

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

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

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、代表性を有する現代日本語書き言葉コーパスの構築を計画している [3]。この計画は、現代日本語の書き言葉を対象としたバランストコーパスを構築する初めての試みであり、語彙調査、文字調査、文法研究などの言語学的研究、辞書編纂や教育などの産業的な応用など、さまざまな要請に対応する大規模汎用コーパスとしての役割が期待されている。コーパスサイズは1億語強が想定されている。現在、コーパスデザインについて検討を進めているところであるが、そこで問題になるのは、(1) 現代日本語の書き言葉の総体をどのように捉えるか、(2) どのようにサンプルを選べば母集団を適切に代表する集合が得られるか、という2点である。

我々はこのコーパスの構築に向けて、実際に産出された書き言葉の総体を捉えるための基礎調査を行い、そこからバランスよくサンプルを採取する、という方針を立てた。そこで本稿では、書き言葉の総体を把握し、そこから母集団を定義することを目的とした基礎調査について報告する。以下では調査の概要について述べ、この書き言葉コーパスの母集団をどのように定めるかについて述べる。なお、コーパス構築計画の概要については山崎他 (2006) を参照されたい。

2 書き言葉の代表性をどう捉えるか

はじめに、代表性を有する書き言葉コーパスを構築する上で、代表性 (representativeness) をどのように実現するかという点について述べる。あるサンプルが母集団を適切に代表している状態を作るためには、母集団の実態になるべく近似した部分集合を抽出する必要がある。バランストコーパスに即して言えば、現代日本語の書き言葉 (の一部) を母集団として定義し、必要に応じて層別化などの操作を施した上で、ランダムにサンプルを抽出することが必要となる。これによって、現代日本語の書き言葉全体の縮図となるような、代表性を有するコーパスが実現されると考える。

我々は、書き言葉の代表性を、2つの観点から捉えることにした。それは「生産実態に基づく代表性」およ

び「流通実態に基づく代表性」という2つである。前者は生産される書き言葉の総体を捉えるための観点であり、後者は実際に分布している書き言葉の有様を捉えるための観点である。

生産実態に基づく代表性は、ある期間に生産された全ての書き言葉を母集団とし、そのサンプルを抽出することによって実現させる。書き言葉の生産量を表す尺度として「文字数」を想定し、新聞、書籍、雑誌など発行された書き言葉に含まれる文字の総数を推計することにより、どのような種類の書き言葉がどれだけ生産されたかを見極めることにした。そして、得られた各メディアにおける総文字数の比を、コーパスの構成比率に用いることにした。このためには、対象期間内に生産された全ての書き言葉を把握し、そこに含まれる文字数を推計することが必要になる。

一方、流通実態に基づく代表性は、ある時点における書き言葉の分布状況を反映するようにサンプルを抽出することによって実現させる。世の中に広く分布している書き言葉を社会的な需要の高いものと見なし、それを母集団とすることにより、社会的に需要の高い書き言葉を抽出することを意図している。書き言葉の分布状況を捉えるにはさまざまな観点が考えられるが、ここでは「東京都内の図書館で共通に所蔵されている書籍」という基準について考える。このためには、都内の図書館における書籍の所蔵状況について把握することが必要になる。

我々の書き言葉コーパスは、以上の2種類の代表性に基づき、2種類のサブコーパスから構成される。2種類のサブコーパスには、それぞれの代表性を実現するための異なる母集団が定義されることになる。

以下では、まず、書き言葉の総体の捉え方について論じ、次いで生産実態を捉えるための調査、流通実態を捉えるための調査のそれぞれについて示す。

3 書き言葉の総体をどう捉えるか

一口に書き言葉の総体と言っても、書き言葉は実に多様な形で実現される。一般的に想起される書き言葉の種類としては、例えば、以下のようなものが考えられる。

新聞、書籍、雑誌、教科書、白書、広報誌、タウン誌、電話帳、取扱説明書、広告、パンフレット、看板、商品パッケージ、作文、手紙、日記、字幕、WWW ページ...

