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

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

独立行政法人 国立国語研究所

1　導入

現代のコーパス言語学は 1960 年代の Brown Corpus の誕生に端を発すると言われる。その後，世界各地でさまざまなコーパスが作られてきているが，日本国内に目を向けると，コーパス整備という点ではかなりの遅れを取っていると言わざるを得ない。特に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は全く未整備の状態で，著作権の切れた文学作品や一部の新聞社による CD-ROM 新聞記事テキストなどが書き言葉コーパスとして用いられているという状況である。

さて，コーパスがコーパスであるために満たすべき条件にはいくつかのものが考えられるが，その一つに「代表性（representativeness）」という考え方がある。コーパスは，母集団（書き言葉の総体等）に含まれる多様性をよりよく代表する集合になるように設計されることが望ましい，とする考え方である。Brown Corpus が意義を持ったのは，それが代表性という考え方に基づいてデザインされた初めてのサンプルコーパスであったという点にある。

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 [4]。このコーパスは，語彙調査，文字調査，文法研究などの言語学的研究だけでなく，辞書編纂や教育への応用など，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。現在，コーパスデザインの検討を進めているが，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。

このうち (1) の問題については，丸山他 (2006) で述べた。本稿では (2) の問題，すなわち，母集団を適切に代表する集合を得るためのサンプリング手法について検討する。言語表現を対象としたサンプリング手法として，現在我々が計画している手法について説明し，それがどのような特性を持つかについて検討する。

2　2 種類のサブコーパスと母集団

はじめに，現在我々が計画している書き言葉コーパスの構成について述べておく。

この書き言葉コーパスは，「生産実態に基づくサブコーパス」と「流通実態に基づくサブコーパス」という 2 種類のサブコーパスから構成される。前者は生産される書き言葉の総体を捉えるためのコーパスであり，コーパスサイズは 1,000 万語が計画されている。一方，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コーパスであり，1 億語の規模が計画されている。個々のサンプルの抽出単位は，生産実態に基づくサブコーパスでは 1,000 文字という固定長の単位が，また流通実態に基づくサブコーパスでは「章」や「記事」など言語的なまとまりを持つ範囲（ただし 10,000 文字を超えない範囲）という可変長の単位が，それぞれ想定されている。以下では，前者を「固定長コーパス」，後者を「可変長コーパス」と呼び分けることにする。

固定長コーパスでは，書き言葉の生産実態を捉えることを目的として，2001〜2005 年の 5 年間に生産された全ての書き言葉を理想的な調査対象集団とする。実際には，出版目録や年鑑等によってその抽出枠（sampling frame）を規定することのできる「刊行物（新聞・書籍・雑誌等）」を母集団として定める。母集団の確定には，『雑誌新聞総かたろぐ』（メディア・リサーチ・センター発行）や，国立国会図書館の蔵書目録（J-BISC, JAPAN/MARC(S)）等を用いる [3]。現在，サンプリングに用いる抽出枠として電子的なリストを整備中であるが，以下ではこのリストが完成していることを想定して議論を進める。このリストには，各メディア（新聞・書籍・雑誌）ごとに，タイトル（新聞名，書籍・雑誌タイトル），ページ数，判型などの書誌情報が含まれているものとする。

一方，可変長コーパスでは，世の中に広く分布している書き言葉を捉えることを目的として，東京都内の公共図書館で共通に所蔵されている書籍を母集団として定める。共通蔵書の分布を知るための資料としては，東京都立中央図書館が取りまとめている「ISBN 総合目録」を利用する。現在，どこまでの範囲の書籍を収録対象とするかについて検討している段階であるが，以下では抽出枠としての共通書籍リストが完成しているものとして議論を進める。リストには，各書籍ごとに，タイトル，ページ数，判型などの書誌情報が含まれているものとする。

3　サンプリングの基本方針と手順

以下では，今回実施するサンプリングの基本方針と抽出単位，および具体的なサンプリング手順について示す。

3.1　基本方針

サンプリングは，層化抽出法に準じて実施する。各メディア（新聞・書籍・雑誌）の違いを母集団を構成する層として捉え，各層からランダムにサンプルを抽出することにより，母集団の構成になるべく近似した部分集合を取り出すことを試みる。

各層から抽出する個々のサンプルの数は，比例割当法に基づき，各層ごとに推計した総文字数の比に応じて割り当てる[1]。今仮に，1 年間に生産される総文字数が，新聞 4 億字，書籍 10 億字，雑誌 6 億字と推計されたとすると，1,000 万語の固定長コーパスは，新聞 200 万語，書籍 500 万語，雑誌 300 万語分のサンプルによって構成されることになる。

