表 2.2 の数値を下回ることになる。特に雑誌については，一定量のサンプルを取得した後，特定の出版社が出版した雑誌のすべてについて利用を拒否する旨の連絡が来たケースもあった。雑誌の達成率が他のメディアに比べて若干高いのは，その分を補正したことによる。

表 2.3: サンプリングの設計時におけるサンプル数と語数の試算，およびその最終結果（出版 SC 全体）

	ジャンル	設計時					最終結果					
		S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	達成率
書籍	0. 総記	425	250,000	2.5%	3,900	975,000	363	213,529	2.4%	3,902	833,197	85.4%
	1. 哲学	674	396,471	4.0%	3,900	1,546,235	610	358,824	4.0%	4,155	1,490,930	90.5%
	2. 歴史	1,117	657,059	6.6%	3,900	2,562,529	926	544,706	6.1%	4,493	2,447,545	82.9%
	3. 社会科学	3,222	1,895,294	19.0%	3,900	7,391,647	2,721	1,600,588	17.9%	4,495	7,194,570	84.5%
	4. 自然科学	1,316	774,118	7.7%	3,900	3,019,059	1,119	658,235	7.4%	4,021	2,646,734	85.0%
	5. 技術工学	1,199	705,294	7.1%	3,900	2,750,647	1,008	592,941	6.6%	4,127	2,447,023	84.1%
	6. 産業	570	335,294	3.4%	3,900	1,307,647	480	282,353	3.2%	4,366	1,232,742	84.2%
	7. 芸術	846	497,647	5.0%	3,900	1,940,824	728	428,235	4.8%	4,225	1,809,129	86.1%
	8. 言語	231	135,882	1.4%	3,900	529,941	198	116,471	1.3%	4,001	466,008	85.7%
	9. 文学	2,426	1,427,059	14.3%	3,900	5,565,529	2,557	1,504,118	16.8%	5,070	7,625,880	105.4%
	n. 記録なし	578	340,000	3.4%	3,900	1,326,000	502	295,294	3.3%	4,564	1,347,602	86.9%
	小計	12,604	7,414,118	74.1%	—	28,915,059	11,212	6,595,294	73.8%	—	29,541,361	89.0%
雑誌	1. 総合	1,927	1,133,529	11.3%	3,000	3,400,588	1,786	1,050,588	11.8%	3,914	4,111,719	92.7%
	2. 教育	228	134,118	1.3%	3,000	402,353	193	113,529	1.3%	4,163	472,600	84.6%
	3. 政治	119	70,000	0.7%	3,000	210,000	114	67,059	0.8%	3,105	208,197	95.8%
	4. 産業	29	17,059	0.2%	3,000	51,176	25	14,706	0.2%	2,258	33,200	86.2%
	5. 工業	381	224,118	2.2%	3,000	672,353	323	190,000	2.1%	4,159	790,200	84.8%
	6. 厚生	47	27,647	0.3%	3,000	82,941	42	24,706	0.3%	2,897	71,569	89.4%
	小計	2,730	1,606,471	16.1%	—	4,819,412	2,483	1,460,588	16.4%	—	5,687,485	91.0%
新聞	全国紙	628	369,412	3.7%	1,000	369,412	550	323,529	3.6%	1,069	345,956	87.6%
	ブロック紙	337	198,235	2.0%	1,000	198,235	305	179,412	2.0%	903	162,057	90.5%
	地方紙	702	412,941	4.1%	1,000	412,941	635	373,529	4.2%	954	356,351	90.5%
	小計	1,666	980,588	9.8%	—	980,588	1,490	876,471	9.8%	—	864,364	89.4%
合計		17,000	10,000,000	100%	—	34,715,059	15,185	8,932,353	100%	—	36,093,211	89.3%

表 2.4: サンプリングの設計時におけるサンプル数と語数の試算，およびその最終結果（図書館 SC 全体）

	ジャンル	設計時					最終結果					
		S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	達成率
書籍	0. 総記	263	154,706	2.1%	3,900	603,353	249	146,471	2.2%	4,108	601,669	94.7%
	1. 哲学	617	362,941	4.9%	3,900	1,415,471	560	329,412	5.0%	4,452	1,466,585	90.8%
	2. 歴史	1,321	777,059	10.5%	3,900	3,030,529	1,133	666,471	10.1%	4,587	3,056,778	85.8%
	3. 社会科学	2,356	1,385,882	18.7%	3,900	5,404,941	2,195	1,291,176	19.5%	4,427	5,716,463	93.2%
	4. 自然科学	797	468,824	6.3%	3,900	1,828,412	663	390,000	5.9%	4,315	1,682,878	83.2%
	5. 技術工学	828	487,059	6.6%	3,900	1,899,529	690	405,882	6.1%	3,983	1,616,570	83.3%
	6. 産業	444	261,176	3.5%	3,900	1,018,588	380	223,529	3.4%	4,274	955,392	85.6%
	7. 芸術	1,070	629,412	8.5%	3,900	2,454,706	897	527,647	8.0%	4,107	2,167,036	83.8%
	8. 言語	252	148,235	2.0%	3,900	578,118	217	127,647	1.9%	3,348	427,326	86.1%
	9. 文学	4,076	2,397,647	32.3%	3,900	9,350,824	3,765	2,214,706	33.5%	5,063	11,212,003	92.4%
	n. 記録なし	583	342,941	4.6%	3,900	1,337,471	493	290,000	4.4%	3,968	1,150,711	84.6%
合計		12,607	7,415,882	100%	—	28,921,941	11,242	6,612,941	100%	—	30,053,412	89.2%

表 2.5: サンプリングの設計時におけるサンプル数と語数の試算，およびその最終結果（出版 SC，2001 年）

	ジャンル	設計時					最終結果					
		S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	達成率
書籍	0. 総記	99	58,235	0.6%	3,900	227,118	83	48,824	0.5%	3,902	190,511	83.8%
	1. 哲学	134	78,824	0.8%	3,900	307,412	116	68,235	0.8%	4,155	283,521	86.6%
	2. 歴史	244	143,529	1.4%	3,900	559,765	203	119,412	1.3%	4,493	536,557	83.2%
	3. 社会科学	659	387,647	3.9%	3,900	1,511,824	557	327,647	3.7%	4,495	1,472,758	84.5%
	4. 自然科学	249	146,471	1.5%	3,900	571,235	211	124,118	1.4%	4,021	499,071	84.7%
	5. 技術工学	280	164,706	1.6%	3,900	642,353	234	137,647	1.5%	4,127	568,059	83.6%
	6. 産業	126	74,118	0.7%	3,900	289,059	108	63,529	0.7%	4,366	277,367	85.7%
	7. 芸術	177	104,118	1.0%	3,900	406,059	150	88,235	1.0%	4,225	372,760	84.7%
	8. 言語	58	34,118	0.3%	3,900	133,059	52	30,588	0.3%	4,001	122,386	89.7%
	9. 文学	460	270,588	2.7%	3,900	1,055,294	470	276,471	3.1%	5,070	1,401,707	102.2%
	n. 記録なし	67	39,412	0.4%	3,900	153,706	62	36,471	0.4%	4,564	166,437	92.5%
	小計	2,553	1,501,765	15.0%	—	5,856,882	2,246	1,321,176	14.8%	—	5,891,134	88.0%
雑誌	1. 総合	371	202,941	2.0%	3,000	608,824	345	202,941	2.3%	3,914	794,257	93.0%
	2. 教育	47	27,059	0.3%	3,000	81,176	46	27,059	0.3%	4,163	112,640	97.9%
	3. 政治	23	14,706	0.1%	3,000	44,118	25	14,706	0.2%	3,105	45,657	108.7%
	4. 産業	6	2,941	0.0%	3,000	8,824	5	2,941	0.0%	2,258	6,640	83.3%
	5. 工業	91	35,294	0.4%	3,000	105,882	60	35,294	0.4%	4,159	146,786	65.9%
	6. 厚生	9	2,353	0.0%	3,000	7,059	4	2,353	0.0%	2,897	6,816	44.4%
	小計	547	285,294	2.9%	—	855,882	485	285,294	3.2%	—	1,112,797	88.7%
新聞	全国紙	126	74,118	0.7%	1,000	74,118	110	64,706	0.7%	1,069	69,191	87.3%
	ブロック紙	67	39,412	0.4%	1,000	39,412	61	35,882	0.4%	903	32,411	91.0%
	地方紙	140	82,353	0.8%	1,000	82,353	128	75,294	0.8%	954	71,831	91.4%
	小計	333	195,882	2.0%	—	195,882	299	175,882	2.0%	—	173,434	89.8%

表 2.6: サンプリングの設計時におけるサンプル数と語数の試算，およびその最終結果（出版 SC，2002 年）

	ジャンル	設計時					最終結果					
		S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	達成率
書籍	0. 総記	94	55,294	0.6%	3,900	215,647	82	48,235	0.5%	3,902	188,215	87.2%
	1. 哲学	139	81,765	0.8%	3,900	318,882	123	72,353	0.8%	4,155	300,630	88.5%
	2. 歴史	223	131,176	1.3%	3,900	511,588	185	108,824	1.2%	4,493	488,980	83.0%
	3. 社会科学	662	389,412	3.9%	3,900	1,518,706	569	334,706	3.7%	4,495	1,504,487	86.0%
	4. 自然科学	263	154,706	1.5%	3,900	603,353	223	131,176	1.5%	4,021	527,455	84.8%
	5. 技術工学	259	152,353	1.5%	3,900	594,176	219	128,824	1.4%	4,127	531,645	84.6%
	6. 産業	112	65,882	0.7%	3,900	256,941	94	55,294	0.6%	4,366	241,412	83.9%
	7. 芸術	176	103,529	1.0%	3,900	403,765	151	88,824	1.0%	4,225	375,245	85.8%
	8. 言語	50	29,412	0.3%	3,900	114,706	42	24,706	0.3%	4,001	98,850	84.0%
	9. 文学	477	280,588	2.8%	3,900	1,094,294	525	308,824	3.5%	5,070	1,565,736	110.1%
	n. 記録なし	122	71,765	0.7%	3,900	279,882	108	63,529	0.7%	4,564	289,922	88.5%
	小計	2,577	1,515,882	15.2%	—	5,911,941	2,321	1,365,294	15.3%	—	6,112,579	90.1%
雑誌	1. 総合	383	224,118	2.2%	3,000	672,353	381	224,118	2.5%	3,914	877,136	99.5%
	2. 教育	46	25,294	0.3%	3,000	75,882	43	25,294	0.3%	4,163	105,294	93.5%
	3. 政治	25	14,706	0.1%	3,000	44,118	25	14,706	0.2%	3,105	45,657	100.0%
	4. 産業	6	3,529	0.0%	3,000	10,588	6	3,529	0.0%	2,258	7,968	100.0%
	5. 工業	81	39,412	0.4%	3,000	118,235	67	39,412	0.4%	4,159	163,911	82.7%
	6. 厚生	10	7,647	0.1%	3,000	22,941	13	7,647	0.1%	2,897	22,152	130.0%
	小計	551	314,706	3.1%	—	944,118	535	314,706	3.5%	—	1,222,119	97.1%
新聞	全国紙	126	74,118	0.7%	1,000	74,118	110	64,706	0.7%	1,069	69,191	87.3%
	ブロック紙	67	39,412	0.4%	1,000	39,412	61	35,882	0.4%	903	32,411	91.0%
	地方紙	140	82,353	0.8%	1,000	82,353	125	73,529	0.8%	954	70,148	89.3%
	小計	333	195,882	2.0%	—	195,882	296	174,118	1.9%	—	171,751	88.9%

表 2.7: サンプリングの設計時におけるサンプル数と語数の試算，およびその最終結果（出版 SC，2003 年）

	ジャンル	設計時					最終結果					
		S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	達成率
書籍	0. 総記	87	51,176	0.5%	3,900	199,588	72	42,353	0.5%	3,902	165,262	82.8%
	1. 哲学	132	77,647	0.8%	3,900	302,824	125	73,529	0.8%	4,155	305,518	94.7%
	2. 歴史	227	133,529	1.3%	3,900	520,765	188	110,588	1.2%	4,493	496,910	82.8%
	3. 社会科学	680	400,000	4.0%	3,900	1,560,000	575	338,235	3.8%	4,495	1,520,352	84.6%
	4. 自然科学	282	165,882	1.7%	3,900	646,941	244	143,529	1.6%	4,021	577,125	86.5%
	5. 技術工学	253	148,824	1.5%	3,900	580,412	215	126,471	1.4%	4,127	521,934	85.0%
	6. 産業	115	67,647	0.7%	3,900	263,824	94	55,294	0.6%	4,366	241,412	81.7%
	7. 芸術	175	102,941	1.0%	3,900	401,471	153	90,000	1.0%	4,225	380,215	87.4%
	8. 言語	41	24,118	0.2%	3,900	94,059	35	20,588	0.2%	4,001	82,375	85.4%
	9. 文学	503	295,882	3.0%	3,900	1,153,941	511	300,588	3.4%	5,070	1,523,983	101.6%
	n. 記録なし	130	76,471	0.8%	3,900	298,235	117	68,824	0.8%	4,564	314,083	90.0%
	小計	2,625	1,544,118	15.4%	—	6,022,059	2,329	1,370,000	15.3%	—	6,129,170	88.7%
雑誌	1. 総合	388	201,765	2.0%	3,000	605,294	343	201,765	2.3%	3,914	789,653	88.4%
	2. 教育	49	18,235	0.2%	3,000	54,706	31	18,235	0.2%	4,163	75,910	63.3%
	3. 政治	24	12,353	0.1%	3,000	37,059	21	12,353	0.1%	3,105	38,352	87.5%
	4. 産業	6	3,529	0.0%	3,000	10,588	6	3,529	0.0%	2,258	7,968	100.0%
	5. 工業	72	36,471	0.4%	3,000	109,412	62	36,471	0.4%	4,159	151,679	86.1%
	6. 厚生	9	6,471	0.1%	3,000	19,412	11	6,471	0.1%	2,897	18,744	122.2%
	小計	548	278,824	2.8%	—	836,471	474	278,824	3.1%	—	1,082,306	86.5%
新聞	全国紙	126	74,118	0.7%	1,000	74,118	109	64,118	0.7%	1,069	68,562	86.5%
	ブロック紙	67	39,412	0.4%	1,000	39,412	62	36,471	0.4%	903	32,943	92.5%
	地方紙	140	82,353	0.8%	1,000	82,353	123	72,353	0.8%	954	69,025	87.9%
	小計	333	195,882	2.0%	—	195,882	294	172,941	1.9%	—	170,530	88.3%

表 2.8: サンプリングの設計時におけるサンプル数と語数の試算，およびその最終結果（出版 SC，2004 年）

	ジャンル	設計時					最終結果					
		S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	達成率
書籍	0. 総記	81	47,647	0.5%	3,900	185,824	68	40,000	0.4%	3,902	156,081	84.0%
	1. 哲学	151	88,824	0.9%	3,900	346,412	139	81,765	0.9%	4,155	339,737	92.1%
	2. 歴史	232	136,471	1.4%	3,900	532,235	190	111,765	1.3%	4,493	502,196	81.9%
	3. 社会科学	665	391,176	3.9%	3,900	1,525,588	553	325,294	3.6%	4,495	1,462,182	83.2%
	4. 自然科学	281	165,294	1.7%	3,900	644,647	236	138,824	1.6%	4,021	558,203	84.0%
	5. 技術工学	224	131,765	1.3%	3,900	513,882	186	109,412	1.2%	4,127	451,534	83.0%
	6. 産業	120	70,588	0.7%	3,900	275,294	104	61,176	0.7%	4,366	267,094	86.7%
	7. 芸術	172	101,176	1.0%	3,900	394,588	149	87,647	1.0%	4,225	370,275	86.6%
	8. 言語	45	26,471	0.3%	3,900	103,235	38	22,353	0.3%	4,001	89,436	84.4%
	9. 文学	517	304,118	3.0%	3,900	1,186,059	548	322,353	3.6%	5,070	1,634,330	106.0%
	n. 記録なし	146	85,882	0.9%	3,900	334,941	121	71,176	0.8%	4,564	324,820	82.9%
	小計	2,634	1,549,412	15.5%	—	6,042,706	2,332	1,371,765	15.4%	—	6,155,888	88.5%
雑誌	1. 総合	391	208,235	2.1%	3,000	624,706	354	208,235	2.3%	3,914	814,977	90.5%
	2. 教育	43	24,706	0.2%	3,000	74,118	42	24,706	0.3%	4,163	102,846	97.7%
	3. 政治	22	14,118	0.1%	3,000	42,353	24	14,118	0.2%	3,105	43,831	109.1%
	4. 産業	5	2,941	0.0%	3,000	8,824	5	2,941	0.0%	2,258	6,640	100.0%
	5. 工業	71	45,294	0.5%	3,000	135,882	77	45,294	0.5%	4,159	188,376	108.5%
	6. 厚生	9	4,706	0.0%	3,000	14,118	8	4,706	0.1%	2,897	13,632	88.9%
	小計	541	300,000	3.0%	—	900,000	510	300,000	3.4%	—	1,170,301	94.3%
新聞	全国紙	126	74,118	0.7%	1,000	74,118	112	65,882	0.7%	1,069	70,449	88.9%
	ブロック紙	67	39,412	0.4%	1,000	39,412	61	35,882	0.4%	903	32,411	91.0%
	地方紙	140	82,353	0.8%	1,000	82,353	127	74,706	0.8%	954	71,270	90.7%
	小計	333	195,882	2.0%	—	195,882	300	176,471	2.0%	—	174,131	90.1%

表 2.9: サンプリングの設計時におけるサンプル数と語数の試算，およびその最終結果（出版 SC，2005 年）

	ジャンル	設計時					最終結果					
		S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	達成率
書籍	0. 総記	65	38,235	0.4%	3,900	149,118	58	34,118	0.4%	3,902	133,128	89.2%
	1. 哲学	119	70,000	0.7%	3,900	273,000	107	62,941	0.7%	4,155	261,524	89.9%
	2. 歴史	192	112,941	1.1%	3,900	440,471	160	94,118	1.1%	4,493	422,902	83.3%
	3. 社会科学	557	327,647	3.3%	3,900	1,277,824	467	274,706	3.1%	4,495	1,234,790	83.8%
	4. 自然科学	240	141,176	1.4%	3,900	550,588	205	120,588	1.4%	4,021	484,880	85.4%
	5. 技術工学	183	107,647	1.1%	3,900	419,824	154	90,588	1.0%	4,127	373,851	84.2%
	6. 産業	97	57,059	0.6%	3,900	222,529	80	47,059	0.5%	4,366	205,457	82.5%
	7. 芸術	145	85,294	0.9%	3,900	332,647	125	73,529	0.8%	4,225	310,633	86.2%
	8. 言語	37	21,765	0.2%	3,900	84,882	31	18,235	0.2%	4,001	72,961	83.8%
	9. 文学	468	275,294	2.8%	3,900	1,073,647	503	295,882	3.3%	5,070	1,500,124	107.5%
	n. 記録なし	113	66,471	0.7%	3,900	259,235	94	55,294	0.6%	4,564	252,340	83.2%
	小計	2,216	1,303,529	13.0%	—	5,083,765	1,984	1,167,059	13.1%	—	5,252,590	89.5%
雑誌	1. 総合	395	213,529	2.1%	3,000	640,588	363	213,529	2.4%	3,914	835,696	91.9%
	2. 教育	43	18,235	0.2%	3,000	54,706	31	18,235	0.2%	4,163	75,910	72.1%
	3. 政治	24	11,176	0.1%	3,000	33,529	19	11,176	0.1%	3,105	34,700	79.2%
	4. 産業	5	1,765	0.0%	3,000	5,294	3	1,765	0.0%	2,258	3,984	60.0%
	5. 工業	65	33,529	0.3%	3,000	100,588	57	33,529	0.4%	4,159	139,447	87.7%
	6. 厚生	9	3,529	0.0%	3,000	10,588	6	3,529	0.0%	2,897	10,224	66.7%
	小計	541	281,765	2.8%	—	845,294	479	281,765	3.2%	—	1,099,961	88.5%
新聞	全国紙	126	74,118	0.7%	1,000	74,118	109	64,118	0.7%	1,069	68,562	86.5%
	ブロック紙	67	39,412	0.4%	1,000	39,412	60	35,294	0.4%	903	31,880	89.6%
	地方紙	140	82,353	0.8%	1,000	82,353	132	77,647	0.9%	954	74,076	94.3%
	小計	333	195,882	2.0%	—	195,882	301	177,059	2.0%	—	174,518	90.4%