書き言葉の総体というものを観念的に考えたとき、上記は全てその対象となるであろう。つまり、誰かによって何らかの形で生産された書き言葉は全て、書き言葉の総体の一部を構成することになる。

しかしながら、バランスよくサンプルを採取するためには、調査対象に含まれる要素をあらかじめ数え上げ、全ての要素が等確率で抽出されるように母集団を構成する必要がある。ところが上記の書き言葉は、本質的には有限であるものの、その全てを把握することは実質的に不可能であり、あくまでも理想的な対象集団 (universe) でしかない。

そこで、出版目録や年鑑等でその抽出枠 (sampling frame) を規定することのできる「刊行物」を、書き言葉コーパスの第一次的な母集団 (population) とすることにした。上記の例で言えば、新聞、書籍、雑誌、教科書、白書がこれに該当する。一方、生産が厳密な形で管理されていない書き言葉 (広告、看板等) や個人的に生産された書き言葉 (手紙、日記等) は、我々のコーパスの収録対象外とした。また、言語コーパスとしての特性を考慮し、漫画、画集、写真集、図説など言語表現が主体でないものや、人名録、電話帳、データ集など言語コーパスとして不適切な内容を持つものは、収録対象外とした。さらに、実作業上の便宜を考慮し、同人誌や非売品の書籍など流通ルートに乗らない特殊な刊行物も収録対象外とした。

4 書き言葉の生産実態に関する基礎調査

以下では、書き言葉の生産実態に関する調査について述べる。ここで言う生産実態とは、ある期間に発行された全ての刊行物の異なりを指す。すなわち、発行部数の多寡は考慮に入れず、どれだけの種類の書き言葉が生産されているかという点を重視する。対象期間は、2001～2005年の5年間とした。この期間に発行された刊行物がコーパスへの収録候補となる。

現在検討しているのは、新聞・書籍・雑誌の母集団の確定手順、および各メディアの総文字数の推計方法である。例として、以下では新聞の母集団の確定手順、および総文字数の推計について詳しく示す。

4.1 新聞の生産実態と母集団の確定

新聞は日常よく目にする書き言葉であるが、国内にはいわゆる全国4大紙 (朝日・読売・毎日・日経) のほかにも実にさまざまな新聞が存在しており、その総体を把握するのは容易ではない。そこで、『雑誌新聞総かたるぐ』(メディア・リサーチ・センター発行、以下

『総かたるぐ』)を利用して調査を行うことにした。『総かたるぐ』は、新聞・雑誌・通信・要覧などを網羅的に収集した目録であり、国内における定期刊行物の総目録として利用できる¹。

ここで、『総かたるぐ』2005年版に掲載されている新聞 (一般紙・スポーツ紙) のタイトルを数えてみると、全国紙が10タイトル、ブロック紙が11タイトル、道府県紙・ローカル紙が275タイトル、スポーツ紙が17タイトル、それぞれ記載があった。これが2005年現在、国内で発行されている新聞のほぼ全容、つまり新聞の生産実態ということになる。

しかし、この中には「月刊新聞」「旬刊新聞」「年10回刊新聞」という刊期のものや、発行部数が1,000部に満たないローカル紙なども含まれている。これらはいわゆる新聞というメディアの刊行形態から外れており、さらに現物の入手が難しくサンプリング作業に問題を引き起こす恐れがある。そこで、新聞タイトルを何らかの条件によって絞り込んだ方がよいと判断した。

現在、「社団法人日本新聞協会に加盟している新聞社が発行している日刊新聞であること (ただし英字紙を除く)」という条件によって絞り込みを行うことを検討している。この条件を確認するための資料として、『全国新聞ガイド 2005年版』(日本新聞協会発行)を用いた。これにより、対象紙は合計103タイトルにまで絞り込まれた。この手順で選定される新聞タイトルの一覧を図1に示す。

全国紙 (9紙): 朝日新聞, 読売新聞, 毎日新聞, 産経新聞, 日本経済新聞, 日経金融新聞...

ブロック紙 (7紙): 北海道新聞, 中日新聞, 西日本新聞, 東京新聞, 夕刊フジ, 内外タイムス...

