『現代日本語書き言葉均衡コーパス』のサンプル収録方法

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
独立行政法人　国立国語研究所

1. はじめに

代表性を備えた均衡コーパスを構築することは，統計学で言う標本調査において標本（サンプル）を抽出することに相当する。母集団の実態を適切に代表するように標本を抽出するためには，母集団の厳密な定義，適切な抽出枠の設計，そして均質的なサンプリングが求められる。

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（以下BCCWJ ）』という１億語規模の均衡コーパスを構築している。そこで求められるのは，種々の書き言葉を対象に，ゆれのない手続きにより斉一的なサンプルを抽出する作業である。しかしながら，実際の書き言葉は様々な構造的・論理的体裁を伴うものであり，そこから均質にサンプルを取り出すためには，一定の方針と作業上の詳細な取り決めが必要となる。

そこで本稿では，BCCWJ に収録するサンプルを取得する基準と手続きについて示す。サンプルとして収録すべき要素の基準を示した上で，実際のサンプリングの方法と，そこで生じる問題のありかを明らかにする。

2. 標本調査の対象としてのコーパス

標本調査とは，母集団から無作為に抽出した標本を分析することによって，直接には知り得ない母集団の性質を推定しようとする調査法である。例えば，日本国内の成人3,000 人を無作為に抽出して支持政党を問う世論調査は，標本調査の例である。母集団が備える性質をより正確に推定するためには，抽出した標本が母集団の実態を適切に代表している必要がある。この性質は，代表性と呼ばれる。世論調査の例で言えば，男女比，年齢比，居住地域などを考慮して，国内の成人の総体を適切に代表する3,000 人が標本として選ばれることが望ましい。

このことを，言語調査のためにコーパスを構築するという事例に置き換えてみよう。目的は，現代日本語の書き言葉の実態を，文字・語彙・文法・文体など種々の側面から調査することとする。この場合，母集団は現代日本語の書き言葉の総体であり，そこから抽出される標本（＝コーパス）は，母集団に対する代表性を備えていなければならない。ここで問題となるのは，(1)書き言葉の母集団をどう定義するか，(2)標本の母集団に対する代表性をどのように保証するか，(3)個々のサンプルとしてどのような単位を採用し，それをどのように取り出すか，という３点である。

BCCWJ の設計にあたり，我々はこれらの点を詳細に検討してきた。まず母集団である現代日本語の書き言葉の総体は，「現代日本語書き言葉の総文字数調査」によって推計した文字数によって定義した。そして，書籍・雑誌・新聞といったメディアごとに下位のジャンルを区分し，それぞれに含まれる推計総文字数を比例割当することにより，各ジャンルから取得するサンプル数を管理している。これにより，母集団に対して統計的に厳密な代表性が保証される。さらに，個々のサンプルの取得方法についても，サンプル取得の原則を定め，ゆれのない手続きにより均質なサンプリングを実施している。

BCCWJ の母集団の定義や代表性を保証する構成比率の算出方法については，すでに別のところで述べた（山崎(2006)，丸山・秋元(2007,2008)）。以下では，上記の３つ目の問題，すなわち，取得する個々のサンプルとして，どのような単位を採用し，それをどのように取り出すか，という点について詳しく述べる。

3. 書き言葉のサンプル収録 ― その単位と基準

3.1 固定長サンプルと可変長サンプル

BCCWJ では，「統計的に厳密な言語調査」「文体研究・テキスト研究」という 2 つの研究用途に応えるために，次の２種類のサンプル単位を採用している。

- 固定長サンプル：一律，1,000 文字。
- 可変長サンプル：章や節など，言語的なまとまり。ただし，最大で１万字を超えない範囲。

固定長サンプルは，母集団（＝推計総文字数）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。一方，可変長サンプルは，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。

いずれのサンプルも，母集団に含まれる全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，その文字を基準点として一定の手続きにより抽出する。これは，概念的に言えば，母集団に含まれる全ての文字を１次元上に配置した上で，ランダムに指定された１文字を基準点として，1,000 文字の範囲，およびその文字を含む「章」「節」などの言語的なまとまりを持つ範囲を，それぞれ固定長サンプル・可変長サンプルとして同時に抽出

する，ということである。

問題は，母集団に含まれる全ての文字を 1 次元上に配置するという操作をどのように近似するかという点である。より具体的には，書籍・雑誌・新聞などの印刷物 ―そこには多様な物理的・論理的な構造が含まれる― からどのように 1 次元の文字列を認定し，2 種類のサンプルをゆれのない手続きで抽出するか，という問題である。