表 2.10: サンプリングの設計時におけるサンプル数と語数の試算，およびその最終結果（図書館 SC，1986 年–1990 年）

	ジャンル	設計時					最終結果					
		S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	達成率
書籍	0. 総記	34	20,000	0.3%	3,900	78,000	32	18,824	0.3%	4,108	77,323	94.1%
	1. 哲学	92	54,118	0.7%	3,900	211,059	81	47,647	0.7%	4,452	212,131	88.0%
	2. 歴史	200	117,647	1.6%	3,900	458,824	171	100,588	1.5%	4,587	461,350	85.5%
	3. 社会科学	304	178,824	2.4%	3,900	697,412	282	165,882	2.5%	4,427	734,416	92.8%
	4. 自然科学	106	62,353	0.8%	3,900	243,176	88	51,765	0.8%	4,315	223,368	83.0%
	5. 技術工学	92	54,118	0.7%	3,900	211,059	77	45,294	0.7%	3,983	180,400	83.7%
	6. 産業	62	36,471	0.5%	3,900	142,235	56	32,941	0.5%	4,274	140,795	90.3%
	7. 芸術	167	98,235	1.3%	3,900	383,118	141	82,941	1.3%	4,107	340,638	84.4%
	8. 言語	39	22,941	0.3%	3,900	89,471	35	20,588	0.3%	3,348	68,924	89.7%
	9. 文学	726	427,059	5.8%	3,900	1,665,529	628	369,412	5.6%	5,063	1,870,156	86.5%
	n. 記録なし	137	80,588	1.1%	3,900	314,294	115	67,647	1.0%	3,968	268,421	83.9%
合計		1,959	1,152,353	15.5%	—	4,494,176	1,706	1,003,529	15.2%	—	4,577,921	87.1%

表 2.11: サンプリングの設計時におけるサンプル数と語数の試算，およびその最終結果（図書館 SC，1991 年–1995 年）

	ジャンル	設計時					最終結果					
		S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	達成率
書籍	0. 総記	58	34,118	0.5%	3,900	133,059	57	33,529	0.5%	4,108	137,731	98.3%
	1. 哲学	149	87,647	1.2%	3,900	341,824	125	73,529	1.1%	4,452	327,363	83.9%
	2. 歴史	322	189,412	2.6%	3,900	738,706	287	168,824	2.6%	4,587	774,312	89.1%
	3. 社会科学	562	330,588	4.5%	3,900	1,289,294	525	308,824	4.7%	4,427	1,367,263	93.4%
	4. 自然科学	186	109,412	1.5%	3,900	426,706	158	92,941	1.4%	4,315	401,048	84.9%
	5. 技術工学	166	97,647	1.3%	3,900	380,824	139	81,765	1.2%	3,983	325,657	83.7%
	6. 産業	90	52,941	0.7%	3,900	206,471	76	44,706	0.7%	4,274	191,078	84.4%
	7. 芸術	271	159,412	2.1%	3,900	621,706	226	132,941	2.0%	4,107	545,987	83.4%
	8. 言語	59	34,706	0.5%	3,900	135,353	49	28,824	0.4%	3,348	96,493	83.1%
	9. 文学	1,055	620,588	8.4%	3,900	2,420,294	968	569,412	8.6%	5,063	2,882,661	91.8%
	n. 記録なし	148	87,059	1.2%	3,900	339,529	123	72,353	1.1%	3,968	287,094	83.1%
合計		3,066	1,803,529	24.3%	—	7,033,765	2,733	1,607,647	24.3%	—	7,336,688	89.1%

表 2.12: サンプリングの設計時におけるサンプル数と語数の試算，およびその最終結果（図書館 SC，1996 年–2000 年）

	ジャンル	設計時					最終結果					
		S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	達成率
書籍	0. 総記	81	47,647	0.6%	3,900	185,824	80	47,059	0.7%	4,108	193,307	98.8%
	1. 哲学	194	114,118	1.5%	3,900	445,059	192	112,941	1.7%	4,452	502,829	99.0%
	2. 歴史	371	218,235	2.9%	3,900	851,118	321	188,824	2.9%	4,587	866,042	86.5%
	3. 社会科学	705	414,706	5.6%	3,900	1,617,353	692	407,059	6.2%	4,427	1,802,183	98.2%
	4. 自然科学	247	145,294	2.0%	3,900	566,647	205	120,588	1.8%	4,315	520,347	83.0%
	5. 技術工学	257	151,176	2.0%	3,900	589,588	212	124,706	1.9%	3,983	496,685	82.5%
	6. 産業	135	79,412	1.1%	3,900	309,706	113	66,471	1.0%	4,274	284,103	83.7%
	7. 芸術	324	190,588	2.6%	3,900	743,294	266	156,471	2.4%	4,107	642,622	82.1%
	8. 言語	76	44,706	0.6%	3,900	174,353	66	38,824	0.6%	3,348	129,970	86.8%
	9. 文学	1,143	672,353	9.1%	3,900	2,622,176	1,086	638,824	9.7%	5,063	3,234,060	95.0%
	n. 記録なし	153	90,000	1.2%	3,900	351,000	132	77,647	1.2%	3,968	308,101	86.3%
合計		3,686	2,168,235	29.2%	—	8,456,118	3,365	1,979,412	29.9%	—	8,980,250	91.3%

表 2.13: サンプリングの設計時におけるサンプル数と語数の試算，およびその最終結果（図書館 SC，2001 年–2005 年）

	ジャンル	設計時					最終結果					
		S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	S 数	固定長 S 語数	構成比	可変長 S 平均字数	可変長 S 語数	達成率
書籍	0. 総記	90	52,941	0.7%	3,900	206,471	80	47,059	0.7%	4,108	193,307	88.9%
	1. 哲学	182	107,059	1.4%	3,900	417,529	162	95,294	1.4%	4,452	424,262	89.0%
	2. 歴史	428	251,765	3.4%	3,900	981,882	354	208,235	3.1%	4,587	955,075	82.7%
	3. 社会科学	785	461,765	6.2%	3,900	1,800,882	696	409,412	6.2%	4,427	1,812,601	88.7%
	4. 自然科学	258	151,765	2.0%	3,900	591,882	212	124,706	1.9%	4,315	538,115	82.2%
	5. 技術工学	313	184,118	2.5%	3,900	718,059	262	154,118	2.3%	3,983	613,828	83.7%
	6. 産業	157	92,353	1.2%	3,900	360,176	135	79,412	1.2%	4,274	339,415	86.0%
	7. 芸術	308	181,176	2.4%	3,900	706,588	264	155,294	2.3%	4,107	637,790	85.7%
	8. 言語	78	45,882	0.6%	3,900	178,941	67	39,412	0.6%	3,348	131,939	85.9%
	9. 文学	1,152	677,647	9.1%	3,900	2,642,824	1,083	637,059	9.6%	5,063	3,225,126	94.0%
	n. 記録なし	145	85,294	1.2%	3,900	332,647	123	72,353	1.1%	3,968	287,094	84.8%
合計		3,896	2,291,765	30.9%	—	8,937,882	3,438	2,022,353	30.6%	—	9,158,553	88.2%

2.2 「特定目的 SC」の設計とサンプリングの結果

2.2.1 「特定目的 SC」の設計方針

特定目的 SC の設計方針は，以下のようにまとめられる。

- 「特定目的 SC」には,「出版 SC」や「図書館 SC」の母集団には入らない，あるいは出版・流通という側面からは捉えきれないけれども，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のサンプルを収める。
- 「特定目的 SC」に収録するメディアは,「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」の 9 種類とする。
- サンプルを取得する対象範囲は明確に定めるが,「出版 SC」「図書館 SC」のように母集団を数量的に定義することは必ずしも必要としない。
- 基本的に，可変長サンプルのみを取得する。

このうち,「白書」「教科書」「広報紙」「法律」は公的な性格の強い書き言葉であり，これらの分析により言語政策に関わる基礎資料を提供することが期待できる。「ベストセラー」はあらゆる書籍の中で特に多くの人に読まれたものであり，出版の実態を反映する「出版 SC」の書籍，流通の実態を反映する「図書館 SC」の書籍に対して，一般読者に受容された実態を反映する資料として考えることができる。「Yahoo!知恵袋」「Yahoo!ブログ」はウェブ上の書き言葉であり，そこに見られる文字遣い・言葉遣いを収集することにより，ウェブ上の書き言葉が持つさまざまな変異のありさまを捉えることができる。「韻文」は，短歌・俳句・詩という，通常の書き言葉（いわゆる文章）とは異なるスタイルを持つ書き言葉であり，現代日本語の書き言葉における重要な一部を構成するものとして収録することにした。「国会会議録」は，国会における会議での発言を書き起こしたテキストである。そもそも書き言葉として執筆されたテキストではないものの,「会議録」自体は書き言葉の一種であることから，書き言葉のバリエーションの 1 つとして収録することにした。

また,「出版 SC」や「図書館 SC」ではサンプルの取得元（原本）はすべて印刷物であったが,「特定目的 SC」のうち「Yahoo!知恵袋」「Yahoo!ブログ」「法律」「国会会議録」については，既存の電子データからサンプルを取得した。

2.2.2 サンプリングの最終結果

「特定目的 SC」に収録されたメディアの種類と，その対象期間，取得対象，取得したサンプル数，取得した語数について，表 2.14 に示す。なお，語数は推計値である。

* ベストセラーの対象期間は，出版年ではなく，ベストセラーとして記録された年を表す。

表 2.14: 「特定目的 SC」の構成

メディア	対象期間	取得対象	S 数	可変長 S 語数	取得元媒体
白書	1976 年–2005 年	1,006 冊	1,500	500 万語	印刷物
教科書	2005 年–2007 年	145 冊	483	120 万語	印刷物
広報紙	2008 年	100 自治体	355	400 万語	印刷物
ベストセラー *	1976 年–2005 年	951 冊	1,696	447 万語	印刷物
Yahoo!知恵袋	2004 年–2005 年	3,120,839 質問	91,450	1,000 万語	電子データ
Yahoo!ブログ	2008 年–2009 年	3,463,413 記事	52,680	1,000 万語	電子データ
韻文	1980 年–2005 年	130 冊	253	15 万語	印刷物
法律	1976 年–2005 年	718 法律	348	100 万語	電子データ
国会会議録	1976 年–2005 年	32,925 会議	159	500 万語	電子データ

第3章　各メディアにおけるサンプリングの手順と結果

3.1　サンプリングが完了したサンプルの一覧

本章では，各メディアで実施したサンプリングの手順と結果について示す。はじめに，サンプリングの作業が完了したサンプルの種類と数を，表 3.1 に示す。

表 3.1: サンプリングが完了したサンプルの一覧

SC	メディア	対象期間	母集団	S 数	可変長 S 語数	取得元媒体
出版 SC	書籍	2001 年–2005 年	約 485 億文字	11,212	2,954 万語	印刷物
	雑誌	2001 年–2005 年	約 105 億文字	2,483	569 万語	印刷物
	新聞	2001 年–2005 年	約 64 億文字	1,490	86 万語	印刷物
図書館 SC	書籍	1986 年–2005 年	479 億文字	11,242	3,005 万語	印刷物
特定目的 SC	白書	1976 年–2005 年	1,006 冊	1,500	500 万語	印刷物
	教科書	2005 年–2007 年	145 冊	483	120 万語	印刷物
	広報紙	2008 年	100 自治体	355	400 万語	印刷物
	ベストセラー	1976 年–2005 年	951 冊	1,696	371 万語	印刷物
	Yahoo!知恵袋	2004 年–2005 年	約 312 万質問	91,450	1,000 万語	電子データ
	Yahoo!ブログ	2008 年–2009 年	約 346 万記事	52,680	1,000 万語	電子データ
	韻文	1980 年–2005 年	130 冊	253	15 万語	印刷物
	法律	1976 年–2005 年	718 法律	348	100 万語	電子データ
	国会会議録	1976 年–2005 年	32,925 会議	159	500 万語	電子データ

なお，出版 SC・図書館 SC の設計に関する詳細については，丸山・秋元 (2006,2007) を，サンプリングの手順については柏野ほか (2009)，丸山ほか (2011) を参照されたい。

3.2　出版 SC「書籍」

概要

- 出版 SC「書籍」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。
- サンプリングの結果，取得したのは，11,212 サンプルである。

母集団の定義

- 「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」を調べるため，国立国会図書館に所蔵されている書籍を調査した。「納本制度」により，国内で発行されるすべての書籍は国立国会図書館に納本されることになっているためである。
- 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 年間に発行された書籍の冊数・ページ数を調査した。
- 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を除外した。その結果，2001 年から 2005 年の間に発行された「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値を算出し，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 2 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 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。

  発行年（5 層）: 書籍の発行年である 2001 年から 2005 年までの，5 分類。

- NDC で層別した母集団の各層について，構成比率を図 3.1 に示す。

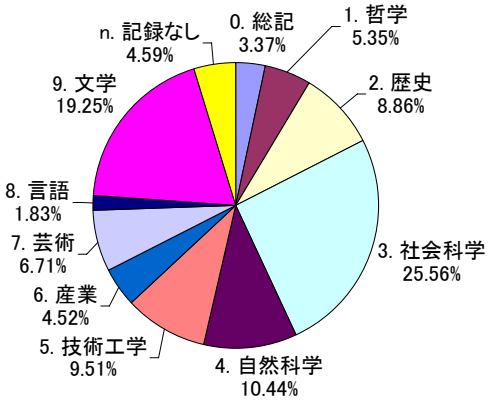

図 3.1: 母集団の構成比率（出版 SC「書籍」，NDC 別）

サンプリング方法

- 母集団の構成比率を，55 の各層から取得する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。
- 取得した 11,212 サンプルについて，NDC ごとの内訳を，図 3.2 に示す。

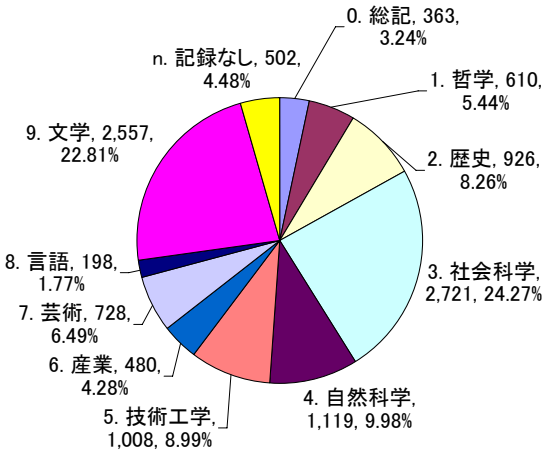

図 3.2: 取得したサンプルの構成比率（出版 SC「書籍」，NDC 別）

3.3　出版 SC「雑誌」

概要

- 出版 SC「雑誌」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。
- サンプリングの結果，取得したのは，2,483 サンプルである。

母集団の定義

- 「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雑誌」を,「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発行した定期刊行物」と定義した。これらが，いわゆる「雑誌」として想起される定期刊行物におおむね合致すると判断したためである。
- 『雑誌新聞総かたろぐ』（メディア・リサーチ・センター発行）から，対象出版社が 5 年間に発行した定期刊行物に関する書誌情報を抽出した。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を除外した。その結果，2001 年から 2005 年の間に発行された「雑誌」は，異なりで 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』で雑誌タイトルごとに分類されている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値を算出し，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

- 上記定義した母集団を，以下の 2 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』で雑誌タイトルごとに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。

  発行年（5 層）：雑誌の発行年である 2001 年から 2005 年までの，5 分類。

- ジャンルで層別した母集団の各層について，構成比率を図 3.3 に示す。

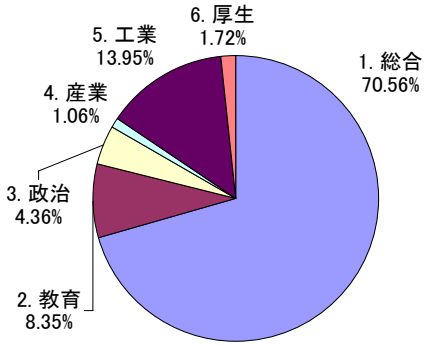

図 3.3: 母集団の構成比率（出版 SC「雑誌」，ジャンル別）

サンプリング方法

- 母集団の構成比率を，30 の各層から取得する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。
- 取得した 2,483 サンプルについて，ジャンルごとの内訳を，図 3.4 に示す。

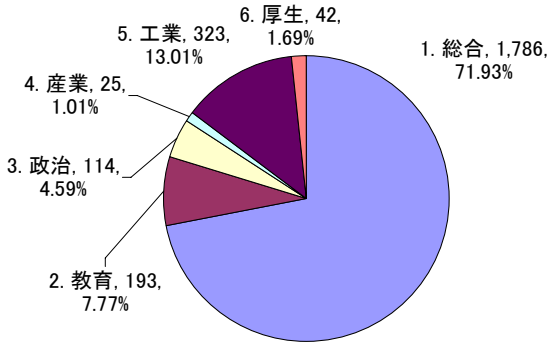

図 3.4: 取得したサンプルの構成比率（出版 SC「雑誌」，ジャンル別）

3.4　出版 SC「新聞」

概要

- 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。
- サンプリングの結果，取得したのは，1,490 サンプルである。

母集団の定義

- 「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,「全国紙・ブロック紙・有力な地方紙」の集合と定義した。そこで,『全国新聞ガイド』（社団法人日本新聞協会発行）において「全国紙」「ブロック紙」として記載されている日刊新聞」に加え，日本各地の有力な地方紙をリスト化した。この結果，以下の 16 タイトルが同定された。

  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

- 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，2001 年から 2005 年の間に発行された「新聞」は，異なりで 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，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文字を実測した。これを 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 2 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。

- 新聞タイトルで層別した母集団の各層について，構成比率を図 3.5 に示す。

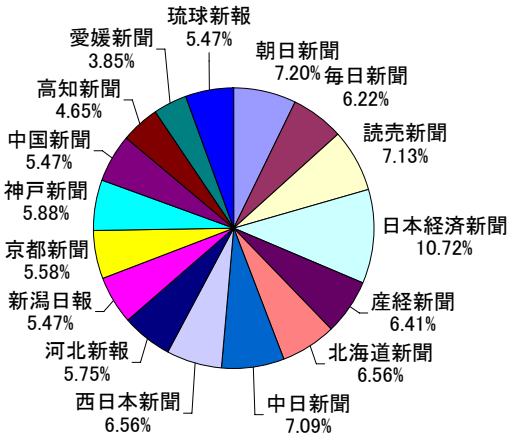

図 3.5: 母集団の構成比率（出版 SC「新聞」，タイトル別）

サンプリング方法

- 母集団の構成比率を，80 の各層から取得する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。この際,「日本経済新聞」「愛媛新聞」については，著作権処理の都合から，採録対象から除外した。
- 取得した 1,490 サンプルについて，ジャンルごとの内訳を，図 3.6 に示す。

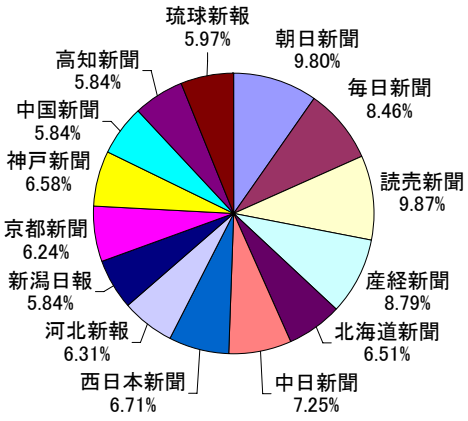

図 3.6: 取得したサンプルの構成比率（出版 SC「新聞」，タイトル別）

3.5　図書館 SC「書籍」

概要

- 図書館 SC「書籍」は，1986 年から 2005 年までの 20 年間に発行された書籍のうち，公立図書館で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。
- サンプリングの結果，取得したのは，11,242 サンプルである。

母集団の定義

- 「1986 年から 2005 年までの 20 年間に発行された書籍のうち，東京都内のより多くの公共図書館で共通に所蔵されている書籍」を定義するため，東京都立中央図書館で取りまとめられている「ISBN 総合目録」を集計した。
- 出版 SC「書籍」の部分と母集団からの抽出比およびサンプルサイズを揃えるため，母集団のサイズは，推計総文字数が出版 SC「書籍」とほぼ等しくなるように定めることにした。
- 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 ページを対象とすれば，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 2 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 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。

  発行年（20 層）：書籍の発行年である 1986 年から 2005 年までの，20 分類。

- NDC で層別した母集団の各層について，構成比率を図 3.7 に示す。

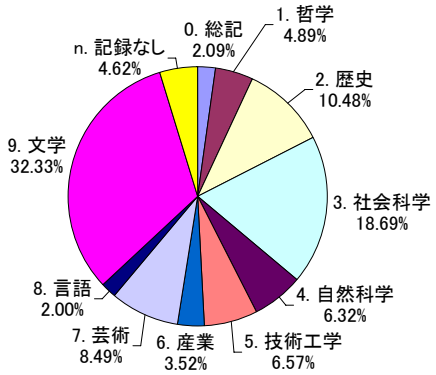

図 3.7: 母集団の構成比率（図書館 SC「書籍」，NDC 別）

サンプリング方法

- 母集団の構成比率を，220 の各層から取得する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。
- 取得した 11,242 サンプルについて，NDC ごとの内訳を，図 3.8 に示す。

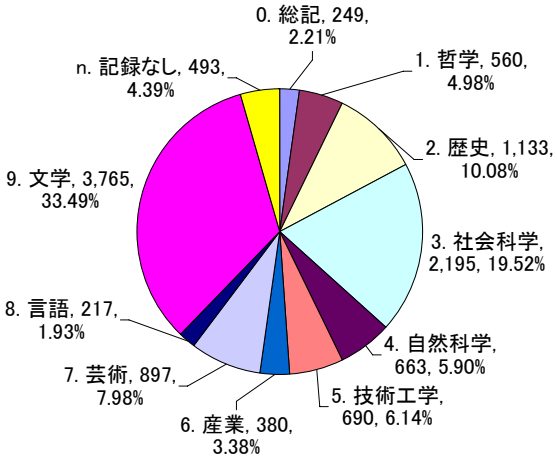

図 3.8: 取得したサンプルの構成比率（図書館 SC「書籍」，NDC 別）

3.6　特定目的 SC「白書」

概要

- 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。
- サンプリングの結果，取得したのは，1,500 サンプルである。

