『現代日本語書き言葉均衡コーパス』におけるサンプリングの原理と運用

国立国語研究所内部報告書（LR-CCG-10-01）

『現代日本語書き言葉均衡コーパス』におけるサンプリングの原理と運用

丸山 岳彦・山崎 誠・柏野 和佳子・佐野 大樹・秋元 祐哉・
稲益 佐知子・田中 弥生・大矢内 夢子

平成23年2月

国立国語研究所内部報告書（LR-CCG-10-01）

『現代日本語書き言葉均衡コーパス』における サンプリングの原理と運用

丸山 岳彦
山崎 誠
柏野 和佳子
佐野 大樹
秋元 祐哉
稲益 佐知子
田中 弥生
大矢内 夢子

平成23年2月

大規模汎用日本語データベースの構築とその活用に関する調査研究

目 次

はじめに

2006年度に『現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese; 以下 BCCWJ）』の構築が開始されてから，5年が経過した。コーパス本体の構築を担う「データ班」では,「サンプリング」「著作権処理」「電子化」「形態論情報」という4つのサブグループに分かれて，BCCWJの構築を分担して進めてきた。サンプリングを担当した我々のグループ（SSG; サンプリングサブグループ）では，これまでに5冊の報告書を刊行し，サンプリングの設計から実作業の手順まで，一連の流れを示してきた。

本報告書は，2008年度に発行した報告書の内容を受ける形で，我々が実施してきたサンプリング作業の考え方を示すものである。BCCWJの中でも中核的な部分を成す書籍のサンプルを中心的に取り上げ，書き言葉をどのように把握し，そこからどのような基準と手順でサンプルを抽出してきたのか，その原理について述べる。これまでに報告してきた内容と一部重複する部分もあるが，これまでに述べることができなかった雑誌や新聞のサンプリングも含めて，まとめて報告することにする。

第I部ではBCCWJに含まれるサンプリングの設計について示す。第II部では書籍を対象として，サンプリングの原理と運用について示す。第III部では，雑誌・新聞を対象とした場合にどのような問題が生じるかについて示す。

謝辞

BCCWJのサンプリング作業を実施するにあたり，以下の各機関・各社より多大なご協力をいただきました。記して感謝申し上げます。

大阪市立中央図書館，オリオン書房，学習研究社，国立国会図書館，
埼玉県立浦和図書館，埼玉県立久喜図書館，埼玉県立熊谷図書館，
自治大学校図書室，　小学館，　湘北短期大学図書館，　高原書店，
立川市図書館，　東京都立多摩図書館，　東京都立中央図書館，
東京都立日比谷図書館，　日本図書館協会，　八王子市図書館，
一橋大学附属図書館，　ヤフー株式会社，　横浜市中央図書館

（五十音順）

第I部

BCCWJにおけるサンプリングの設計

第1章　BCCWJの基本理念と構成

本章の概要：　本章では，BCCWJ の構築にあたって我々が実施したサンプリングの基本理念および方針を述べる。以下，BCCWJ を構築する上での基本方針，ならびに BCCWJ の内部構成について確認した後，BCCWJ を構成する各サブコーパス・各メディアについて，母集団の定義や層別の方法，構成比率の算出方法とその結果などについて示す。

1.1　BCCWJ 構築の基本理念

BCCWJ の構築計画が開始されたのは，2006 年度であった。当時，山崎ほか (2006) では，BCCWJ 構築計画の基本理念が，次の 4 点にまとめられていた。

(1) **現代日本語の縮図となるコーパス**
これまで研究所が行ってきた語彙調査の手法を生かし，コーパスがその母集団の統計的な縮図になるよう設計する。それにより，母集団における言語的諸特性の分布が縮図において過不足なく再現でき，母集団における分布を高い精度で推測できるようになる。

(2) **汎用的な目的に供するコーパス**
言語研究（語彙・文法・文字）以外にも，応用面として，辞書編集や言語政策，日本語教育などでも使えることを意図し，多様な日本語の姿を捉えることができるよう設計する。また，言語変化に対応するためには，同じ設計のコーパスを繰り返し構築するなど定点観測的な工夫も必要である。

(3) **公開可能なコーパス**
収録する著作物の利用許諾を得て，公開を目指す。インターネット上からの簡易検索のほか，共起条件を指定できる検索ツールなどもあわせて提供する。

(4) **既存のコーパスとの調和**
解析単位の仕様を『CSJ』に合わせ，短単位，長単位の 2 種類の解析を行う。

これらの基本理念のうち，(1) と (2) はサンプリングに関わる理念である。また，(3) は著作権処理，(4) は形態論情報の付与に関わる理念である。(1) については，メディアごとに母集団を厳密に定義して，層別ランダムサンプリングを実施することにより実現した。(2) については，サンプリングの際，固定長サンプル・可変長サンプルという 2 種類のサンプルを取得することにより，統計的な研究から文章研究までに対応できるサンプル抽出を実現した。

1.2　BCCWJ を構成する 3 つのサブコーパス

次に，BCCWJ の内部構成について確認しておく。BCCWJ の内部構成を，図 1.1 に示す。

図 1.1: BCCWJ の内部構成

各サブコーパス（以下，SC）の概要を，以下に述べる。

1.2.1　出版 SC

出版 SC は，書き言葉の出版・生産という側面に着目する SC である。2001 年から 2005 年の間に国内で出版されたすべての**書籍・雑誌・新聞**に含まれる文字の総体を母集団として，ランダムサンプリングによって得られる約 3,500 万語分のデータを収める。書き言葉が実際に出版された結果を，文字数という量的側面からできる限り忠実に反映することで，5 年間における書き言葉の出版に関するありさまを捉えることを目的とする。

1.2.2　図書館 SC

図書館 SC は，書き言葉の流通・流布の実態という側面に着目する SC である。東京都内の公立図書館に所蔵されている**書籍**（ただし 1986 年から 2005 年の 20 年間に発行されたもの）を対象として，ランダムサンプリングによって得られる約 3,000 万語分のデータを収める。書き言葉（書籍）が世の中に流通している状態を公立図書館の所蔵状況によって近似的に把握し，世の中に広く行き渡っている書き言葉のありさまを捉えることを目的とする。

1.2.3 特定目的 SC

特定目的 SC は，生産・流通という側面からは捉えきれない，あるいは，出版 SC・図書館 SC の母集団には入らないけれども，書き言葉の研究を遂行する上で必要と思われる種類の書き言葉を収める SC である。**白書**，**教科書**，**広報紙**，**ベストセラー**，**Yahoo!知恵袋**，**Yahoo!ブログ**，**韻文**，**法律**，**国会会議録**を対象として，約 3,500 万語分のデータを収める。収録対象期間はメディアによって異なる。

1.3 BCCWJ を構成する 2 種類のサンプル

上記に挙げた 3 つの SC は,「**固定長サンプル**」「**可変長サンプル**」という 2 種類のサンプルによって構成する。

- **固定長サンプルの設計方針**：
  統計的に厳密な言語調査に耐え得る設計にする。
- **可変長サンプルの設計方針**：
  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。

1.3.1 固定長サンプル

「固定長サンプル」は，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，その文字を始点として 1,000 文字目までの範囲を抽出するサンプルである。全ての文字に対して等確率を与えるために，母集団に含まれる文字の総数をあらかじめ推計しておく必要がある。母集団（＝推計された総文字数）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。また，母集団の層的かつ量的な構造を忠実に反映する点で，統計的な代表性を備えた均衡コーパスとしての性格を強く持つ。

1.3.2 可変長サンプル

「可変長サンプル」は，固定長サンプルと同様，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 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 万字を上限とする）を抽出するサンプルである。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。

なお，可変長サンプルは，3つのSCの全てに対して提供される。一方，固定長サンプルは，統計的な言語調査を行なう可能性の高いSC，すなわち，出版SC，図書館SC，および，特定目的SCの一部（白書）に対して提供される。

第2章　出版SC・図書館SCのサンプリングの設計

本章の概要：　本章では，BCCWJ で実施したサンプリングの基本方針を述べる。BCCWJ における内部構成のうち，標本調査という性格を特に強く持つのは，出版 SC・図書館 SC の 2 つである。これらについては，母集団の数量的な定義，抽出枠・抽出方法の決定，母集団のリスト化，サンプリングの基準と手順などが，コーパスデザインの段階で厳密に設計されている。そこで以下では，出版 SC・図書館 SC におけるサンプリングの設計について示す。

2.1　基本方針

まず，出版 SC・図書館 SC におけるサンプリングの基本方針を述べる。出版 SC・図書館 SC において実施したサンプリングは，基本的に，図 2.1 に示す方針に基づく。

調査目的：文字・表記研究，語彙研究，文法研究，語義記述，変異研究，辞書編纂，教材開発，言語処理，言語政策など，種々の調査・研究の目的に供する。

調査対象：現代日本語の書き言葉を対象とする。特に，出版 SC では 2001 年から 2005 年に出版された書籍・雑誌・新聞を，図書館 SC では 1986 年から 2005 年に出版された書籍を，それぞれ対象とする。

母集団：文字数によって母集団を定義する。

抽出枠：母集団をメディア・ジャンル・発行年によって層別する。各層に含まれる文字数の比を各層から抽出する標本サイズに比例割当する。

抽出方法：無作為抽出法とする。

抽出単位，標本サイズ，標本数：「固定長サンプル」「可変長サンプル」の 2 種類を抽出単位とする。出版 SC において 1,000 万語分の固定長サンプルを抽出することを基準として，全体の構成比を算出する。

抽出対象：現代日本語で書かれた表現を抽出対象とする。

図 2.1: 出版 SC・図書館 SC におけるサンプリングの基本方針

以下，各項目について詳しく述べる。

2.2　調査目的

BCCWJ は，文字・表記研究，語彙研究，文法研究，語義記述，変異研究，辞書編纂，教材開発，言語処理，言語政策など，多様な研究目的に利用される汎用コーパスとして構築されることが想定されている。すなわち，単独の言語調査のために構築されるものではなく，汎用的な目的に供されるためのコーパスであるということである。

国民が政権を支持するかどうかを問う世論調査を考えた場合，そこで抽出される標本は，ある時点における政権の支持率を調査するという目的のためだけに利用されるものである。これに対して，大規模な言語コーパスは，通常，特定の調査目的のためだけに構築されるというものではない。むしろ，比較的長期間にわたって，言語研究のさまざまな用途に利用されることがあらかじめ想定されていると言ってよい。

このうち出版 SC は，出版に関わる書き言葉の主要なメディアである書籍・雑誌・新聞について，2001 年から 2005 年までに出版された総体を母集団としてサンプリングを実施することにより，出版の実態を反映する資料を提供するものである。出版という行為を書き言葉の生産力と結び付けて考えることにより，短期間のうちにどのような書き言葉が生産され，世の中に発信されるのか，そのありさまを捉えることを目的とする。

また，図書館 SC は，公立図書館での蔵書を母集団としてサンプリングを実施することにより，流通・流布の実態を反映する資料を提供するものである。書き言葉（書籍）が世の中に流通している状態を公立図書館の所蔵状況によって近似的に把握し，世の中に広く行き渡っている書き言葉のありさまを捉えることを目的とする。

2.3　調査対象

BCCWJ に収録する対象は,「現代日本語の書き言葉」である。「現代日本語」の範囲や定義についてはさまざまな考え方があり得るが，我々は「**明治初年（1868 年）以降に書かれた日本語**」を現代日本語と定義した。具体的には，出版 SC では，比較的短期間に出版された書き言葉の実態を知るという目的から，2001 年から 2005 年までに出版された「書籍」「雑誌」「新聞」という 3 種類のメディアを調査対象とすることにした。また，図書館 SC では，比較的長期間にわたって図書館に収蔵されている書籍を対象とするという目的から，1986 年から 2005 年までに出版された「書籍」を調査対象とした。

これらが「現代日本語の書き言葉」として十全な調査対象であるとは必ずしも言い切れないが，現代日本語の書き言葉を構成する主たるメディア（媒体）であるという点から，また，すぐ後に述べる母集団を数量的に定義する可能性という点から，これらのメディアを調査対象として定めた。

2.4 母集団

書籍・雑誌・新聞の母集団は，文字数により定義した。これは，書き言葉を構成する最も基本的な要素は文字であるという見方に基づく。さらに，文字量によって母集団の量的な構造を定義し，その構成比を用いた層化抽出によって，母集団から精密なサンプルを抽出するという方針に立つものである。各メディアの文字数は，所定の期間に発行された書籍・雑誌・新聞に含まれる文字数を推計するための調査「現代日本語書き言葉の文字数調査」を実施し，その結果によって定めた（調査の詳細は，丸山・秋元 (2007,2008) を参照）。

2.4.1 書籍（出版 SC）の母集団

2001 年から 2005 年の間に国内で出版されたすべての書籍に含まれる文字の総体。ただし，漫画・写真集・楽譜・地図のように言語表現が主体でないもの，1 冊が 40 ページ以下の書籍などを除く。「現代日本語書き言葉の文字数調査」の結果，48,539,925,351 文字と推計された。

2.4.2 雑誌（出版 SC）の母集団

2001 年から 2005 年の間に発行された『雑誌新聞総かたろぐ』（メディア・リサーチ・センター発行）に記載のある雑誌タイトルのうち，当該の 5 年間に社団法人日本雑誌協会に加盟していた出版社が発行していたすべての雑誌タイトルに含まれる文字の総体。ただし，新聞，要覧，漫画，非日本語による定期刊行物などを除く。「現代日本語書き言葉の文字数調査」の結果，10,515,681,634 文字と推計された。

2.4.3 新聞（出版 SC）の母集団

2001 年から 2005 年の間に発行された，社団法人日本新聞協会発行『全国新聞ガイド』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，および日本各地の有力な地方紙に含まれる文字の総体。「現代日本語書き言葉の文字数調査」の結果，6,416,070,114 文字と推計された。