3.2 サンプリングの基本方針

書籍・雑誌・新聞などの紙面からサンプルを均質的に抽出するためには，紙面を構成する諸要素のうち，どの要素を抽出し，どの要素を抽出しないのかを前もって決めておかなければならない。このためには，書き言葉が持つ構造をあらかじめ体系的に把握しておいた上で，個別の事例に対処していく必要がある。

以下では，書籍における書き言葉の構造を，「冊子」「版面」「文字」という 3 つの側面によって段階的に捉え，その中からコーパスに収録するサンプルとして抽出する部分と，それが満たすべき基準について示す。

「冊子」

書籍の冊子は通常，本文と呼ばれる書籍の実質的本体（以下，「冊本体」と呼ぶ）以外に，「とびら」「献辞」「凡例」「口絵」「序文」「目次」などの「前付」の要素や，「付録」「索引」「奥付」「参考文献」「後書き」などの「後付」の要素など，複数の要素から構成されている。基本的にサンプリングの対象とするのは，「冊本体」に加え，一定の文章量のある「序文」と「後書き」とする。これ以外の要素や「広告」などは，原則，サンプリングの対象としない。

「版面」

書籍の版面はおよそ図 1 に示すような要素から構成される。そのうち，「大見出し」「脇見出し」「リード」「中見出し」「小見出し」「本文」「コラム」「キャプション」「注」は，サンプリングの対象とする。一方，文字を主体としないフィギュア（「図」「写真」等）は，サンプリングの対象から外す。また，文字が主体であっても一方向に読めないフィギュア（「表」等）も，基本的にサンプリングの対象外とする。また，実質的内容をもたない「ノンブル（頁）」「柱（版面の余白部に示された見出し）」も対象外とする。

「文字」

書籍に印字されている文字のうち，「仮名」「漢字」「数字」「アルファベット」はサンプリングの対象とする。一方，「句読点・疑問符・感嘆符」「括弧・その他記号」などは，入力はするが，固定長サンプル 1,000 文字の対象とはしない。この区別は，純粋な言語表現を構成する文字種に限定して 1,000 文字を取得することにより，より精密な文字調査や語彙調査を実現しようという意図による。

さらに，上記までの基準に加えて，本文部分に含まれる言語表現そのものに関する条件として，「現代日本語として書かれたもの」という条件を設ける。したがって，以下のような「非現代日本語」の表現がまとまって出現した場合，その部分はサンプリングの対象から外す。

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

ただし，「彼は Thank you とだけ言った。」という例のように，一連の本文中に非現代日本語が混じっている場合は，除外しない。上記の各表現がサンプリングの対象外となるのは，ページや章，あるいは書籍全体が非現代日本語で構成されている場合，または，典型的には，前後に改行を伴い，主たる文からインデントされてブロック形式で引用されているような場合である。

以上，書籍の書き言葉の構造を段階的に捉え，それぞれの段階に応じて構成要素ごとに設けているサンプル収録のための基準を示した。

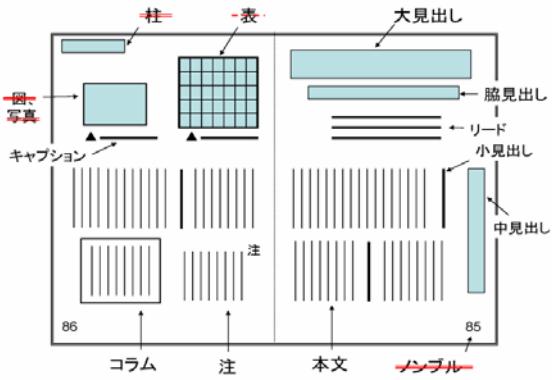

図 1 版面の構成要素の例

4. サンプル作成の手順と問題

3 章に示した基準にのっとり，サンプルとして収録すべき文字列を指定するサンプル作成の実作業は，①対象頁の指定，②範囲の指定，③対象外要素の排除指定，④対象要素の確定と読み順の指定と，段階的に収録対象を絞り込んでいく手順で行っている。以下，そのサンプル作成の作業段階を追って，段階ごとに生じる作業上の問題のありかを明らかにする。