対象データの定義

- 「1976 年から 2005 年までの 30 年間に発行されたすべての白書」は，以下のように同定した。まず，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（正確には，2001 年から 2005 年の間に,『官報』の付録である『官報資料版』の目次に『白書』『青書』『年次報告』として掲載されたタイトルから，重複などを省いたものである)。
- これらの白書について,『日本白書総攬』（丸善プラネット，1997 年）や国立国会図書館蔵書検索システムなどを用いて，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，別タイトルとせず，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。
- 調査の結果，合計で 40 タイトル，1,006 冊の白書が同定され，これらを特定目的 SC「白書」の対象データとして定義した。

層別方法

- 上記で定義した対象データを，以下の 2 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）：白書の発行年である 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

- 全体で約 500 万語分のサンプルを取得することとした。1 期から 6 期のそれぞれから 250 サンプルずつを選び，全体で 1,500 サンプルを取得することを計画した。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。
- 取得し 1,500 サンプルについて，ジャンルごとの内訳を，図 3.9 に示す。

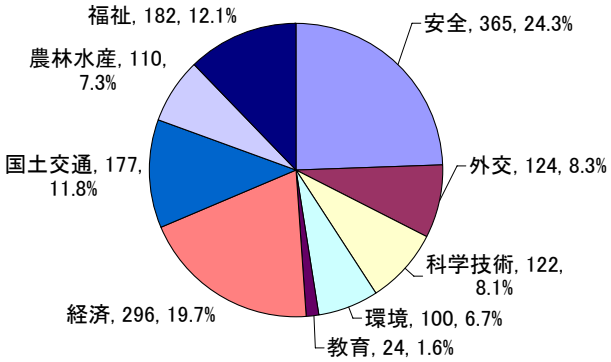

図 3.9: 取得したサンプルの構成比率（特定目的 SC「白書」，ジャンル別）

3.7　特定目的 SC「教科書」

概要

- 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。
- サンプリングの結果，取得したのは，483 サンプルである。

対象データの定義

- 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 冊の教科書が同定された。これらを，特定目的 SC「教科書」の対象データとして定義した。

層別方法

- 上記で定義した対象データを，以下の 2 つの基準により，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。

※ ただし,「外国語」は中学校と高等学校のみ,「情報」は高等学校のみ,「生活」は小学校のみとなる。

- また，対象データとなった教科書に印刷されている総文字数を推計したところ，7,859,456 文字という結果を得た。
- 教科で層別した対象データの各層について，構成比率を図 3.10 に示す。

サンプリング方法

- 対象データの構成比率を，25 の各層から取得するサンプル数に比例割当した。

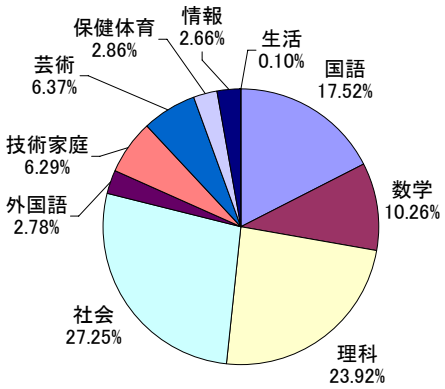

図 3.10: 母集団の構成比率（特定目的 SC「教科書」，教科別）

- 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。(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)。
- 取得した 483 サンプルについて，教科ごとの内訳を，図 3.11 に示す。

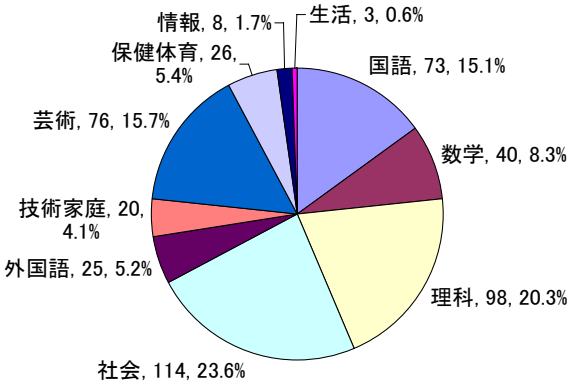

図 3.11: 取得したサンプルの構成比率（特定目的 SC「教科書」，教科別）

3.8　特定目的 SC「広報紙」

概要

- 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。
- サンプリングの結果，取得したのは，355 サンプルである。

対象データの定義

- 対象を「地方自治体で 2008 年に発行された広報紙」と定めた。人口構成比などを考慮し，全国から 100 の自治体（区市町村）をサンプリングし，そこで 2008 年度に発行された広報紙を対象データとして定義した。
- 100 自治体で 2008 年に発行された広報紙を入手した。Web 上から PDF ファイルで入手したものもあるが，自治体から現物を取り寄せた場合もあった。

層別方法

- 上記で定義した対象データを，以下の基準により，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。
- 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，355 サンプルを取得した。地域ごとの内訳を，図 3.12 に示す。

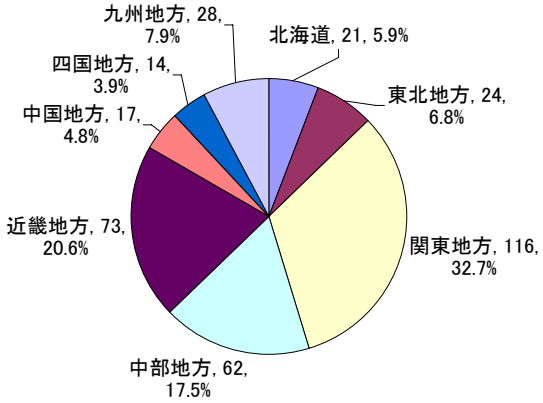

図 3.12: 取得したサンプルの構成比率（特定目的 SC「広報紙」, 地域別）

3.9　特定目的SC「ベストセラー」

概要

- 特定目的 SC「ベストセラー」では，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。
- サンプリングの結果，取得したのは，1,696 サンプルである。

対象データの定義

- 1976 年から 2005 年までの 30 年間において，各年のベストセラーとして 20 位までに挙げられた書籍を対象とした。
- 『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ところ，951 冊が同定された。これらを，特定目的 SC 「ベストセラー」の対象データとして定義した。
- なお，1971 年に出版された本が 1976 年のベストセラーになるなど，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，実際に取得できたサンプル数は 1,696 サンプルにとどまった。
- 取得した 1,696 サンプルについて，NDC ごとの内訳を，図 3.13 に示す。

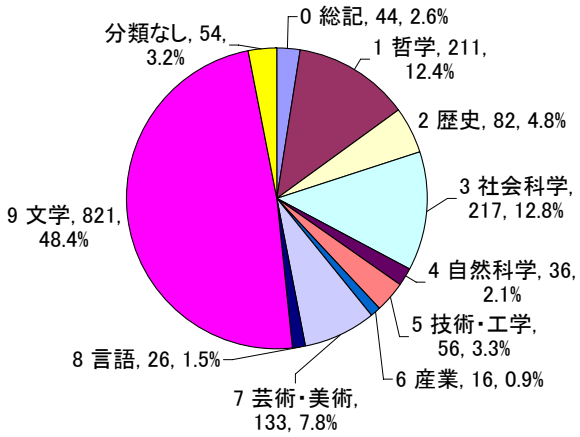

図 3.13: 取得したサンプルの構成比率（特定目的 SC「ベストセラー」, NDC 別）

3.10　特定目的 SC「Yahoo!知恵袋」

概要

- 特定目的 SC「Yahoo!知恵袋」は，参加者同士で知識を教えあうことを目的とした 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。
- サンプリングの結果，取得したのは，91,450 サンプルである。

対象データの定義

- ヤフー株式会社より提供された「Yahoo!知恵袋」のデータには，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」の対象データとして定義した。

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，以下のように，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 階層に分かれる。小カテゴリには，固有の「カテゴリ番号」が付いている。

大カテゴリ		中カテゴリ		小カテゴリ	カテゴリ番号
エンターテインメントと趣味	＞	おもちゃ，ホビー	＞	おもちゃ	2078523513
エンターテインメントと趣味	＞	ゲーム	＞	オンラインゲーム	2078297515
ビジネス，経済とお金	＞	保険，税金，年金	＞	保険	2078297810

- このうち，小カテゴリによって，対象データ全体を合計 279 の層に層別した。

サンプリング方法

- 対象データから，1 つの質問とそれに対する 1 つの回答（「ベストアンサー」と呼ばれる，質問者が「もっとも納得，満足した回答」として選んだ回答）の組を抽出して 1 サンプルとすることにした。例を以下に示す。

> **質問**：竹で編んだ矢の入れ物の名前はなんですか？
>
> **回答**：箙ですね。「えびら」と読みます。念のため参考 URL に写真を入れておきます。

- 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 91,450 サンプルを取得した。大カテゴリごとのサンプル数を，図 3.14 に示す。

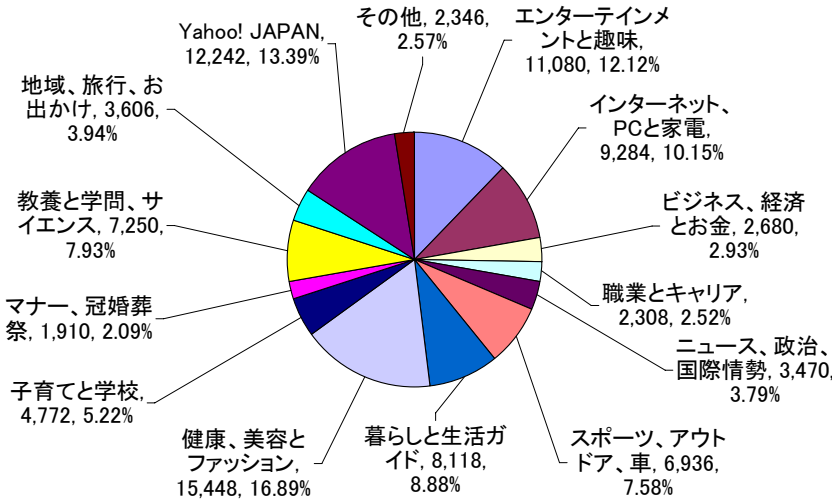

図 3.14: 取得したサンプルの構成比率（特定目的 SC「Yahoo!知恵袋」，大カテゴリ別）

3.11　特定目的 SC「Yahoo!ブログ」

概要

- 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。
- サンプリングの結果，取得したのは，52,680 サンプルである。

対象データの定義

- ヤフー株式会社より提供された「Yahoo!ブログ」の元データには，合計 3,463,413 の記事が含まれていた。これらを，特定目的 SC「Yahoo!ブログ」の対象データとして定義した。
- なお，元データは，以下の条件を満たすものとなっている。
  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ヶ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. 1 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，以下のように，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 階層に分かれる。小カテゴリには，固有の「カテゴリ番号」が付いている。

大カテゴリ		中カテゴリ		小カテゴリ	カテゴリ番号
生活と文化	＞	祝日，記念日，年中行事	＞	クリスマス	555000540
生活と文化	＞	祝日，記念日，年中行事	＞	誕生日	555000549
家庭と住まい	＞	住まい	＞	ガーデニング	555002691

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。大カテゴリごとのサンプル数を，図 3.15 に示す。

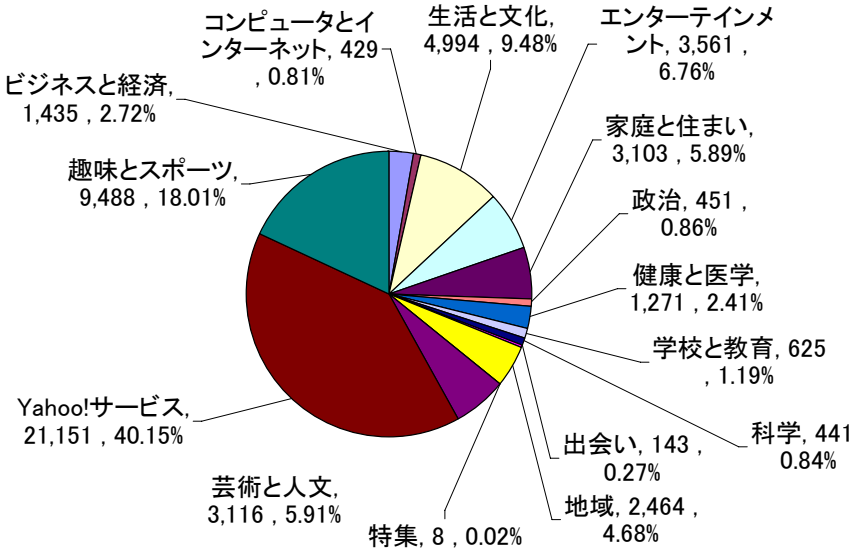

図 3.15: 取得したサンプルの構成比率（特定目的 SC「Yahoo!ブログ」，大カテゴリ別）

3.12　特定目的 SC「韻文」

概要

- 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。
- サンプリングの結果，取得したのは，253 サンプルである。

対象データの定義

- サンプルを取得する対象について，日本文藝家協会と協議した結果，以下の作品を対象とすることとした。

  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

- なお,『現代短歌全集』は昭和 34 年（1959 年）から昭和 63 年（1988 年）の間に発表された歌集,『増補現代俳句大系』は昭和 25 年（1950 年）から昭和 54 年（1979 年）の間に発表された句集を集めたものである。BCCWJ に収録された他のメディアの対象期間から外れることになるが，韻文という文学作品が持つ特性を考慮し，これらを対象とすることとした。

層別方法

- 上記で定義した対象データは,「短歌」「俳句」「詩」という 3 種類の区別以外，層別は実施しなかった。

サンプリング方法

- 著作権処理の結果，対象データのうち，60 の歌集，92 の句集，101 の詩集を利用できることになった。
- 短歌・俳句・詩からそれぞれ 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。
- 取得した 253 サンプルについて，内訳を図 3.16 に示す。

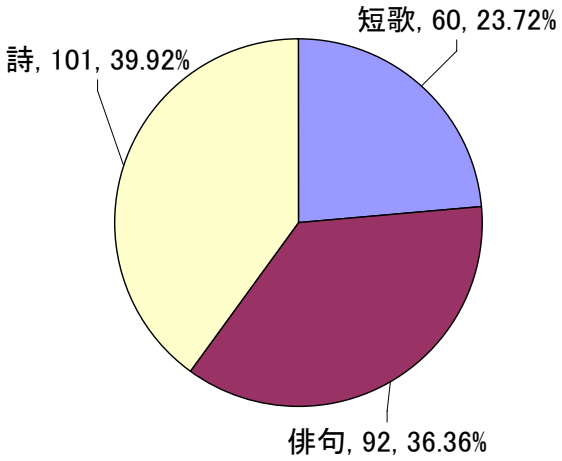

図 3.16: 取得したサンプルの構成比率（特定目的 SC「韻文」）

3.13　特定目的 SC「法律」

概要

- 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。
- サンプリングの結果，取得したのは，348 サンプルである。

対象データの定義

- 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。これらをダウンロードし，特定目的 SC 「法律」の対象データとして定義した。

層別方法

- 上記で定義した対象データを，以下の基準により，合計 6 層に層別した。

> **公布年（6 層）**：法律の公布年である 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

- 「法令データ提供システム」では，法務省『日本現行法規』に基づいて法律が 50 のジャンルに分類されている（「事項別分類」）が，事前の層別には用いなかった。
- 公布年で層別した対象データの各層について，構成比率を図 3.17 に示す。

サンプリング方法

- 1 期から 6 期のそれぞれから 30 万文字ずつを取得し，全体で 180 万文字，約 100 万語分のサンプルを取得することとした。
- 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。

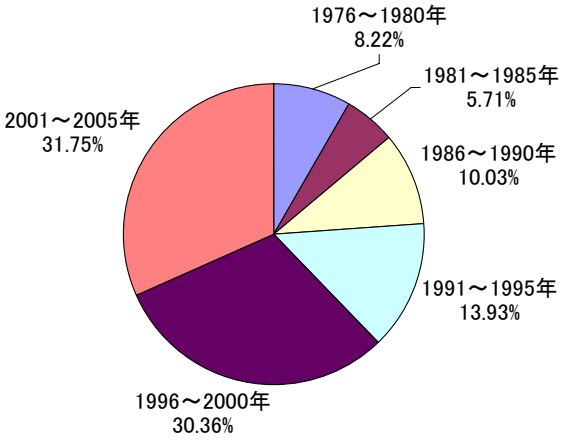

図 3.17: 母集団の構成比率（特定目的 SC「法律」，公布年別）

- 取得した 348 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構成比率（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**	348

3.14　特定目的 SC「国会会議録」

概要

- 特定目的 SC「国会会議録」は，1976 年から 2005 年までの 30 年間における国会での「国会会議録」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。
- サンプリングの結果，取得したのは，159 サンプルである。

対象データの定義

- 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国立国会図書館より受領し，これらを特定目的 SC「国会会議録」の対象データとした。
- 対象データのうち,「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

- 上記で定義した対象データを，以下の 3 つの基準により，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）：会議の開催された年である 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。

常任委員会：会議名末尾に「委員会」が付くもの。ただし，末尾が「特別委員会」「小委員会」のものは除く。

特別委員会：会議名末尾に「特別委員会」が付くもの。

本会議：会議名が「本会議」であるもの。

その他：上記以外のすべての会議。「小委員会」「分科会」「調査会」「公聴会」「審査会」「互選会」「打合会」など。

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ために，159 の会議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，159 サンプルに比例割当して，各層から取得するサンプル数を算出した。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。
- 取得した 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 3.18 に示す。

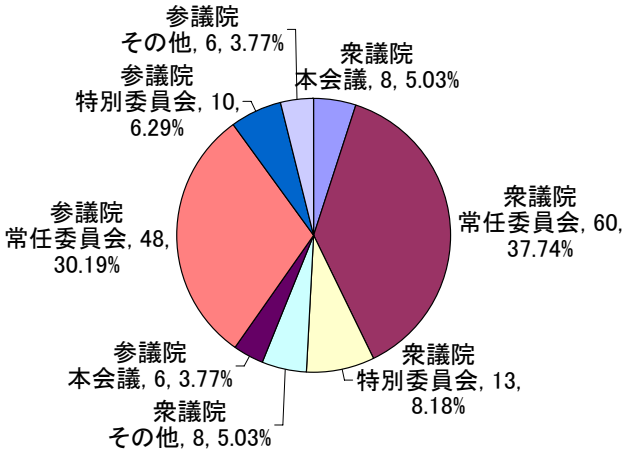

図 3.18: 取得したサンプルの構成比率（特定目的 SC「国会会議録」，開催院・会議種別）

第II部

書誌情報の設計と実装

第4章　BCCWJの書誌情報

4.1　均衡コーパスにおける書誌情報の役割

一般に，均衡コーパスとは，さまざまなメディアやジャンルから抽出されたサンプルの集合体と見なすことができる。ある均衡コーパスがどのようなメディアやジャンルのサンプルを含むかは，そのコーパスがどのような設計に基づいているかに依存するが，どのような設計であっても，そこに含まれている各サンプルの出自が明示されていることが望ましい。均衡コーパスを検索した結果を分析したり解釈したりする際，その結果が幅広いメディアを通して一般的に観察される現象なのか，あるいは（例えば）「雑誌」に特有な現象なのか，といった違いを捉えるためには，各サンプルの出自を表わす「書誌情報」が必要不可欠である。

BCCWJ の構築過程においては，サンプリングの作業と並行して，各サンプルの出自を示す「**書誌情報データベース**」を整備してきた。BCCWJ の利用者は，この書誌情報データを参照することにより，BCCWJ を構成するすべてのサンプルの出自と属性を知ることができる。

厳密な手順で取得された大量のサンプルを，その書誌情報と関連づけて利用することにより，コーパスの分析結果が現代日本語書き言葉のどの位相に位置づけられるものであるかを明確にすることができるわけである。このような利点は，例えば Web をコーパスとして用いる方法論では得ることのできないものであり，均衡コーパスとしての BCCWJ が持つ意義を最大限に特徴づけるものであると言える。

4.2　書誌情報データベースの構成

BCCWJ で提供される書誌情報データベースは，以下のデータ群から構成される。なお，ここで示す構成は 2011 年 1 月時点のものであり，BCCWJ の最終的な公開時には異なる構成を取る可能性がある。

- **書誌情報データ**（`Bibliography.txt`）：各サンプルを抽出した原本に関する情報。
- **サンプル情報データ**（`Sample.txt`）：各サンプルの ID や抽出状況に関する情報。
- **人名録データ**（`Directory.txt`）：各サンプルの著者や著作権者などに関する情報。
- **サンプル著者対応情報データ**（`Sample_author.txt`）

次節以降では，各データの設計および構成について示す。

第5章　書誌情報データ（Bibliography.txt）

5.1　書誌情報データの概要

書誌情報データ（Bibliography.txt）は，サンプルが抽出された出典元（原本）に関する情報を表わすデータである。BCCWJ に収録されたサンプルの書誌情報が，表 5.1 に示す 15 列（タブ区切り）によって表現されている。