2.4.4 書籍（図書館 SC）の母集団

1986 年から 2005 年の間に国内で出版されたすべての書籍のうち，2007 年の時点で東京都内の公立図書館で共通に所蔵されていたすべての書籍に含まれる文字の総体。ただし，漫画・写真集・楽譜・地図のように言語表現が主体でないもの，1 冊が 40 ページ以下の書籍などを除く。出版 SC の書籍に含まれる総文字数とほぼ等しくなるように調整した結果，都内 13 自

治体以上の公立図書館で共通に所蔵されていた書籍に含まれる総文字数は，47,877,656,072 文字と推計された。

2.5　抽出枠

書き言葉のメディアとして，書籍・雑誌・新聞という別を設けたが，これらをさらに，以下の基準によって層別することにした。

- 抽出枠 (1)「ジャンル・発行形態」
- 抽出枠 (2)「発行年」

2.5.1　書籍の抽出枠

書籍は,「日本十進分類法（NDC)」および「発行年」という基準によって，母集団を層別した。NDC については，表 2.1 に示すように，国立国会図書館が書籍のタイトルごとに付与した NDC の 1 桁目による 10 分類，および NDC が付与されていない場合（「記録なし」）の，合計 11 種類に層別した。発行年については，出版 SC では，2001 年から 2005 年までの 5 年間によって 5 層に，図書館 SC では，1986 年から 2005 年までの 20 年間によって 20 層に，それぞれ層別した。

表 2.1: NDC による書籍の 11 分類

0. 総記	2. 歴史	4. 自然科学	6. 産業	8. 言語	n. 記録なし
1. 哲学	3. 社会科学	5. 技術工学	7. 芸術	9. 文学	

2.5.2　雑誌の抽出枠

雑誌は,「分野」および「発行年」という基準によって母集団を層別した。分野については，表 2.2 に示すように,『雑誌新聞総かたろぐ』（メディア・リサーチ・センター発行）において分類されている「分野」の情報により，6 種類に分類した。発行年については，2001 年から 2005 年までの 5 年間によって 5 層に層別した。

表 2.2:『雑誌新聞総かたろぐ』による雑誌の 6 分類

1. 総合	3. 政治・経済・商業	5. 工業
2. 教育・学芸	4. 産業	6. 厚生・医療

2.5.3 新聞の抽出枠

新聞は「紙種および新聞タイトル」および「発行年」という基準によって母集団を層別した。紙種については，表 2.3 に示すように「全国紙・ブロック紙・地方紙」の別，およびその下位に位置づけられる 16 種の新聞のタイトルによって層別した。発行年については，2001 年から 2005 年までの 5 年間によって 5 層に層別した。

表 2.3: 新聞の分類

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

上記の結果，総文字数によって定義された母集団は，表 2.4 のように層別された（新聞の抽出枠 (1) は，新聞タイトルによれば 16 分類となる）。

表 2.4: 母集団の層別

メディア・SC	抽出枠 (1)	抽出枠 (2)	合計層数
書籍（出版 SC）	11 分類	5 分類	55 層
雑誌（出版 SC）	6 分類	5 分類	30 層
新聞（出版 SC）	3 分類	5 分類	15 層
書籍（図書館 SC）	11 分類	20 分類	220 層

抽出枠 (1) による分類と総文字数の分布を，出版 SC・図書館 SC の別に，表 2.5，2.6 に示す。

2.6 抽出方法

母集団からの標本抽出の方法は，層別無作為抽出法によることとした。すなわち，母集団を層ごとにリスト化し，各リストを構成する抽出単位の全てに通し番号を付してランダマイズし，その結果の並びを優先順位と見なして，順に抽出単位を取得していくことにした。

ここで，母集団を抽出単位（個々のサンプル）ごとにリスト化する必要があるが，文字によって定義されている母集団をどのようにリスト化してランダマイズするか，という技術的な問題がある。母集団に含まれる文字をすべてリスト化してランダマイズすることは，原理的には可能であるが，現実的には不可能である。そこで，何らかの方法により，これに近似する結果を得なくてはならない。

これを実現するための手段として，次のような方法を採用した。まず，母集団に含まれる全てのページを各層ごとにリスト化し，それらをランダマイズして優先順位を付した。さらに，

表 2.5: 推計総文字数の分布（出版 SC）

層		総文字数	構成比
書籍	0. 総記	1,636,414,548	2.50%
	1. 哲学	2,597,610,813	3.97%
	2. 歴史	4,301,204,340	6.57%
	3. 社会科学	12,408,321,943	18.95%
	4. 自然科学	5,069,594,034	7.74%
	5. 技術工学	4,615,929,967	7.05%
	6. 産業	2,196,387,437	3.35%
	7. 芸術	3,258,432,447	4.98%
	8. 言語	888,800,128	1.36%
	9. 文学	9,341,275,486	14.27%
	n. 記録なし	2,225,954,208	3.40%
書籍 小計		48,539,925,351	74.14%
雑誌	1. 総合	7,421,447,806	11.34%
	2. 教育・学芸	877,875,592	1.34%
	3. 政治・経済・商業	456,459,405	0.70%
	4. 産業	110,640,958	0.17%
	5. 工業	1,468,293,360	2.24%
	6. 厚生・医療	180,964,513	0.28%
雑誌 小計		10,515,681,634	16.07%
新聞	全国紙	2,417,622,461	3.69%
	ブロック紙	1,296,592,154	1.98%
	地方紙	2,701,855,499	4.13%
新聞 小計		6,416,070,114	9.80%
合計		65,471,677,100	100%

表 2.6: 推計総文字数の分布（図書館 SC）

層	総文字数	構成比
0. 総記	1,003,528,880	2.01%
1. 哲学	2,343,849,711	4.90%
2. 歴史	5,010,749,621	10.47%
3. 社会科学	8,946,058,392	18.69%
4. 自然科学	3,028,276,363	6.33%
5. 技術工学	3,149,144,051	6.58%
6. 産業	1,690,150,481	3.53%
7. 芸術	4,057,291,256	8.47%
8. 言語	956,625,910	2.00%
9. 文学	15,485,091,056	32.34%
n. 記録なし	2,206,890,351	4.61%
合計	47,877,656,072	100%

ランダムに選ばれたページの中に印刷されている文字の中から 1 文字をランダムに指定し，この 1 文字を，抽出単位を取り出すための基準点（「**サンプル抽出基準点**」）として利用することにした。このような 2 段階の抽出（ページの無作為抽出，文字の無作為抽出）によって，母集団に含まれる全ての文字をリスト化し，そこからランダムに 1 文字を抽出することに近似させることにした（母集団のリスト化とサンプルの抽出手順の詳細は，丸山・秋元 (2008) の第 3 章 2 節を参照）。

2.7 抽出単位，標本サイズ，標本数

抽出単位は，先に述べた「固定長サンプル」「可変長サンプル」の 2 種類とした。母集団の中からランダムに指定された 1 文字を「サンプル抽出基準点」として，そこから固定長サンプルと可変長サンプルを同時に取得することにした。固定長サンプルは，サンプル抽出基準点として指定された文字から数え始めて 1,000 文字目までの範囲を抽出するものである[1]。可変長サンプルは，サンプル抽出基準点を含む言語的まとまり（章，節など）のうち，1 万字を上限とする最大の範囲を見定め，その範囲を抽出するものである。

なお，1,000 字・1 万字という文字の数え方は，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**」のみによってカウントすることとした。「句読点・疑問符・感嘆符」「括弧・その他記号」などは，サンプルの範囲に含まれる要素として収録はするけれども，固定長サンプル 1 字，可変長サンプルの上限 1 万字として数える対象とはしないことにした。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標本を抽出することにより，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ものである（カウント対象となる文字の詳細については，第 5 章を参照）。

また，サンプル抽出基準点の位置によっては，すでに取得した部分がもう一度取得されてしまう可能性がある。すでに取得済みのページの直前のページにサンプル抽出基準点が当たった場合などが，これに該当する。特に統計的な研究に用いる固定長サンプルの場合，取得するサンプルに重複が含まれていることは設計上望ましくない。そこで，このようなサンプルの重複は一切認めず，仮に同じ部分が取得されそうになった場合は，そのサンプル抽出基準点を破棄することとした。

全体の標本サイズ（コーパスサイズ）は，出版 SC における固定長サンプルの合計を 1,000 万語とすることを前提として，そこから全体を算出することにした。1,000 万語という数値は，文字調査や語彙調査などの統計的な言語調査に十分耐え得るサイズとして経験的に判断したものである。さらに，1,000 字の固定長サンプルを 1,000 万語分収集するために，1 語を平均 1.7 文字で構成されるものと試算して，抽出すべきサンプル数を 17,000 サンプルと算出した。

[1] 実際には，サンプル抽出基準点が含まれる文の文頭，およびサンプル抽出基準点から数えて 1,000 文字目が含まれる文の文末までが合わせて抽出される。

各層から抽出するサンプル数は，各層を構成する総文字数を用いた比例割当によって算出した。これにより，出版 SC として抽出する 17,000 サンプルの内訳が算出できる。すなわち，多くの文字数が含まれている層からはより多くのサンプルが，少ない文字数しか含まれていない層からは少ないサンプルが，それぞれ抽出されることになる。

さらに，図書館 SC から抽出するサンプル数は，出版 SC における書籍のサンプル数と一致させることにした。これにより，ほぼ等しいサイズの母集団から，同一の抽出比によって，同じサイズの標本が抽出できることになる。このような設計により，出版された書籍の実態を代表する部分と，図書館に所蔵されている書籍の実態を代表する部分とを比較し，両者の特徴の違いを厳密に検討できるようにした。

出版 SC と図書館 SC から抽出されるサンプル数を，表 2.7，2.8 に示す。

表 2.7: サンプル構成比（出版 SC）

層		構成比	サンプル数
書籍	0. 総記	2.50%	425
	1. 哲学	3.97%	674
	2. 歴史	6.57%	1,117
	3. 社会科学	18.95%	3,222
	4. 自然科学	7.74%	1,316
	5. 技術工学	7.05%	1,199
	6. 産業	3.35%	570
	7. 芸術	4.98%	846
	8. 言語	1.36%	231
	9. 文学	14.27%	2,426
	n. 記録なし	3.40%	578
書籍 小計		74.14%	12,604
雑誌	1. 総合	11.34%	1,927
	2. 教育・学芸	1.34%	228
	3. 政治・経済・商業	0.70%	119
	4. 産業	0.17%	29
	5. 工業	2.24%	381
	6. 厚生・医療	0.28%	47
雑誌 小計		16.06%	2,730
新聞	全国紙	3.69%	628
	ブロック紙	1.98%	337
	地方紙	4.13%	702
新聞 小計		9.80%	1,666
合計		100%	17,000

表 2.8: サンプル構成比（図書館 SC）

層	構成比	サンプル数
0. 総記	2.01%	264
1. 哲学	4.90%	617
2. 歴史	10.47%	1,319
3. 社会科学	18.69%	2,355
4. 自然科学	6.33%	797
5. 技術工学	6.58%	829
6. 産業	3.53%	445
7. 芸術	8.47%	1,068
8. 言語	2.00%	252
9. 文学	32.34%	4,077
n. 記録なし	4.61%	581
合計	100%	12,604

2.8 抽出対象

抽出対象としてサンプルに含めるのは，原則として，「**現代日本語で書かれた表現**」とした。実際の印刷紙面上にある現代日本語の表現を，一定の基準と手順で抽出していくことにより，サンプルを抽出することにした。

一見，目の前に書かれている現代日本語の表現を取り出すことは簡単な作業のように思われるが，実際には非常に詳細な規則と判断基準が必要になり，かつ事例ごとに柔軟な判断が求められる場合が多い。例えば，カタログのような様式の印刷紙面上にある文字列のうち，どの部分をどのような順序で抽出していけばよいか，日本語と外国語が混じった文章，数式や化学式などが混じった文章をどう扱うか，表組みのように複雑な構造を持つ部分をどう扱うか，などといった問題に直面するのである。このような問題に対処しながら，均質的な手順でサンプルを抽出するのは，簡単なことではない。

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」「前書き」など，さまざまな要素から構成されている。それらの要素は，漢字で書かれていたり，仮名で書かれていたり，アルファベットで書かれていたり，記号で表現されていたりする。書き言葉の印刷紙面からサンプルを抽出するためには，印刷紙面を構成する要素のうち，どの要素をどのように抽出し，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。言い換えれば，書き言葉の多様な構造はどのように一元的に把握できるか，さらに言えば，**さまざまな体裁を持つ書き言葉の実体から，1 次元の文字列（1 個以上の文字の連鎖）をどのように取り出すか**，という問題について，考えておく必要があるのである。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かなければならない。

以上に示した設計をもとに，3 万冊以上におよぶ書籍・雑誌・新聞などを手に取り，サンプリングの実作業を継続してきた。その中で目指してきたのは，**揺れのない手続きによる，斉一なサンプリング作業**という点に尽きる。原本によって，あるいは作業者によって，サンプリングの結果に違いが生じることのないよう，常に安定した作業結果が得られるように努めてきた。その上で必要となったのが，サンプリングの作業を進める上での「原理」であった。すなわち，書き言葉というものがどのような構造をしており，どのようにそれを把握し，そしてどの部分をどの順に取り出すか，という一連の過程を明示化することである。

そこで続く第 II 部では，我々がサンプリング作業に従事しながら規定してきた，サンプリングの原理について示す。対象としては，BCCWJ の主たる部分を構成する「書籍」を取り上げる。書籍の構造を把握し，そこから可変長サンプル・固定長サンプルを取りだすための原理と運用について述べていくことにする。

