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—

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

独立行政法人 国立国語研究所

1　導入

国立国語研究所では現在、『現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese; 以下 BCCWJ）』の構築を進めている。BCCWJ は、過去 30 年間に渡るさまざまな書き言葉を収録する、1 億語規模のバランストコーパスである [2]。

BCCWJ を構成する 3 つのサブコーパス（以下 SC）の 1 つ「生産実態 SC」は、2001 年から 2005 年に発行された書籍・雑誌・新聞の生産実態を反映するコーパスである。我々は、この 5 年間に生産された書籍・雑誌・新聞の生産実態を調査し、複数の基準で層化した母集団を定義した上で、各層に含まれる「総文字数」を推計することにした。そして、この総文字数の比を用いて生産実態 SC の構成比率を決定することとした [1]。これは、書き言葉の生産力という側面を文字数という総量によって近似的に把握するという見方に拠るものであり、これによって統計的な代表性を備えたコーパスを実現する。

そこで問題となるのは、(a) 5 年間に生産された書き言葉の総体をどのように把握し、そこから定義される母集団をどのように層化するか、(b) 各層の総文字数をどのように推計するか、という 2 点である。本稿では、5 年間に生産された書き言葉の中から我々が母集団をどのように定義・層化し、そこに含まれる総文字数をどのように推計したかについて報告する。

2　書き言葉の総体の把握と母集団の定義

以下では、書き言葉の生産実態を捉えるための調査方法と結果、そして母集団の定義、層化方法について述べる。

2.1　書籍の生産実態と母集団の定義

まず、書籍の生産実態に関する調査と母集団の定義について述べる。2001 年から 2005 年の間に国内で発行された書籍の生産実態を把握するために、我々は国立国会図書館の蔵書目録を典拠として用いることにした。国立国会図書館法により、国内で刊行される出版物は全て国立国会図書館に納入される。そこで、国立国会図書館の蔵書目録データ「J-BISC」をデータベース化し、国内で発行された書籍の書誌情報を網羅的に収めたリストを作成した。その結果、2001 年から 2005 年までに発行された書籍の全冊数は 561,514 冊、合計 112,504,115 ページであった。

ただしこの中には、漫画、写真集、人名録のように言語表現が主体でないものや、極端にページ数の少ないもの、非流通資料や非売品など、我々の意図に照らして不適切な内容をもつものも多く含まれている。そこで、コーパスの収録対象を絞り込むための条件を記述した「適切性条件」を設定し、約 56 万冊の書籍を絞り込むことにした。適切性条件の例[1]と、その条件によって除外される冊数（複数該当）を表 1 に示す。

表 1: 「適切性条件」による書籍の絞り込み

除外条件	除外数
40 ページ以下の書籍	101,417
ページ数の記録がない書籍	67,424
官公庁刊行物のうち非流通物	47,185
学習試験図書	37,780
電子資料、地図資料など	32,635
漫画	24,128
写真集・図画集	10,367

この絞り込みによって、全体の冊数は 317,117 冊、74,911,520 ページに絞り込まれた。この結果を、生産実態 SC「書籍」の母集団と定義した。

次に、この母集団を「発行年」および「日本十進分類法（NDC）」という基準で層化した。「発行年」は 2001 から 2005 年の 5 分類、「NDC」は J-BISC に付与されている NDC（1 桁目）の 11 分類（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、null（情報なし））とした。書籍の母集団を「NDC」で層化した際の冊数とページ数を、表 2 に示す。これにより、5 年間に発行された書籍の実態を、冊数・ページ数の側面から把握することができた。

[1] 適切性条件は、実際には 1,440 行から成る SQL 文である。

表 2: 書籍の母集団

NDC	冊数	ページ数
0	11,132	2,859,793
1	18,067	4,529,329
2	24,624	6,449,172
3	62,986	16,059,116
4	28,745	6,771,958
5	31,377	6,681,335
6	15,332	3,298,313
7	25,387	5,153,531
8	5,211	1,196,840
9	73,716	18,888,278
null	20,540	3,023,855
合計	317,117	74,911,520