4.1 対象頁の指定

はじめに，サンプル抽出基準点があたった頁が，書籍の実質的本体である「冊本体」に位置する対象頁であるか，それより前の「前付」か，後の「後付」に位置する頁で

あるかを判断する。また，「前付」「後付」に位置する場合は，その頁が文章としての体裁をとっていて，文章量のある類型と見なせる対象頁であるか否かを判断する。

ここで問題になる点は大きく2点ある。まず，「前付」「冊本体」「後付」の分割がし難い場合があることである。「前付」の判断には，はじめに「目次」を用いる。「目次」，及びその前にあるものはすべて「前付」と認定する。よって，「目次」より前に「とびら」があれば，それは「前付」に位置する「標題紙」とみて対象頁にしない。しかし，「目次」と「冊本体」の間に「とびら」があれば，それは冊本体の「中とびら」とみて，対象頁とする。この冊本体の「中とびら」がある場合は，「前付」の境界を，「目次」からこの「中とびら」の直前にずらして考える。つまり，「目次」より後ろであっても，「中とびら」より前に存在するものがあれば，それも「前付」になる。ところが，「目次」がない，「中とびら」がないなど，この認定手順を適用できない体裁をもつ書籍も少なくはなく，個別判断が必要になる。また，「後付」は，「前付」以上に個別判断が必要になる。「奥付」以降は確実に「後付」と判断できるが，それ以外は内容，レイアウト，類型から判断せざるを得ないためである。

2点目の問題は，「前付」「後付」に位置するもののうち，文章量のある類型と認める際の判断のゆれである。原則は，「序文」か「後書き」であれば対象とする，というものであるが，「口絵，凡例，登場人物紹介，献辞，用語解説，文献解題，年譜，取扱説明，ことわりがき，付録」などについては個別の判断が必要になる場合が少なくない。

4.2 範囲の指定

続く作業は，固定長サンプルと可変長サンプルの範囲指定である。固定長サンプルの場合は，入力対象となる部分に留意して1,000字を含む部分を範囲指定すればよいため，ここでは取り上げない。

可変長サンプルは，1万字を超えない範囲で，章や節など，言語的なまとまりを取得しようとするものである。可変長サンプルの範囲を指定するために，第一にサンプル抽出基準点の「著者」を確認し，同一著者が書いた論理的構造の範囲を把握する。単著の冊本体に抽出点があたった場合は，その冊本体全文を「可変長サンプル範囲」とみなすことになる。しかしながら，上限1万字という字数制限があるため，冊本体全文を取得することは少ない。多くの場合，1 万字におさまる，サンプル抽出基準点を含む，章や節の1階層を「完結した構造」として取得する。1万字におさまる適当な章や節がない場合は，章や節相当とみなせる論理的なまとまりを取得する。また，「前付」「後付」に抽出点が定まる場合は，冊本体に及ばずにその位置において完結した構造を取得する。

以上に該当せず，完結した構造が取得できない場合が2つある。1つは，章題などにサンプル抽出基準点があたりその該当章が 1 万字を超えるような場合である。この時は，章題に加え，直近の 1 つ下の階層を取得する。もう1つは，1 万字を超える文章に，論理的な区切りを認めることができず，1 つ下の階層が取得できない場合である。その場合は，当該の章や節の冒頭から1万字を取得する。

範囲の指定における問題の一つは，取得可能な論理的なまとまりを捉える判断が，しばしば困難なことである。

例えば，図 2 は，最終章である 4 章の直後にある「結論」という部分に，サンプル抽出基準点があたった場合である。この「結論」を含む冊本体の全体をとろうとすると，1万字制限を超えてしまう。物理的な位置は4章の中であるが，書籍全体の結論であるため，4 章の下位に含めることもし難い。この例は，レイアウト上，結論部分だけが取得可能な最大の論理的なまとまりであると判断し，可変長範囲は「結論」部分のみとした。このように，内容にも踏み込んだ判断もしばしば必要になる。

図2 物理的に4章の下位に位置づけられている「結論」

さらに，範囲指定作業においては，著者認定，作品集における個別の作品認定，構造把握のための見出し・区切りの認定が必要であり，そこにも問題が多く存在する。

例えば，著者に関しては，共著の場合，分担明記の有無によって，著者と範囲の見方が変わる。対談，座談，インタビューの場合，いずれも原則は共著扱いであるが，インタビュアーの発言が引用的であれば，著者はインタビュイーのみとみる場合がある。料理レシピの場合，料理人が著者の場合もあれば，料理人とは別に編者がいて，そちらを著者とみる場合もある。著者の肩書き，記名のスタイルも様々あり，イニシャル，絵文字などの表示のみから適宜著者認定をしなければならない場合も多い。