表 5.1: 書誌情報データを構成する列

1. 書誌 ID （Bib_ID）	サンプルを抽出した原本に対して付された ID。
2. タイトル （Title）	原本のタイトル。
3. 副題 （Subtitle）	原本の副題（サブタイトル）。
4. 巻号 （Number）	原本の巻号。
5. 責任表示 （Bib_author）	原本の責任表示（著者，編者，監修者など）。
6. 出版者 （Publisher）	原本の出版者（出版社）。
7. 出版年 （Year）	原本の出版年。
8. ISBN（ISBN）	原本に付された ISBN（国際標準図書番号）。
9. 判型 （Size）	原本のサイズ。
10. ページ数 （Pages）	原本のページ数。
11. ジャンル (1) （Genre_1）	原本のジャンルに関する情報 (1)。
12. ジャンル (2) （Genre_2）	原本のジャンルに関する情報 (2)。
13. ジャンル (3) （Genre_3）	原本のジャンルに関する情報 (3)。
14. ジャンル (4) （Genre_4）	原本のジャンルに関する情報 (4)。
15. 責任表示 ID （Bib_author_ID）	原本の責任表示に対応する ID。

書誌情報データに格納された情報の例を，表 5.2 に示す。「書籍」「雑誌」「新聞」「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」というメディアの違いに応じて，書誌情報が記載されている。実際には 15 列のタブ区切りテキストだが，ここでは折り返して表示している。

表 5.2: 書誌情報データの例

メディア	Bib_ID	Title	Subtitle	Number	Bib_author	Publisher
書籍	BK_20126734	ペルシャの幻術師	—	—	司馬遼太郎 | 著	文藝春秋
雑誌	PM_00070308	論座	—	2003 年 8 月号	—	朝日新聞社
新聞	PN_01030302	朝日新聞	朝刊	2003/3/2	—	朝日新聞社
白書	WR_00000003	わが外交の近況	昭和 51 年版（上）	—	外務省	大蔵省印刷局
教科書	TB_01000009	国語　五上　銀河	—	—	宮地裕 | ほか著	光村図書出版
広報紙	PR_14212017	広報あつぎ	—	2008 年 17 号	—	神奈川県厚木市
Yahoo!知恵袋	YC_00297502	Yahoo!知恵袋	—	—	—	Yahoo!
Yahoo!ブログ	YB_00002691	Yahoo!ブログ	—	—	—	Yahoo!
韻文	VE_93066308	山村暮鳥詩集	—	—	山村暮鳥 ‖ 著	思潮社
法律	LA_S63HO108	消費税法	—	昭和六十三年十二月三十日法律第百八号	—	—
国会会議録	MD_02010001	国会会議録	—	第 154 回国会	—	—

メディア	Year	ISBN	Size	Pages	Genre_1	Genre_2	Genre_3	Genre_4	Bib_author_ID
書籍	2001	4167105926	16cm	368	9 文学	913	0193		00070104
雑誌	2003		A5 判	260	1 総合			月刊	
新聞	2003		ブランケット判	37	全国紙				
白書	1976				外交				
教科書	2006				国語	小	5		00045734
広報紙	2008				関東地方	神奈川県			
Yahoo!知恵袋	2005				子育てと学校	子育て，出産	子育ての悩み		
Yahoo!ブログ	2008				家庭と住まい	住まい	ガーデニング		
韻文	1991	4783708665	19cm	160	詩				00093767
法律	1988				23_国税				
国会会議録	2002				衆議院	常任委員会	環境委員会		

5.2 書誌情報データの定義

以下では，書誌情報データ（Bibliography.txt）の各列に記載された情報の定義について記述する。

5.2.1 書誌 ID

書誌 ID（Bib_ID）列は，各サンプルを取得した原本に対して一意に付された ID を表わす。

例

- 「BK_20000563」（書籍）
- 「PM_00010409」（雑誌）
- 「PN_01010202」（新聞）
- 「WR_00000001」（白書）
- 「TB_01000001」（教科書）
- 「PR_01103001」（広報紙）
- 「YC_00297787」（Yahoo!知恵袋）
- 「YB_00000549」（Yahoo!ブログ）
- 「VE_00010001」（韻文）
- 「LA_S51HO042」（法律）
- 「MD_00297787」（国会会議録）

1・2 桁目（BK，PM，PN，WR，TB，PR，YC，YB，VE，LA，MD）は，メディアの違いを表わす。区切り記号の「_」以降の 8 桁の数字は，原本を一意に同定するための番号を表わす。

以下では，各メディアにおける書誌 ID の構造について解説する。

「書籍」の書誌 ID

「書籍」の書誌 ID は，以下の構造を持つ。

- BK_20000215 ～ BK_99131275，BK_XXXXXX02 ～ BK_XXXXXX40，BK_7501115D ～ BK_8900620D

1・2 桁目	「BK」	「書籍（Book）」であることを表す。
3 桁目	「_」	区切り記号。
4～11 桁目		原本に付された一意の ID。

※ 4～11 桁目の ID は，国立国会図書館の「全国書誌番号」に対応している。

※ 4 桁目が「X」で始まる ID は，2005 年 10 月時点で「全国書誌番号」が存在しなかったため，その代替として我々のサブグループで独自に付与した ID である。

※ 11 桁目に「D」が付されているものは，全国書誌番号が上下巻に対して 1 つしか振られていないため，下巻の最終桁を「D」に変更したものである。

「雑誌」の書誌 ID

「雑誌」の書誌 ID は，以下の構造を持つ。

- PM_00010120 ～ PM_12590109

1・2 桁目	「PM」	「雑誌（Magazine)」であることを表す。
3 桁目	「_」	区切り記号。
4～7 桁目		同一タイトルの雑誌に付された一意の ID。
8～9 桁目		発行年。
10～11 桁目		その発行年における号数。

※ 4～7 桁目（0001～1259）は，雑誌の母集団に含まれる 1,259 タイトルに対して，我々のサブグループで独自に付与した ID である。例えば,「0001」は『ＡＥＲＡ』に,「0002」は『ＡＳＡＨＩパソコン』に，それぞれ対応している。

※ 2001 年から 2005 年の間にタイトルの改題があった場合や，異なるタイトルを持つ増刊号が発行された場合，同じ ID の中で異なるタイトル表示が生じることがある。

- 1229 『Ｙｏｍｉｕｒｉ　Ｗｅｅｋｌｙ』『Ｙｏｍｉｕｒｉ　Ｗｅｅｋｌｙ臨時増刊』

※ 2001 年から 2005 年の間にタイトルは継続されたものの出版社が変更されたケースがあった。この場合，同じタイトルだが異なる ID を持つことがある。

※ 8～9 桁目（01～05）は，発行年（2001 年から 2005 年）の下 2 桁を表す。

※ 10～11 桁目（01～52）が例えば「11」の場合，月刊誌ではその年の 11 月号が，週刊誌では「11 号」という号数表示を持つ冊が，それぞれ収録されている。実際の巻号表示に関する情報は,「**巻号 (Number)**」列で表わされる。

「新聞」の書誌 ID

「新聞」の書誌 ID は，以下の構造を持つ。

- PN_01010125 ～ PN_31041101

1・2 桁目	「PN」	「新聞（Newspaper)」であることを表す。
3 桁目	「_」	区切り記号。
4～5 桁目		新聞タイトル・朝夕刊の別を表す ID。
6～7 桁目		発行年。
8～11 桁目		発行日。

※ 4～5 桁目（01～31）は，新聞の母集団に含まれる 16 タイトル，および朝夕刊の別について，我々のサブグループで独自に付与した ID である。例えば,「01」は『朝日新聞』の朝刊に,「31」は『琉球新報』の夕刊に，それぞれ対応する。ID とタイトルの対応については，79 ページの 5.3.3 を参照。

※ 6〜7 桁目（01〜05）は，発行年（2001 年から 2005 年）の下 2 桁を表わす。

※ 8〜11 桁目（0101〜1231）は，新聞の発行日（1 月 1 日から 12 月 31 日）を 4 桁で表す。

「白書」の書誌 ID

「白書」の書誌 ID は，以下の構造を持つ。

- WR_00000001 〜 WR_00001006

1・2 桁目	「WR」	「白書」であることを表す。
3 桁目	「_」	区切り記号。
4〜11 桁目		原本に付された一意の ID。

※ 4〜11 桁目の ID は，白書の母集団に含まれる 1,006 冊に対して，国立国語研究所で独自に付与した ID である。

「教科書」の書誌 ID

「教科書」の書誌 ID は，以下の構造を持つ。

- TB_01000001 〜 TB_91000002

1・2 桁目	「TB」	「教科書（TextBook)」であることを表す。
3 桁目	「_」	区切り記号。
4 桁目		教科。 「0」= 国語　「3」= 社会　「6」= 芸術　「9」= 生活 「1」= 数学　「4」= 外国語　「7」= 保健体育 「2」= 理科　「5」= 技術家庭　「8」= 情報
5 桁目		学校。 「1」= 小学校　「2」= 中学校　「3」= 高校
6〜11 桁目		教科・学校ごとに分類された教科書の通し番号。

「広報紙」の書誌 ID

「広報紙」の書誌 ID は，以下の構造を持つ。

- PR_01103001 〜 PR_47209008

1・2 桁目	「PR」	「広報紙（Public Relations)」であることを表す。
3 桁目	「_」	区切り記号。
4〜8 桁目		自治体に付された一意の ID。
9〜11 桁目		その自治体における号数。

※ 4〜8 桁目の ID は，総務省「全国地方公共団体コード」の上 5 桁に対応している。

※ 10〜11 桁目（01〜36）が例えば「11」の場合，2008 年にその自治体で 11 冊目に発行された広報紙を指す。

「Yahoo!知恵袋」の書誌 ID

「Yahoo!知恵袋」の書誌 ID は，以下の構造を持つ。

- YC_00297287 〜 YC_00585157

1・2 桁目	「YC」	「Yahoo!知恵袋（Yahoo! Chiebukuro）」であることを表す。
3 桁目	「_」	区切り記号。
4〜11 桁目		「Yahoo!知恵袋」の小カテゴリごとに付された一意の ID。

※ BCCWJ には，異なりで 130 の小カテゴリが収録されている。

※ Yahoo!知恵袋の小カテゴリについては，81 ページの 5.3.5 を参照。

「Yahoo!ブログ」の書誌 ID

「Yahoo!ブログ」の書誌 ID は，以下の構造を持つ。

- YB_00000075 〜 YB_00023084

1・2 桁目	「YB」	「Yahoo!ブログ（Yahoo! Blog）」であることを表す。
3 桁目	「_」	区切り記号。
4〜11 桁目		「Yahoo!ブログ」の小カテゴリごとに付された一意の ID。

※ BCCWJ には，異なりで 316 の小カテゴリが収録されている。

※ Yahoo!ブログの小カテゴリについては，84 ページの 5.3.6 を参照。

「韻文」の書誌 ID

「韻文」の書誌 ID は，以下の構造を持つ。

- VE_00010001 〜 VE_99099368

1・2 桁目	「VE」	「韻文（Verse）」であることを表す。
3 桁目	「_」	区切り記号。
4〜11 桁目		原本に付された一意の ID。

※ 4〜11 桁目の ID は，詩の場合，国立国会図書館の「全国書誌番号」に対応している。短歌・俳句の場合，4〜7 桁目が短歌（0001）と俳句（0002）の別を表し，8〜11 桁目が個々の歌集・句集に独自に付与した ID を表す。

「法律」の書誌 ID

「法律」の書誌 ID は，以下の構造を持つ。

- LA_S51HO042 ～ LA_H17HO124

1・2 桁目	「LA」	「法律（Law)」であることを表す。
3 桁目	「_」	区切り記号。
4～6 桁目		法律の公布年。
7～8 桁目		「法律（HO)」であることを表す。
4～11 桁目		法令番号。

※ 4～11 桁目の ID は，Web 上の「法令データ提供システム」においてその法律が表示される HTML ファイル名に相当する。

「国会会議録」の書誌 ID

「国会会議録」の書誌 ID は，以下の構造を持つ。

- MD_00010004 ～ MD_99060001

1・2 桁目	「MD」	「国会会議録（Minutes of the Diet)」であることを表す。
3 桁目	「_」	区切り記号。
4～5 桁目		開催年。
6～7 桁目		会議種別。
		「01」= 衆議院・常任委員会　「05」= 参議院・常任委員会
		「02」= 衆議院・特別委員会　「06」= 参議院・特別委員会
		「03」= 衆議院・本会議　「07」= 参議院・本会議
		「04」= 衆議院・その他　「08」= 参議院・その他
8～11 桁目		会議種別ごとの会議に付された一意の ID。

※ 4～5 桁目（76～05）は，会議の開催年（19<u>76</u> 年から 20<u>05</u> 年）の下 2 桁を表わす。

※ 国会会議録の会議種別と会議名称の詳細については，91 ページの 5.3.8 を参照。

5.2.2　タイトル

タイトル（Title）列は，原本のタイトルを表わす。

例

- 「ファンの心をときめかせた世界の映画ベストセレクション」（書籍）
- 「塩狩峠; 道ありき」（書籍）
- 「週刊朝日」（雑誌）
- 「北海道新聞」（新聞）
- 「情報通信白書」（白書）
- 「こくご　一上　かざぐるま」（教科書）
- 「広報あげお」（広報紙）
- 「Yahoo!知恵袋」（Yahoo!知恵袋）
- 「Yahoo!ブログ」（Yahoo!ブログ）
- 「谷川俊太郎詩集」（韻文）
- 「民事保全法」（法律）
- 「国会会議録」（国会会議録）

※ 書籍のうち，特にアンソロジーなどでは，1 冊に複数のタイトルが併記されることがある。その場合，タイトル間は「;」で区切られる。

※ 「Yahoo!知恵袋」「Yahoo!ブログ」「国会会議録」の場合，タイトルはそれぞれ 1 通りとなる。

5.2.3　副題

副題（Subtitle）列は，原本の副題・サブタイトルを表わす。

例

- 「伝説の呼び屋・永島達司の生涯」（書籍）
- 「朝刊」（新聞）
- 「平成 4 年版」（白書）
- 「サラダ記念日」（韻文）

※ 「新聞」の場合,「朝刊」「夕刊」の別が表わされる。

※ 「韻文」の場合，短歌は歌集のタイトル，俳句は句集のタイトルが表される。詩は，副題の情報は付与されない。

※ 「雑誌」「教科書」「広報紙」「Yahoo!知恵袋」「Yahoo!ブログ」「法律」「国会会議録」には，副題の情報は付与されない。

5.2.4 巻号

巻号（Number）列は，原本の巻号・巻次に関する情報を表わす。

例

- 「第 6 巻」（書籍）
- 「3(神の星編)」（書籍）
- 「2002 年 4 月 15 日号（第 15 巻第 16 号，通巻 750 号)」（雑誌）
- 「サンデー毎日臨時増刊（第 80 巻第 49 号，通巻 4467 号)」（雑誌）
- 「2001/10/24」（新聞）
- 「2008 年 12 号」（広報紙）
- 「第 17 巻 (昭和 55 年〜昭和 63 年)」（韻文）
- 「平成元年六月二十八日法律第五十八号」（法律）
- 「第 154 回国会」（国会会議録）

※ 「雑誌」の場合，背表紙などに記載されている巻号表示の情報が表わされる。あるタイトルの増刊号であることの情報が表わされることもある。

※ 「新聞」の場合，発行日が表わされる。

※ 「韻文」の場合，短歌・俳句のみ，原本の巻号情報が表される。

※ 「法律」の場合，法律の公布日および法令番号が表される。

※ 「国会会議録」の場合，開催国会の回次が表わされる。

※ 「白書」「教科書」「Yahoo!知恵袋」「Yahoo!ブログ」には，巻号の情報は付与されない。

5.2.5　責任表示

責任表示（Bib_author）列は，原本の責任表示（著者，編者，監修者など）の情報を表わす。

例

- 「司馬遼太郎 | 著」（書籍）
- 「七田眞，七田厚 | 著」（書籍）
- 「高橋貞巳 | 監修 ; 三菱総合研究所 | 著」（書籍）
- 「カフカ | 著 ; 池内紀 | 訳」（書籍）
- 「ロナルド・A. モース | 編著 ; 日下公人 | 監修 ; 時事通信社外信部 | ほか訳」（書籍）
- 「経済産業省 ; 厚生労働省 ; 文部科学省」（白書）
- 「宮地裕 | ほか著」（教科書）

※ 責任表示列に記載された人名・組織名は，サンプルに含まれる文章の書き手と同一であるとは限らない。103 ページの**人名録データ**（Directory.txt）での記述を参照。

※ 「書籍」「教科書」では，複数の人名などが役割ごとに併記されることがある。その場合，人名・組織名が「;」で区切られる。同一役割の場合は,「, 」で区切られる。

※ 「白書」では，複数の組織名が「;」で区切られることがある。

※ 「雑誌」「新聞」「広報紙」「Yahoo!知恵袋」「Yahoo!ブログ」「法律」「国会会議録」には，責任表示の情報は付与されない。

5.2.6　出版者

出版者（Publisher）列は，原本の出版者（出版社）を表わす。

例

- 「岩波書店」（書籍）
- 「日本図書刊行会 ; 近代文芸社 (発売)」（書籍）
- 「マガジンハウス」（雑誌）
- 「株式会社朝日新聞社」（新聞）
- 「大蔵省印刷局」（白書）
- 「光村図書出版株式会社」（教科書）
- 「北海道札幌市東区」（広報紙）
- 「Yahoo!」（Yahoo!知恵袋，Yahoo!ブログ）
- 「筑摩書房」（韻文）

※ 書籍では，複数の出版者が役割ごとに併記されることがある。その場合，出版者間は「;」で区切られる。

※ 「Yahoo!知恵袋」「Yahoo!ブログ」の場合，出版者は「Yahoo!」の 1 通りとなる。

※ 「法律」「国会会議録」には，出版者の情報は付与されない。

5.2.7 出版年

出版年（Year）列は，4 桁の数字で，原本が出版された年を表わす。

例

- 「2001」

※ 「Yahoo!知恵袋」の場合，出版年は「2005」の 1 通りとなる。

※ 「広報紙」「Yahoo!ブログ」の場合，出版年は「2008」の 1 通りとなる。

※ 「国会会議録」の場合，出版年は会議が開催された年を表わす。

5.2.8 ISBN

ISBN（ISBN）列は，原本に付された ISBN（国際標準図書番号）を表わす。

例

- 「4889916687」

※ ISBN は，2007 年以降，13 桁の規格になっているが，ここでは 10 桁の旧規格となる。

※ 「書籍」「韻文」以外のメディアには，ISBN の情報は付与されない。

5.2.9 判型

判型（Size）列は，原本の大きさを表わす。

例

- 「20cm」（書籍）
- 「A4 変型判」（雑誌）
- 「ブランケット判」（新聞）
- 「23cm」（韻文）

※ 「書籍」「雑誌」「新聞」「韻文」以外のメディアには，判型の情報は付与されない。

5.2.10　ページ数

ページ数（Pages）列は，原本のページ数を表わす。

例

- 「222」

※ 「書籍」「雑誌」「新聞」「韻文」以外のメディアには，ページ数の情報は付与されない。

5.2.11　ジャンル (1)〜(4)

ジャンル (1)〜(4)（Genre_1〜Genre_4）列は，原本のジャンルに関連した情報を表わす。

表 5.3: ジャンル情報の例

メディア	ジャンル (1)	ジャンル (2)	ジャンル (3)	ジャンル (4)
書籍	9 文学	913	0193	
雑誌	1 総合	一般	総合誌	週刊
新聞	全国紙			
白書	外交			
教科書	国語	小	3	
広報紙	東北地方	青森県		
Yahoo!知恵袋	子育てと学校	子育て，出産	子育ての悩み	
Yahoo!ブログ	家庭と住まい	ペット，動物	犬	
韻文	短歌			
法律	35_金融・保険			
国会会議録	衆議院	常任委員会	文教委員会	

以下では，各メディアにおけるジャンルの分類について解説する。

「書籍」のジャンル情報

「書籍」のジャンル情報は，以下のように構成されている。

ジャンル (1)	「NDC（日本十進分類法）第 9 版」第 1 次区分（類） + 分類名
ジャンル (2)	「NDC（日本十進分類法）第 9 版」第 3 次区分（目）
ジャンル (3)	C コード

ジャンル (1) 「書籍」のジャンル (1) 列には，国立国会図書館で付与された「NDC（日本十進分類法）第 9 版」の第 1 次区分（類）を表す数値と，その分類名が記載されている。

例

- 「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」
- 「分類なし」

※ 「分類なし」は，2005 年 10 月時点で国立国会図書館で NDC が付与されていなかった場合に相当する。

ジャンル (2) 「書籍」のジャンル (2) 列には，国立国会図書館で付与された「NDC（日本十進分類法）第 9 版」の第 3 次区分（目）を表す数値が記載されている。

例

- 「002」 〜 「992」,「(空文字)」

※ NDC の第 2 次区分（綱，左から 2 桁目まで）の分類（綱目表）は，74 ページの 5.3.1 を参照。第 3 次区分（目，左から 3 桁目まで）までの詳細な分類（要目表）については，『日本十進分類法 新訂 9 版』（日本図書館協会）などを参照。

※ 空文字は，2005 年 10 月時点で国立国会図書館で NDC が付与されていなかった場合に相当する。

ジャンル (3) 「書籍」のジャンル (3) 列には,「C コード（図書分類コード)」が記載されている。

例

- 「0000」 〜 「9979」,「(空文字)」

「C コード」は日本図書コードの一部で，4 桁の数値で構成される。左から 1 桁目は「販売対象コード」で，対象読者を表わす。2 桁目は「発行形態コード」で，発行形態を表わす。3・4 桁目は「内容コード」で，書籍の内容を表わす。

※ 「C コード」の 1 桁目「販売対象コード」の分類を，以下に示す。

「0」＝ 一般　「3」＝ 専門　「6」＝ 学参 I（小中）　「9」＝ 雑誌扱い
「1」＝ 教養　「4」＝ （欠番）　「7」＝ 学参 II（高校）
「2」＝ 実用　「5」＝ 婦人　「8」＝ 児童

※ 「C コード」の 2 桁目「発行形態コード」の分類を，以下に示す。

「0」＝ 単行本	「3」＝ 全集・双書	「6」＝ 図鑑	「9」＝ コミック
「1」＝ 文庫	「4」＝ ムック・その他	「7」＝ 絵本	
「2」＝ 新書	「5」＝ 事・辞典	「8」＝ 磁性媒体など	

※ 「C コード」の 3・4 桁目「内容コード」の分類は，76 ページの 5.3.1 を参照。

※ 空文字は，C コードの情報が入手できなかった場合に相当する。

「雑誌」のジャンル情報

「雑誌」のジャンル情報は，以下のように構成されている。

ジャンル (1)	大ジャンル
ジャンル (2)	中ジャンル
ジャンル (3)	小ジャンル
ジャンル (4)	刊行形態

このうち,「雑誌」のジャンル (1) 列には,「大ジャンル」の情報が，雑誌タイトルごとに記載されている。

例

- 「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」

※ 「大ジャンル」の情報は,『雑誌新聞総かたろぐ』の記載に基づく。

※ 「中ジャンル」「小ジャンル」については，77 ページの 5.3.2 を参照。

また,「雑誌」のジャンル (4) 列には，雑誌タイトルの「刊行形態」が記載されている。

例

- 「月刊」「週刊」「隔週刊」「隔月刊」「月 2 回刊」「年刊」「季刊」...

