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—

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

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、明治期から現代に至る日本語を体系的に収集し、言語コーパスとして整備する計画を進めている [4]。「KOTONOHA 計画」と称されるこのコーパス整備計画では、時代の別、書き言葉・話し言葉の別、発行媒体（メディア）の別など、多角的な視点から「現代日本語」の姿を総合的に把握・記述するための言語資源を構築することを目的とする。これまでに国立国語研究所を中心に構築されてきた『日本語話し言葉コーパス（CSJ）』（2004 年公開）や『太陽コーパス』（2005 年公開）も、KOTONOHA 計画の一部を成すサブコーパスとして位置づけられることになる。

2006 年 4 月からは、『現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese; 以下 BCCWJ と記す）』の構築を開始した。このコーパスは、1976 年から 2005 年までの 30 年間に生産された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り、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、工学への応用など、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、コーパスデザインに関する検討、サンプリング・構造化・アノテーションなどのコーパス構築手法に関する検討、およびその実践を進めている。

さて、実際の書き言葉を広く見渡すと、それらはさまざまな体裁（論理的な構造）を伴って実現されている。BCCWJ に収録するサンプルを実際の書き言葉から収集するためには、あらゆるメディア・ジャンルの書き言葉が持つ体裁を一元的に捉え、そこに現れている言語表現を一次元の構造として把握する必要がある。そこで本稿では、BCCWJ のサンプリング手法に関して、我々がどのような見方に基づいて実際の書き言葉からサンプルを抽出しているか、その方針と基準について述べる。書き言葉の構造を、その実態に即して段階的に捉え、サンプルとしてコーパスに収録する部分を絞り込んでいく手続きを示す。

2　BCCWJ におけるサンプリングの方針

2.1　サンプルの仕様

まず、BCCWJ で採用するサンプルの仕様について示す[1]。我々は、BCCWJ に収録されるサンプルが備えるべき条件として、以下の方針を立てた。

- 統計的に厳密な言語調査に耐え得るよう、母集団からの抽出比率を重視した設計にする。
- 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。

これらの方針に対処するため、「固定長サンプル」「可変長サンプル」という 2 種類のサンプルを設計した。固定長サンプルでは、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、ランダムに抽出した 1 文字を基準として 1,000 文字を取得する。母集団からの抽出比率が明らかである点で、語彙表や漢字表の作成、書き言葉の実態調査などに適している。一方、可変長サンプルでは、やはり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、ランダムに抽出した 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 万字を超えない範囲）を取得する。文章としてのまとまりを重視しているため、テキストの論理構造の把握や文脈の分析などに適している。

実作業では、母集団（を層化した各層）に含まれるすべてのページに対して等確率を与えた上でランダムに 1 ページを抽出し、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、固定長サンプルでは 1,000 文字の範囲を、可変長サンプルではその文字を含む「章」「記事」などの範囲を、それぞれサンプルとして抽出する。理念的に言えば、母集団に含まれる書き言葉をすべて一次元上に配置した上で、ランダムに指定された点を基準点とする一定の範囲を取得していくということである。

BCCWJ 全体は 3 つのサブコーパスから構成されるが、このうち固定長サンプルを格納するサブコーパスについては、母集団のサイズが文字の量によって把握

[1] BCCWJ の設計全般については、山崎他 (2007) を参照。

されている。さらに、一定の手続きにより定義された母集団に含まれる総文字数を推計した上で、それらの比によってコーパスの構成比率を決定するという設計になっている[2]。この設計方針によって、特に固定長サンプルは、母集団から一定の抽出比率に基づいてサンプルが抽出されることになる。これは、母集団の量的構造をコーパスに適切に反映させるための設計方針であり、代表性を備えたバランストコーパスである BCCWJ の大きな特徴となっている。

以上から、BCCWJ におけるサンプル抽出とは、「等確率を与えられた文字の集合（文字列）からランダムに選ばれた 1 文字を基準点として、そこから一定範囲（1,000 文字、または言語的な構造のまとまり）を抽出する作業」と捉えることができる。

2.2 書き言葉の一元的な把握

以上で述べたような方針に基づいて実際にサンプリングを行う場合、まず考えなければならないのは、書き言葉の構造をどのように一元的に見なすか、という問題である。言い換えれば、多様な構成要素を含む書き言葉の版面（印刷紙面）から、一次元上の文字列をどのように抽出するか、という問題である。