第II部

書籍におけるサンプリングの原理と運用

第3章　書籍の構造とサンプリングの原理

本章の概要：　本章では，サンプリングを実施するにあたって必要となる，サンプリングの原理について述べる。書き言葉の構造をどのように把握するか，その中からどの部分をサンプリングの対象とするか，という点について，書き言葉の代表的なメディアであり，かつ最も多様な体裁を持つ書籍を例に取り，その具体的な内実について示す。

以下，3.1 節では，書籍の構造を捉える見方と，そこから書き言葉をサンプリングするという作業の本質を述べる。3.2 節では，書籍の印刷紙面がどのような要素から構成されているかを定義する。その上で，各要素をサンプリングの対象とするか否かについて示す。3.3 節では，書籍を物理的に構成する諸要素を定義し，各層にどのような要素が分布しているかを示す。その上で，各要素をサンプリングの対象とするか否かについて示す。3.4 節では，特に可変長サンプルの範囲を決定するために設けた「理想範囲」「完結構造」という 2 つの観点を示し，それらがサンプリングの範囲とどのように関わるかについて示す。

3.1　書籍の構造をどう捉えるか

書籍に含まれる書き言葉の実体は，紙面の上に印刷された 1 つ 1 つの文字によって構成される。この中から一定範囲の部分をサンプルとして抽出するためには，印刷紙面上にある文字列のうち，どの部分をどのような判断基準によって抽出対象とするかを定めなければならない。そこで，複数の観点によって書き言葉の構造を把握し，抽出する部分を定義することにする。

書籍に含まれる書き言葉がどのような構造を持っているか，それらのどの部分をサンプリングの対象とすべきか，という 2 点を特定するために，ここでは，以下の 3 つの観点から書籍の構造を捉える。

1. 書籍の紙面構成に関わる要素
2. 書籍の階層的な成立に関わる要素
3. 同一著者の執筆範囲，および完結性

3.1.1　書籍の紙面構成に関わる要素

書籍の紙面上に印刷された文字には，レイアウトやサイズ，紙面構成上の扱いなどによって，「本文」「見出し」「注」「表」「目次」「前書き」「後書き」「索引」「柱」「ノンブル」「奥付」「表紙タイトル」などの役割が与えられている。これらを,「紙面構成に関わる要素」と呼ぶことにする。

ここで,「本文」「見出し」「注」などの諸要素を，読み手がどのように区別しているのか，という問題について考えてみたい。これらの要素の区別は，一見，自明であるように思われるが，しかしながら，ある言語表現がどのような構成に関わる要素であるのかは，印刷紙面上に明示されているわけではない。むしろ，印刷紙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。ある言語表現が,「本文」の要素として書かれているのか,「見出し」の要素として書かれているのか,「脚注」の要素として書かれているのかは，実際の出現形式や文脈に応じて，読み手が主体的に判断しているわけである。

先にも述べたように，書籍の中から固定長サンプル・可変長サンプルという 2 種類のサンプルを抽出するという作業は，概念的に言えば，紙面上に印刷してあるすべての文字を 1 次元に配置して，そこから当該の範囲を抽出していく作業であると言える。作業者は，書籍の物理的な構成に関わる要素，または紙面構成に関わる要素の中から，一定の基準に従って，1 次元の文字の連鎖を抽出しなければならない。

そのためにまず必要となるのが，2 種類のサンプルを抽出するための基準となる「サンプル抽出基準点」を取得することである。サンプル抽出基準点は，ランダムに指定されたページからランダムに取得される 1 文字であるが，これを取得するためには，サンプルを取得してよい範囲をまず定義しておく必要がある。すなわち，書籍の紙面構成に関わる要素のうち，どの要素をサンプルに収録する対象として選択し，どの要素をサンプルに収録しない対象として排除するのかを前もって定義しておかなければ，当該のページに含まれる文字列からサンプル抽出基準点を取得してよいか否かを判断することができないわけである。

そこで，紙面構成に関わる要素にはどのような要素があり，そのうちどの要素からサンプル抽出基準点を取得してよいかを定義した。これらの詳細については，3.2 節で示す。

3.1.2　書籍の階層的な成立に関わる要素

紙面構成に関わる要素よりも大きな視点として，書籍という物理的な印刷物がどのような要素によって成立しているか，という見方がある。例えば，1 冊の書籍を構成する要素を考えてみた場合，いわゆる本文部分の外側には，目次や口絵，奥付などがあり，さらに表紙がある。ケースやカバーがある場合や，付録としてポスターや CD-ROM が添付されていることもある。

逆に，本文部分の内側についても，ノンブル（ページ番号），フィギュア（図表など），数式や化学式，キャプションなど，書籍の物理的・論理的な構成に関わるさまざまな要素がある。

サンプリングを実施するためには，書籍の成立を支えるこれらの要素のうち，どの部分を対象としてサンプルに含めるのか，逆に，どの部分はサンプルに含めないのか，といった規則を定めておく必要がある。そこで，書籍の構造を階層的に成立するものと捉え，各層に含まれる文字をサンプリングの対象とするか否かを判断することにする。ある文字列が，書籍を成立させる階層のどこに位置づけられるかによって，その文字をサンプリングの対象とするか否かを決めるのである。この見方により，固定長サンプルとして抽出する「1,000 文字」や，可変長サンプルの上限である「1 万字」の範囲も決められることになる。

これらの判断基準を定めるために，書籍という印刷物の成立を階層的に把握し，その中からサンプルとして収録する範囲を定めた。この詳細については，3.3 節で示す。

3.1.3 同一著者の執筆範囲，および完結性

上記の 2 点とはさらに別の観点として，可変長サンプルを取得する範囲をどう定めるか，という視点がある。この際書籍に含まれる文章の著者，および作品としての完結性を考慮する。

可変長サンプルとは,「言語的な構造のまとまり（「章」や「節」など，ただし 1 万字を上限とする)」を抽出するサンプルであるが，そのまとまりの認定には,「著者」の異同が大きく関与する。すなわち，同一の著者が同一のテーマのもとに執筆した文章全体を，可変長サンプルで取得する理想的な「言語的な構造のまとまり」と見なすのである。

同一著者による同一テーマの書籍，例えば小説の単行本であれば，1 冊全体を完結した構造を持つ範囲と見なし，その全体を可変長サンプルとして取得することが理想的である。ただし，その全体が 1 万字を超える場合は，サンプル抽出基準点の位置に応じて,「第 5 章」や「第 3 章第 2 節」といった部分的な構造を取得することになる。つまり，可変長サンプルの取得とは，対象となる書籍に含まれる「理想的な範囲」を見定め，そこから 1 万字の上限を超えない範囲にまで対象を狭めていく作業であると言うことができる。

このようなサンプル抽出の範囲に関して,「理想範囲」「完結構造」という 2 つの視点を導入する。この詳細については，3.4 節で示す。

以下では，書籍の構造を捉えるための 3 つの視点,「書籍の紙面構成に関わる要素」「書籍の階層的な成立に関わる要素」「同一著者の執筆範囲，および完結性」の 3 点について，具体例も交えて詳しく述べる。

3.2　書籍の構造 (1) — 書籍の紙面構成に関わる要素

3.2.1　書籍の紙面構成

前節で述べたように，紙面上に印刷された文字には，書籍の紙面構成を支えるための役割が与えられている。これらを,「紙面構成に関わる要素」と呼ぶ。ここでは，紙面構成に関わる要素を，図 3.1 のような形で把握する。

書籍				
表表紙	**前付**	**冊本体**	**後付**	**裏表紙**
	口絵	中扉	付録	
	標題紙	見出し	索引	
	献辞	本文	後書き	
	前書き	注	奥付	
	目次	フィギュア	広告	
	凡例	キャプション		
		ノンブル		
		柱		

図 3.1: 書籍の紙面構成に関わる要素

以下では，これらの各要素についてその定義を示す[1]。

書籍： 文字などが書き込まれたページをひとまとめに冊子の形に綴じ付けたもの。「図書」「本」などともいう。

表紙： 書籍などの印刷物の中身を保護・保持するための外装。開きはじめの側を表(おもて)表紙といい，その反対側の部分を裏表紙という。

前付(まえづけ)： 冊本体の前に付けられているひとまとまりの部分のことで，口絵，標題紙，献辞，前書き，目次などからなる。

口絵： 標題紙の前に入っている別刷りの図版。

標題紙(ひょうだいし)： 通常，前付の冒頭にあって，その出版物の最も完全な書誌的情報を提供しているページのこと。書籍のタイトルのほか，責任表示，版次，出版地，出版者，出版年の全部または一部などが記載される。

献辞(けんじ)： 著者が先輩・友人・家族などに対して，その著書を捧げることを表明したことば。

前書き： 本文に先立って，著者が著述の動機や追想などを記した文章。序，序文，序言，はしがき，前言，などともいう。

[1] 定義の大半は，日本図書館協会用語委員会編『図書館用語集 三訂版』から抜粋，あるいは一部改変して用いた。

目次： 本文内容を一覧し，検索できるようにした部分。編・章・節などの見出しや論文名・記事の題名・著者名を，普通は記載順に列挙し，それぞれに本文の該当ページ数を付ける。

凡例： 書籍の目的や方針，記号の意味や約束事などを示したもの。

冊本体（さつほんたい）： 書籍の実質的な内容の主体をなす部分で,「前付」に続く部分。書籍の中身のうち,「前付」と「後付」を除いた部分を指していう。書誌学的には「本文（ほんぶん）」という用語が適切であるが，以下の「本文」と区別するために，ここでは「冊本体」と呼ぶことにする。

中扉： 目次より後にあり，それ以降の部分のタイトルなどを記載したページ。

本文（ほんぶん）： 冊本体の中でも，主になっている部分。一般的に文章の形で記述され，書籍の実質的な中身を表す。

見出し： 本文の各編・章・節などに付けられた題名。

注： 本文に対する注釈や説明。注記ともいう。巻末または各章末に一括して記される場合（巻末・章末注）と，各ページ内に記される場合（脚注など）がある。

フィギュア： 本文中に含まれている写真や図など，言語表現以外の内容が主たる対象となっている部分。このうち，写真，イラスト，漫画，図解，グラフなどを総称して特に「**フィギュア本体**」と呼ぶことにする。また，フィギュア本体の近くに配置されてそのフィギュア本体に対して解説を加える部分のことを，特に「**キャプション**」と呼ぶことにする。

ノンブル： 1 ページごとに順を追って入れてある数字のこと。

柱（はしら）： ページの欄外（上下・左右）に書かれた，書名や章節名，あるいは見出しなどの部分。

後付（あとづけ）： 冊本体の後に続くひとまとまりの部分のこと。付録・索引・後書き・奥付などからなる。

付録： 冊本体を補うために巻末に付される関連論文，解説，図表，資料などを指していう。後付以外の位置に綴じ込まれたポスターや葉書，巻末に添付された CD-ROM，工作材料やおもちゃなどが添付されている場合なども含む。

索引： ある特定の情報を示す語句などを一定の順序に配列し，その情報の所在を指示するもの。

後書き： 書籍の末尾に著者が付ける文章。「前書き」とほぼ同じ性質を持つ。

奥付： 書籍の末尾，最終ページ，時には裏表紙の内側などに，著者・編者・訳者などの名，書名，出版者，印刷者，印刷・発行の年月日，版次，価格，著作権その他の出版上の条件などを表示した部分。

広告：商品の内容を消費者に伝達・宣伝するための部分。書籍の場合，同じ出版者が出版している他の書籍を宣伝する部分が巻末に付されることがある。

3.2.2　サンプル抽出基準点の取得に関する原則と判断

上記で定義した書籍の紙面構成に関わる要素は，書籍に含まれる書き言葉がどのような役割を果たすかを整理する上で，基礎的な概念となる。サンプリングの実作業においては，ランダムに指定されたある 1 文字をサンプル抽出基準点として取得してよいか否かを判断する基準として，これらの要素の区別を用いる。各要素の区別とサンプル抽出基準点の取得の可否については，以下に示す原則を採用する。

サンプル抽出基準点の取得に関する原則：

- 「冊本体」に分類される要素は，サンプル抽出基準点を取得する対象としてよい。
- 「前付」「後付」に分類される要素のうち，一定の文章量を備えているものについては，サンプル抽出基準点を取得する対象としてよい。典型的には，「前書き」「後書き」がこれに該当する。
- 「前付」「後付」に分類される要素のうち，「口絵」「標題紙」「献辞」「目次」「凡例」「付録」「索引」「奥付」「広告」は，基本的に，サンプル抽出基準点を取得する対象とはしない。

この原則を定めることで，サンプル抽出基準点を取得してよい範囲を明確に定義することができる。より具体的には，ランダムに選ばれたある 1 ページからサンプル抽出基準点を取得できるか否かを判定する際，まずはそのページが「前付」「冊本体」「後付」のどこに含まれるのかを判断する。「冊本体」であれば，サンプル抽出基準点を取得できるページであると判定してよい。「前付」「後付」の場合は，そのページが「前書き」「後書き」に含まれていれば，やはりサンプル抽出基準点を取得できるページであると判定してよい。それ以外の要素に該当した場合には，原則，そこからサンプル抽出基準点を取得できないものと見なす。無論，当該のページが「冊本体」に位置していたとしても，そのページが白紙だった場合や，図やグラフ，写真しか掲載されていないページだった場合は，そこからサンプル抽出基準点を取得することはできない[2]。

さて，実際のサンプリング作業においては，NDC および発行年によって層別された各層に含まれる全ページに優先順位がランダムに振られ，その順に現物の書籍を手にとって指定されたページを開けていくことになる。当該のページからサンプル抽出基準点を取得できるか否か