表 3: 雑誌の母集団

分野	冊数	ページ数
1.総合	38,383	7,163,989
2.教育	5,456	983,224
3.政治・経済・商業	3,168	469,282
4.産業	599	115,172
5.工業	7,101	1,493,800
6.厚生・医療	1,072	189,488
合計	55,779	10,414,955

2.2 雑誌の生産実態と母集団の定義

次に、雑誌の生産実態に関する調査と母集団の定義について述べる。「雑誌」という概念を、定期的に刊行される冊子（定期刊行物）という観点から捉えると、いわゆる月刊誌や週刊誌だけでなく、学会誌や紀要などの冊子や、極めて限られた地域コミュニティだけで流通している冊子など、実に多様な様式のものが入ってきてしまう。そこで、「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る出版社が発行した定期刊行物」という条件で絞り込むことにした。日本雑誌協会の加盟社には有名な出版社が多く含まれることから、いわゆる「雑誌」として想起される定期刊行物の範囲を捉えるのに適切であると判断した。

まず、2001 年から 2005 年における雑誌協会加盟社のリストを作成した。該当数は異なりで 102 社である。さらに、国内で刊行される雑誌・新聞・要覧などを網羅的に収集した目録『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、対象出版社が各年に発行した定期刊行物に関する書誌情報を抽出した。その際、新聞・通信、コミック、要覧、非日本語による定期刊行物は除外した。その結果、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、合計 55,779 冊、10,414,955 ページとなった。この結果を、生産実態 SC「雑誌」の母集団と定義した。

次に、この母集団を「発行年」「分野」という基準で層化した。「発行年」は 2001 年から 2005 年の 5 分類、分野は『雑誌新聞総かたろぐ』で分類されている 6 分類（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）とした。雑誌の母集団を「分野」で層化した際の冊数とページ数を、表 3 に示す。これにより、5 年間に発行された雑誌の実態を、冊数・ページ数の側面から把握することができた。

2.3 新聞の生産実態と母集団の定義

最後に、新聞の生産実態に関する調査と母集団の定義について述べる。雑誌と同様、「新聞」という概念にも、全国紙、地方紙、スポーツ紙、専門紙、タウン紙など、実に幅広い様式が存在する。これらを全て収録対象とすると、研究用途上あるいは実務上の障害を生じさせる可能性が高い。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」という条件で絞り込むことにした。さらに、この条件ではカバーできない各地域の有力な地方紙も取り入れ、日本全国で発行されている新聞の集合となるよう調整した。その結果、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[2]、1,198,189 ページとなった。この結果を、生産実態 SC「新聞」の母集団と定義した。新聞 16 タイトルを、図 1 に示す。

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

図 1: 新聞の母集団となるタイトル一覧

次に、この母集団を「発行年」「紙種」という基準で層化した。「発行年」は 2001 年から 2005 年の 5 グループ、「紙種」は全国紙・ブロック紙・地方紙の別、および 16 種の新聞のタイトルとした。新聞の母集団を「紙種」で層化した際の冊数とページ数を、表 4 に示す。

表 4: 新聞の母集団

紙種	冊数	ページ数
全国紙	15,950	426,472
ブロック紙	9,570	248,300
地方紙	24,105	523,417
合計	49,625	1,198,189