書き言葉は、それが実現されている文書中において、「本文」「見出し」「注」「ルビ」「目次」など、さまざまな論理構造に関与している。書き言葉の版面からサンプルを抽出するためには、版面を構成する要素のうち、どの要素をサンプルとして抽出し、どの要素を抽出しないのかをあらかじめ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、個別の事例について対処していく必要がある。

また、版面上に現れる「本文」「見出し」「注」などの要素をどのように区別するかということも問題となる。これらの要素の区別は一見自明的であるように思われるが、しかしながら、論理的な構造に関する情報が文書中に明示的に表示されているわけではない。むしろ、版面上のある言語表現が「見出し」であり、別の言語表現が「本文」であることは、意識的であれ無意識的であれ、読み手が能動的に読み取っている情報である。

BCCWJ では、サンプルは最終的に XML 形式で電子化され、「本文」「見出し」「キャプション」などの要素に対して文書構造情報を表すタグが付与されることになる [3]。しかしながら、サンプリングの段階においては、あらゆる実現様式・あらゆる論理構造に関わる書き言葉を一元化し、順序立てて、一次元構造の文字列として把握する必要がある。

[2] 母集団に含まれる総文字数の推計方法とその結果については、秋元他 (2006) を参照。

そこで以下では、書き言葉の版面からどの部分をサンプリングの対象と認定するかについて、我々が採用している基準を示す。

3 書き言葉の構造とサンプリング対象

以下では、書籍を例として書き言葉の物理的・論理的な構造について示し、書籍に含まれる書き言葉のどの部分がサンプリングの対象となるかについて詳述する。書籍の構造を、「形態」「版面」「本文」「文字」という 4 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される部分と、それが満たすべき基準について示す。

3.1 書籍の「形態」を構成する要素

初めに、書籍の「形態」という側面について示す。1 冊の書籍の形態は、おおむね、図 1 のように分類できる。

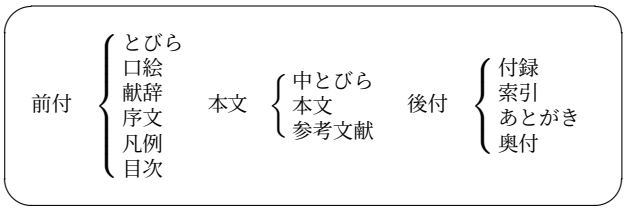

図 1: 書籍の形態に関する分類

このうち、主として文章表現によって実現されるのは、「序文」「本文」「あとがき」である。そこで、これらのカテゴリに相当する部分はサンプリングの対象とする。「中とびら」は章立てを表す要素の一つと考え、やはりサンプリングの対象とする。「とびら」「凡例」「目次」「参考文献」「索引」「奥付」には現代日本語が現れるものの、箇条書き・リストであったり図的な扱いであったりするために、書き言葉コーパスに収録する対象としてはふさわしくない。そこで、これらのカテゴリに相当する部分はサンプリングの対象外とする。「口絵」は言語表現ではないため、対象外とする。「献辞」「付録」は、文章表現によって構成される場合とそうでない場合があるため、収録対象とするか否かは個別に判断する。

3.2 書籍の「版面」を構成する要素

次に、書籍の「版面」という側面について示す。書籍の版面は、おおむね、図 2 のような構成要素から成り立っている。

このうち、主として文章表現によって実現されるのは、「大見出し」「脇見出し」「リード」「中見出し」「小見出し」「本文」「コラム」である。そこで、これらのカテゴリに相当する部分はサンプリングの対象とする。「キャプション」「注」は文章表現によって実現される場合とそうでない場合（キャプションが「19,800 円」の

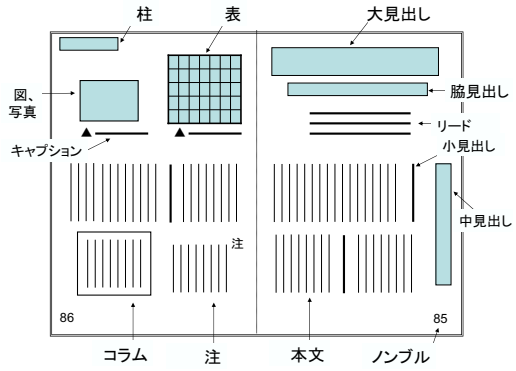

図 2: 書籍の版面に関する分類

みである場合、注が「丸山 (2006) 参照。」のみである場合など）があるが、これらは一括してサンプリングの対象とする。「図、写真」は言語表現ではないため、サンプリングの対象から外す。仮に図・写真の中に言語表現が含まれていても、それが図・写真の範囲内にあるものであれば、一括してサンプリングの対象から外す。「柱」「ノンブル」は書籍のメタ的な構造に関わる部分であるため、サンプリングの対象から外す。「表」は、基本的にはサンプリングの対象外とするが、その内部に文章表現を含み、そのページ全体が表によって成立しているような場合は、サンプリングの対象とする。