[2] なお，当該のページのほとんどが「固有名詞」「数字」の羅列である場合は，例外的に，そのページを回避し，サンプル抽出基準点を取得しないこととする。

を判断するには，そのページが紙面構成に関わる要素のどこに分類されるかを確認し，上記の「原則」に照らして判断すればよい。

この際，当該のページが「前付」「後付」に相当する場合，一定の文章量を備えている要素であることが，サンプル抽出基準点を取得するページとして同定するための条件となる。その典型は,「前書き」と「後書き」である。

一方，サンプル基準点を取得するページの対象としないもののうち,「前付」に位置するものには「口絵，標題紙，献辞，目次，凡例」などがある。また,「後付」に位置するものには「付録，索引，奥付」などがある。原則として，これらは文章量の少ない要素であると見なし，サンプル抽出基準点を取得するページとはしない。

ただし，上記の要素のうち「口絵」「献辞」「凡例」「付録」（あるいはそれに類似した要素）については，一定の文章量を備えていることがある。そのような場合，そこに書かれている文章はサンプリングの対象にしてよいと考える。例えば,「口絵」に一定量の文章が付されている場合，それを本文と見なすこともできる。「献辞」に長い文章が載っている場合は,「前書き」あるいは「後書き」に代わるものと考えることもできる。また，古典全集における「凡例」などは，それ自身が独立した 1 章を成すものと見てもよい。さらに,「付録」としてまとまった量の文章が掲載されていることもある。このような場合，その冊において当該個所が占める役割を考慮した上で，サンプル基準点を取得するページの対象にするという判断を個別に下してもよい。

ただし,「広告」は書籍の主たる内容ではないため，例え文章量があってもサンプル抽出基準点を取得するページとはしない。

3.3 書籍の構造 (2) — 書籍の階層的な成立に関わる要素

3.3.1 書籍を構成する諸要素の階層構造

次に，書籍の構造を捉えるための 2 つ目の視点として,「書籍の階層的な成立に関わる要素」を取り上げる。ここでは，書籍という印刷物の実体を，図 3.2 に示すような 7 段階の階層によって成立するものと捉えることにする。「第 0 層」から「第 6 層」へと階層が深くなるにしたがって，書籍の外側から内側へと構成要素が細分化され，サンプルに収録する範囲や条件が絞り込まれていくことになる。

以下，各階層に位置づけられる物理的・言語的要素について述べる。

第 0 層： 物理的実体

第 0 層は，印刷物・出版物としての書籍が持つ「物理的な実体」そのものを指す。書籍のケース，帯，カバーなども含めた，手に取って見ることができる書籍の全体に相当する。

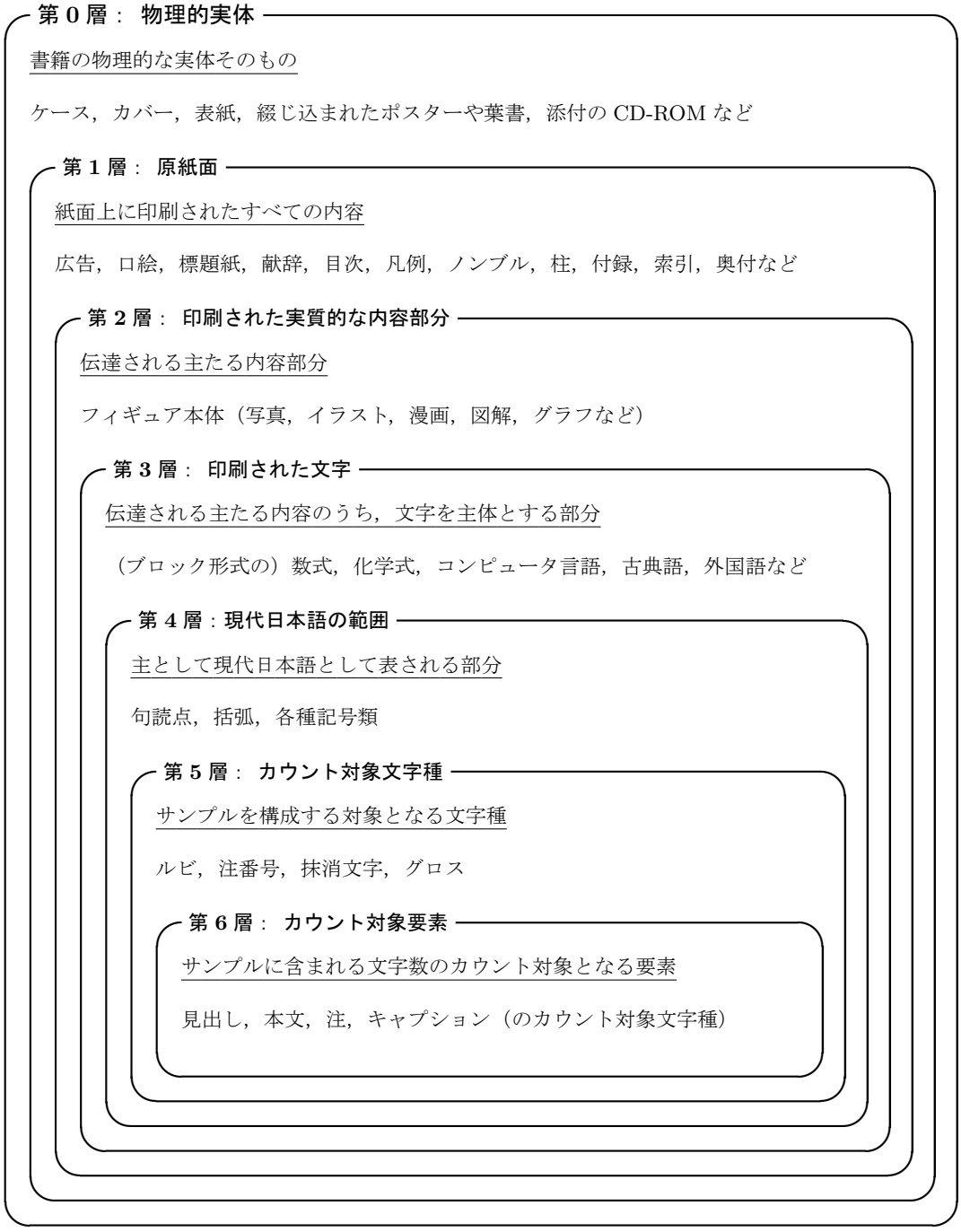


図 3.2: 書籍の階層的な成立に関わる要素

第 1 層： 原紙面

第 1 層は，書籍の「紙面上に印刷されたすべての内容」を指す。表紙の内側に綴じられて印刷された紙面の集合であり，第 0 層の物理的実体のうち，本のケース，カバー，表紙や，綴じ込まれたポスターや葉書，添付の CD-ROM などの要素を除外した残りの部分に相当する。

第 2 層：印刷された実質的な内容部分

第 2 層は，当該の書籍によって伝達される主たる内容に関わる部分を指す。第 1 層の「原紙面」のうち，口絵，標題紙，献辞，目次，凡例，ノンブル，柱，付録（参考資料として付された統計図表のまとまりなど），索引，奥付，広告などは,「伝達される主たる内容」以外の要素と見なして排除し，残った部分を第 2 層とする。

第 3 層：印刷された文字

第 3 層は，第 2 層で示した「伝達される主たる内容」のうち，印刷された文字を主体とする部分を指す。実際には，第 2 層の中から「フィギュア本体」を排除して残った部分に相当する。フィギュア本体と文字が重なっている場合，フィギュア本体が主たる要素であれば，文字の部分もあわせて排除する。逆に，文字の部分が主たる要素であれば，それらは残す。

第 4 層：現代日本語の範囲

第 4 層は，第 3 層で示した「印刷された文字」のうち，主に現代日本語として表されている範囲を指す。第 3 層の中から，ひとまとまりの形（ブロック形式）で記述される数式や化学式，コンピュータ言語，外国語や古典語などを除外した部分に相当する。

第 5 層：カウント対象文字

第 5 層は，第 4 層で示した「現代日本語の範囲」のうち「仮名」「漢字」「数字」「アルファベット」で表記された文字を指す。固定長サンプルを構成する 1,000 字，・可変長サンプルを構成する最大 1 万字としてカウントするのは，これらの文字種である。句読点，括弧，各種記号類などの文字は，カウント対象とならないため，第 5 層には含めない。

第 6 層：カウント対象要素

第 6 層は，第 5 層で示したカウント対象文字種のうち，実際に固定長サンプルを構成する 1,000 字，可変長サンプルを構成する最大 1 万字としてカウントされる文字の集合を指す。典型的には,「見出し」「本文」「注」「キャプション」を構成する文字に相当する。「ルビ」「注番号」「抹消文字」「グロス」などの要素は，カウント対象とならないため，第 6 層には含めない。

3.3.2　サンプル構成要素の排除と取得に関する原則

以上で示した書籍の構造の階層的な把握にしたがって，以下では，実際の印刷紙面からサンプルの範囲に含める要素を絞り込んでいく原則について示す。以下では，サンプルの範囲に含める要素を「サンプル構成要素」と呼ぶことにする。

サンプリングの原理的な考え方として，書籍の実体を手に取った後，そこから不要な要素を順次排除していくことによって，サンプル構成要素の範囲を絞り込んでいくものとする。具体的には，第 0 層から第 6 層へと進んでいくことによって，書籍を構成する要素が徐々に削ぎ落とされていくわけである。その原則を，以下のように定める。

サンプル構成要素の排除と取得に関する原則：

(1) 第 0 層から第 3 層までに位置づけられる要素は，サンプルの範囲から排除する。

(2) 第 4 層から第 6 層までに位置づけられる要素は，サンプルの範囲に含めてよい。

この原則により，第 0 層から第 3 層までに位置づけられる構成要素は，サンプルの範囲からは排除されることになる。それゆえ，例えば，目次の部分はサンプルには含まれず，またノンブルやブロック形式の非現代日本語の部分はサンプル抽出基準点とはなり得ない。

逆に，第 4 層以上の要素として残った部分は，サンプルの範囲に含まれる資格を備える。さらに，第 4 層に含まれる要素のうち，句読点や記号類，またはルビや注番号などの要素を排除し，最後まで残った第 6 層の要素が，固定長サンプル（1,000 字），可変長サンプル（最大 1 万字）としてカウントされる対象に認定されるわけである。サンプリングの作業者は，書籍の現物を手に取り，指定されたページの印刷紙面を見てその構成を確認し，上記の原則に基づいて排除すべき対象要素を判断し，残った部分から固定長サンプル・可変長サンプルの範囲を抽出するのである。

3.3.3　原則の運用と判断基準 ―フィギュアの処理

書籍の構造を階層的に捉えた上で,「サンプル構成要素の排除と取得に関する原則」を設けることにより，サンプル範囲から排除される要素，サンプル範囲として取得される要素の範囲を定めることができた。しかしながら，この原則に則って作業を進めていくと，紙面上のある表現を階層構造におけるどの要素として把握すべきか，その判断に迷う例が少なからず見つかる。そこで以下では，この原則を適用する際に問題となる具体的な事例と，その判断基準を示す。

判断に迷うケースの最たる例は，フィギュアと文字が併存する場合の扱いについてである。一見フィギュア本体に見える要素の内部に，文字列が多く含まれている場合に，その文字列をサンプル構成要素と見なしてよいか否か，ということである。

これに対して,「一見フィギュア本体に見える要素であっても，その内部にある言語表現を一方向に読み進めることができれば，フィギュア本体とは見なさず，排除の対象とはしない」という原則を定める。その根本にあるのは,「**印刷紙面上に現れた文字列は，それが現代日本語として一方向に読み進められる限り，できるだけサンプルとして収録する**」という方針である。

このことを，(1)「フローチャート」，(2)「表」という 2 つを例として説明する。まず，図 3.3，3.4 のようなフローチャートの例を見てみよう。これらのフローチャートに含まれる文字列が，サンプルの範囲から排除される要素になるかどうかを考えることにする。

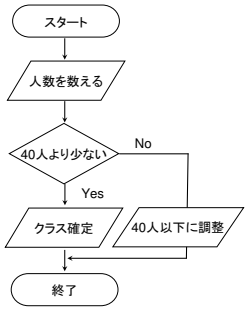


図 3.3: 分岐型フローチャートの例

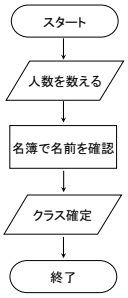


図 3.4: 直線型フローチャートの例

図 3.3 に示したフローチャートに含まれる文字列はサンプル範囲から排除されるが，図 3.4 のフローチャートに含まれる文字列はサンプル範囲に含まれるものと考える。ここで判断基準となるのは,「サンプル構成要素の排除と取得に関する原則」ではなく，むしろ「**そこに書かれている文字列を 1 方向に読み進めることができるかどうか**」という点である。

図 3.3 のような形をしたフローチャート（**分岐型**）は，途中に 2 方向以上の分岐を持つ立体的な構造を取っているため，中に書かれている言語表現を一方向に読み進めることができない。一方，図 3.4 のように途中での分岐を持たないフローチャート（**直線型**）は，フローチャートの形式を取ってはいるものの，中に書かれている文字列を一方向に読み進めることができる。先に述べたように，サンプリングを行なう作業者は，印刷紙面に現れるあらゆる要素から 1 次元の文字の連鎖を取得する。このことを制約として考えると，分岐型のフローチャートは，それが図式化されていて一方向に読み進めることができない以上，1 次元の文字列を取り出すことができず，サンプル範囲からは排除せざるを得ない。しかしながら，直線型のフローチャートは，例えそれが図式化されているものであっても，一方向に読み進めることができる以上，サンプルの範囲から排除する理由はないと考えるのである。