※ 「刊行形態」の情報は,『雑誌新聞総かたろぐ』の記載に基づく。

「新聞」のジャンル情報

「新聞」のジャンル情報は，以下のように構成されている。

ジャンル (1)	配達エリア

ジャンル (1) 「新聞」のジャンル (1) 列には，その新聞タイトルが配達される範囲の違いによって，以下の分類が記載されている。

例
- 「全国紙」「ブロック紙」「地方紙」

※ 配達エリアに関する情報は,『全国新聞ガイド』の記載に基づく。

※ 新聞の各タイトルとジャンルの対応については，79 ページの 5.3.3 を参照。

「白書」のジャンル情報

「白書」のジャンル情報は，以下のように構成されている。

ジャンル (1)	ジャンル名

ジャンル (1) 「白書」のジャンル (1) 列には，9 種類のジャンル名が記載されている。

例
- 「安全」「外交」「科学技術」「環境」「教育」「経済」「国土交通」「農林水産」「福祉」

※ 白書のジャンルは，各タイトルの内容に応じて，我々のサブグループで独自に分類したものである。

※ 白書のタイトルとジャンル名との対応については，80 ページの 5.3.4 を参照。

「教科書」のジャンル情報

「教科書」のジャンル情報は，以下のように構成されている。

ジャンル (1)	教科
ジャンル (2)	学校
ジャンル (3)	学年

ジャンル (1) 「教科書」のジャンル (1) 列には，教科の別が記載されている。

例
- 「国語」「数学」「理科」「社会」「外国語」「技術家庭」「芸術」「生活」「保健体育」「情報」

ジャンル (2)　「教科書」のジャンル (2) 列には，学校の別が記載されている。

例
- 「小学校」「中学校」「高校」

ジャンル (3)　「教科書」のジャンル (3) 列には，学年の別が記載されている。

例
- 「1」「2」「3」「4」「5」「6」「(空文字)」

※「学年」の情報は，小学校・中学校の場合にのみ記載される。高校の場合は空文字になる。

「広報紙」のジャンル情報

「広報紙」のジャンル情報は，以下のように構成されている。

ジャンル (1)	地域
ジャンル (2)	都道府県

ジャンル (1)　「広報紙」のジャンル (1) 列には，当該の自治体の地域が記載されている。

例
- 「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」

ジャンル (2)　「教科書」のジャンル (2) 列には，当該の自治体の都道府県が記載されている。

例
- 「北海道」～「沖縄県」

「Yahoo!知恵袋」のジャンル情報

「Yahoo!知恵袋」のジャンル情報は，以下のように構成されている。

ジャンル (1)	「質問」が投稿された大カテゴリ名
ジャンル (2)	「質問」が投稿された中カテゴリ名
ジャンル (3)	「質問」が投稿された小カテゴリ名

「Yahoo!知恵袋」のジャンル情報には，質問が投稿されたカテゴリの名称が記載されている。投稿されたカテゴリは，14 種類の「大カテゴリ」，59 種類の「中カテゴリ」，および 130 種類の「小カテゴリ」という 3 階層のカテゴリで構成される。「大カテゴリ」の分類を，以下に示す。

例

「エンターテインメントと趣味」	「インターネット，PC と家電」
「ビジネス，経済とお金」	「職業とキャリア」
「ニュース，政治，国際情勢」	「スポーツ，アウトドア，車」
「暮らしと生活ガイド」	「健康，美容とファッション」
「子育てと学校」	「マナー，冠婚葬祭」
「教養と学問，サイエンス」	「地域，旅行，お出かけ」
「Yahoo! JAPAN」	「その他」

※ 「Yahoo!知恵袋」の中カテゴリ・小カテゴリについては，81 ページの 5.3.5 を参照。

「Yahoo!ブログ」のジャンル情報

「Yahoo!ブログ」のジャンル情報は，以下のように構成されている。

ジャンル (1)	「記事」が投稿された大カテゴリ名
ジャンル (2)	「記事」が投稿された中カテゴリ名
ジャンル (3)	「記事」が投稿された小カテゴリ名

「Yahoo!ブログ」のジャンル情報には，ブログの記事が投稿されたカテゴリの名称が記載されいる。投稿されたカテゴリは，15 種類の「大カテゴリ」，54 種類の「中カテゴリ」，および 316 種類の「小カテゴリ」という 3 階層のカテゴリで構成される。「大カテゴリ」の分類を，以下に示す。

例

「ビジネスと経済」	「コンピュータとインターネット」	「生活と文化」
「エンターテインメント」	「家庭と住まい」	「政治」
「健康と医学」	「学校と教育」	「科学」
「出会い」	「地域」	「特集」
「芸術と人文」	「Yahoo!サービス」	「趣味とスポーツ」

※ 「Yahoo!ブログ」の中カテゴリ・小カテゴリについては，84 ページの 5.3.6 を参照。

「韻文」のジャンル情報

「韻文」のジャンル情報は，以下のように構成されている。

ジャンル (1)	「短歌」「俳句」「詩」

「法律」のジャンル情報

「法律」のジャンル情報は，以下のように構成されている。

ジャンル (1)	分類名

ジャンル (1)　「法律」のジャンル (1) 列には，43 種類のジャンル名が記載されている。

例

- 「04_国家公務員」

※ ジャンルの詳細については，90 ページの 5.3.7 を参照。

「国会会議録」のジャンル情報

「国会会議録」のジャンル情報は，以下のように構成されている。

ジャンル (1)	開催院
ジャンル (2)	会議種別
ジャンル (3)	委員会名称

ジャンル (1)　「国会会議録」のジャンル (1) 列には，2 種類の「開催院」の別が記載されている。

例

- 「衆議院」「参議院」

1

ジャンル (2)　「国会会議録」のジャンル (2) 列には，4 種類の「会議種別」が記載されている。

例

- 「常任委員会」「特別委員会」「本会議」「その他」

ジャンル (3) 「国会会議録」のジャンル (3) 列には，59 種類の「会議名称」が記載されている。

※ 会議名称の一覧，および会議種別との対応については，91 ページの 5.3.8 を参照。

5.2.12 責任表示 ID

責任表示 ID（Bib_author_ID）列は，責任表示（Bib_author）列に記載されている人名・組織名などに対して付された ID である。記載されている ID は，**人名録データ**（Directory.txt）の「人名 ID （Directory_ID）」列に記載された ID に対応している。詳しくは，103 ページの 7.2.1 を参照。

例

- 「00685074」 （書籍）
- 「00254659 ; 00184422」 （書籍）
- 「00113880 ; 00166885 ; 00124738」（教科書）
- 「00037561」（韻文）

※ 「責任表示 ID」列は，8 桁の数値（右詰，0 埋め）で表示されている。

※ 複数の責任表示が併記されている場合，責任表示 ID は「;」で区切られる。

※ 別途提供される人名録データ（Directory.txt）に含まれるのは，各サンプルに含まれる文章の著者または著作権者のみとなる。そのため，書誌情報データの責任表示 ID で記載された ID が，著者情報データには記載されていない場合がある。

※ 「書籍」「教科書」「韻文」以外のメディアには，責任表示 ID の情報は付与されない。

5.3　ジャンル情報の詳細

5.3.1　「書籍」のジャンル情報の詳細

「書籍」の NDC（第 2 次区分）

「書籍」のジャンル (2) 列に記載された 3 桁の「NDC（日本十進分類法)」のうち，第 2 次区分（左から 2 桁目まで）の分類（綱目表）を，以下に示す。第 3 次区分（左から 3 桁目まで）までの詳細な分類（要目表）については,『日本十進分類法新訂 9 版』（日本図書館協会）などを参照。

00	**総記**
01	図書館. 図書館学
02	図書. 書誌学
03	百科事典
04	一般論文集. 一般講演集
05	逐次刊行物
06	団体
07	ジャーナリズム. 新聞
08	叢書. 全集. 選集
09	貴重書. 郷土資料. その他の特別コレクション

10	**哲学**
11	哲学各論
12	東洋思想
13	西洋哲学
14	心理学
15	倫理学. 道徳
16	**宗教**
17	神道
18	仏教
19	キリスト教

20	**歴史**
21	日本史
22	アジア史. 東洋史
23	ヨーロッパ史. 西洋史
24	アフリカ史
25	北アメリカ史
26	南アメリカ史
27	オセアニア史. 両極地方史
28	伝記
29	**地理. 地誌. 紀行**

30	**社会科学**
31	政治
32	法律
33	経済
34	財政
35	統計
36	社会
37	教育
38	風俗習慣. 民俗学. 民族学
39	国防. 軍事

40	**自然科学**
41	数学
42	物理学
43	化学
44	天文学. 宇宙科学
45	地球科学. 地学
46	生物科学. 一般生物学
47	植物学
48	動物学
49	**医学. 薬学**

50	**技術. 工学**
51	建設工学. 土木工事
52	建築学
53	機械工学. 原子力工学
54	電気工学. 電子工学
55	海洋工学. 船舶工学. 兵器
56	金属工学. 鉱山工学
57	化学工業
58	製造工業
59	**家政学. 生活科学**

60	**産業**
61	農業
62	園芸
63	蚕糸業
64	畜産業. 獣医学
65	林業
66	水産業
67	商業
68	運輸. 交通
69	通信事業

70	**芸術. 美術**
71	彫刻
72	絵画. 書道
73	版画
74	写真. 印刷
75	工芸
76	音楽. 舞踊
77	演劇. 映画
78	**スポーツ. 体育**
79	**諸芸. 娯楽**

80	**言語**
81	日本語
82	中国語. その他の東洋の諸言語
83	英語
84	ドイツ語
85	フランス語
86	スペイン語
87	イタリア語
88	ロシア語
89	その他の諸言語

90	**文学**
91	日本文学
92	中国文学. その他の東洋文学
93	英米文学
94	ドイツ文学
95	フランス文学
96	スペイン文学
97	イタリア文学
98	ロシア・ソヴィエト文学
99	その他の諸文学

「書籍」の C コード（内容コード）

「書籍」のジャンル (3) 列に記載された 4 桁の「C コード」のうち，3・4 桁目に記載された「内容コード」の分類を，以下に示す。なお，C コードの「内容コード」は，NDC の第 2 次区分（左から 2 桁目まで）の分類と必ずしも一致するわけではない。

00	総記
01	百科事典
02	年鑑・雑誌
04	情報科学
10	哲学
11	心理（学）
12	倫理（学）
14	宗教
15	仏教
16	キリスト教
20	歴史総記
21	日本歴史
22	外国歴史
23	伝記
25	地理
26	旅行
30	社会科学総記
31	政治-含む国防軍事
32	法律
33	経済・財政・統計
34	経営
36	社会
37	教育
39	民族・風習
40	自然科学総記
41	数学
42	物理学
43	化学
44	天文・地学
45	生物学
47	医学・歯学・薬学
50	工学・工学総記
51	土木
52	建築
53	機械
54	電気
55	電子通信
56	海事
57	採鉱・冶金
58	その他の工業
60	産業総記
61	農林業
62	水産業
63	商業
65	交通・通信
70	芸術総記
71	絵画・彫刻
72	写真・工芸
73	音楽・舞踊
74	演劇・映画
75	体育・スポーツ
76	諸芸・娯楽
77	家事
79	コミックス・劇画
80	語学総記
81	日本語
82	英米語
84	ドイツ語
85	フランス語
87	各国語
90	文学総記
91	日本文学総記
92	日本文学詩歌
93	日本文学、小説・物語
95	日本文学、評論、随筆、その他
97	外国文学小説
98	外国文学、その他

5.3.2 「雑誌」のジャンル情報の詳細

「雑誌」のジャンル情報の詳細

「雑誌」のジャンル (1)〜(3) 列には，6 種類の「大ジャンル」，27 種類の「中ジャンル」，71 種類の「小ジャンル」という 3 階層のカテゴリが記載されている。その一覧を，以下に示す。

大ジャンル	中ジャンル	小ジャンル
1_総合	総記／マスコミ	総記
		マスコミ（新聞・放送）
		出版・読書・図書館
		出版情報・書評
	一般	一般週刊誌
		総合誌
		女性週刊誌
		婦人誌
		読み物
		東京都／タウン・地域誌
		関東地方／タウン・地域誌
		近畿地方／タウン・地域誌
	家庭／生活	生活情報
		ファッション
		料理・栄養
		住居・インテリア
		育児・家庭教育
	児童	少年
		少女
	娯楽／芸能	ヤング
		テレビ・ラジオ・芸能・映画
	レジャー／趣味	レジャー
		旅行・観光
		趣味の乗り物
		釣り・狩猟
		写真・カメラ
		家庭園芸
		ホビー・クラフト・日曜大工
		模型・無線・コンピュータゲーム
		音楽・オーディオ
		囲碁・将棋
		ペット

大ジャンル	中ジャンル	小ジャンル
1_総合（続き）	スポーツ	スポーツ一般・陸上競技
		アウトドア・海／山
		球技
		ゴルフ
		武道・格闘技
2_教育・学芸	教育	教育技術
	学習／語学	小・中学生
		高校・大学生
	文学／芸術	文学文芸総合
		大衆文芸
		俳句
		短歌
		芸術・美術
	人文科学	宗教
	社会科学	歴史一般
	自然科学	自然科学一般
		地球宇宙科学
3_政治・経済・商業	政治／外交	国会行政
		海外情勢外交
	経済／経営	経営／経済
	金融／財政	金融財政
	商業／消費者	広告宣伝・ＰＲ
	国勢／民力	国勢／民力
		所得・物価・消費
4_産業	農林水産	農業経営
	食料／食品	醸造業
	運輸／通信	海事・海運・港湾
5_工業	工業一般	公害・環境保全
	建設／土木	建設一般
	機械	機械一般
		自動車・オートバイ・自転車
	電気機／電子	家電・弱電・照明
		エレクトロニクス
		コンピュータ／情報処理
		電波・電気通信
6_厚生・医療	厚生	福祉
	医学	医学総合
		家庭医学・健康

5.3.3 「新聞」のジャンル情報の詳細

「新聞」の書誌 ID の 4〜5 桁目で表わされる ID（01〜31）は，新聞の母集団に含まれる 16 タイトル，および朝夕刊の別について，我々のサブグループで独自に付与した ID である。各タイトルに対応づけられたジャンル（配達エリア）との対応関係は，以下のようになっている。

ID	タイトル	朝夕刊	配達エリア	ID	タイトル	朝夕刊	配達エリア
01	朝日新聞	朝刊	全国紙	17	河北新報	朝刊	地方紙
02	朝日新聞	夕刊	全国紙	18	河北新報	夕刊	地方紙
03	毎日新聞	朝刊	全国紙	19	新潟日報	朝刊	地方紙
04	毎日新聞	夕刊	全国紙	20	新潟日報	夕刊	地方紙
05	読売新聞	朝刊	全国紙	21	京都新聞	朝刊	地方紙
06	読売新聞	夕刊	全国紙	22	京都新聞	夕刊	地方紙
09	産経新聞	朝刊	全国紙	23	神戸新聞	朝刊	地方紙
10	産経新聞	夕刊	全国紙	24	神戸新聞	夕刊	地方紙
11	北海道新聞	朝刊	ブロック紙	25	中国新聞	朝刊	地方紙
12	北海道新聞	夕刊	ブロック紙	26	中国新聞	夕刊	地方紙
13	中日新聞	朝刊	ブロック紙	27	高知新聞	朝刊	地方紙
14	中日新聞	夕刊	ブロック紙	28	高知新聞	夕刊	地方紙
15	西日本新聞	朝刊	ブロック紙	30	琉球新報	朝刊	地方紙
16	西日本新聞	夕刊	ブロック紙	31	琉球新報	夕刊	地方紙

※ ID「07」「08」「29」は，著作権処理の都合上，欠番となった。

5.3.4 「白書」のジャンル情報の詳細

「白書」のジャンル (1) 列には，白書のタイトルおよび内容によって分類した 9 種類のジャンル名（「安全」「外交」「科学技術」「環境」「教育」「経済」「国土交通」「農林水産」「福祉」）が記載されている。各ジャンルと白書のタイトルは，以下のように対応している。

ジャンル	白書タイトル
安全	警察白書
	原子力安全白書
	原子力白書
	交通安全白書
	公害紛争処理白書
	消防白書
	犯罪白書
	防衛白書 / 日本の防衛
	防災白書
外交	外交青書 / わが外交の近況
	政府開発援助（ODA）白書 / 我が国の政府開発援助
科学技術	科学技術白書
	情報通信白書 / 通信白書
環境	環境白書
	循環型社会白書
教育	文部科学白書 / 我が国の文教施策
経済	エネルギー白書
	ものづくり白書 / 製造基盤白書
	経済財政白書 / 経済白書
	公益法人白書
	地方財政白書
	中小企業白書
	通商白書
	独占白書 / 独占禁止白書
	労働経済白書 / 労働白書

ジャンル	白書タイトル
国土交通	観光白書
	国土交通白書 / 運輸白書 / 建設白書
	首都圏白書
	土地白書 / 国土利用白書
農林水産	食料・農業・農村白書 / 農業白書
	森林・林業白書 / 林業白書
	水産白書 / 漁業白書
福祉	厚生労働白書 / 厚生白書
	高齢社会白書
	国民生活白書
	少子化社会白書
	障害者白書
	人権教育・啓発白書
	青少年白書
	男女共同参画白書

※ 「防衛白書 / 日本の防衛」のように,「/」で区切られている白書タイトルは，1976 年から 2005 年までの間にタイトルの変更があったことを表わす。

5.3.5 「Yahoo!知恵袋」のジャンル情報の詳細

「Yahoo!知恵袋」のジャンル (1)〜(3) 列には，14 種類の「大カテゴリ」，59 種類の「中カテゴリ」，および 130 種類の「小カテゴリ」という 3 階層のカテゴリがそれぞれ記載されている。大カテゴリ・中カテゴリ・小カテゴリの一覧を，以下に示す。

大カテゴリ	中カテゴリ	小カテゴリ
エンターテインメントと趣味	ゲーム	ゲーム
		オンラインゲーム
		トレーディングカード
	テレビ、ラジオ	テレビ、ラジオ
		CM
		ラジオ
	映画	映画
	音楽	音楽
		楽器
		邦楽
		洋楽
	芸能人、タレント	芸能人、タレント
		あの人は今
		話題の人物
	占い、超常現象	占い、懸賞
	本、雑誌、コミック	本、雑誌、コミック
		コミック
		雑誌
インターネット、PC と家電	インターネット	インターネット
	パソコン、周辺機器	パソコン、周辺機器
	家電、AV 機器	家電、AV 機器
		オーディオ
	携帯電話、モバイル	携帯電話、モバイル
ビジネス、経済とお金	家計、貯金	家計、貯金
		ローン
		家計、節約
		貯金
	株と経済	株と経済
		株式
		経済、景気
	企業と経営	企業と経営
		会計、経理、財務
		会社情報、業界動向
		企業法務、知的財産
		起業
	保険、税金、年金	保険、税金、年金
		税金
		年金
		保険