3.3 書籍の「本文」を構成する要素

次に、書籍の版面を構成する要素のうち、「本文」部分そのものの構成について示す。本文部分は、おおむね、図 3 のような構成要素から成り立っている。

- 本文
- 引用文
- ブロック引用
- リスト
- ルビ、グロス
- 注番号、添え字

図 3: 本文の構成に関する分類

ここで言う「本文」は、いわゆる「地の文」を指す。「引用文」は、主として発言を引用する部分を指す。前後に改行を伴い、通常は括弧（「...」）で囲まれる部分である。発言ではない言語表現を引用する場合も同様に扱う。括弧で囲まれて引用されていても、それが地の文に含まれる格好であれば、引用文とは見なさない。「ブロック引用」は、前後に改行を伴い、本文部分からインデントされる形で引用されている部分を指す。「リスト」は、箇条書きの形になっている部分を指す。これらの要素は、基本的にすべてサンプリングの対象とする。

3.4 書籍の「文字」を構成する要素

最後に、文字の側面について述べる。サンプリングの対象となった部分に含まれる文字は、JIS X 0213:2004 に依拠してすべて電子化されることになる[3]。ただし、固定長サンプルとして 1,000 文字を取得する際、文字種の違いによって 1,000 文字としてカウントする対象にするか否かを区別している。これは、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、研究用途上の要請によるものである。

(1) に固定長サンプル 1,000 文字のカウント対象とする文字種、(2) にカウント対象としない文字種を示す。

(1) a. 仮名文字（平仮名・片仮名・変体仮名）
b. 漢字
c. 準仮名・漢字（「ー」「々」「ゝ」等）
d. 数字（アラビア数字・ローマ数字）
e. アルファベット（ローマ字・ギリシャ文字）

(2) a. 句読点類（「、」「。」「,」「.」「…」「・」「:」「;」等）
b. 疑問符、感嘆符（「？」「！」等）
c. 括弧類（「(」「「」「{」「＜」「《」「【」「〔」等）
d. 線記号類（「‐」「—」「〜」等）
e. 矢印類（「→」「↓」「⇒」「⇔」等）
f. 算術記号類（「＋」「−」「×」「÷」「＝」「±」等）
g. 通貨・単位記号類（「£」「$」「¥」「%」「‰」等）
h. 音符類（「♪」等）
i. 絵文字（携帯電話の絵文字など）
j. その他記号類（「○」「▲」「※」「#」「&」「☆」等）

3.5 BCCWJ 収録サンプルとしての条件

上記までの諸基準に加えて、本文部分に含まれる言語表現そのものに関する条件が設けてある。それは、BCCWJ が「現代日本語書き言葉」のコーパスである以上、サンプルとして収録する言語表現は現代日本語として書かれたものでなければならないという条件である。したがって、以下のような表現が出現した場合、その部分はサンプリングの対象から外す。

(3) a. 非日本語（英語、フランス語、中国語等）
b. 非現代日本語（明治元年より前に書かれた日本語）
c. 非言語（数式、化学式等）

ただし、本文（地の文）に現れる非日本語・非現代日本語までをサンプリングの対象外とすると、言わば「穴の開いた」サンプルが抽出されることになり、研究用途上、好ましくない。そこで、これらの表現が地の文に現れている場合はサンプリングの対象として認めることにする。(3) の各表現がサンプリングの対象外となるのは、典型的には、ブロック引用の形で現れた場合である。

[3] 再現できない漢字や記号などは、タグによって記述される [3]。

以上、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基準を立てることによって、サンプリングの対象を絞り込んでいく手続きを示した。さまざまな様式・体裁を持つ書き言葉からサンプルを抽出するためには、以上で示したような、書き言葉の実態に即した分析的な手順が必要となる。