これと同様のことが,「表」にも言える。

図 3.5 に示したのは，行見出しと列見出しを備えた表（いわゆるクロス表）であり，表の中でも典型的なものである。このような構成を持つ表は，全体が図式化されており，そこに含まれている文字列に対して一方向に読み順を定めることができないものと見なす。そこで，サン

	明日	明後日
東京	晴れでしょう	晴れでしょう
大阪	曇でしょう	雨でしょう
福岡	雨でしょう	曇でしょう

図 3.5: 行列見出しを備えた表の例

日本のお酒：	日本酒
ドイツのお酒：	ビール
フランスのお酒：	ワイン

図 3.6: 2 列から構成される表の例

プル範囲からは排除する対象と判断する。

一方，図 3.6 に示したのは，行見出しと列見出しを備えず，2 つの列から構成される表である。このような形で構成される表は,「日本のお酒 → 日本酒，ドイツのお酒 → ビール，フランスのお酒 → ワイン」という具合に，全体の構成を崩すことなく，一方向に読み進めることができる。このような表は，全体が図式化されているとは判断せず，サンプル範囲に含めることとする。つまり，そこから 1 次元の文字列を取り出すことができる対象である以上，サンプル範囲から排除する理由はない。

このように考えると,「フローチャート」「表」については，サンプルの範囲から排除されるものと排除されるものとが区別できるわけである。分岐型のフローチャートや，行列見出しを備える表は，**一方向の読み順を定められないという点において図式化されていると**考え，先に定義した「フィギュア本体」に相当するものと見なし，第 2 層に属する要素として排除される対象と見なすのである。一方，直線型のフローチャートや，行列見出しを持たない 2 列の表は，一方向に読み進められるという点において，サンプルに収録する対象から排除される理由はないと考えるのである。

以上に示した，フローチャートや表の形状によってサンプル範囲から排除されるかどうかを決定するという方針は，印刷紙面上から 1 次元の文字列を取り出すという，書き言葉におけるサンプリングの原理に基づくものである。いかに周囲が罫線で囲まれていても，あるいは，いかに他のフィギュア本体と形状が似通っていたとしても，表面的なレイアウトのありようではなく，そこに含まれている文字列をどのように読むことができるか（1 次元に展開できるか否か）というあり方によって，サンプリングの対象とするかどうかが決まるわけである。

3.4　書籍の構造 (3) — 同一著者の執筆範囲，および完結性

3.4.1　「理想範囲」と「完結構造」

可変長サンプルとは，前述の通り，母集団の中からランダムに指定した 1 文字（サンプル抽出基準点）を含む言語的な構造のまとまりを抽出するサンプルである。言語的な構造のまとまりとは,「章」や「節」など，その文書を論理的に構成する単位を指す。ただし，可変長サンプルとして抽出するサイズの上限は，1 万字とする。文字数の少ない書籍の場合は，1 冊まるごとが 1 つの可変長サンプルを構成することもある。

可変長サンプルを抽出する際，文字数としての上限は 1 万字と定めたが，それ以外にも可変長サンプルの範囲を決めるための条件がある。この条件は,「**理想範囲**」と「**完結構造**」という 2 点によって定められる。これらはサンプリング作業を進める上で独自に定めた用語であり，以下のように定義される。

理想範囲：当該文書のうち，同一著者によって同一テーマのもとに書かれた範囲の全体

完結構造：当該文書のうち，その文書を論理的に構成する部分的なまとまり（「章」「節」など）が過不足なくサンプルとして抽出された構造

「理想範囲」は，同じ著者が同じテーマについて書いた文書の範囲全体を指す。例えば 1 人の著者による小説の単行本の場合，そこに含まれる冊本体全体が「理想範囲」となる。1 人の著者による著作を集めた個人全集や短編集であれば，そこに含まれる個々の著作や短編それぞれが「理想範囲」となる。複数の著者が寄稿した論文集の場合は，収められた各論文がそれぞれ「理想範囲」となる。

また，ある文書の構成要素（「章」「節」など）の範囲が，抽出された可変長サンプルの範囲と一致する場合，そのサンプルの「完結構造」は「完結している」と見なす。後述するように，可変長サンプルの中に章や節の一部しか含まれないケースが発生することがあるが，その場合，そのサンプルの「完結構造」は「一部完結」または後述する「冒頭 1 万字」であると見なす。

可変長サンプルを抽出する際，1 万字を上限とするという条件以外に,「理想範囲」と「完結構造」という 2 つをあわせて考える。すなわち，可変長サンプルを抽出する場合，**可能な限り「理想範囲」全体を抽出する，ただしその範囲が 1 万字を超える場合は，できるだけ「完結構造」が「完結」になる範囲を抽出する**，ということである。

単行本の小説の場合，その冊本体全体を可変長サンプルとして抽出するのが理想だが，その範囲が 1 万字を超える場合は，不完全な「理想範囲」にはなるけれども，その下位の構成要素（例えば「第 4 章」）を抽出する。この場合,「第 4 章」というまとまりが過不足なく抽出されている点で,「完結構造」は「完結」であると見なす。一方，複数の著者が寄稿した論文集から 1 編の論文全体が可変長サンプルとして取得できた場合，そのサンプルの「理想範囲」は「完全」,「完結構造」は「完結」であると見なす。

3.4.2 「理想範囲」と「完結構造」の組み合わせ

以下では,「理想範囲」と「完結構造」の組み合わせにどのようなパタンが生じうるかについて示す。

「完全・完結」　上述したように，同じ著者が同じテーマで執筆した文書の全体が 1 万字以内である場合,「理想範囲」は「完全」,「完結構造」は「完結」の可変長サンプルが取得できることになる。文書全体が 1 万字で収まるということは，比較的短い文書が該当する。例えば，分担執筆の論文集に含まれる論文が取得できた場合や，個人全集・短編集に含まれる 1 作品がまるごと取得できた場合，あるいは，単行本の小説から，その小説の著者とは別の著者による「解説」全体が取得できた場合などが相当する。

「不完全・完結」　一方，同じ著者が同じテーマで執筆した文書の全体が 1 万字を超える場合は，その文書を構成するまとまりのうち，例えば「4 章」や「第 3 章第 2 節」というように，できるだけ完結したまとまりで，かつ 1 万字に最も近い範囲を抽出することになる。この場合,「理想範囲」は「不完全」,「完結構造」は「完結」の可変長サンプルが抽出されることになる。特に単著の書籍のように，1 人の著者が同じテーマでまとまった量の文章を著している場合,「不完全・完結」の可変長サンプルが抽出されることが多い。

「不完全・一部完結」　例えば，第 3 章の見出しを構成する文字が，サンプル抽出基準点として指定されることがある。この場合，第 2 章全体が 1 万字以内であれば，第 3 章全体が可変長サンプルとなり,「完結構造」は「完結」となるが，第 3 章全体が 1 万字を超える場合は，第 3 章の見出し部分に加えて，第 3 章のすぐ下にあるまとまり（例えば第 3.1 節）を抽出することになる。この場合，第 3 章の見出し部分が抽出されているものの，第 3 章全体が抽出されているわけではない。このように抽出された可変長サンプルの「完結構造」は,「一部完結」と見なす。

「不完全・冒頭 1 万字」　さらに，サンプル抽出基準点を含む 1 万字に収まる範囲内に，適当な論理的なまとまりを認定できない場合がある。例えば，章や節の構造が存在せず，文章のみが延々と書き進められていく小説や，章や節に含まれる文字数が非常に多い場合などが該当する。このような場合は，サンプル抽出基準点を含む最小の論理的構造の冒頭から 1 万字目までを可変長サンプルとして取得する。このように取得された可変長サンプルは,「理想範囲」を「不完全」，完結構造を「冒頭 1 万字」と呼ぶ。

以上で述べた通り，可変長サンプルの「理想範囲」と「完結構造」の組み合せには,「完全・完結」,「不完全・完結」「不完全・一部完結」「不完全・冒頭 1 万字」という 4 つのパタンが存在する。「不完全・完結」や「完全・一部完結」などの組み合わせは，それぞれの定義上，あり得ない。例として，ある書籍の「第 3 章」の中にサンプル抽出基準点があった場合を仮定すると，4 つのパタンは次のように分類することができる。

- 第 3 章がある著者による個別の著作であり，かつ 1 万字以内である。
  → 理想範囲「完全」，完結構造「完結」
- 第 3 章より上の範囲が同一著者の同一著作であり，第 3 章全体が 1 万字以内に収まる。
  → 理想範囲「不完全」，完結構造「完結」
- 第 3 章より上の範囲が同一著者の同一著作であり，第 3 章全体は 1 万字を超える。サンプル抽出基準点が第 3 章の章見出しにあり，第 3 章第 1 節が 1 万字以内に収まる。
  → 理想範囲「不完全」，完結構造「一部完結」
- 第 3 章より上の範囲が同一著者の同一著作であり，第 3 章全体は 1 万字を超える。その下には論理的なまとまりがない。
  → 理想範囲「不完全」，完結構造「冒頭 1 万字」

以上のパタンを図示すると，次の図 3.7 のようになる。

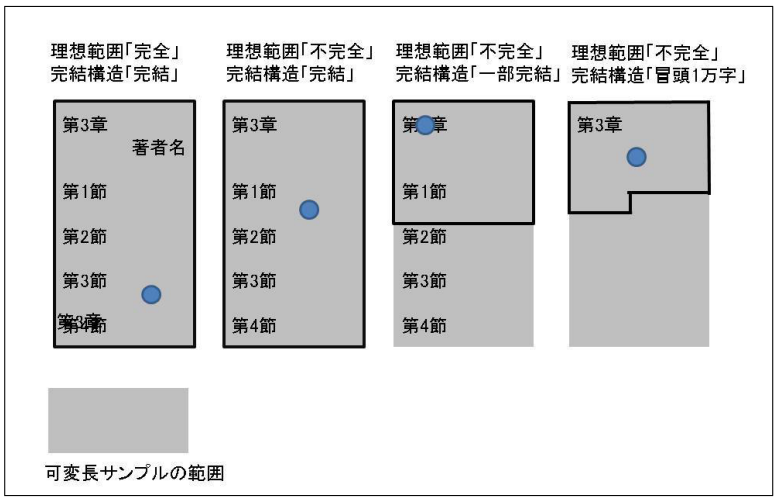


図 3.7: 「理想範囲」と「完結構造」の関係パタン

3.4.3 「理想範囲」の認定に関わる問題と判断基準

以下では,「理想範囲」を認定する際に問題となる事例とその判断基準について，実際の例を挙げながら示す。

複数の著者が存在する場合

前述のように，1 人の著者が書いた小説の単行本であれば,「同一著者によって同一テーマのもとに書かれた範囲の全体」として，そこに含まれる冊本体全体を「理想範囲」と見なす。一方，複数の著者が関与している場合（共著）は，各著者があるテーマのもとに著した範囲が「理想範囲」となる。実際の紙面上に分担執筆の範囲が明記されていれば，その情報を手掛かりとして各著者の「理想範囲」を捉えることができる。なお，分担執筆の範囲が実際の紙面上（目次など）に明記されていなければ，各著者の執筆範囲が不明である以上，冊本体の全体を共著者全員による 1 つの「理想範囲」と捉えざるを得ない。

「対談，座談，インタビュー」の扱い

対談，座談，インタビューなどの場合は，原則として，発話者一人ひとりを単独の著者としては見なさない。つまり，個々の発話を「理想範囲」とせずに，対談，座談，インタビュー全体を，複数の発話者による共著の「理想範囲」として扱う。

なお，文章の終わりに，対談のホスト役に当たる人物が単独で記述した部分があることがある。そのような場合も，文章としての類型を優先させ，当該の部分も含めて対談全体を 1 つの「理想範囲」とする。単独で記述した部分だけを独立した著者による「理想範囲」として認めることはしない。

「あらすじ」の扱い

「本文」の前に「あらすじ」が示されていることがある。場合によっては,「あらすじ」は「本文」と別の著者による文章である可能性も否定できない。しかしながら,「あらすじ」は，その性質上,「本文」に付随するものと見なす。したがって,「あらすじ」の部分だけを「理想範囲」と認めることはしない。

「往復書簡」の扱い

さて，ここまでの例とは逆に，あるまとまりが 1 つの「理想範囲」とならない類型もある。例えば，往復書簡である。往復書簡は，往と復の組み合わせで成立する，共著による 1 つの「理想範囲」であるという考え方もできなくはないが，ここでは個々の書簡を，個々の著者による個別の「理想範囲」を構成するものと考える。

「Q&A」の扱い

次に，Q&A 形式の場合を考える。Q&A 形式については，A の著者が Q を引用して自らの見解を述べるものと考える。つまり，A の方を「地の文」と見るわけである。Q だけ，あるい

は A だけで「理想範囲」とすることはしない。また，Q&A の組み合わせを「理想範囲」として区切ることもしない。単独の著者による単行本で，随所に第 3 者による Q が入り込むような場合があるが，このような Q は引用の範囲と見なす。よって，理想範囲の認定に影響はない。原則にのっとって「冊本体」以下の「理想範囲」を認定すればよい。

ただし，Q&A 集のように，A の著者が複数いて，各 Q&A の執筆分担が明記されている場合は，個々の Q&A を 1 つの「理想範囲」とする。

複数の翻訳者がいる場合

ある文書が複数の著者によって書かれている場合，個々の著者による理想範囲を認めるために，文章をどこまで分割して捉えるべきかが問題になることがある。例えば，翻訳書において，原著者 1 人に対し，訳者が各章で異なる場合がある。このような場合は，各章それぞれを，原著者と訳者の共著による「理想範囲」として捉える。

「個人全集」「短編集」などの扱い

ある著者が個別に執筆・発表した作品を，後年になってまとめた「個人全集」「短編集」などは，その冊全体が「同一のテーマのもとに書かれた範囲」とは見なし難い。そこで，その全集・短編集に含まれる個々の作品それぞれを「理想範囲」とする。