県紙・ローカル紙 (77紙): 北陸中日新聞, 日刊県民福井, 釧路新聞, 十勝毎日新聞, 苫小牧民報...

スポーツ紙 (10紙): 中日スポーツ, 東京中日スポーツ, 西日本スポーツ, サンケイスポーツ...

専門紙 (5紙): 電波新聞, 日本繊維新聞, 日本海軍新聞, 日刊水産経済新聞, 日本農業新聞

図1: 新聞の母集団リスト (一部)

現在、以上のような手順に基づき、2001～2005年の5年間に発行された新聞について選定作業を進めている。この選定作業によって最終的に得られる新聞タイトルの集合が、生産実態に基づくサブコーパスにおける新聞コーパスの母集団となる。

4.2 新聞の総文字数の推計

次に、書き言葉の生産力を表す尺度として、新聞に含まれる総文字数を推計した結果について示す。2003

¹ 『総かたるぐ』2005年度版には、22,559タイトルの刊行物、13,251社の発行社に関する情報が記載されている。

表 1: 新聞 1 日分に含まれる文字数 (実測値)

紙種		本文	見出し	データ	株価欄	ラテ欄	柱	1 日合計
朝日	朝刊	87,948	4,075	12,101	0	21,020	2,232	127,376
	夕刊	57,170	2,780	8,015	23,232	15,359	1,269	107,825
読売	朝刊	94,159	3,739	6,700	62,134	25,678	2,082	194,492
	夕刊	53,965	2,217	3,811	25,060	11,433	961	97,447
毎日	朝刊	93,098	4,342	11,739	52,595	28,209	1,973	191,956
	夕刊	36,620	1,596	17,478	0	8,984	823	65,501
日経	朝刊	126,157	5,094	63,476	101,473	22,298	2,283	320,781
	夕刊	51,465	2,448	24,455	57,963	13,006	1,366	150,703

年に発行された朝日新聞・毎日新聞・読売新聞・日本経済新聞について、ランダムに選んだ 1 日分 (朝夕刊別) の新聞に含まれる文字数を、本文 (キャプション含む)、見出し、データ (図表など)、株価欄、ラジオ・テレビ欄、柱の別に、人手で数えた²。結果を表 1 に示す。

表 1 を見ると、朝日新聞の朝刊、毎日新聞の夕刊は株価欄のない曜日が選ばれているため、株価欄のある日に比べて合計値が小さくなっている。この差を修正するために、表 1 の実測値が印刷されている新聞の面積を測り、1 平方センチメートルあたりに入る文字数を面種ごとに割り出した。さらに、ランダムに選んだ 1 週間分の朝夕刊の面積を面種ごとに調べ、先に求めた面積あたりの文字数を係数として掛け合わせて、1 週間分の新聞に印刷されている文字数を推計した。さらに、休刊日数を勘案して、1 年分の新聞に印刷されている文字数を推計した。結果を表 2 に示す。

表 2: 新聞 1 年分の総文字数 (推計値)

紙種		1 日 (平均)	1 週間分	1 年分
朝日	朝刊	181,659	1,271,614	64,488,996
	夕刊	98,588	591,526	27,900,310
読売	朝刊	175,412	1,227,881	62,271,108
	夕刊	103,098	618,590	29,176,828
毎日	朝刊	158,016	1,106,113	56,095,731
	夕刊	83,816	502,896	23,719,928
日経	朝刊	280,615	1,964,305	99,618,325
	夕刊	134,287	805,723	38,003,268
合計		1,215,491	8,088,648	401,274,493

この結果から、4 大紙を合計すると 1 年間に約 4 億字が生産されていると推計される。今後、この推計方法をブロック紙・地方紙などにも適用していくことにより、新聞の母集団に含まれる総文字数を推計する予定である。