大カテゴリ	中カテゴリ	小カテゴリ
職業とキャリア	資格、習い事	資格、習い事
		資格
		専門学校、職業訓練
	就職、転職	就職、転職
		就職活動
		退職、入社手続き
	派遣、アルバイト、パート	派遣、アルバイト、パート
		アルバイト、フリーター
		パート
		派遣
	労働問題、働き方	労働問題、働き方
		失業、リストラ
		労働条件、給与、残業
		労働問題
ニュース、政治、国際情勢	ニュース、事件	ニュース、事件
		事件、事故、流行
		話題のことば
	政治、社会問題	政治、社会問題
スポーツ、アウトドア、車	アウトドア	アウトドア
		キャンプ
		釣り
	スポーツ	スポーツ
		オリンピック
		サッカー
		ダイビング、サーフィン
		格闘技、武術
		野球
	バイク	バイク
	自動車	自動車
		新車
		中古車
暮らしと生活ガイド	ショッピング	ショッピング
		これ、探してます
	ボランティア、環境問題、国際協力	ボランティア、環境問題、国際協力
	家事、住宅	家事、住宅
		家事
		不動産、引越し
	公共施設、役所	公共施設、役所
		美術館、博物館、図書館
		役所、手続き
	福祉、介護	福祉、介護
	法律、消費者問題	法律、消費者問題
		消費者問題
		法律相談

大カテゴリ	中カテゴリ	小カテゴリ
暮らしと生活ガイド（続き）	料理、グルメ、レシピ	お酒、ドリンク
		レシピ、調理法
		飲食店、デパ地下
		料理、食材
		料理、グルメ、レシピ
健康、美容とファッション	コスメ、美容	コスメ、美容
		エステ、マッサージ
		コスメ、化粧品
	ファッション	ファッション
	メンタルヘルス	カウンセリング、治療
		ストレス
		心の悩み、相談
	健康、病気、ダイエット	健康、病気、ダイエット
		ダイエット
		病気、症状、ヘルスケア
	恋愛相談、人間関係の悩み	恋愛相談、人間関係の悩み
子育てと学校	子育て、出産	子育て、出産
		子どもの病気とトラブル
		子育ての悩み
		妊娠、出産
	受験、進学	受験、進学
	小・中学校、高校	小・中学校、高校
	大学、留学	大学、留学
		大学
		留学
	幼児教育、幼稚園、保育園	幼児教育、幼稚園、保育園
マナー、冠婚葬祭	マナー	マナー
		あいさつ、てがみ、文例
	冠婚葬祭	冠婚葬祭
		結婚
		葬儀
	祭りと年中行事	祭りと年中行事
教養と学問、サイエンス	一般教養	一般教養
	芸術、文学、歴史	芸術、文学、歴史
	言葉、語学	言葉、語学
	数学、サイエンス	数学、サイエンス
	天気、天文、宇宙	天気、天文、宇宙
	動物、植物、ペット	動物、植物、ペット
地域、旅行、お出かけ	海外	海外
	交通、地図	交通、地図
	国内	国内
		花火大会
Yahoo! JAPAN	Yahoo!オークション	Yahoo!オークション
	Yahoo!サービス	Yahoo!サービス
	Yahoo!知恵袋	Yahoo!知恵袋
その他	アダルト	アダルト
	ギャンブル	ギャンブル

5.3.6 「Yahoo!ブログ」のジャンル情報の詳細

「Yahoo!ブログ」のジャンル (1)〜(3) 列には，15 種類の「大カテゴリ」，54 種類の「中カテゴリ」，および 316 種類の「小カテゴリ」という 3 階層のカテゴリがそれぞれ記載されている。大カテゴリ・中カテゴリ・小カテゴリの一覧を，以下に示す。

大カテゴリ	中カテゴリ	小カテゴリ
ビジネスと経済	金融と投資	通貨、為替
		株式
		保険
		貯蓄、預金
		銀行
		不動産
		その他金融と投資
	雇用	就職
		転職
		アルバイト
		人材派遣
		失業、無職
		その他雇用
	ビジネス	会社経営
		起業
		その他ビジネス
	職種	事務職
		営業職
		技術職
		企画職
		専門職
		公務員
		その他職種
	経済	景気
		国際経済
		その他経済
コンピュータとインターネット	インターネット	ホームページ
		ネットサービス
		その他インターネット
	コンピュータ	ソフトウェア
		パソコン
		周辺機器
		Windows
		Macintosh
		その他コンピュータ
		UNIX
生活と文化	祝日、記念日、年中行事	クリスマス
		正月
		誕生日
		バレンタインデー
		花火
		ホワイトデー
		花見
		エイプリルフール
		その他祝日、記念日、年中行事

大カテゴリ	中カテゴリ	小カテゴリ
生活と文化 （続き）	グルメ、ドリンク	レシピ
		飲食店
		食べ物
		飲み物
		菓子、デザート
	環境問題	その他環境問題
		省エネ
		自然保護
		リサイクル
		ごみ問題
		地球温暖化
	事件・事故	事件
		事故
		防犯
	災害	火災
		地震
		台風
		火山活動
		その他災害
	文化活動	宗教
		ボランティア活動
		祭りと伝統
		その他文化活動
	季節	冬
		秋
		夏
		春
エンターテインメント	映画	俳優、女優
		その他映画
		映画祭
		映画レビュー
		映画監督
	テレビ	アナウンサー
		コマーシャル
		その他テレビ
		ドラマ番組
		バラエティ番組
	音楽	その他音楽
		音楽祭
		洋楽
		邦楽
		音楽レビュー
		ミュージシャン
	占い	心理テスト、性格診断
		タロット占い
		星占い
		血液型占い
		風水
		その他占い
	芸能人、タレント	男性
		女性
		グループ
	超常現象	幽霊、心霊
		都市伝説
		UFO
		超能力
		その他超常現象
	テーマパーク	ディズニーリゾート
		ユニバーサル・スタジオ・ジャパン
		遊園地
		その他テーマパーク

大カテゴリ	中カテゴリ	小カテゴリ
家庭と住まい	住まい	ガーデニング
		修理とリフォーム
		住居
		インテリア
	ペット、動物	昆虫
		観賞魚、水草
		鳥
		ウサギ
		ハムスター
		犬
		猫
		その他ペット
	家庭電化製品	オーディオ
		季節家電
		映像機器
		調理器具
		その他家電
	家庭	家計
		育児
		家族
		家庭環境
政治	政界と政治活動	政党、団体
		選挙
		政界
		地方自治
		軍事
		国会
		その他政界と政治活動
	国際情勢	中東情勢
		アジア情勢
		アフリカ情勢
		アメリカ情勢
		ヨーロッパ情勢
		オセアニア情勢
		その他国際情勢
健康と医学	美容と健康	フィットネス
		スキンケア
		ボディケア
		ネイルケア
		ダイエット
	病気、症状	その他美容と健康
		子どもの病気
		メンタルヘルス
		生活習慣病
		アレルギー
		その他の病気
		花粉症
学校と教育	学校	小学校
		中学校
		高校
		専門学校
		大学
		その他学校
		受験
	教育	習いごと
		幼児教育
		社会教育
		その他教育

大カテゴリ	中カテゴリ	小カテゴリ
科学	社会科学	人類学と考古学
		経済学
		心理学
		政治学
		法学
		その他社会学
	自然科学	化学
		工学
		物理学
		天文学
		気象学
		生物学
		その他自然科学
出会い	恋愛	失恋
		遠距離
		アドバイス
		片思い
		初恋
		その他恋愛
	結婚	離婚
		結婚式
		見合い
		再婚
		その他結婚
		婚約、結納
地域	日本	北海道
		青森県
		岩手県
		宮城県
		秋田県
		山形県
		福島県
		東京都
		神奈川県
		埼玉県
		千葉県
		茨城県
		栃木県
		群馬県
		山梨県
		新潟県
		長野県
		富山県
		石川県
		福井県
		愛知県
		岐阜県
		静岡県
		三重県
		大阪府
		兵庫県
		京都府
		滋賀県
		奈良県
		和歌山県
		島根県
		岡山県
		広島県
		山口県

大カテゴリ	中カテゴリ	小カテゴリ
地域 （続き）	日本 （続き）	徳島県
		香川県
		愛媛県
		高知県
		福岡県
		佐賀県
		長崎県
		熊本県
		大分県
		宮崎県
		鹿児島県
		沖縄県
	世界の地方	アジア
		アフリカ
		オセアニア
		北アメリカ
		中東
		ヨーロッパ
		ラテンアメリカ
特集	趣味とスポーツ	CLUB KEIBA
芸術と人文	芸術、アート	イラストレーション
		絵画
		写真
		工芸
		書道
		その他芸術、アート
	文学	ノンフィクション、エッセイ
		小説
		詩
		俳句、川柳
		短歌
		その他文学
		伝記、自伝
	デザイン	ファッション
		工業デザイン
		建築デザイン
		その他デザイン
	舞台、演劇	観劇
		伝統芸能
		その他舞台、演劇
	人文科学	倫理学
		哲学
		歴史
		その他人文科学
Yahoo!サービス	Yahoo!ブログ	練習用
	Yahoo!オークション	出品
		落札
		ウォッチリスト
		Yahoo!オークションストア
	Yahoo!ゲーム	その他 Yahoo!ゲーム
	Yahoo!アバター	アバター作成
	Yahoo!スポーツ	ファンタジーサッカー
	Yahoo!ショッピング	Yahoo!ショッピングストア

大カテゴリ	中カテゴリ	小カテゴリ
趣味とスポーツ	スポーツ	野球
		サッカー
		ゴルフ
		テニス
		格闘技
		モータースポーツ
		スキー
		スノーボード
		マリンスポーツ
		その他スポーツ
		陸上競技
		バスケットボール
		オリンピック
		バレーボール
		ラグビー
		卓球
	レジャー	旅行
		釣り
		登山
		散歩
		キャンプ
		その他レジャー
	趣味	読書
		漫画、コミック
		アニメーション
		ゲーム
		おもちゃ
		カラオケ
		携帯電話
		その他趣味
	乗り物	鉄道、列車
		自動車
		オートバイ
		その他乗り物
		飛行機
		自転車
	ギャンブル	パチンコ、パチスロ
		競馬
		宝くじ
		その他ギャンブル

5.3.7 「法律」のジャンル情報の詳細

「法律」のジャンル (1) 列には，データの取得元である「法令データ提供システム」で採用されている，法務省『日本現行法規』に基づく法律のジャンルが記載されている。一覧を以下に示す。

01_憲法	19_災害対策	35_金融・保険
02_国会	20_建築・住宅	37_陸運
03_行政組織	21_財務通則	38_海運
04_国家公務員	23_国税	39_航空
05_行政手続	24_専売・事業	40_貨物運送
07_地方自治	25_国債	42_郵務
08_地方財政	26_教育	43_電気通信
09_司法	27_文化	44_労働
10_民事	28_産業通則	45_環境保全
11_刑事	29_農業	46_厚生
12_警察	30_林業	47_社会福祉
14_国土開発	31_水産業	49_防衛
15_土地	32_鉱業	50_外事
16_都市計画	33_工業	
17_道路	34_商業	

5.3.8 「国会会議録」のジャンル情報の詳細

「国会会議録」のジャンル (2)〜(3) 列には，4 種類の会議種別（「常任委員会」「特別委員会」「本会議」「その他」）と会議名称が，それぞれ記載されている。会議種別と会議名称は，以下のように対応している。

会議種別	会議名称
本会議	本会議
常任委員会	安全保障委員会
	運輸委員会
	科学技術委員会
	外交防衛委員会
	外務委員会
	環境委員会
	議院運営委員会
	経済産業委員会
	決算委員会
	決算行政監視委員会
	建設委員会
	厚生委員会
	厚生労働委員会
	行政監視委員会
	国土・環境委員会
	国土交通委員会
	財政・金融委員会
	財政金融委員会
	社会労働委員会
	商工委員会
	総務委員会
	大蔵委員会
	地方行政委員会
	逓信委員会
	内閣委員会
	農林水産委員会
	文教委員会
	法務委員会
	予算委員会

会議種別	会議名称
特別委員会	ロッキード問題に関する調査特別委員会
	安全保障特別委員会
	沖縄及び北方問題に関する特別委員会
	科学技術振興対策特別委員会
	個人情報の保護に関する特別委員会
	交通安全対策特別委員会
	公害対策及び環境保全特別委員会
	国会等の移転に関する特別委員会
	国旗及び国歌に関する特別委員会
	国際平和協力等に関する特別委員会
	災害対策特別委員会
	世界貿易機関設立協定等に関する特別委員会
	政治倫理の確立及び公職選挙法改正に関する特別委員会
	青少年問題に関する特別委員会
	物価等対策特別委員会
	物価問題等に関する特別委員会
その他	議院運営委員会庶務小委員会
	憲法調査会
	国民生活・経済に関する調査会
	国民生活・経済に関する調査特別委員会高齢化社会検討小委員会
	産業・資源エネルギーに関する調査会
	少子高齢社会に関する調査会
	文教委員会入試問題に関する小委員会
	予算委員会公聴会
	予算委員会第三分科会
	予算委員会第四分科会
	予算委員会第五分科会
	予算委員会第六分科会
	予算委員会第八分科会

第6章　サンプル情報データ（Sample.txt）

6.1　サンプル情報データの概要

サンプル情報データ（Sample.txt）は，BCCWJ 収録された各サンプルの ID や抽出状況に関する情報を表わす。表 6.1 に示す 4 列（タブ区切り）によって表現されている。

表 6.1: サンプル情報データを構成する列

1. サンプル ID （Sample_ID）	各サンプルに対して一意に付された ID。
2. 書誌 ID （Bib_ID）	各サンプルを抽出した原本に対して付された ID。
3. サンプル抽出基準点 ページ （Sampling_page）	「サンプル抽出基準点」を取得したページ。
4. サンプル抽出基準点 座標 （Sampling_point）	「サンプル抽出基準点」を取得した交点。

サンプル情報データの例を，以下に示す。

	Sample_ID	Bib_ID	Sampling_page	Sampling_point
出版・書籍	PB10_00047	BK_20205918	163	5D
出版・雑誌	PM11_00053	PM_10550109	76	9F
出版・新聞	PN1a_00013	PN_01010225	4	6C
図書館・書籍	LBa1_00004	BK_86049602	230	2H
白書	OW6X_00009	WR_00000066	285	4C
教科書	OT01_00008	TB_01000002	31	8A
広報紙	OP00_00001	PR_01103001		
ベストセラー	OB0X_00001	BK_75079014	358	4D
Yahoo!知恵袋	OC01_00001	YC_00297514		
Yahoo!ブログ	OY01_00005	YB_00010571		
韻文	OV0X_00001	VE_00010001		
法律	OL3X_00072	LA_H01HO058		
国会会議録	OM11_00001	MD_80010001		

6.2　サンプル情報データの定義

以下では，サンプル情報データ（Sample.txt）の各列に記載された情報の定義について記述する。

6.2.1　サンプル ID

サンプル ID（Sample_ID）列は，各サンプルに対して一意に付された ID を表わす。

例

- 「PB10_00047」（出版 SC「書籍」）
- 「PM11_00053」（出版 SC「雑誌」）
- 「PN1a_00013」（出版 SC「新聞」）
- 「LBa1_00004」（図書館 SC「書籍」）
- 「OW6X_00009」（特定目的 SC「白書」）
- 「OT01_00008」（特定目的 SC「教科書」）
- 「OP01_00008」（特定目的 SC「広報紙」）
- 「OB0X_00001」（特定目的 SC「ベストセラー」）
- 「OC01_00001」（特定目的 SC「Yahoo!知恵袋」）
- 「OY01_00005」（特定目的 SC「Yahoo!ブログ」）
- 「OV0X_00001」（特定目的 SC「韻文」）
- 「OL1X_00001」（特定目的 SC「法律」）
- 「OM11_00001」（特定目的 SC「国会会議録」）

左から 1 桁目（P，L，O）は SC の違いを表わす。2 桁目（B，M，N，W，T，P，C，Y，V，L）は，各 SC 内におけるメディアの違いを表わす。3・4 桁目の意味は，1・2 桁目の違いによって異なる意味を持つ。区切り記号の「_」以降の 5 桁の数字は，サンプルの取得順位を表わす。

以下では，各メディアにおけるサンプル ID の構造について解説する。

出版 SC「書籍」のサンプル ID

出版 SC「書籍」のサンプル ID は，以下の構造を持つ。

- PB10_00001 ～ PB5n_00141

1 桁目	「P」	出版 SC（Publication）に所属することを表す。
2 桁目	「B」	書籍（Book）のサンプルであることを表す。
3 桁目	「1～5」	出版年を表す。 「1」＝ 2001 年　「3」＝ 2003 年　「5」＝ 2005 年 「2」＝ 2002 年　「4」＝ 2004 年
4 桁目	「0～9,n」	当該書籍に付された NDC（日本十進分類法）の第 1 次区分を表す。 「0」＝総記　「4」＝自然科学　「8」＝言語 「1」＝哲学　「5」＝技術・工学　「9」＝文学 「2」＝歴史　「6」＝産業　「n」＝分類なし 「3」＝社会科学　「7」＝芸術・美術
5 桁目	「_」	区切り記号。
6～10 桁目		各出版年・各 NDC におけるサンプルの取得順位を表す。

出版 SC「雑誌」のサンプル ID

出版 SC「雑誌」のサンプル ID は，以下の構造を持つ。

- PM11_00002 ～ PM56_00004

1 桁目	「P」	出版 SC（Production）に所属することを表す。
2 桁目	「B」	雑誌（Magazine）のサンプルであることを表す。
3 桁目	「1～5」	出版年を表す。 「1」＝ 2001 年　「3」＝ 2003 年　「5」＝ 2005 年 「2」＝ 2002 年　「4」＝ 2004 年
4 桁目	「1～6」	当該雑誌に付されたジャンルを表す。 「1」＝総合　「4」＝産業 「2」＝教育・学芸　「5」＝工業 「3」＝政治・経済・商業　「6」＝厚生・医療
5 桁目	「_」	区切り記号。
6～10 桁目		各雑誌タイトル・各出版年におけるサンプルの取得順位を表す。

出版 SC「新聞」のサンプル ID

出版 SC「新聞」のサンプル ID は，以下の構造を持つ。

- PN1a_00001 ～ PN5o_00021

1 桁目	「P」	出版 SC（Publication）に所属することを表す。
2 桁目	「N」	新聞（Newspaper）のサンプルであることを表す。
3 桁目	「1～5」	出版年を表す。 「1」＝ 2001 年　「3」＝ 2003 年　「5」＝ 2005 年 「2」＝ 2002 年　「4」＝ 2004 年
4 桁目	「a～o」	新聞タイトルを表す。 「a」＝朝日新聞　「f」＝中日新聞　「k」＝神戸新聞 「b」＝毎日新聞　「g」＝西日本新聞　「l」＝中国新聞 「c」＝読売新聞　「h」＝河北新報　「m」＝高知新聞 「d」＝産経新聞　「i」＝新潟日報　「o」＝琉球新報 「e」＝北海道新聞　「j」＝京都新聞
5 桁目	「_」	区切り記号。
6～10 桁目		各新聞タイトル・各出版年におけるサンプルの取得順位を表す。

図書館 SC「書籍」のサンプル ID

図書館 SC「書籍」のサンプル ID は，以下の構造を持つ。

- LBa0_00002 ～ LBtn_00025

1 桁目	「L」	図書館 SC（Library）に所属することを表す。
2 桁目	「B」	書籍（Book）のサンプルであることを表す。
3 桁目	「a～t」	出版年を表す。 「a」＝ 1986 年　「h」＝ 1993 年　「o」＝ 2000 年 「b」＝ 1987 年　「i」＝ 1994 年　「p」＝ 2001 年 「c」＝ 1988 年　「j」＝ 1995 年　「q」＝ 2002 年 「d」＝ 1989 年　「k」＝ 1996 年　「r」＝ 2003 年 「e」＝ 1990 年　「l」＝ 1997 年　「s」＝ 2004 年 「f」＝ 1991 年　「m」＝ 1998 年　「t」＝ 2005 年 「g」＝ 1992 年　「n」＝ 1999 年
4 桁目	「0～9,n」	当該書籍に付された NDC（日本十進分類法）の第 1 次区分を表す。 「0」＝総記　「4」＝自然科学　「8」＝言語 「1」＝哲学　「5」＝技術・工学　「9」＝文学 「2」＝歴史　「6」＝産業　「n」＝分類なし 「3」＝社会科学　「7」＝芸術・美術
5 桁目	「_」	区切り記号。
6～10 桁目		各出版年・各 NDC におけるサンプルの取得順位を表す。

特定目的 SC「白書」のサンプル ID

特定目的 SC「白書」のサンプル ID は，以下の構造を持つ。

- OW1X_00000 ～ OW6X_03369

1 桁目	「O」	特定目的 SC に所属することを表す。
2 桁目	「W」	白書（White Paper）のサンプルであることを表す。
3 桁目	「1～6」	出版時期を表す。
		「1」＝第 1 期（1976～1980 年）　「2」＝第 2 期（1981～1985 年）
		「3」＝第 3 期（1986～1990 年）　「4」＝第 4 期（1991～1995 年）
		「5」＝第 5 期（1996～2000 年）　「6」＝第 6 期（2001～2005 年）
4 桁目	「X」	ダミー記号。
5 桁目	「_」	区切り記号。
6～10 桁目		各出版時期におけるサンプルの取得順位を表す。