この場合,「個人全集」「短編集」であるか否かを判定することが必要になるが，書籍のタイトルなどに「〜集」とある場合は,「個人全集」「短編集」に相当するものと見なし，作品単位で「理想範囲」を認定する。また，タイトルなどに「〜集」がない場合でも，各作品が過去に執筆したものの再録であることが明示されていれば，各作品を 1 つの「理想範囲」とする。

その書籍の構成上，複数の作品を束ねる部立てや章立てが施されている場合でも，各作品がそれぞれ「理想範囲」を構成するものと考える。無論，サンプル抽出基準点が章の見出しの文字列に当たった場合は，冊本体が理想範囲となり，適切な範囲の可変長サンプルを取得することになる。

印刷紙面上の位置と論理的な位置

ある単著の書籍の中で，第 1 章，第 2 章... と続いてきた章立ての最後に,「結論」という独立した章が設けられることがある。この「結論」の本文中にサンプル抽出基準点が当たった場合，冊全体が「理想範囲」となるが，冊全体が 1 万字に収まらなければ,「結論」の部分のみを抽出することになる（「理想範囲」は「不完全」,「完結構造」は「完結」）。

ところが，実際の印刷紙面上では,「結論」の章が独立して設けられず，最終章の最後に取りこまれるような体裁を取っているものがある。例えば，最終章である 4 章の直後に「結論」と題された文章が置かれ，その中にサンプル抽出基準点が当たった場合を考える。この「結論」

がレイアウト上は 4 章の中に取り込まれているように見える場合，それが 4 章の結論なのか，1 冊全体の結論なのか，印刷紙面上の体裁からは判断しにくい。

この「結論」が，内容を読む限り，4 章の結論ではなく，1 冊全体に対する結論となっている場合には，冊全体は 1 万字を超えるため,「結論」の部分のみをサンプルとして抽出する。逆に，もしこの「結論」が 4 章に対する結論である場合は，4 章全体がサンプルとして抽出されることになる。そして 4 章全体が 1 万字を超える場合は，4 章の下位階層にある「結論」の部分がサンプルとして抽出されることになる。

このように，章や節の構造を把握するには，内容にも踏み込んだ判断がしばしば必要になる。

3.4.4　「完結構造」の認定に関わる問題と判断基準

「完結構造」は，図 3.7 でも示したように，当該のサンプルの範囲が完結した構造と範囲を備えるか否かに関する見方である。この構造の完結性を把握するには，その冊本体における章立てや節などの論理構造を把握し，当該のサンプルがどのような範囲に渡っているかを考えればよい。

ところが小説などの場合は，章や節に相当する見出しがなく,「一」「二」... のような単なる連番で表された見出しによって全体が並列的に構造化されているものがある。また，段落間に挿入された空行，記号，イラスト，線などで，区切りの位置が示されるだけの場合もある[3]。このような場合，どこからどこまでの範囲を完結した構造と認めるかについては，その冊全体の構造や体裁を考慮した上での判断が必要となる。

以上，本章では，書籍の構造を捉えるための 3 つの視点,「書籍の紙面構成に関わる要素」「書籍の階層的な成立に関わる要素」「同一著者の執筆範囲，および完結性」という 3 点について示した。「書籍の紙面構成に関わる要素」の区別は，サンプル抽出基準点を取得する可能性に関わる。「書籍の階層的な成立に関わる要素」の区別は，サンプルの範囲から除外する要素の指定，またはサンプルを構成する要素の指定に関わる。そして「同一著者の執筆範囲，および完結性」は，可変長サンプルを抽出する範囲の上限，およびそこで取得された範囲の構造のあり方に関わる。これら 3 つの視点から書籍を構造的に把握することにより，書籍に含まれる書き言葉をサンプリングするための原理を定めた。

[3] ただし，引用前後の空行は区切り位置とは見なさないなど，その印刷紙面が取っている体裁に留意する必要がある。

第4章　可変長サンプルの抽出

本章の概要：　本章では，書籍から可変長サンプルを抽出する一連の手続きと，その基準について示す。第3章では書籍の構造を3つの視点から把握する見方を示したが，特に「書籍の階層的な成立に関わる要素」による区別をもとに，サンプルの範囲から排除する要素を指定し，サンプルの範囲に含める要素を指定する手順について，具体的な事例を交えながら示す。

4.1　可変長サンプルを抽出する原理

はじめに，書籍から可変長サンプルを抽出する原理について述べる。第1章で提示した可変長サンプルの定義を，一部再掲しておく。

> **可変長サンプル**：「可変長サンプル」は，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，ランダムに指定した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1万字を上限とする）を抽出するサンプルである。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。

ここで，実際の書籍の印刷紙面から「言語的な構造のまとまり（「章」や「節」など，ただし1万字を上限とする）」の範囲を特定するためには，書籍という書き言葉の実体に対する構造的な把握と，そこに配置された各要素の扱いに対する統一的な判断基準が求められる。このうち前者については，第3.3節で，書籍を構成する諸要素が「第0層」から「第6層」に渡って分布することを見た。また後者については,「サンプル構成要素の排除と取得に関する原則」として，第0層から第3層までに位置づけられる要素はサンプルの範囲から排除する，第4層から第6層までに位置づけられる要素はサンプルの範囲に含めてよい，と定めたことを述べた。

実際のサンプリング作業では，眼前にある印刷紙面からどの部分を抽出するかではなく，**どの部分をサンプルから排除するか**という点に重点が置かれる。すなわち，実際に手に取った書籍のうち，第0層から第3層までに位置づけられる要素を段階的に排除していくことで，可変長サンプルの範囲が特定されるのである。その点から見れば，手に取った書籍がどのような構造を持つものかを把握し，原則に従って第0層から第3層までの要素を排除していくこと，そして残った要素に入力順を与えて1次元の文字列を取りだすこと，この2点がサンプリングの作業原理であるということになる。

以下では，書籍を構成する各層ごとに，どのような要素が排除される対象となるのかについ

て，実際の例を挙げながら見ていく。その後，最終的に残った要素がどのような形でサンプルとして抽出されるのかについて示す。

4.2　サンプル範囲から排除される要素の特定

「サンプル構成要素の排除と取得に関する原則」を拡張する形で，サンプル範囲から排除される要素に関する原則を，以下のように定義する。

排除原則 1：第 0 層に属する要素，すなわち，書籍の構成上，実質的な内容とは見なせない要素（本のケース，カバー，表紙，綴じ込まれたポスターや葉書，添付の CD-ROM など）は，サンプル範囲から排除する。

排除原則 2：第 1 層に属する要素，すなわち，冊本体の構成上，実質的な内容とは見なせない要素（口絵，標題紙，献辞，目次，凡例，ノンブル，柱，付録，索引，奥付，広告など）は，サンプル範囲から排除する。

排除原則 3：第 2 層に属する要素，すなわち，文字を主体としないフィギュア要素（写真，イラスト，漫画，図解，グラフなど）は，サンプル範囲から排除する。

排除原則 4：図式化されていて，一方向に読み進められない文字列の集合（分岐型のフローチャート，行列見出しを備える表など）は，第 2 層に属する要素と見なし，サンプル範囲から排除する。

以下，特に第 2 層の要素について，具体例を挙げながら見ていく。なお，第 1 層の要素の具体例については 3.2.2 節で述べたので，ここでは割愛する。

4.2.1　第 2 層の要素

写真

図 4.1 は，写真内に文字列があるが，あくまでもその文字列は写真の一部であるため，文字列を含む写真ごと排除対象となる例である。一方，図 4.2 は，地の部分が写真であり，その上に印字された文字列が配置されている。これらの文字列は，写真の一部ではなく，主体的な言語表現である。よって，当該の文字列は収録対象となる例である。

実際には，印字された文字列であっても，写真の中の対象物と一体化しており，主体的な言語表現として取り出しにくい場合がある。そのような場合は，前後のレイアウトとも照らし合わせ，サンプル内で統一がとれるよう，総合的に判断した。

また，本文中に画像を取りこんで表示したものを,「写し込み」と呼び，写真の下位類型として考える。例えば，DVD などのパッケージ，書籍の表紙などの画像，コンピュータのキャプチャ画面などである（図 4.3，図 4.4，図 4.5）。パソコンソフトで作成したスライド画面をそ

図 4.1: 写真の一部に文字列（看板の文字）

図 4.2: 写真の上に文字列

のまま貼りつけたようなものも，この延長で考える。これらの写し込みの中に文字列が含まれていても，それらはすべて「フィギュア」の内部にある文字列と捉え，排除する対象とする。

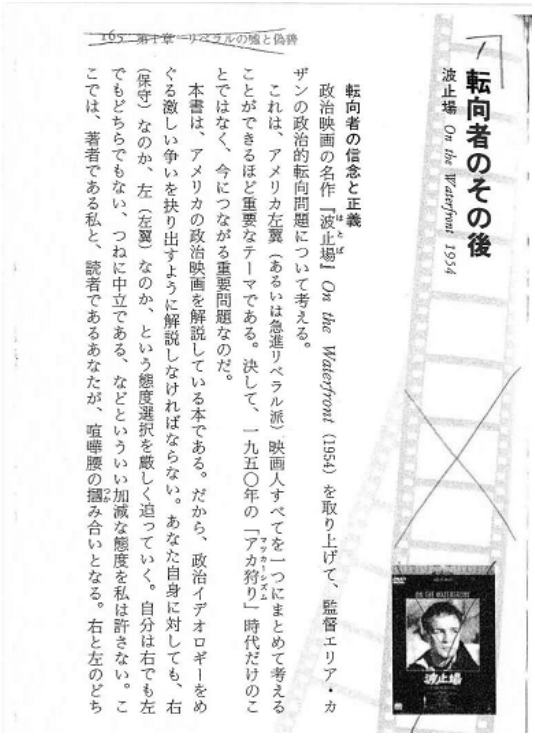


図 4.3: パッケージ

また，ある文書形式を示すためにもとの形態を残したままで書式を貼りつけたと考えられる例も，この「写し込み」として捉える。例えば，婚姻届，確定申告の書類などが冊本体に写し込まれている場合である。図 4.6 は，明細書の書式が写し込まれた例である。また，新聞記事（図 4.7）や週刊誌の記事がそのままの体裁で転載されている場合も「写し込み」として捉える。いずれも，サンプルからは排除される対象となる。

イラスト・漫画

イラストは，典型的な場合には文字列を含まないため，写真と同様に,「フィギュア」の典型例と言える。イラストの内部に文字列が含まれていても，その文字列ごと排除対象とする（図 4.8)。一方，漫画は文字列を含むことが多い類型である。しかしながら，漫画は視覚表現と言語表現の併存によって初めて成り立つメディアであり，文字列のみを抽出したところで，それが十分な言語表現を成すとは言いがたい。そこで，漫画はイラストと同様，文字列ごと排除対象とする。なお，1 冊が丸ごと漫画である「漫画本」は，そもそも母集団を定義する際に除外されている。

図 4.4: 書籍の表紙

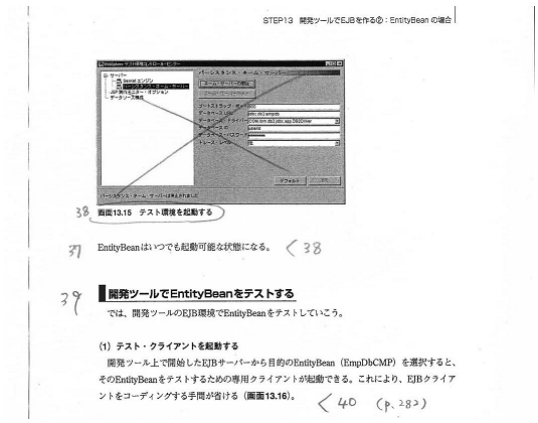


図 4.5: コンピュータのキャプチャ画面

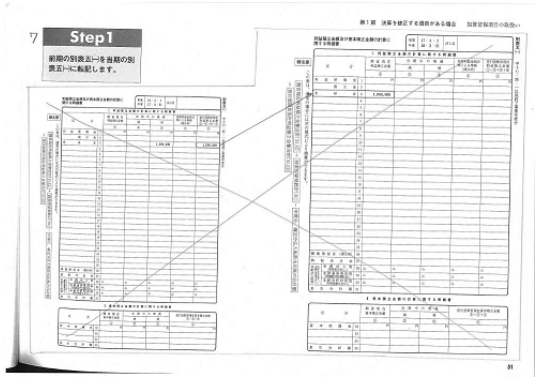


図 4.6: 文書形式を示す書式

図 4.7: 新聞記事

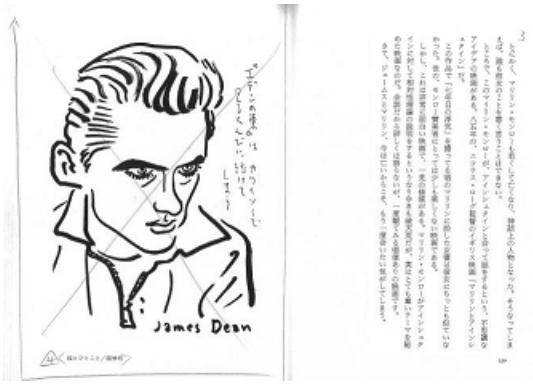


図 4.8: イラスト（中に文字列有り）

図解

図 4.9 に示すような引き出し線のついた文字列が「フィギュア」の細部を指し示しているものを図解の典型と考える。図説とも言われるものである。引き出し線によって文字列が「フィギュア」と結ばれていることをもって，文字列は「フィギュア」に含まれる一部であり，文字列よりも「フィギュア」が主体であると考える。図 4.10 のような場合も同様に考え，引き出し線によって結ばれる文字列は「フィギュア」に含まれるものと捉える。ただし，例外的に，図 4.11 のように引き出し線で結ばれる文字列が，当該のサンプルにおいて章節構造を持つ本文に相当していると見なせる場合は図解とは考えず，図のみを「フィギュア」として排除対象とし，文字列部分はサンプリング対象とする（柏野ほか (2009) を参照）。この「フィギュア」の図解の類型として,「地図，スポーツのポジション図，棋譜，碁譜，牌図」などを扱う。これらはいずれも文字列を含むものであるが，その配置などに意味があることを重視し，図解の類型とするものである。よって，いずれも文字列を含めて排除対象とする。図 4.12〜図 4.16 にそれらの例を順に示す。