4.3 書籍・雑誌の母集団と総文字数推計

以下、書籍・雑誌の調査について簡単に述べておく。

書籍の生産実態の把握には、国立国会図書館の蔵書目録を手がかりとすることを検討している。国立国会図書館法により国内で刊行された出版物は全て国立国

²新聞テキストを収めた CD-ROM は、記事の重複や欠落が多数あるため、生産された文字数の正確な測定には不向きであると判断した。

会図書館に納入されることになっており、この納本制度によって生産された書籍のほぼ全容を捉えることができる。また、国立国会図書館の蔵書目録として J-BISC が市販されており、電子的なリストとして活用できる。

ただし、国会図書館には一般の書籍だけでなく、官公庁出版物などの非流通資料や非売品なども多く所蔵されており、その全てを対象とすることは問題がある。そこで、ISBN が付与されている書籍のみを抽出し、母集団とすることを検討している。対象期間 (2001~2005 年) に発行された ISBN 付きの書籍数は、現在のところおよそ 40~50 万冊になると見込まれている。

次に、雑誌の生産実態については、新聞と同様『総かたろぐ』を利用する予定である。ただし『総かたろぐ』にはいわゆる雑誌と種々の定期刊行物が区別なく列挙されており、中にはごく限られた範囲でしか流通していない定期刊行物も多く含まれている。ここでも新聞と同様、何らかの条件によって絞り込みを行うことが適切であると判断した。

現在、「社団法人日本雑誌協会に加盟している出版社が発行している定期刊行物であること (ただし英字紙を除く)」という条件によって絞り込みを行うことを検討している。これによって、『総かたろぐ』2005 年版に記載されている 18,331 タイトルを、1,078 タイトルにまで絞り込むことができる。同様の調査を 2001~2005 年に拡大して行い、対象期間内に発行された雑誌の選定と母集団の確定を行う予定である。

さらに、書籍・雑誌ともに、母集団の確定が済み次第、そこに含まれる総文字数の推計を行う予定である。これによって得られる新聞・書籍・雑誌の総文字数の比率を、生産実態に基づくサブコーパスにおける各メディアの構成比率として採用する予定である。

5 書き言葉の流通実態に関する基礎調査

以下では、書き言葉の流通実態に関する調査について述べる。ここで言う流通実態とは、ある時点における書き言葉の分布状況そのものを指している。書き言葉の分布を捉える観点にはさまざまあり得るが、ここでは「東京都内の図書館で共通に所蔵されている書籍」という基準について検討する。図書館は書籍に関する公

535,865	目黒区	322,985	杉並区	197,224	小金井市
453,016	江東区	317,199	港区	193,479	青梅市
441,520	足立区	317,151	立川市	186,514	東久留米市
440,829	調布市	314,349	荒川区	180,236	昭島市
440,346	町田市	310,845	品川区	180,073	東大和市
416,217	八王子市	303,871	文京区	178,150	あきる野市
413,609	大田区	300,666	武蔵野市	178,116	国立市
404,082	府中市	296,547	新宿区	166,001	羽村市
388,527	江戸川区	294,225	豊島区	152,997	福生市
377,975	練馬区	285,204	渋谷区	147,646	狛江市
373,324	板橋区	274,115	日野市	136,631	武蔵村山市
361,546	葛飾区	268,511	三鷹市	121,957	千代田区
355,857	墨田区	264,338	西東京市	112,666	稲城市
353,878	北区	262,273	国分寺市	112,401	瑞穂町
352,713	中野区	254,113	東村山市	70,317	日の出町
329,566	多摩市	244,409	中央区	53,132	奥多摩町
326,412	小平市	234,294	台東区		

図 2: 都内 22 区+29 市町村が所蔵する書籍数

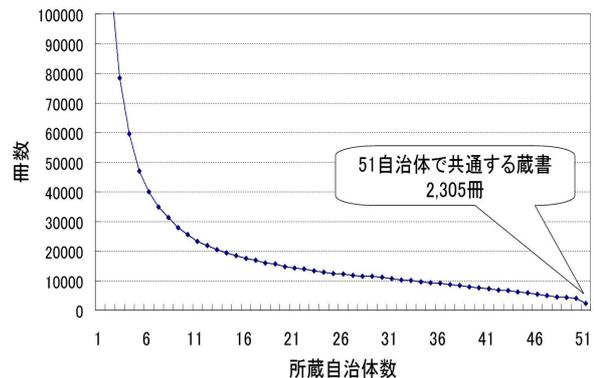


図 3: 51 自治体における共通蔵書の分布

共サービスを提供する機能を持つ機関であり、図書館の所蔵書籍は、生産された全ての書籍とは異なり、国民にとって需要が高いものが選定されていると考えられる。さらに、複数の図書館で共通に所蔵されている書籍ほどこの傾向が強いと見ることができる。