以上のように、生産実態 SC に含まれる書籍・雑誌・新聞の実態を冊数・ページ数の側面から把握し、母集団を定義した。

3 母集団に含まれる総文字数の推計

以下では、母集団の各層に含まれる総文字数を推計し、生産実態 SC の構成比率を求める具体的な手順について述べる。

[2] この場合の 1 冊は、1 つの朝刊または夕刊を指す。

表 5: 総文字数推計基準表 （書籍）

	≦15cm	16cm	17cm	18cm	19cm	20cm	21cm	22cm	23cm	24cm	25cm	26cm	27cm	28cm	29cm	30cm	31cm≧	null
0	*410.8*	*442.0*	*473.3*	530.8	557.8	389.8	660.9	502.2	591.2	467.6	*723.3*	1026.4	*785.8*	*817.1*	*848.3*	569.4	*910.8*	*629.9*
1	544.0	418.5	*473.3*	542.2	497.0	529.6	743.9	655.8	*660.8*	*692.1*	*723.3*	*754.6*	*785.8*	*817.1*	*848.3*	*879.6*	*910.8*	*675.1*
2	516.0	434.0	*473.3*	466.6	471.9	529.8	652.9	747.1	*660.8*	*692.1*	*723.3*	1234.6	851.6	*817.1*	*848.3*	*879.6*	*910.8*	*700.6*
3	413.8	*442.0*	*473.3*	456.0	602.5	603.0	618.6	885.9	*660.8*	*692.1*	*723.3*	1272.6	*785.8*	*817.1*	*848.3*	1674.1	*910.8*	*757.6*
4	*410.8*	*442.0*	290.8	457.0	469.1	631.6	680.8	801.1	*660.8*	*692.1*	*723.3*	1159.7	473.6	*817.1*	*848.3*	580.4	*910.8*	*650.0*
5	430.4	396.0	*473.3*	340.8	409.9	539.0	672.8	815.5	*660.8*	860.8	*723.3*	702.4	544.8	*817.1*	636.2	962.5	*910.8*	*641.0*
6	*410.8*	*442.0*	*473.3*	558.8	513.9	511.6	448.3	1192.7	*660.8*	*692.1*	*723.3*	1254.6	1123.2	*817.1*	*848.3*	441.5	*910.8*	*707.2*
7	503.6	492.8	*473.3*	451.2	573.4	654.5	667.1	638.8	*660.8*	*692.1*	*723.3*	318.5	*785.8*	*817.1*	955.8	1510.8	778.8	*688.1*
8	600.6	345.0	*473.3*	522.2	632.0	561.6	910.4	1052.0	*660.8*	*692.1*	*723.3*	506.6	*785.8*	*817.1*	*848.3*	*879.6*	*910.8*	*701.3*
9	435.2	487.2	*473.3*	482.8	447.8	501.4	753.3	585.3	*660.8*	*692.1*	*723.3*	*754.6*	*785.8*	*817.1*	*848.3*	*879.6*	*910.8*	*661.1*
null	*410.8*	*442.0*	*473.3*	*504.5*	*535.8*	*567.0*	*598.3*	*629.5*	*660.8*	*692.1*	*723.3*	*754.6*	*785.8*	*817.1*	*848.3*	*879.6*	*910.8*	*660.8*

3.1 文字数計測作業の概要

まず、文字数計測作業の概要について述べる。母集団を構成する各層を、判型などの観点からさらに下位の層に分割し、ランダムサンプリングによって各層から複数のページをサンプルとして抽出した[3]。次に、抽出した各ページに含まれる文字数を実測し、層ごとに 1 ページあたりの平均文字数を求め、これを「キャラクタ密度（character density）」とした。その後、キャラクタ密度を係数として各層の 5 年間の総ページ数に掛け合わせることにより、各層に含まれる総文字数、および母集団全体の総文字数を推計した。

文字数を計測する範囲は本文部分に限定し、広告は全て対象外とした。本文部分に現れた文字要素については、文章内の論理構造の別（本文、図表、脚注、キャプション、ルビ、柱など）や、文字種の別（かな・カナ・漢字・記号・外国語・絵文字など）を問わず、全て計測対象とした。

3.2 書籍の文字数計測

以下、書籍の文字数計測について詳述する。2003 年に発行された全書籍（65,719 冊、15,544,357 ページ）を対象に、11 種類の「NDC」と、18 種類の「判型」によって行列を作成し、各「NDC」内で 10%以上の構成比をもつ判型を求めた。次にこの判型に該当する書籍をランダムに抽出し、さらに各冊から 5 ページをランダムに抽出し、人手または OCR によって文字数を計測して、その冊の平均キャラクタ密度を求めた。実際に計測したのは、227 冊（2003 年全体の 0.345%）、1,135 ページ（2003 年全体の 0.073 ‰）であった。