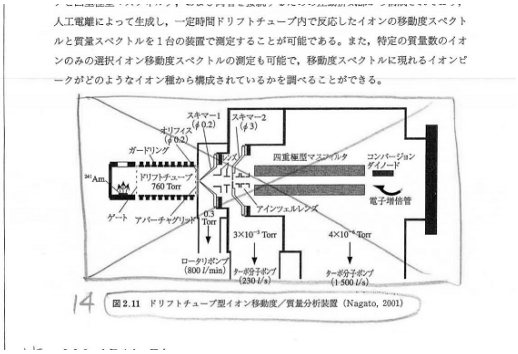


図 4.9: 引き出し線付き図解 その 1

グラフ

グラフの典型例は，棒グラフ，折れ線グラフ，円グラフである。文字列がグラフ上に示されることが多くあるが，それらの文字列はグラフに含まれる補助的なものであり，主体はフィギュアであると考える。よって，図解同様，文字列を含めて排除対象とする。典型的なものを図 4.17〜図 4.19 に示す。

なお，文字列と文字列とが矢印で結ばれており，なおかつ，二方向以上に分岐，もしくは二方向以上から収束しているものを「分岐型フローチャート」と呼ぶ。「分岐型フローチャート」については分岐や収束があるゆえに，文字列を一方向に読むことができないことを根拠に文字列が図式化されている「フィギュア」の類型の 1 つと考える（この類型については 3.3.3 で述べた）。典型例は図 4.20 である。また，図 4.21 のようなものも同じ類型と捉える。逆に，文

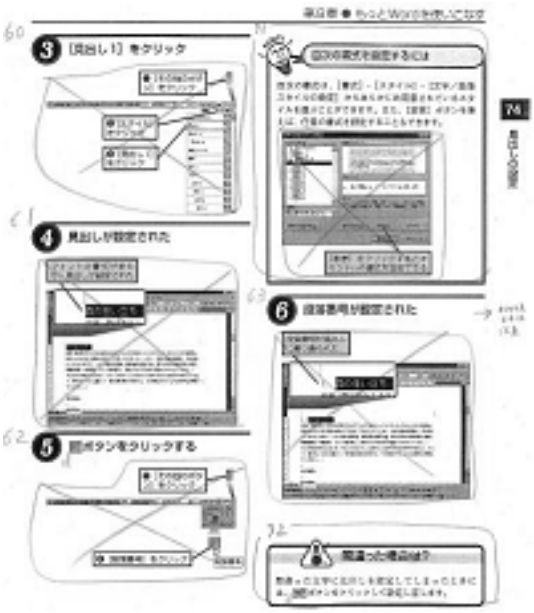


図 4.10: 引き出し線付き図解 その 2

図 4.11: 引き出し線付き文字列部分は本文

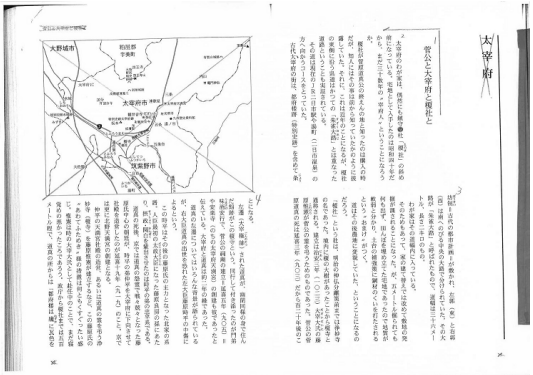


図 4.12: 地図

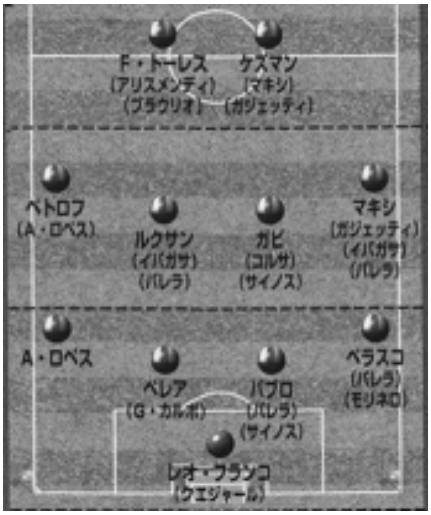


図 4.13: スポーツポジション図

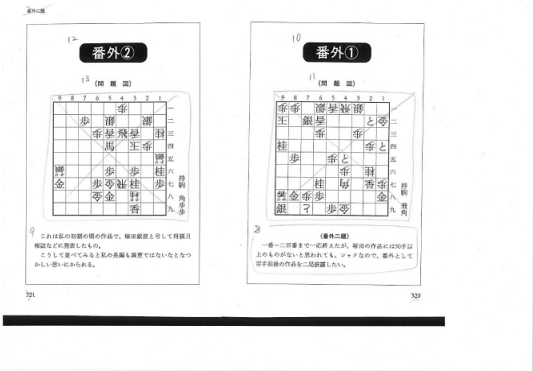


図 4.14: 棋譜

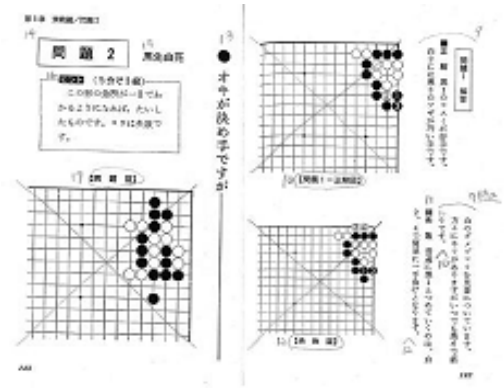


図 4.15: 碁譜

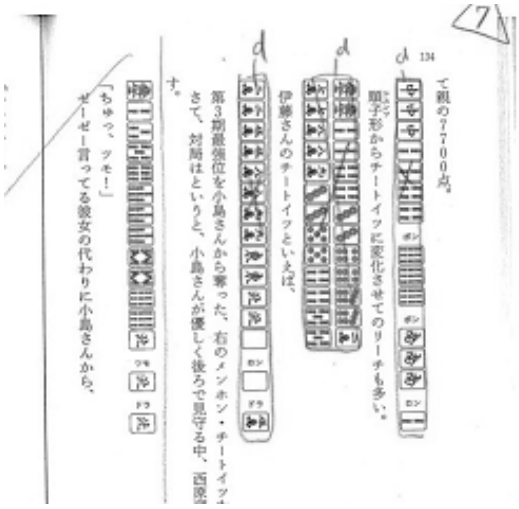


図 4.16: 牌図

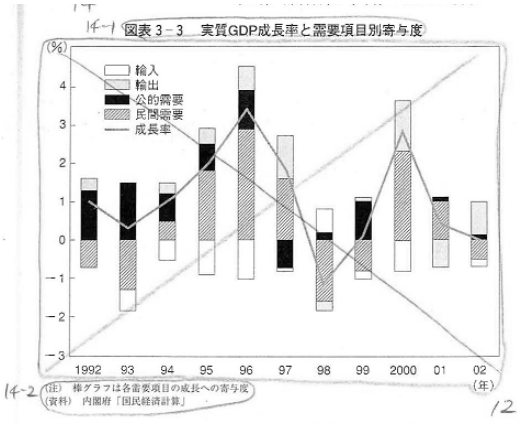


図 4.17: 棒グラフ

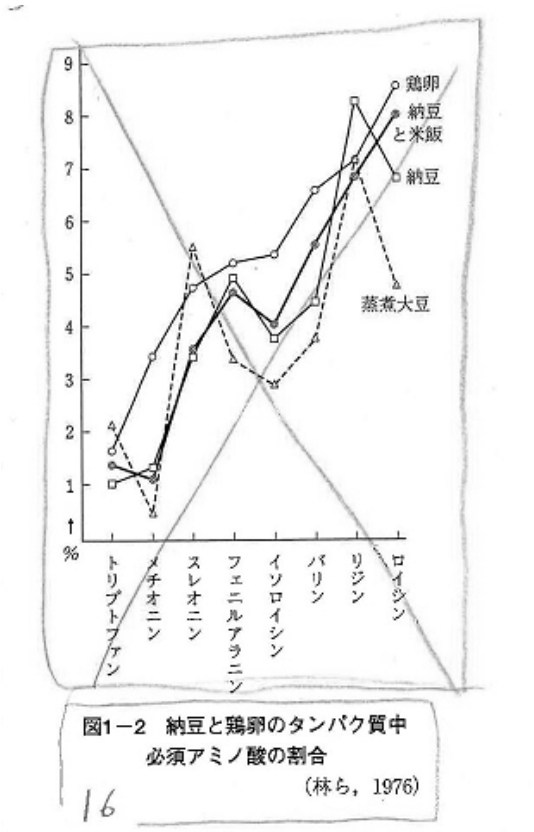


図 4.18: 折れ線グラフ

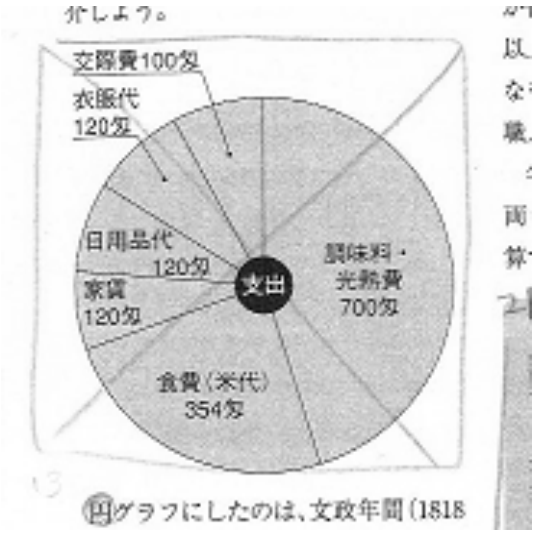


図 4.19: 円グラフ

字列が矢印で結ばれているものに分岐や収束がなく，一方向に読むことができるものは「直線型フローチャート」と呼び,「分岐型フローチャート」の類型としては扱わず，サンプリング対象とする。例えば，図 4.22 は本文中に矢印で結ばれるチャートのような記述があるが，一方向に読むことが十分可能なため，排除対象とはしない。また，図 4.23 についても文字列そのものは一方向に読むことが十分可能であるため，このようなものも排除対象とはしない。

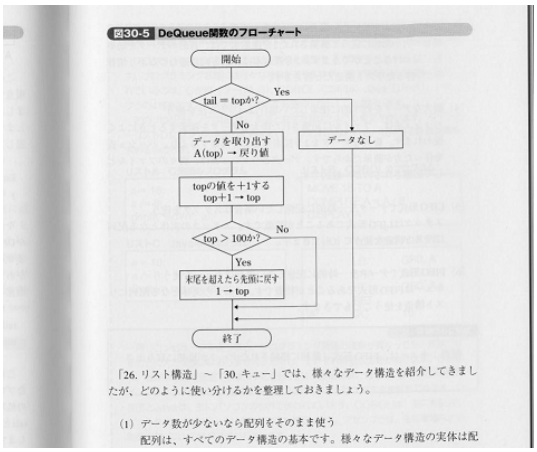


図 4.20: 分岐型フローチャート その 1

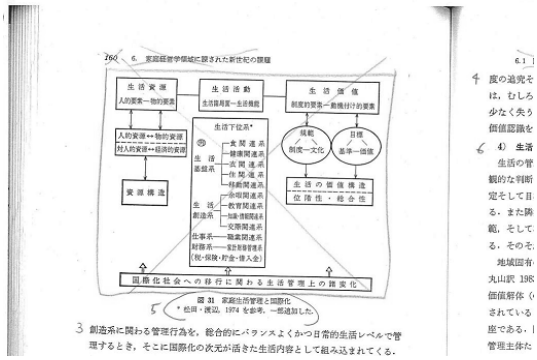
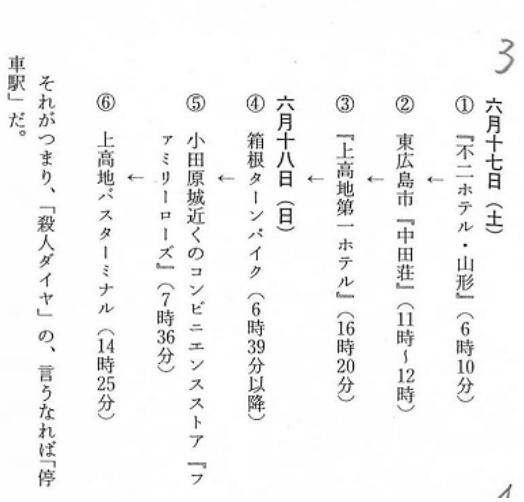


図 4.21: 分岐型フローチャート その 2

表

図 4.24 に示すような「行列見出しを備えた表」を「表」の典型と考える。このような表は先述の通り，文字列を一方向に読むことができない。そのことを根拠に文字列が図式化されている「フィギュア」の類型の 1 つと考える。