4 「原サンプル」の作成

最後に、我々が実際どのような形でサンプルを抽出・作成しているかについて触れておく。

サンプリングを担当する作業者は、実際の書籍を手に取り、あらかじめ定められた手続き[4]と本稿で示した基準に従って、固定長サンプル・可変長サンプルを抽出する。その結果は、書籍の該当部分をコピーした紙面の上に転記される。この紙媒体を、「原サンプル」と呼ぶ。原サンプルの一部を、図 4 に示す。

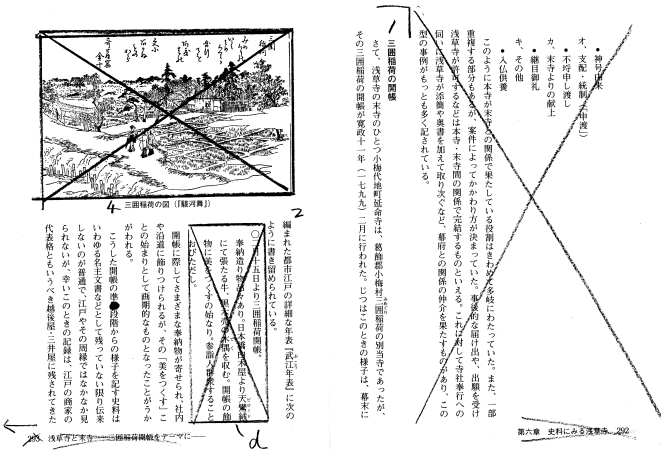

図 4: 原サンプルの例（一部）

図 4 は見開き 2 ページ分の原サンプルの一部である。このうち、左ページに転記された「●」の記号は、「サンプル抽出基準点」としてランダムに選ばれた 1 文字の位置を示す[5]。作業者は、サンプル抽出基準点から、3.4 節で示した基準に即して 1,000 文字目までを数え、固定長サンプルの範囲を確定する。また、サンプル抽出基準点を含む言語的まとまりのうち、10,000 文字を超えない最大の範囲を求め、可変長サンプルの範囲を確定する。図 4 では、見開き右ページの左側に、可変長サンプルの開始点を示す記号「⌉」が付与されている。

これらの範囲指定に加えて、サンプルを一次元の文字列として把握するために、どのような順序で読み進めていくかを指示する連番が付与される。図 4 では図のキャプションに「4」という数字が付与されているが、これは次ページ以降にある「3」の後にキャプション部分を読み込むことを指示するものである。さらに、柱やノンブル、ブロック引用として引用されている非現代日本語の部分は、サンプリングの対象外となるので、「×」の記号が付けられる。可変長サンプルに含まれない本文部分（右ページの大半）もまた、「×」記号によって削除される。

このようにして作成された原サンプルは、抽出された固定長・可変長サンプルを電子化する際に用いられる。

5 まとめ

本稿では、BCCWJ におけるサンプリング手法として、我々が実際の書き言葉からどのようにしてサンプルを抽出しているか、その方針について述べた。書き言葉が持つさまざまな体裁（論理的な構造）を一元的に把握し、サンプルとなる部分を絞り込んでいくための基準を示した。書き言葉の構造を段階的に分解し、各段階に関わる構成要素に対して基準を設けることにより、書き言葉の版面から一定の手続きでサンプルを抽出することができる。

現代では、膨大な量の電子テキストを簡単に入手できるという背景もあり、書き言葉コーパスを作るという作業は比較的単純であるかのように見える。しかしながら、書き言葉の現実的なありさまを精密に捉え、その実態を記述していこうとする限り、既存の電子テキストを集積しただけでは「書き言葉コーパス」たり得ない。そのような集積は、現代日本語の書き言葉の姿を反映するものではないからである。書き言葉の実態を捉えるためのコーパスを構築するためには、綿密な調査に基づく母集団の定義とコーパスデザイン、そしてきめの細かいサンプリングの実践が求められる。

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

[1] 秋元祐哉他 (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—. 本論文集所収.

[2] 丸山岳彦他 (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』. 言語処理学会.

[3] 間淵洋子他 (2006) 代表性を有する書き言葉コーパスの電子化フォーマットについて. 『言語処理学会 第 12 回大会 発表論文集』. 言語処理学会.

[4] 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』. 国立国語研究所.

[4] サンプリングの手順の詳細については、丸山他 (2006) を参照。

[5] 図 4 では塗りつぶされているが、実際には赤く塗られている。