また，各メディアによって構成される層を，その特性によってさらに下位の層に分類する。例えば新聞は，その流通範囲および性格の違いによって「全国紙」「ブロック紙」「地方紙」「スポーツ紙」「専門紙」という 5 つの層に分ける。また書籍は，各書籍に対して付与されている NDC（日本十進分類法）の分類[2]に基づいて，「総記」「哲学」「歴史」「社会科学」「自然科学」「工学」「産業」「芸術」「言語」「文学」という 10 の層に分ける。

それぞれの層の内部は流通範囲や内容といった点で均質な性格を持っており，かつ各層相互は異質な性格を持っている。このような層化抽出を行うことにより，母集団が内包する多様性からなるべく偏りなくサンプルを抽出しようとするのが，今回のサンプリングの基本方針である。

3.2　抽出単位

さて，実際にサンプルを抽出するには，多くの方法的な選択肢があり得る。従来，国語研究所が行ってきた漢字調査・語彙調査においては，主としてエリアサンプリングが採用されてきた。例えば「現代新聞の漢字調査」の中では，新聞 1 ページを 30 のエリア（縦 15 段×横 2 段）に分割し，乱数表によって該当エリアを選び，結果として母集団全体の 1/60 の面積に含まれる言語表現が選ばれるようなサンプリング方法を取っている [1]。

今回実施するサンプリングでは，エリアサンプリングではなく，いわば「文字サンプリング」とでも呼ぶべき方法を取る。これは，エリアサンプリングのように言語を載せる媒体（面積）の物理的な計測に基づいてサンプルを抽出するのではなく，あくまでも言語表現そのものが持つ絶対量，すなわち文字の総量に基づいてサンプルを抽出することを意図するものである。

理想的には，各層に含まれる全ての文字に対して等確率を与え，ランダムに 1 文字を指定することにより，その文字を基準点とする一定範囲のサンプルを抽出することを目指す。実際にはこれに近似した方法として，全てのページに対して等確率を与えた上で，ランダムに 1 ページを抜き出し，さらにそのページに含まれる 1 文字をランダムに指定するという方法を取る。指定された 1 文字をサンプル抽出基準点として，固定長コーパスでは 1,000 文字という固定長の範囲を[3]，可変長コーパスではその文字を含む「章」「記事」などの言語的なまとまりを持つ範囲を，それぞれサンプルとして抽出する。

3.3　サンプリング手順

以下，サンプリングの具体的な作業手順を示す。まず固定長コーパスについては，以下の手順によって作業を進める。

1. 2001〜2005 年の間に発行された全ての書き言葉のうち，出版目録や年鑑等によってその全体を規定することのできる「刊行物（書籍・雑誌・新聞等）」を調査し，母集団を確定する。
2. 母集団を新聞・書籍・雑誌等のメディア別に層別化する。さらに各層の内部を，発行形態・ジャンル・NDC 等によって層別化する。
3. 対象期間（5 年間）に生産された総文字数を，各層について推計する。
4. 総文字数（推計値）の比に応じて，各メディアが固定長コーパス全体に占める割合（構成比率）を決定する。そこから，各メディア・各層から抽出するサンプルの個数を算出する。
5. 各層の総ページ数の値をシャッフルし，全てのページに優先順位を割り当てる。優先順位の高いページから順に，以下の手順でサンプルを抽出する。
6. 該当ページに 10 × 10 の座標枠を当て，1 つの交点をランダムに指定する。
7. 指定された点に最も近い文字を，サンプル抽出基準点として定める。
8. サンプル抽出基準点を含む文の文頭に移動し，そこから 1,000 文字目の文字を含む文の末尾までを抽出し，サンプルとして確定する。
9. 白紙や全面広告などのページが当たった場合は無効とし，次候補のページからサンプルを抽出する。
10. 必要なサンプル数が獲得されるまで，6〜9 の手順を繰り返す。

[1] 各メディアにおける総文字数の推計方法については，丸山他 (2006) 参照。

[2] J-BISC の中で各書籍に付与された NDC コードを採用する。

[3] ただし，後述するように，文頭・文末の整形を行う。そのため，厳密に 1,000 文字が抽出されるわけではない。

一方，可変長コーパスについては，以下の手順を踏む。

1. 「ISBN 総合目録」を用いて，東京都内の公共図書館で共通に所蔵されている書籍のリストを作成する。
2. 得られたリストを NDC で層別化し，各層が可変長コーパス全体に占める割合（構成比）を決定する。そこから，各層から抽出するサンプルの個数を見積もる。
3. 各層の総ページ数の値をシャッフルし，全てのページに優先順位を割り当てる。優先順位の高いページから順に，以下の手順でサンプルを抽出する。
4. 該当ページに 10 × 10 の座標枠を当て，1 つの交点をランダムに指定する。
5. 指定された点に最も近い文字を，サンプル抽出基準点として定める。
6. サンプル抽出基準点を含む「章」「記事」「段落」など言語的なまとまりのうち，10,000 文字を超えない最大の範囲を抽出し，サンプルとして確定する。
7. 白紙や全面広告などのページが当たった場合は無効とし，次候補のページからサンプルを抽出する。
8. 必要なサンプル数が獲得されるまで，4〜7 の手順を繰り返す。