特定目的 SC「教科書」のサンプル ID

特定目的 SC「教科書」のサンプル ID は，以下の構造を持つ。

- OT01_00002 ～ OT91_00009

1 桁目	「O」	特定目的 SC に所属することを表す。
2 桁目	「T」	教科書（TextBook）のサンプルであることを表す。
3 桁目	「0～9」	教科を表す。
		「0」＝ 国語　「5」＝ 技術家庭
		「1」＝ 数学　「6」＝ 芸術
		「2」＝ 理科　「7」＝ 保健体育
		「3」＝ 社会　「8」＝ 情報
		「4」＝ 外国語　「9」＝ 生活
4 桁目	「1～3」	学校を表す。
		「1」＝ 小学校　「2」＝ 中学校　「3」＝ 高校
5 桁目	「_」	区切り記号。
6～10 桁目		各教科・学校におけるサンプルの取得順位を表す。

特定目的 SC「広報紙」のサンプル ID

特定目的 SC「広報紙」のサンプル ID は，以下の構造を持つ。

- OP00_00001 ～ OP99_00003

1 桁目	「O」	特定目的 SC に所属することを表す。
2 桁目	「P」	広報紙（Public Relation）のサンプルであることを表す。
3・4 桁目	「00～99」	対象となった 100 自治体の通し番号を表す。
5 桁目	「_」	区切り記号。
6～10 桁目		各自治体から取得したサンプルの取得順位を表す。

特定目的 SC「ベストセラー」のサンプル ID

特定目的 SC「ベストセラー」のサンプル ID は，以下の構造を持つ。

- OB0X_00001 ～ OB6X_00257

1 桁目	「O」	特定目的 SC に所属することを表す。
2 桁目	「B」	ベストセラー（Best-seller）のサンプルであることを表す。
3 桁目	「0～6」	出版時期を表す。 「0」＝第 0 期（1975 年以前）　「4」＝第 4 期（1991～1995 年） 「1」＝第 1 期（1976～1980 年）「5」＝第 5 期（1996～2000 年） 「2」＝第 2 期（1981～1985 年）「6」＝第 6 期（2001～2005 年） 「3」＝第 3 期（1986～1990 年）
4 桁目	「X」	ダミー記号。
5 桁目	「_」	区切り記号。
6～10 桁目		各出版時期におけるサンプルの取得順位を表す。

特定目的 SC「Yahoo!知恵袋」のサンプル ID

特定目的 SC「Yahoo!知恵袋」のサンプル ID は，以下の構造を持つ。

- OC01_00001 ～ OC15_01173

1 桁目	「O」	特定目的 SC に所属することを表す。
2 桁目	「C」	Yahoo!知恵袋（*Chiebukuro*）のサンプルであることを表す。
3・4 桁目	「01～15」	質問が投稿された大カテゴリ ID を表す。 「01」＝「エンターテインメントと趣味」 「02」＝「インターネット，PC と家電」 「03」＝「ビジネス，経済とお金」 「04」＝「職業とキャリア」 「05」＝「ニュース，政治，国際情勢」 「06」＝「スポーツ，アウトドア，車」 「08」＝「暮らしと生活ガイド」 「09」＝「健康，美容とファッション」 「10」＝「子育てと学校」 「11」＝「マナー，冠婚葬祭」 「12」＝「教養と学問，サイエンス」 「13」＝「地域，旅行，お出かけ」 「14」＝「Yahoo! JAPAN」 「15」＝「その他」
5 桁目	「_」	区切り記号。
6～10 桁目		各大カテゴリにおけるサンプルの取得順位を表す。

※ 大カテゴリ ID の「07（コンピュータテクノロジー）」は，「Yahoo! 知恵袋」の元データに十分な量のデータがなく，サンプルが取得できなかったため，欠番になっている。

特定目的 SC「Yahoo! ブログ」のサンプル ID

特定目的 SC「Yahoo! ブログ」のサンプル ID は，以下の構造を持つ。

- OY01_00005 ～ OY15_09456

1 桁目	「O」	特定目的 SC に所属することを表す。
2 桁目	「Y」	Yahoo! ブログ（Blog）のサンプルであることを表す。
3・4 桁目	「01～15」	記事が投稿された大カテゴリ ID を表す。 「01」＝「ビジネスと経済」 「02」＝「コンピュータとインターネット」 「03」＝「生活と文化」 「04」＝「エンターテインメント」 「05」＝「家庭と住まい」 「06」＝「政治」 「07」＝「健康と医学」 「08」＝「学校と教育」 「09」＝「科学」 「10」＝「出会い」 「11」＝「地域」 「12」＝「特集」 「13」＝「芸術と人文」 「14」＝「Yahoo!サービス」 「15」＝「趣味とスポーツ」
5 桁目	「_」	区切り記号。
6～10 桁目		各大カテゴリにおけるサンプルの取得順位を表す。

特定目的 SC「韻文」のサンプル ID

特定目的 SC「韻文」のサンプル ID は，以下の構造を持つ。

- OV0X_00001 ～ OV2X_00108

1 桁目	「O」	特定目的 SC に所属することを表す。
2 桁目	「V」	韻文（Verse）のサンプルであることを表す。
3 桁目	「0～2」	韻文の種類を表す。 「0」＝短歌　「1」＝俳句　「2」＝詩
4 桁目	「X」	ダミー記号。
5 桁目	「_」	区切り記号。
6～10 桁目		サンプルの取得順位を表す。

特定目的 SC「法律」のサンプル ID

特定目的 SC「法律」のサンプル ID は，以下の構造を持つ。

- OL1X_00001 ～ OL6X_00066

1 桁目	「O」	特定目的 SC に所属することを表す。
2 桁目	「L」	法律（Law）のサンプルであることを表す。
3 桁目	「1～6」	法律の公布年を表す。
		「1」＝第 1 期（1976～1980 年）　「2」＝第 2 期（1981～1985 年）
		「3」＝第 3 期（1986～1990 年）　「4」＝第 4 期（1991～1995 年）
		「5」＝第 5 期（1996～2000 年）　「6」＝第 6 期（2001～2005 年）
4 桁目	「X」	ダミー記号。
5 桁目	「_」	区切り記号。
6～10 桁目		各期におけるサンプルの取得順位を表す。

特定目的 SC「国会会議録」のサンプル ID

特定目的 SC「国会会議録」のサンプル ID は，以下の構造を持つ。

- OM11_00001 ～ OM68_00001

1 桁目	「O」	特定目的 SC に所属することを表す。
2 桁目	「M」	国会会議録（Minutes of the Diet）のサンプルであることを表す。
3 桁目	「1～6」	会議の開催時期を表す。
		「1」＝第 1 期（1976～1980 年）　「4」＝第 4 期（1991～1995 年）
		「2」＝第 2 期（1981～1985 年）　「5」＝第 5 期（1996～2000 年）
		「3」＝第 3 期（1986～1990 年）　「6」＝第 6 期（2001～2005 年）
4 桁目	「1～8」	会議の開催院・会議種別を表す。
		「1」＝衆議院・常任委員会　「5」＝参議院・常任委員会
		「2」＝衆議院・特別委員会　「6」＝参議院・特別委員会
		「3」＝衆議院・本会議　「7」＝参議院・本会議
		「4」＝衆議院・その他　「8」＝参議院・その他
5 桁目	「_」	区切り記号。
6～10 桁目		開催時期，開催院・会議種別におけるサンプルの取得順位を表す。

6.2.2　書誌 ID

書誌 ID（Bib_ID）列は，各サンプルを取得した原本に対して一意に付された ID を表わす。記載されている ID は，**書誌情報データ**（Bibliography.txt）の「書誌 ID（Bib_ID）」列に記載された ID に対応している。詳しくは，57 ページの 5.2.1 を参照。

6.2.3　サンプル抽出基準点 ページ

サンプル抽出基準点 ページ（Sampling_page）列は,「サンプル抽出基準点」（6.2.4 参照）を含むページ番号を表わす。

※ サンプル抽出基準点 ページの情報は,「書籍」「雑誌」「新聞」「白書」「教科書」にのみ付与される。

6.2.4 サンプル抽出基準点 座標

サンプル抽出基準点 座標（Sampling_point）列は,「サンプル抽出基準点」を同定する際,サンプル抽出基準点 ページ内でランダムに指定されたある 1 点（交点）を表わす。

※ 横軸に 0〜9，縦軸に A〜J という目盛りを配置した 10×10 のマス目を準備し，それを印刷した透明なシートを実際のページに当て，ランダムに指定された交点（「3E」など）に最も近接している文字を「サンプル抽出基準点」として指定した。このサンプル抽出基準点をもとに，サンプルを取得した。

※ サンプル抽出基準点 座標の情報は,「書籍」「雑誌」「新聞」「白書」「教科書」にのみ付与される。

第7章　人名録データ（Directory.txt）

7.1　人名録データの概要

人名録データ（Directory.txt）は，各サンプルに含まれる文章を実際に執筆した著者のほか，編集者，監修者，翻訳者，著作権保持者などの人名・組織名などの情報を表わす。Directory.txt では，著者・著作権者に関する情報が以下の 4 列によって表現されている。

1. 人名 ID （Directory_ID）　人物に対して一意に付された ID。
2. 人名 （Name）　人物の氏名・組織名。
3. 性別（Sex）　性別。
4. 生年（BirthYear）　生年（10 年単位）。

人名録データの例を，以下に示す。

Directory_ID	Name	Sex	BirthYear
634	会田　雄次	男	1910
98948	アントニオ猪木	男	1940
153494	群　ようこ	女	1950
840303	厚生労働省労働基準局		
2000130	山と渓谷社		
2502212	NHK「プロジェクト X」制作班		

7.2　人名録データの定義

以下では，人名録データ（Directory.txt）の各列に記載された情報の定義について記述する。

7.2.1　人名 ID

人名 ID（Directory_ID）列は，個人名・組織名に対して付された一意の ID を表わす。

例

- 「4078」　（「阿刀田高」に付された ID）
- 「31535」　（「木下順二」に付された ID）
- 「2505106」　（「（株）共同通信社　出版本部　編集部」に付された ID）

※ 人名 ID には，国立国会図書館の「著者名典拠 ID」を利用している。そこに登録がない人名・組織名については，我々のサブグループで独自に番号を付与している。

7.2.2　人名

人名（Name）列は，人物の氏名・組織名などを表わす。

例

- 「夏目　漱石」
- 「読売新聞社　東京本社」

※ 人名の漢字表記は，国立国会図書館の典拠データを元にしている。ただし，著作権処理の過程で要望があった場合は，表記を修正している。一部コンピュータで表現できない文字は,「**=**」に置き換え，その文字の一般的な形を（　）で補記している。以下の「**=**」は,「己」の部分が「巳」になった文字を代用している。

– （元表記）　泡坂　妻夫　→　（修正後表記）　**=**（泡）坂　妻夫

7.2.3　性別

性別（Sex）列は，著者の性別を表わす。

※ 性別の情報は，原則として本人から得られた回答を記載しているが，国立国会図書館の典拠データなどから補足・記載しているものもある。また，組織の場合には，記載していない。

7.2.4　生年

生年（BirthYear）列は，人物の生年を表わす。ただし,「1950」「1960」のように，西暦の 10 年単位でまとめてある。

※ 生年の情報は，原則として本人から得られた回答を記載しているが，国立国会図書館の典拠データなどから補足・記載しているものもある。また，組織の場合には，記載していない。

第8章　サンプル著者対応情報データ（Sample_author.txt）

8.1　サンプルと著者の対応関係

各サンプルと著者･著作権者との対応関係は，**サンプル著者対応情報データ**（Sample_author.txt）で表わされる。Sample_author.txt では，サンプルと著者・著作権者との対応関係が以下の2列によって表現されている。

1. サンプル ID （Sample_ID）　サンプルに対して一意に付された ID。
2. 人名 ID（Directory_ID）　著者・著作権者に対して一意に付された ID。

サンプルと著者・著作権者の対応関係の例を，以下に示す。

Sample_ID	Directory_ID
PB10_00022	107107
PM43_00020	303855
LBl9_00073	327382
LBl9_00073	556836

※ 1つのサンプルを複数の著者が執筆している場合，上記の LBl9_00073 のように，同じサンプル ID に対して複数の異なる人名 ID が関係づけられる。

※ 「Yahoo!知恵袋」「Yahoo!ブログ」「法律」「国会会議録」には，サンプル著者対応情報データは付与されない。

8.2　サンプル著者対応情報データの定義

以下では，サンプル著者対応情報データ（Sample_author.txt）の各列に記載された情報の定義について記述する。

8.2.1　サンプル ID

サンプル ID（Sample_ID）列は，各サンプルに対して一意に付された ID を表わす。記載されている ID は，**サンプル情報データ**（Sample.txt）の「サンプル ID（Sample_ID）」列に記

載された ID に対応している。詳しくは，94 ページの 6.2.1 を参照。

8.2.2　人名 ID

人名 ID（Directory_ID）列は，人物に対して付された ID を表わす。記載されている ID は，**人名録データ**（Directory.txt）の「人名 ID（Directory_ID）」列に記載された ID に対応している。詳しくは，103 ページの 7.2.1 を参照。

第9章　書誌情報データの運用と拡張

9.1　書誌情報データベースの構築

以上までで,「書誌情報データ」「サンプル情報データ」「人名録データ」「サンプル著者対応情報データ」という 4 つから構成される書誌情報データを示した。これらを結合することにより，書誌情報データベースを構築することができる。

まず，各データについて列名の一覧を，図 9.1 に示す。* の付いた列名は，主キーであることを示す。これらの各データは，Microsoft Access，MySQL，SQL Server などのリレーショナルデータベースに取り込むことによって，柔軟に運用することができる。例として，これら 4 つのデータを，サンプル ID を主キーとして結合したデータ「**統合データ**」を，図 9.2 に示す。

このような形で書誌情報を 1 つに取りまとめておき，Microsoft Excel などの表計算ソフトに読み込むと，ある条件を満たすサンプル ID を同定したり，サンプルごとの書誌情報リストを作ったりするなど，書誌情報データの柔軟な運用が可能になる。

書誌情報データ（Bibliography.txt）

1.	Bib_ID*	書誌 ID*
2.	Title	タイトル
3.	Subtitle	副題
4.	Number	巻号
5.	Bib_author	責任表示
6.	Publisher	出版者
7.	Year	出版年
8.	ISBN	ISBN
9.	Size	判型
10.	Pages	ページ数
11.	Genre_1	ジャンル (1)
12.	Genre_2	ジャンル (2)
13.	Genre_3	ジャンル (3)
14.	Genre_4	ジャンル (4)
15.	Bib_author_ID	責任表示 ID

サンプル情報データ（Sample.txt）

1.	Sample_ID*	サンプル ID*
2.	Bib_ID	書誌 ID
3.	Sampling_page	サンプル抽出基準点 ページ
4.	Sampling_point	サンプル抽出基準点 座標
5.	Status	著作権許諾状況
6.	Core	コアデータ フラグ

人名録データ（Directory.txt）

1.	Directory_ID*	人名 ID*
2.	Name	人名
3.	Sex	性別
4.	BirthYear	生年

サンプル著者対応情報データ（Sample_author.txt）

1.	Sample_ID*	サンプル ID*
2.	Directory_ID*	人名 ID*

図 9.1: 書誌情報データの構成

統合データ

1.	Sample_ID*	サンプル ID*
2.	Bib_ID	書誌 ID
3.	Title	タイトル
4.	Subtitle	副題
5.	Number	巻号
6.	Bib_author	責任表示
7.	Publisher	出版者
8.	Year	出版年
9.	ISBN	ISBN
10.	Size	判型
11.	Pages	ページ数
12.	Genre_1	ジャンル (1)
13.	Genre_2	ジャンル (2)
14.	Genre_3	ジャンル (3)
15.	Genre_4	ジャンル (4)
16.	Bib_author_ID	責任表示 ID
17.	Sampling_page	サンプル抽出基準点 ページ
18.	Sampling_point	サンプル抽出基準点 座標
19.	Status	著作権許諾状況
20.	Core	コアデータ フラグ
21.	Author_ID*	著者 ID*
22.	Author	著者名
23.	Sex	性別
24.	BirthYear	生年

図 9.2: 書誌情報データの結合例（「統合データ」）

9.2　書誌情報データベースの拡張

最後に，書誌情報データの拡張について触れておく。上記のような形で整備した書誌情報データベースは，研究の目的や用途に応じて新規にデータを追加することで，拡張していくことができる。

例えば,「出版 SC」は 2001 年から 2005 年に出版されたすべての書籍・雑誌・新聞からランダムに取得したサンプルで構成されるが，この中には 2003 年に出版された夏目漱石『吾輩は猫である』の文庫から取得したサンプルが含まれる。出版のありさまを捉える設計上，この結果は正しいものであるが，検索結果を分析する際，研究の目的によっては好ましくない結果が得られることになる。そこで,「出版年」とは別に,「初出年」という情報をサンプルごとに付与することが考えられる。サンプル ID と初出年の値を組み合わせたテーブル（上記の例で言えば,「PB39_00742 ， 1905」という 2 列の表）を作成してデータベースに組み込むことによって,「初出情報」という検索条件を新規に加えることができることになる。

また，現時点のデータでは，各サンプル（記事）を実際に執筆した「著者」に関する情報は，厳密には存在しない。「サンプル著者対応データ」で示した関係は，サンプルと著者または著作権者に関するものであり，サンプル内に含まれる文章を実際に執筆した人物の対応が取れているわけではない。そこで,「実著者」に関する情報をサンプルごとに（あるいは「記事」ごとに）付与してデータベースに組み込むことにより，実際の著者を手掛かりにして検索を実施することが可能になる。さらに，サンプルに含まれる文章の「難易度」を何らかの方法で判定し，その結果をサンプル ID と組にしてテーブル化しておくことにより，サンプルの難易度に基づいた検索や分類が可能になる。

文書構造タグのような，サンプル本体に埋め込まれたアノテーション情報とは別に，ここで示したような外部アノテーションとしての書誌情報データを豊富に付与していくことにより，BCCWJ をより柔軟に検索・運用することが可能になる。このような情報の付与とデータの拡張は，今後の課題である。

第III部

資料編

第10章　研究成果一覧

第 III 部では，特定領域研究「日本語コーパス」の「データ班」においてサンプリングを担当した我々のグループ（SSG; サンプリングサブグループ）で，この 5 年間に発表してきた研究成果をまとめる。

特定領域「日本語コーパス」研究成果報告書

[1] 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.

[2] 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.

[3] 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.

[4] 佐野大樹, 丸山岳彦, 山崎誠, 柏野和佳子, 秋元祐哉, 稲益佐知子, 田中弥生, 大矢内夢子 (2009). 『語彙密度を利用した『現代日本語書き言葉均衡コーパス』テクスト分類の試み』, 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-08-02), 特定領域研究「日本語コーパス」データ班.

[5] 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.

[6] 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』,

特定領域研究「日本語コーパス」平成 22 年度研究成果報告書 (JC-D-10-01), 特定領域研究「日本語コーパス」データ班.

[7] 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』に含まれるサンプルおよび書誌情報の設計と実装』, 特定領域研究「日本語コーパス」平成 22 年度研究成果報告書 (JC-D-10-02), 特定領域研究「日本語コーパス」データ班.

特定領域「日本語コーパス」全体会議・公開ワークショップ

[8] 山崎誠, 丸山岳彦, 柏野和佳子, 山口昌也, 間淵洋子, 高田智和, 小椋秀樹, 森本祥子, 大和淳 (2006). 現代日本語書き言葉均衡コーパスの現状（データ班：代表性を有する現代日本語書籍コーパスの構築）. 『特定領域「日本語コーパス」平成 18 年度全体会議予稿集』. 9-16.

[9] 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』. 79-88.

[10] 山崎誠, 小椋秀樹, 柏野和佳子, 高田智和, 間淵洋子, 丸山岳彦, 森本祥子, 山口昌也, 大和淳 (2007). 平成 18 年度研究進捗状況報告：データ班（代表性を有する現代日本語書籍コーパスの構築）. 『特定領域研究「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』. 25-28.

[11] 山崎誠, 小椋秀樹, 小沼悦, 柏野和佳子, 高田智和, 間淵洋子, 丸山岳彦, 森本祥子, 山口昌也 (2007). 平成 19 年度研究進捗状況報告：データ班（代表性を有する現代日本語書籍コーパスの構築）. 『特定領域研究「日本語コーパス」平成 19 年度全体会議予稿集』. 3-8.

[12] 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. 『特定領域研究「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』. 127-136.

[13] 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). 書籍サンプルの多様性. 『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』. 143-152.

[14] 佐野大樹 (2008). 大規模バランストコーパスにおけるテクスト分類 —システミック理論の観点から—. 『特定領域研究「日本語コーパス」平成 20 年度全体会議予稿集』. 83-90.

[15] 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』. 37-46.

[16] 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.

[17] 山崎誠, 小椋秀樹, 小沼悦, 柏野和佳子, 佐野大樹, 高田智和, 冨士池優美, 間淵洋子, 丸山岳彦, 森本祥子, 山口昌也 (2008). 平成 20 年度研究進捗状況報告：データ班（代表性を有する現代日本語書籍コーパスの構築）. 『特定領域研究「日本語コーパス」平成 20 年度全体会議予稿集』. 5-10.

[18] 山崎誠, 小椋秀樹, 小沼悦, 柏野和佳子, 高田智和, 間淵洋子, 丸山岳彦, 森本祥子, 山口昌也 (2008). 平成 19 年度研究進捗状況報告：データ班（代表性を有する現代日本語書籍コーパスの構築）. 『特定領域研究「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』. 65-72.