実測により得られた平均キャラクタ密度の分布から、「NDC」および「判型」に関する回帰直線 y=31.255x+379.5 を得た。結果を図 2 に示す。

この回帰直線を用いて、実測しなかった 10%未満の構成比しかもたない判型について、キャラクタ密度を計算した。これにより完成した行列を、書籍の「総文字数推計基準表」とした。表 5 に示す（表 5 中、イタリック体になっている数値は、回帰直線に基づいて算出された値であることを表す）。

[3] サンプルは 2003 年に発行されたものに限定した。

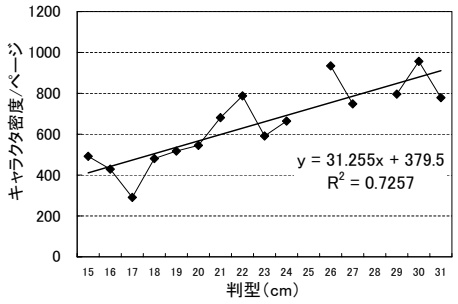


図 2: 判型別のキャラクタ密度と回帰直線（書籍）

この推計基準表を 5 年間のページ数に掛け合わせ、書籍の 5 年間の総文字数を推計したところ、約 485 億文字という結果になった。表 6 に示す。

表 6: 総文字数の推計結果（書籍）

NDC	総文字数	構成比率
0	1,636,414,548	3.371%
1	2,597,610,813	5.351%
2	4,301,204,340	8.861%
3	12,408,321,943	25.563%
4	5,069,594,034	10.444%
5	4,615,929,967	9.510%
6	2,196,387,437	4.525%
7	3,258,432,447	6.713%
8	888,800,128	1.831%
9	9,341,275,486	19.245%
null	2,225,954,208	4.586%
合計	48,539,925,351	100.00%

3.3 雑誌の文字数計測

次に、雑誌の文字数計測について詳述する。雑誌の文字数計測は、書籍の場合とほぼ同一の手順によって行った。2003 年に発行された全雑誌（909 タイトル、11,167 冊、2,095,217 ページ）を対象に、6 種類の「分野」と 5 種類の「判型」によって行列を作り、各「分野」内で 10%以上の構成比をもつ「判型」を求めた。次にこの判型に該当する雑誌の冊をランダムに抽出し、さらに各冊から 5 ページをランダムに抽出し、人手または OCR によって文字数を計測してキャラクタ密度を求めた。実際に計測したのは、53 冊（2003 年全体の 0.475%）、265 ページ（2003 年全体の 0.127 ‰）であった。

実測により得られた値については行列に埋め込み、実測しなかった 10%未満の構成比しかもたない判型については他の分野で計測したその判型のキャラクタ密度の平均値を割り当てた。この結果を、雑誌の「総文字数推計基準表」とした。表 7 に示す（表 7 中、イタリック体になっている数値は、他の分野で計測した平均値に基づいて算出された値であることを表す）。

表 7: 総文字数推計基準表 （雑誌）

	A4系	A5系	AB系	B4系	B5系
1.総合	1022.1	1031.2	1413.7	831.1	807.5
2.教育・学芸	765.4	884.8	*1413.7*	*831.1*	878.1
3.政治・経済・商業	1014.2	*921.4*	*1413.7*	*831.1*	798.2
4.産業	*1093.2*	*921.4*	*1413.7*	*831.1*	506.6
5.工業	973.2	*921.4*	*1413.7*	*831.1*	*767.9*
6.厚生・医療	1273.3	*921.4*	*1413.7*	*831.1*	510.0

この推計基準表を 5 年間のページ数に掛け合わせ、雑誌の 5 年間の総文字数を推計したところ、約 105 億文字という結果になった。表 8 に示す。

表 8: 総文字数の推計結果（雑誌）

分野	総文字数	構成比率
1.総合	7,421,447,806	70.575%
2.教育・学芸	877,875,592	8.348%
3.政治・経済・商業	456,459,405	4.341%
4.産業	110,640,958	1.052%
5.工業	1,468,293,360	13.963%
6.厚生・医療	180,964,513	1.721%
合計	10,515,681,636	100.00%