固定長コーパスと可変長コーパスは，それぞれ想定される母集団が異なっているものの，サンプリングの方法自体はほぼ同一である。すなわち，母集団をある基準によって層別化し，各層のリストを作った上で，ランダムにページを選び，さらにランダムに選んだサンプル抽出基準点をもとに所定の範囲のサンプルを抽出する。

4 書籍のサンプリングの実際

以下では，実際のサンプリング作業の具体例として，書籍から固定長コーパスのサンプルを抽出する手順を追ってみることにする。なお，この手順は試行として行ったものであり，対象期間を 2003 年のみに絞っていること，抽出枠の決定に用いる資料が 3 節で挙げたものと異なっていることを断っておく。

4.1 母集団の確定と層別化，構成比率の決定

1. 『日本書籍総目録 CD-ROM 2003 年版』（日本書籍出版協会）を用いて，2003 年に発行された書籍の総数を調べた。漫画，写真集，画集，楽譜，学習参考書の類は除外し，結果として残った 41,968 タイトルを，書籍の母集団として定めた。

2. 得られた 41,968 タイトルを NDC ごとに 10 の層に分類し，各層に含まれる冊数，総ページ数を求めた（表 1）。このうち総ページ数の比を，全体に対して各層が占める構成比率とした。

3. 1 つの層を巨大な 1 冊の本とみなし，各層ごとに通しページを振った。

表 1: 2003 年発行書籍の層別化と総ページ数の比

ジャンル	冊数		総ページ数	
総記	545	(1.31%)	143,339	(1.43%)
哲学	1,688	(4.05%)	445,675	(4.44%)
歴史	2,683	(6.44%)	681,554	(6.79%)
社会科学	9,244	(22.17%)	2,520,757	(25.10%)
自然科学	3,284	(7.88%)	797,004	(7.94%)
技術工学	2,216	(5.32%)	613,338	(6.11%)
産業	668	(1.60%)	145,776	(1.45%)
芸術	6,681	(16.03%)	1,060,751	(10.56%)
言語	960	(2.30%)	194,218	(1.93%)
文学	13,721	(32.91%)	3,438,677	(34.25%)
合計	41,968	(100%)	10,041,089	(100%)

上記より,「総記」層からは書籍に関する固定長コーパス全体の 1.43%に該当する量のサンプルが，また「文学」層からは 34.25%に該当する量のサンプルが，それぞれ抽出されることになる。

4.2 サンプル抽出範囲の確定

1. 各層に与えられた通しページ番号をシャッフルし，全てのページに対して優先順位を与えた。例えば，上記の「社会科学」層では，1〜2,520,757 の数値全てに対して順位を与えた。

2. 優先順位の高いページから順にサンプルを抽出する。そのページを含む書籍を割り出し，実際に該当するページを特定する。ここでは，以下の書籍の 49 ページ目が当たったものとする。

   『新「ことば」シリーズ 18 伝え合いの言葉』
   国立国語研究所 （国立印刷局）

3. 該当ページの印刷面に対して 10 × 10 の座標軸を当て，ランダムに 1 交点を指定する。ここでは，交点 “4–E” が選ばれたものとする。（図 1）

4. 交点に最も近い文字をサンプル抽出基準点とする。サンプル抽出基準点となった文字を含む文の文頭を起点として，そこから 1,000 文字目の文字を含む文の末尾まで移動する。結果として選ばれた文頭から文末までの範囲をサンプルとして確定する。

図 1 の例に即して言えば，交点 “4–E” に最も近い文字「た」がサンプル抽出基準点として選ばれ,「た」を含む文の文頭「歴史的に見ると...」からサンプルが抽出される。文頭の「歴」の字から数えて 1,000 文字目の文字（この例では数ページ先になる）を含む文の末尾までが，サンプルとして抽出されることになる[4]。