[19] 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 (2009). 『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 (3) —代表性を実現するためのサンプリング手法—. 『特定領域「日本語コーパス」平成 20 年度公開ワークショップ（研究成果報告会）予稿集』. 33-42.

[20] 山崎誠, 小椋秀樹, 小沼悦, 柏野和佳子, 佐野大樹, 高田智和, 間淵洋子, 丸山岳彦, 森本祥子, 山口昌也 (2009). 平成 21 年度研究進捗状況報告：データ班（代表性を有する現代日本語書籍コーパスの構築）. 『特定領域研究「日本語コーパス」平成 21 年度全体会議予稿集』. 3-8.

[21] 山崎誠 (2009). 『現代日本語書き言葉均衡コーパス』における固定長サンプルと可変長サンプルの比較. 『特定領域研究「日本語コーパス」平成 20 年度公開ワークショップ（研究成果報告会）予稿集』. 5-12.

[22] 佐野大樹 (2010). ブログにおける評価の分析 —アプレイザル理論を用いて—. 『特定領域研究「日本語コーパス」平成 21 年度公開ワークショップ（研究成果報告会）予稿集』. 47-54.

[23] 田中弥生 (2010). Yahoo!ブログの文体的特徴 —投稿に使用した機器による比較—. 『特定領域研究「日本語コーパス」平成 22 年度全体会議予稿集』. 73-80.

[24] 田中弥生 (2010). Q&A コミュニティの談話機能と構造 —「Yahoo!知恵袋」を対象に—. 『特定領域研究「日本語コーパス」 平成 21 年度公開ワークショップ（研究成果報告会）予稿集』. 55-62.

[25] 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 (2010). 『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 (4) —コーパスの設計とサンプリングの実際—. 『特定領域研究「日本語コーパス」平成 21 年度公開ワークショップ（研究成果報告会）予稿集』. 37-46.

[26] 山崎誠, 小椋秀樹, 小沼悦, 柏野和佳子, 佐野大樹, 高田智和, 冨士池優美, 間淵洋子, 丸山岳彦, 森本祥子, 山口昌也 (2010). 平成 22 年度研究進捗状況報告：データ班（代表性を有する現代日本語書籍コーパスの構築）. 『特定領域研究「日本語コーパス」平成 22 年度全体会議予稿集』. 3-8.

[27] 山崎誠 (2010). 語の平均使用度数に現れるテキストの特徴. 『特定領域研究「日本語コーパス」平成 21 年度公開ワークショップ（研究成果報告会）予稿集』. 5-14.

[28] 山崎誠 (2010). BCCWJ モニター公開データの利用実態について. 『特定領域研究「日本語コーパス」平成 22 年度全体会議予稿集』. 109-112.

[29] 佐野大樹, 柏野和佳子 (2011). 『現代日本語書き言葉均衡コーパス』における評価表現の分布—『日本語アプレイザル評価表現辞書（態度表現編）』を用いて—. 『特定領域研究「日本語コーパス」平成 22 年度公開ワークショップ（研究成果報告会）予稿集』.

[30] 田中弥生, 佐野大樹 (2011). Yahoo!知恵袋の質問における修辞機能の分布 —修辞ユニット分析を用いて—. 『特定領域研究「日本語コーパス」平成 22 年度公開ワークショップ（研究成果報告会）予稿集』.

[31] 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概要 (5) —サンプリングの最終結果—. 『特定領域「日本語コーパス」平成 22 年度公開ワークショップ（研究成果報告会）予稿集』.

[32] 山崎誠, 小椋秀樹, 小沼悦, 柏野和佳子, 佐野大樹, 高田智和, 冨士池優美, 間淵洋子, 丸山岳彦, 森本祥子, 山口昌也 (2011). 研究活動・成果の総括：データ班（代表性を有する現代日本語書籍コーパスの構築）. 『特定領域研究「日本語コーパス」平成 22 年度公開ワークショップ（研究成果報告会）予稿集』.

[33] 山崎誠 (2011). 多義語における意味の分布. 特定領域研究「日本語コーパス」『特定領域研究「日本語コーパス」平成 22 年度公開ワークショップ（研究成果報告会）予稿集』.

公刊論文, 書籍等

[34] 柏野和佳子 (2006). 書き言葉コーパスで探る日本語のありさま. 『日本語学』 25(9). 18-27. 明治書院.

[35] 丸山岳彦, 田野村忠温 (2007). コーパス日本語学の射程. 『日本語科学』 22. 5-12. 国立国語研究所.

[36] 山崎誠 (2007). 国立国語研究所の言語コーパス整備計画「KOTONOHA」の紹介. 『漢字文献情報処理研究』 8. 180-183. 漢字文献情報処理研究会.

[37] 田中弥生 (2008). 電子コミュニケーションの配慮意識表現 —クチコミサイトへの携帯電話からの投稿にみられる特徴—. 『青山 国際コミュニケー ション研究』 12. 49-69. 青山学院大学国際コミュニケーション学会.

[38] Sano, M. & Thomson, E. A. (2008). Japanese Folk Tales: text structure and evaluative expression. Bridging Discourse: ASFLA 2007 Online Proceedings. 1-17.

[39] 田中弥生 (2009). インターネットの知識検索サービスにおける談話構造の諸相 —Yahoo! 知恵袋の情報要求モデルの検討—. 『ことばと人間』 7. 57-68. 「言語と人間」研究会.

[40] 丸山岳彦 (2009). 作文の文体情報 —『現代日本語書き言葉均衡コーパス』から見えるもの—. 『日本語教育』 140. 26-36. 日本語教育学会.

[41] 丸山岳彦 (2009). 日本語コーパスの現状. 『国文学 解釈と鑑賞 平成 21 年 1 月号（特集 日本語研究とコーパス)』 122-130. 至文堂.

[42] 丸山岳彦 (2009). 「1.2 コーパス」「1.2.1 コーパスの類型」「1.2.2 コーパスの構築」. 言語処理学会編, 『言語処理学事典』. 58-71. 共立出版.

[43] 山崎誠 (2009). 国立国語研究所における諸研究 —語彙調査の系譜を中心にして. 『国文学解釈と鑑賞』 74(1). 183-191. 至文堂.

[44] 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. 人工知能学会.

[45] 山崎誠 (2009). コーパスにみる特許関連文章の特徴. 『Japio 年誌』. 118-121. 日本特許情報機構.

[46] Maruyama, T., Yamazaki, M. & Maekawa, K. (2009). Statistical sampling method used in the Balanced Corpus of Contemporary Written Japanese. Current ISSUES in Unity and Diversity of Languages: Collection of the papers selected from the CIL 18. 3864-3876.

[47] 佐野大樹 (2010). 選択体系機能言語理論を基底とする特定目的のための作文指導方法について. 『専門日本語教育研究』 12. 19-26. 専門日本語教育学会.

[48] 佐野大樹 (2010). ブログにおける評価表現の使い分けの特徴—アプレイザル理論からみた評価基準と表現の直接性/間接性の関係—. 『計量国語学』27(7). 249-269. 計量国語学会.

[49] 佐野大樹 (2010). ブログにおける評価情報の分類と体系化 —アプレイザル理論を用いて—. 『信学技報』 109(390). 37-42. 電子情報通信学会.

[50] 田中弥生 (2010). 質問サイトにおける情報要求モデルと待遇コミュニケーション —「アットコスメ美容事典」の談話機能・談話構造の分析から. 『待遇コミュニケーション研究』 7. 3-48. 待遇コミュニケーション学会.

[51] 田中弥生 (2010). Q&A サイトの「質問—回答」における結束性 —省略の特徴分析—. 『信学技報』 109(390). 7-12. 電子情報通信学会.

[52] 佐野大樹 (2011). 日本語における評価表現の分類体系 〜アプレイザル理論をベースに〜. 『信学技報』110(400). 19-24. 電子情報通信学会.

[53] 佐野大樹, 小磯花絵 (2011 予). 現代日本語書き言葉における修辞ユニット分析の適用性の検証 —「書き言葉らしさ, 話し言葉らしさ」と脱文脈化言語, 文脈化言語の関係—. 『機能言語学研究』 6. 日本機能言語学会.

[54] 佐野大樹 (2011 予). ベストセラーの文体的特徴の検討 —『現代日本語書き言葉均衡コーパス』を用いた一考察—. 『ことばと人間』 8. 「言語と人間」研究会.

[55] 田中弥生, 佐野大樹 (2011). Yahoo!知恵袋における質問の修辞ユニット分析 —脱文脈化-文脈化の程度による分類—. 『信学技報』 110(400). 13-18. 電子情報通信学会.

[56] 田中弥生 (2011 予). 「質問－回答」における待遇表現の特徴 —書籍 QA、WebQA、Yahoo!知恵袋の比較から—. 『待遇コミュニケーション研究』 8. 65-80. 待遇コミュニケーション学会.

[57] 丸山岳彦 (2011 予). 第 10 章 コーパス研究. 益岡隆志編, 『はじめて学ぶ日本語学』. ミネルヴァ書房.

[58] 丸山岳彦 (2011 予). 第 3 章 コーパスデータの処理方法. 荻野綱男, 田野村忠温編, 『現代日本語 IT 講座 第 5 巻 コーパス日本語学 —パソコンでコーパスを活用する—』. 明治書院.

[59] 水澤祐美子, 佐野大樹 (2011 予). 社会的機能に基づくテクスト分類法の構築に向けて —システミック理論の観点から—. 『機能言語学研究』 6.

口頭発表

[60] 丸山岳彦, 柏野和佳子, 山崎誠, 前川喜久雄, 吉田谷幸宏, 稲益佐知子 (2006). 現代日本語の書き言葉に関する生産実態と流通実態 —代表性を有する書き言葉コーパスのための基礎調査—. 『言語処理学会第 12 回年次大会発表論文集』. 444-447.

[61] 丸山岳彦, 柏野和佳子, 山崎誠, 前川喜久雄, 稲益佐知子, 秋元祐哉 (2006). 代表性を有する書き言葉コーパスのサンプリング手法について. 『言語処理学会第 12 回年次大会発表論文集』. 680-683.

[62] 山崎誠, 前川喜久雄, 田中牧郎, 小椋秀樹, 柏野和佳子, 小磯花絵, 間淵洋子, 丸山岳彦, 山口昌也, 秋元祐哉, 稲益佐知子, 吉田谷幸宏 (2006). 代表性を有する現代日本語書き言葉コーパスの設計. 『言語処理学会第 12 回年次大会発表論文集』. 440-443.

[63] 秋元祐哉, 丸山岳彦, 吉田谷幸宏, 山崎誠, 柏野和佳子, 稲益佐知子, 前川喜久雄 (2007). 書き言葉の総量を捉える —書き言葉はどれだけ生産されるのか—. 『言語処理学会第 13 回年次大会発表論文集』. 708-711.

[64] 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』. 704-707.

[65] 山崎誠, 丸山岳彦, 山口昌也, 小椋秀樹, 森本祥子, 柏野和佳子, 佐野大樹, 高田智和, 間淵洋子, 北村雅則, 小木曽智信, 小磯花絵, 冨士池優美, 小沼悦, 田中牧郎, 前川喜久雄 (2007). 現代日本語書き言葉均衡コーパスの設計と検索デモンストレーション. 『日本語学会 2007 年度秋期大会要旨集』. 239-246.

[66] 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). 書籍の生産実態を反映するサンプリング —NDC ごとに取得したサンプルの多様性の分析—. 『言語処理学会第 14 回年次大会発表論文集』. 939-942.

[67] 柏野和佳子 (2008). 書籍の文章の多様性をとらえる観点付与の設計 —『現代日本語書き言葉均衡コーパス』の収録文章を対象に. 『ことば工学研究会資料』 30. 11-32.

[68] 佐野大樹, 丸山岳彦 (2008). システミック文法に基づく書きことばの複雑さ測定. 『言語処理学会第 14 回年次大会 発表論文集』. 1097-1100.

[69] 佐野大樹 (2008). 大規模バランストコーパスにおけるテクスト分類に向けて —語彙密計測からみたコンテクスト情報. 2008 年度日本機能言語学会秋期大会. お茶の水女子大学.

[70] 田中弥生 (2008). 電子コミュニケーションにおける「質問表現」の特徴 —Yahoo!知恵袋を対象に—. 『社会言語科学会第 22 回研究大会論文集』. 114-117.

[71] 田中弥生 (2008). ブログの言語表現にみる対人配慮意識 —媒体差と世代差に注目して—. 『社会言語科学会第 21 回研究大会論文集』. 92-95.

[72] 田中弥生 (2008). クチコミサイトにおける世代別・媒体別言語表現の分析. 『言語処理学会第 14 回年次大会発表論文集』. 911-914.

[73] 田中弥生 (2008). 電子コミュニケーションの質問における発話構造の様相 —「Yahoo! 知恵袋」の質問部分を対象に—. 待遇コミュニケーション学会 2008 年秋季大会.

[74] Sano, M. & Maruyama, T. (2008). Lexical Density in Japanese Texts: classifying text samples in the Balanced Corpus of Contemporary Written Japanese (BCCWJ). Proceedings of 35th International Systemic Functional Congress. 359-364.

[75] Sano, M. & Mizusawa, Y. (2008). Describing Japanese Language and Text: applications of systemic functional theory. Columbia University Teachers College Public Seminar. Columbia University Teachers College.

[76] 柏野和佳子, 丸山岳彦, 稲益佐知子, 秋元祐哉, 田中弥生, 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』のサンプル収録方法. 『言語処理学会第 15 回年次大会発表論文集』. 196-199.

[77] 柏野和佳子, 奥村学 (2009). 和語や漢語のカタカナ表記 —『現代日本語書き言葉均衡コーパス』における使用実態—. 『計量国語学会第五十三回大会予稿集』. 38-43.

[78] 佐野大樹, 水澤祐美子 (2009). Context based register typology における社会意味過程カテゴリの言語的特徴の検討. 2009 年度日本機能言語学会秋期大会. 同志社大学.

[79] 田中弥生 (2009). 電子コミュニケーションにおける情報要求の諸相 —クチコミサイトアットコスメを対象に—. 『社会言語科学会第 23 回研究大会論文集』. 28-31.

[80] 田中弥生 (2009). アットコスメの待遇意識表現. 「言語と人間」研究会 11 月例会. 立教大学.

[81] 山崎誠, 丸山岳彦, 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 (2009). 現代日本語書き言葉均衡コーパスのサンプル長と言語的特徴 —固定長サンプルと可変長サンプルの質的な違い—. 『言語処理学会第 15 回年次大会発表論文集』. 618-621.

[82] 山崎誠 (2009). テキストにおける語の平均使用度数と文体差. 『大規模データ, リンケージ, データマイニングと統計手法』. 新領域融合プロジェクトによる研究会. 47-50.

[83] 山崎誠 (2009). 『現代日本語書き言葉均衡コーパス』の構築と日本語研究の展望. 『韓国日本語学会第 20 会国際学術発表会論文集』. 163-177.

[84] 柏野和佳子, 奥村学 (2010). 国語辞典に「古い」と注記される語の現代書き言葉における使用傾向の調査. 情報処理学会 『人文科学とコンピュータ研究会報告』 88. 59-70.

[85] 柏野和佳子 (2010). 「直接的な語り」という表現スタイルをもつ書籍テキストの人手抽出の試み. 『ことば工学研究会資料』 35. 63-72.

[86] 佐野大樹, 田中牧郎, 丸山岳彦 (2010). 「病院の言葉」の類型の推測とモデル化 —『現代日本語書き言葉均衡コーパス』における語の使用度数を用いた一考察— 『日本言語学会第 140 回大会予稿集』. 370-375.

[87] 佐野大樹, 小磯花絵 (2010). 修辞ユニットを用いた書き言葉の分析 —「書き言葉・話し言葉らしさ」と (脱) 文脈化の関係—. 『社会言語科学会第 25 回研究大会発表論文集』. 182-185.

[88] 佐野大樹, 丸山岳彦 (2010). 評価表現に基づくブログ分類の試み —アプレイザル理論を用いて—. 『言語処理学会第 16 回年次大会発表論文集』. 174-177.

[89] 佐野大樹 (2010). アプレイザル理論を用いた「健康と病いの語り」における感情・評価表現の分析 —症状の変化と「語り」の変化の関係—. 『計量国語学会第五十四回大会予稿集』. 43-48.

[90] 佐野大樹 (2010). 評価表現の分類体系構築の試み —日本語分析におけるアプレイザル理論（attitude について）の再構築—. 2010 年度日本機能言語学会秋期大会.

[91] 田中弥生 (2010). 社会科学と文学の「あとがき」における文体的特徴の相違. 『計量国語学会第五十四回大会予稿集』. 49-54.

[92] 田中弥生 (2010). 異なる媒体における「QA」の文体的特徴 —書籍と Web を比較して—. 『社会言語科学会第 26 回研究大会論文集』. 142-145.

[93] 田中弥生 (2010). Yahoo!ブログにおける待遇表現 —投稿に使用した機器による比較—. 待遇コミュニケーション学会 2010 年秋季大会.

[94] 丸山岳彦 (2010). 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—. 『言語処理学会第 16 回年次大会発表論文集』. 150-153.

[95] 山崎誠 (2010). テキストにおける多義語の意味実現の傾向. 『計量国語学会第五十四回大会予稿集』. 25-30.

[96] 山崎誠 (2010). 『現代日本語書き言葉均衡コーパス』（BCCWJ）を利用した語彙研究の進展. 記念北京日本学研究中心成立 25 周年国際学術検討会論文集（北京日本学研究センター創立 25 周年記念国際シンポジウム論文集）. 16-17.

[97] Kashino, W. & Okumura, M. (2010). An Approach toward Register Classification of Book Samples in the Balanced Corpus of Contemporary Written Japanese. Proc. of PACLIC24. 433-438.

[98] 柏野和佳子, 奥村学 (2011 予). 国語辞典に「古風」と注記される語の使用実態調査 —『現代日本語書き言葉均衡コーパス』を用いて—. 『言語処理学会第 17 回年次大会発表論文集』.

[99] 佐野大樹 (2011 予). 『日本語アプレイザル評価表現辞書（態度表現編）』の構築 —評価の多様性を捉えるための言語資源の開発—. 『言語処理学会第 17 回年次大会発表論文集』.

[100] 佐野大樹 (2011 予). 患者の語りにおける感情表現の使用傾向 —『アプレイザル評価表現辞書（態度表現編）』を用いた乳がん患者・前立腺がん患者の語りの分析—. 『社会言語科学会第 27 回研究大会発表論文集』.

[101] 田中弥生, 佐野大樹 (2011 予). Yahoo!知恵袋における質問と回答の分類 —修辞ユニット分析を用いた脱文脈化-文脈化の程度による検討—. 『社会言語科学会第 27 回研究大会発表論文集』.

[102] 田中弥生, 佐野大樹 (2011 予). 修辞ユニット分析からみた Q&A サイトの言語的特徴. 『言語処理学会第 17 回年次大会発表論文集』.

[103] 山崎誠 (2011 予). 多義語を構成する意味の使用傾向 —品詞と活用形による違い—. 『言語処理学会第 17 回年次大会発表論文集』.

招待講演等

[104] 丸山岳彦 (2006). 現代日本語書き言葉均衡コーパスサンプリング方法について. 現代日本語書き言葉均衡コーパス仕様説明会 NAIST 東京事務所.

[105] 丸山岳彦 (2006). 言語学におけるコーパスの位置づけ —コーパス言語学の現状と今後の展開—. 神奈川大学 言語研究センター 招聘.

[106] 山崎誠 (2006). 国立国語研究所の語彙調査の歴史と課題. 東京大学大学院教育学研究科教育研究創発機構「教育測定・カリキュラム開発（ベネッセコーポレーション講座）」第 12 回研究会.

[107] 丸山岳彦 (2008). 『現代日本語書き言葉均衡コーパス』の設計と構築 Balanced Corpus of Contemporary Written Japanese -its design and compilation-. 韓国 国立国語院 招聘.

[108] 丸山岳彦 (2008). 日本語コーパスの現状と課題 —『現代日本語書き言葉均衡コーパス』を中心に—. 獨協大学 国際教養学部 言語文化学科 招聘.

[109] 山崎誠 (2008). 現代日本語書き言葉均衡コーパスと日本語研究の展開. 東京外国語大学グローバル COE 講演会.

[110] 丸山岳彦 (2009). 日本語コーパスの現状と課題 —『日本語話し言葉コーパス』 および『現代日本語書き言葉均衡コーパス』 を中心に—. 東京大学 文学部 言語学研究室 招聘.

コーパス開発センター （サンプリングサブグループ）

山崎 誠　　（言語資源研究系准教授、コーパス開発センター（兼））
柏野 和佳子　（言語資源研究系准教授、コーパス開発センター（兼））
丸山 岳彦　（言語資源研究系助教、コーパス開発センター（兼））
佐野 大樹　（コーパス開発センタープロジェクト特別研究員）
田中 弥生　（コーパス開発センタープロジェクト特別研究員）
秋元 祐哉　（コーパス開発センタープロジェクト奨励研究員）
大矢内 夢子　（コーパス開発センタープロジェクト奨励研究員）
稲益 佐知子　（派遣社員、マンパワー・ジャパン株式会社）

国立国語研究所内部報告書（LR-CCG-10-02）

『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

平成23年2月25日
執筆者　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
発行者　大学共同利用機関法人 人間文化研究機構 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）

ISBN 978-4-906055-01-2

国立国語研究所