3.4 新聞の文字数計測

最後に、新聞の文字数計測について詳述する。新聞の場合、印刷紙面がほぼ定型であることを考慮し、書籍・雑誌とは異なる方法で推計を行った。まず、「朝日」「毎日」「読売」「日本経済」の全国紙 4 紙の朝夕刊、各 1 日分をランダムに抽出し、各ページに含まれる文字数を人手により計測した。実際に計測したのは、8 冊（全体の 0.081%）、211 ページ（全体の 0.880 ‰）であった（2003 年に発行された新聞は、16 タイトル、9,925 冊、239,638 ページ）。

次に、1 日分の紙面構成を分析し、1cm^2 あたりに含まれる文字数を面種 (いわゆる社会面・政治面など) ごとに算出した。さらに、実測しなかった曜日について面種ごとに面積を計測し、1 週間分の紙面構成を明らかにした。ここで得られた面種ごとの面積の合計と面種ごとの文字数/1cm^2 を掛け合わせることにより、新聞 1 週間分に含まれる総文字数を推計した。これにより、曜日間での文字数の誤差を修正し、平均的な 1 ページあたりの文字数を算出、上記 4 紙のキャラクタ密度とした。また、実測しなかったその他の新聞については、朝日・毎日・読売各紙のキャラクタ密度の平均値を割り当てることにより、各紙のキャラクタ密度とした。このキャラクタ密度を 5 年間の総ページ数と掛け合わせ、新聞の 5 年間の総文字数を推計したところ、約 64 億文字という結果になった。表 9 に示す。

表 9: 総文字数の推計結果（新聞）

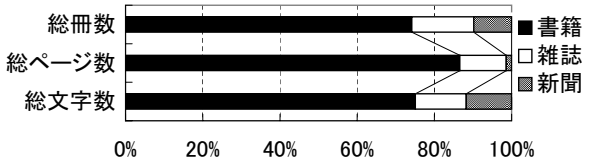
紙種	総文字数	構成比率
全国紙	2,417,622,461	37.681%
ブロック紙	1,296,592,154	20.209%
地方紙	2,701,855,499	42.111%
合計	6,416,070,114	100.00%

3.5 5 年間の書き言葉の総量

以上の手続きにより、2001 年から 2005 年に生産された書籍・雑誌・新聞に含まれる総文字数を推計したところ、全体で約 655 億文字という結果を得た。調査結果全体をまとめたものを、表 10 に示す。

表 10: 2001 年から 2005 年における書き言葉の総量

	書籍	雑誌	新聞	合計
総冊数	317,117	55,779	49,625	422,521
%	75.054%	13.201%	11.745%	100.00%
総ページ数	74,911,520	10,414,955	1,198,189	86,524,664
%	86.578%	12.037%	1.385%	100.00%
総文字数	48,539,925,351	10,515,681,636	6,416,070,114	65,471,677,100
%	74.139%	16.061%	9.800%	100.00%
キャラクタ密度（平均）	648.0	1009.7	5354.8	—



4 まとめ

本稿では、BCCWJ 生産実態 SC における母集団の定義と層化、ならびにそこに含まれる総文字数の推計について、調査結果の報告を行った。本稿で推計された総文字数の比によって、生産実態 SC の構成比率が決められた。

本稿で示されたデータは、BCCWJ 生産実態 SC の設計・構築だけでなく、統計的な言語研究を行う上での資料として重要な手がかりを与えるものである。本稿で示した手法の有効性の検証、ならびにデータ整備を引き続き進めていきたいと考えている。

付記： J-BISC のデータベース化には、国立国会図書館より協力を得た。文字数計測の調査には、東京都立多摩図書館、立川市中央図書館より協力を得た。記して感謝申し上げる。本研究は、文部科学省科学研究費補助金特定領域研究「日本語コーパス」による補助を得た。

参考文献

[1] 丸山他 (2006) 現代日本語の書き言葉に関する生産実態と流通実態.『言語処理学会 第 12 回大会 発表論文集』.

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