[4] なお，可変長コーパスの場合は,「た」を含む言語的なまとまり，すなわち「コラム 2」の全体がサンプルとして抽出される。

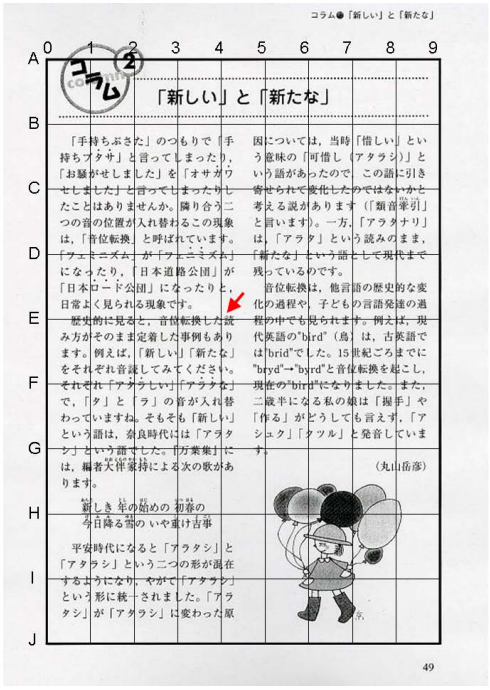

図 1: サンプル抽出基準点の確定

5　提案したサンプリング手法の特性

最後に，本稿で示したようなサンプリングの手法がどのような特性を持つかについて検討する。先述したように，国語研究所がこれまで行ってきた書き言葉の調査では，主としてエリアサンプリングが採用されてきた。ここでは，以上で提示した文字サンプリングが持つ特性を，エリアサンプリングと比較しながら考えてみたい。

母集団の定義　エリアサンプリングでは，調査対象となる全てのページ数を数え上げることにより，母集団の定義を容易かつ正確に行うことができる。一方，文字サンプリングでは，対象となる書き言葉に含まれる総文字数を推計しなければならず，あくまでも推計値としての母集団しか定義することができない。

抽出比の正確さ　エリアサンプリングでは面積比をもとにサンプル抽出を行うので，全体に対するサンプルの抽出比は明確である。一方，文字サンプリングの場合，総文字数を推計した値に対する抽出比しか計算できないため，抽出比の正確さは総文字数推計の精度に依存することになる。

個々のサンプルの画一性　エリアサンプリングでは当該エリアにどれだけの文字（抽出単位は文）が入っているかはあらかじめ予測できず，個々のサンプルの文字数にばらつきが生じる可能性がある。一方，固定長データの文字サンプリングでは 1,000 文字という規定枠があるため，画一的な大きさのサンプルを確実に抽出できる。可変長データの場合は，「章」や「記事」など完結した文書という意味での画一性が実現できる。

サンプルサイズの設計　エリアサンプリングでは，最終的にどれだけの文字数が抽出されるかはサンプリングが終了した後でないと分からず，あらかじめサンプルサイズを設計することは難しい。一方，固定長データの文字サンプリングでは，全体のサンプルサイズや必要なサンプル数をあらかじめ計画することができる。

サンプルの言語的特性　エリアサンプリングと固定長データの文字サンプリングでは，ともにまとまった範囲の言語表現が抽出される。抽出箇所がある程度まとまった文脈を構成している場合，一つの話題に関する語が集中して現れることがあり，語の出現傾向に偏りが生じる可能性がある。このような偏りは，例えば，小規模なサンプルを対象とした語彙調査にとって影響を与えることがある [2]。しかしながら固定長データは 1,000 万語という規模を想定しており，文脈が語彙の分布に与える影響は相対的に小さいものと考えられる。抽出単位をさらに小さくし，抽出箇所をばらつかせることでこのような偏りを小さくすることも考えられるが，作業にかかる負荷が大きくなるため，現在は 1,000 文字という抽出単位が適当であると考えている。

一方，可変長データは「章」や「記事」などの範囲が非画一的な長さで抽出されるため，統計的な厳密さを求める調査にはやや不向きである。ただし，文書として完結した構造を持つため，談話研究やテキスト処理など，文脈の働きを積極的に取り入れる研究に向く。また，語の正確な意味記述には広範囲の文脈の確認が必要となる場合が多々あるが，そのような用途にとっても可変長データは有用であると考えられる。

6　まとめ

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，現在計画しているサンプリングの基本方針と具体的な手順について示した。言語学的に有意味でかつ統計的な分析にも耐え得るサンプリング手法の開発には，試行を重ねながらその有効性を検証していくことが必要であると考えられる。今後，我々の書き言葉コーパスを構築していく過程において，本稿で示した手法の有効性を検証していきたいと考える。

参考文献

[1] 国立国語研究所 (1976)『現代新聞の漢字』. 秀英出版.

[2] 佐竹秀雄 (2001) 研究対象の量とサンプリング.『日本語学臨時増刊号 日本語の計量研究法』**20**(4). 明治書院.

[3] 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. 本予稿集所収.

[4] 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.