さて、東京都内の各図書館（各自治体）で所蔵している書籍の全容は、東京都立中央図書館によって情報が取りまとめられ、「ISBN 総合目録」として電子データ化されている。この目録は本来図書館の相互貸借事務で用いるためのデータで、ISBN を検索キーとしてその書籍がどこに所蔵されているかを知るためのものである。我々は、東京都立図書館からの協力を仰ぎ、このデータを共通蔵書の調査に用いることにした。

今回用いた 2005 年 10 月期の目録では、東京都下の 22 区、29 市町村の蔵書情報がデータ化されている。これら 51 自治体が所蔵する書籍数を、図 2 に示す。

51 自治体で所蔵する書籍数の合計は 14,215,915 タイトルであった³。この異なりを求めると 1,146,418 タイトルとなり、さらに日本語の書籍（ISBN が 4 から始まるレコード）のみを数えると 1,064,186 タイトルとなった。東京都下の公共図書館には、異なりで約 106 万冊の日本語書籍があることになる。

次に、これらの書籍がどれだけ共通に所蔵されているかについて、その分布状況を調べた。結果を図 3 に示す。集計の結果、51 自治体の全てで共通に所蔵されている書籍は 2,305 タイトルとなることが分かった。この 2,305 タイトルについて、ISBN を検索キーとして書誌情報を調べたところ、その大半が児童書またはいわゆる文芸書で占められることが分かった。

これら 2,305 タイトルの中身を見る限り、51 自治体全てで共通に所蔵されている書籍は、ベストセラーとして知られる小説やロングセラーとして知られる児童書など、確かに社会的な需要が高いものが大半である

³厳密に言えば、ISBN が正式に付与され始めたのは 1981 年以降であるため、それ以前に発行された書籍がどのように所蔵されているかについては分からない。それゆえ、ここで示した数字はその自治体の正確な蔵書数とは一致しない可能性がある。

と言えそうである。その一方で、この 2,305 タイトルを母集団と定めてサンプリングを行うと、内容に相当の偏りが生じることが予想される。少なくとも、言語研究をはじめとして多様な用途を想定する大規模汎用コーパスにふさわしいサンプリング結果は得られないと思われる。現在、図 3 に示した共通蔵書の分布の結果をもとに、なるべく偏りが少なく、多様な日本語表現を収集するにはどのような範囲からサンプルを取得すればよいかについて、検討を進めている。

6 まとめと今後の展望

本稿では、代表性を有する現代日本語書き言葉コーパスの構築に向けて、現在我々が進めている調査の概要について示した。日本語の書き言葉を対象としたサンプルコーパスの構築は、これまでまったく手が付けられていない状態であり、どのような情報をもとにどのような設計を行えば適切な代表性が得られるかという問題について、知見の蓄積がほとんどない。

Biber(1993) は、統計的な手法によって計量的な言語分析を行うためには、まず母集団の定義について厳密に検討することが必要であると述べる [1]。我々は、生産実態・流通実態という 2 つの観点により、それぞれ異なる母集団を厳密に定義した上で、日本語の書き言葉の 2 つの側面についてサンプルを得ることを目論んでいる。今後、母集団の厳密な定義とリスト化が完了し次第、実際のサンプリング作業に移る予定である。なお、具体的なサンプリングの手順については、丸山他 (2006) を参照されたい。

参考文献

- [1] Biber, D. (1993) "Representativeness in corpus design." *Literary & linguistic computing* 8(4). Oxford University Press.
- [2] 丸山他 (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。本予稿集所収。
- [3] 山崎他 (2006) 代表性を有する現代日本語書き言葉コーパスの設計。本予稿集所収。