作品集については，個人全集の場合，1 冊を範囲とせず，1 作品を範囲とみる。たとえ，複数の作品を束ねる部

立て，章立てがある場合でも，1 作品 1 範囲とみることに注意が必要になる。

見出しや区切りについては，多様な形式への対応が必要になる。見出しには，「一」「二」などの順番を持つもの，イラストで表されるものなど様々あり，区切り記号にも，記号，絵文字，線，イラストなど様々ある。見出しや区切り記号がない場合は，空行を区切りとして認定する必要が生じるが，一行空行，二行空行などで使い分けのある場合は留意せねばならず，また，引用前後の空行は，区切りとみないよう留意せねばならない，といったことがある。

4.3 対象外要素の排除指定

範囲を定めた次に，その範囲内に含まれる，対象外要素の排除指定をする。排除する要素は，①「ノンブル」や「柱」など，実質的でない部分，②文字を主体としない，あるいは，文字が一方向に読めない「フィギュア」，③現代日本語を主体としないブロック形式部分，である。

排除される「フィギュア」とは，以下のようなものである。まず，一切文字のない写真，イラストなどは，典型的な「フィギュア」であり，問題なく対象外と判断できる。その延長で，多少文字が含まれているイラスト，また，文字はあるがイラストが主体である漫画も，排除対象とする。次に，図解，グラフは，主体が図であると考えられるため，そこに文字があっても対象外とする。しかしながら，フローチャート，表になると，主体は文字列になってくる。これらについては，一方向で読むことの可否を用いて「フィギュア」であるかを判断する。つまり，分岐しているフローチャート，表のうち列見出しのある表は，一方向に読むことができないという理由により，「フィギュア」の一種であると判断し，対象外とするのである。

次に，「現代日本語を主体としないブロック形式」を検討する。この類型に該当する条件は，非日本語，非現代語，非言語のいずれかの要素で構成されていること，現代日本語が混在したとしても，非現代日本語が主体とみなせること，インラインではなくブロック形式であること，の 3 つである。明らかに非現代日本語である要素のみで構成され，インデントされたブロック形式であれば，この 3 条件が揃うため，排除対象であることが瞬時に判断可能である。

以上，「フィギュア」と「現代日本語を主体としないブロック形式」について，典型例やそれに準ずる類型を区別する指針までは定まっている。しかしながら，実作業上は，典型例に外れる事例が数多く出現するため，排除要素を判断する作業の負荷は非常に大きいものになっている。均質的なサンプリング作業を実現するために，実例を整理し，フィギュアの類型，非現代日本語の類型，現代日本語が混在する場合に非現代日本語が主体とみなせる場合の類型，ブロック形式の類型など，書き言葉の多様な体裁・論理構造を体系的に把握することに努めている。

4.4 対象要素の確定と読み順の指定

最後の作業段階は，対象外要素を排除した後に残っている部分が，本当にサンプルとして収録すべき対象部分であるかを確定し，電子テキストとして入力される際の読み順を指定することである。その際に留意すべき点は，適切な論理構造および対象要素が，1 次元の文字列として取得できているかをチェックすることである。

論理構造の把握の要として，見出しの収録は必須である。そこで，通常は収録指示対象外となる，非現代日本語であっても収録指示をしなければならない点に注意が必要である。また，「フィギュア」は 4.3 で述べたとおり，排除対象であるが，「フィギュア」に伴う「キャプション」は収録対象である。その判断にも注意を要する。さらに，注なども見落とすことなく，適当な箇所で読むよう指示することにも留意せねばならない。

5. まとめ

本稿では，BCCWJ におけるサンプリング作業における原則および手順を示し，判断に迷う事例などの問題点を示した。現代では，ウェブから膨大な量の電子テキストを簡単に得ることができる。しかしながら，ウェブ上の電子テキストの集合は，現代日本語の総体を代表する標本たり得ない。現代日本語の実態を捉えるための均衡コーパスを実現するためには，母集団の厳密な定義，適切な抽出枠の設計，そして本稿で示したような，書き言葉の多様な構造を体系的に把握し，一定の作業方針・原則にもとづいた斉一的なサンプリング作業が求められる。

[**謝辞**] 本研究は，文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」（平成 18～22 年度，領域代表者：前川喜久雄）による補助を得た。

[**参考文献**]

山崎誠ほか(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 回年次大会予稿集，pp.440―443.

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―』(LR-CCG-06-02).

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―』(LR-CCG-07-01).