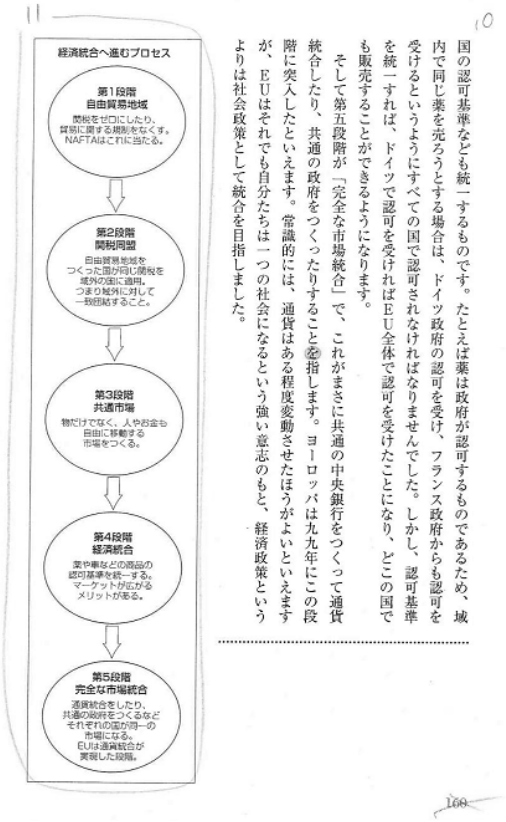
しかしながら，3.3.3 で「一見フィギュア本体に見える要素であっても，その内部にある言語表現を一方向に読み進めることができれば，フィギュア本体とは見なさず，排除の対象とは



六月十七日（土）
① 『不二ホテル・山形』（6時10分）
← ② 東広島市『中田荘』（11時～12時）
← ③ 『上高地第一ホテル』（16時20分）
← 六月十八日（日）
④ 箱根ターンパイク（6時39分以降）
← ⑤ 小田原城近くのコンビニエンスストア『ファミリーローズ』（7時36分）
← ⑥ 上高地バスターミナル（14時25分）

それがつまり、「殺人ダイヤ」の、言うなれば「停車駅」だ。

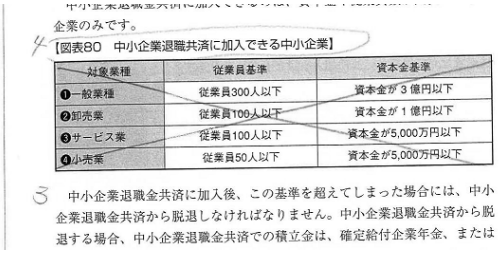
図 4.22: 直線型フローチャート その 1



国の認可基準なども統一するものです。たとえば薬は政府が認可するものであるため、域内で同じ薬を売ろうとする場合は、ドイツ政府の認可を受け、フランス政府からも認可を受けるというようにすべての国で認可されなければなりませんでした。しかし、認可基準を統一すれば、ドイツで認可を受ければEU全体で認可を受けたことになり、どこの国でも販売することができるようになります。

そして第五段階が「完全な市場統合」で、これがまさに共通の中央銀行をつくって通貨統合したり、共通の政府をつくったりすることを指します。ヨーロッパは九九年にこの段階に突入したといえますが、EUはそれでも自分たちは一つの社会になるという強い意志のもと、経済政策というよりは社会政策として統合を目指しました。

図 4.23: 直線型フローチャート その 2


企業のみです。

【図表80　中小企業退職共済に加入できる中小企業】

対象業種	従業員基準	資本金基準
❶一般業種	従業員300人以下	資本金が３億円以下
❷卸売業	従業員100人以下	資本金が１億円以下
❸サービス業	従業員100人以下	資本金が5,000万円以下
❹小売業	従業員50人以下	資本金が5,000万円以下

中小企業退職金共済に加入後、この基準を超えてしまった場合には、中小企業退職金共済から脱退しなければなりません。中小企業退職金共済から脱退する場合、中小企業退職金共済での積立金は、確定給付企業年金、または

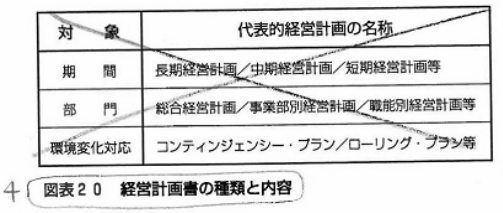
図 4.24: 行列見出しを備えた表

しない」ことに注意が必要であり，その根本には「印刷紙面上に現れた文字列は，それが現代日本語として読み進められる限り，できるだけサンプルとして収録する」という姿勢があると述べた。このことがもっともよく問題になるのは,「表」の認定においてである。

サンプリングする紙面には,「表」，あるいは「表のようなもの」が数多く出現する。それらのうち，一方向に読むことが十分可能である文字列が，ただ罫線で囲んであるだけで，なおかつ,「図表」などと明記されている場合がある。しかし，それらは「図表」とは認定せず，積極的にサンプル構成要素とすべきものであると考える。逆に,「図表」という明記はなく，場合によっては本文中に入り込んでいるようなものでも，一方向に読み進めがたい，図式化された文字列は,「表」と認定し，積極的に排除要素と指定すべきものであるとも考える。

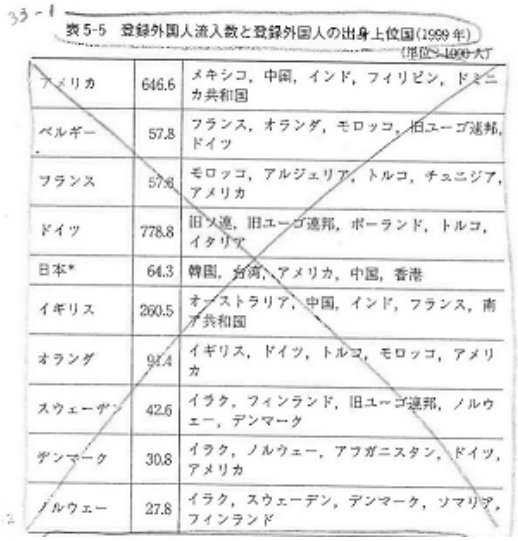
サンプリングする紙面には，典型的な「行列見出しを備えた表」ではない「表のようなもの」が数多く出現し，その判断はしばしば難しい。3.3.3 で既に「行列見出しを備えた表」と「2 列から構成される表」については説明を行った。以下,「表」として認めるものと，認めないものとの判断基準とその適用について，詳細に説明する。

列見出しを備えた 2 列の表，もしくは 3 列以上の表　「表」の典型である,「行列見出しを備えた表」に近く，ほぼ「表」と認めているのは,「行列見出し」のうち，少なくとも「列見出しを備えた 2 列」のもの（図 4.25）である。また,「3 列以上」であるもの（図 4.26）も便宜的に「表」と認めている。以上のものは，少なくとも「列見出し」を備えているということで，あるいは，少なくとも「3 列」はあるということで，図式化された「表」であると考える。


対　象	代表的経営計画の名称
期　間	長期経営計画／中期経営計画／短期経営計画等
部　門	総合経営計画／事業部別経営計画／職能別経営計画等
環境変化対応	コンティンジェンシー・プラン／ローリング・プラン等

図表２０　経営計画書の種類と内容

図 4.25: 列見出しを備えた 2 列表



資5-5 登録外国人流入数と登録外国人の出身上位国(1999年)

(単位:1000人)

アメリカ	646.6	メキシコ，中国，インド，フィリピン，ドミニカ共和国
ベルギー	57.8	フランス，オランダ，モロッコ，旧ユーゴ連邦，ドイツ
フランス	57.0	モロッコ，アルジェリア，トルコ，チュニジア，アメリカ
ドイツ	778.8	旧ソ連，旧ユーゴ連邦，ポーランド，トルコ，イタリア
日本*	64.3	韓国，台湾，アメリカ，中国，香港
イギリス	260.5	オーストラリア，中国，インド，フランス，南ア共和国
オランダ	97.4	イギリス，ドイツ，トルコ，モロッコ，アメリカ
スウェーデン	42.6	イラク，フィンランド，旧ユーゴ連邦，ノルウェー，デンマーク
デンマーク	30.8	イラク，ノルウェー，アフガニスタン，ドイツ，アメリカ
ノルウェー	27.8	イラク，スウェーデン，デンマーク，ソマリア，フィンランド

図 4.26: 3 列表

一方向に読み進められる表　逆に，たとえ紙面上に「図」や「表」と明記されていても，たとえ周りが罫線で囲まれていたとしても,「一方向に読み進められるもの」は「図」や「表」とは認めない。例えば，次の図 4.27 のようなものである。

列見出しを備えない 2 列の表　よって，問題になるのは,「列見出しを備えない 2 列」の場合である。この時に留意しなければならない点は,「本文」中に多々用いられる，いわゆる「箇条書き」との異同である。「箇条書き」は，往々にして，連番，記号，マーク，項目名などの「ラベル」と，項目内容の「アイテム」の「ラベル＋アイテム」の形を取るものであるが，それと「列見出しを備えない 2 列」との差異はあまり大きくないと言える。

罫線で囲んだ「箇条書き」の例として図 4.28 を示す。先の図 4.27 で引いた例と，文字列を一方向に読むという点において，差がないことが確認できる。

そこで,「本文」とは形式的にも文脈的にも区別される「列見出しを備えない 2 列」があった場合，その右列の属性によって,「表」か否かを判断する。通常は,「列見出しを備えない 2 列」は「箇条書き」の「ラベル＋アイテム」であると考え,「表」とは認めずサンプリング対象とする。例えば，次に示す図 4.29 のようなものである。

「列見出しを備えない 2 列」:右列非現代日本語　「列見出しを備えない 2 列」の場合，右列が英語など非現代日本語である場合は，サンプリング対象とはしない。それは,「ラベル＋アイテム」という考え方を「列見出しを備えない 2 列」に適用することによって可能である。

「箇条書き」の「ラベル＋アイテム」においては「アイテム」が主たるサンプル構成要素と考えられる。よって,「アイテム」が排除対象であれば,「ラベル」がたとえサンプリング対象

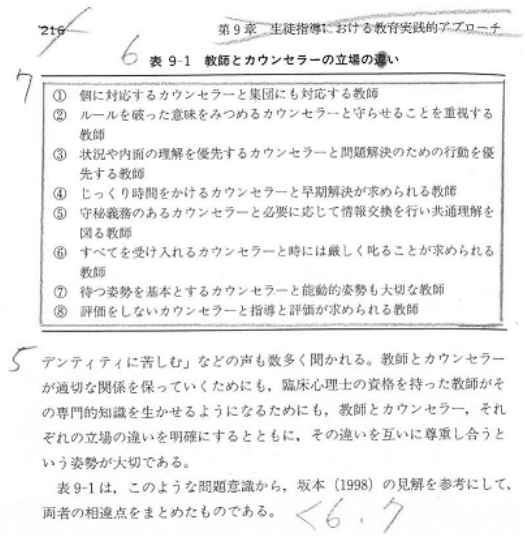
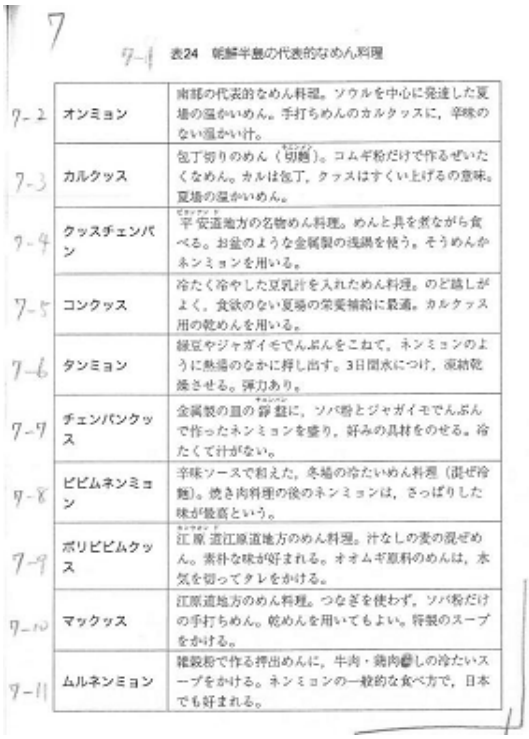


図 4.27: 罫線で囲まれた「一方向に読み進められるもの」

1. 母集団の定義
2. 抽出枠，抽出方法の決定
3. 抽出単位，標本数の決定
4. 母集団のリスト化
5. 標本抽出

図 4.28: 罫線で囲まれた「箇条書き」



7

7-1 表24 朝鮮半島の代表的なめん料理

7-2	オンミョン	南部の代表的なめん料理。ソウルを中心に発達した夏場の温かいめん。手打ちめんのカルクッスに，辛味のない温かい汁。
7-3	カルクッス	包丁切りのめん（切麺）。コムギ粉だけで作るぜいたくなめん。カルは包丁，クッスはすくい上げるの意味。夏場の温かいめん。
7-4	クッスチェンバン	平安道地方の名物めん料理。めんと具を煮ながら食べる。お盆のような金属製の浅鍋を使う。そうめんかネンミョンを用いる。
7-5	コンクッス	冷たく冷やした豆乳汁を入れためん料理。のど越しがよく，食欲のない夏場の栄養補給に最適。カルクッス用の乾めんを用いる。
7-6	タンミョン	緑豆やジャガイモでんぷんをこねて，ネンミョンのように熱湯のなかに押し出す。3日間水につけ，凝結乾燥させる。弾力あり。
7-7	チェンバンクッス	金属製の皿の錚盤に，ソバ粉とジャガイモでんぷんで作ったネンミョンを盛り，好みの具材をのせる。冷たくて汁がない。
7-8	ビビムネンミョン	辛味ソースで和えた，冬場の冷たいめん料理（混ぜ冷麺）。焼き肉料理の後のネンミョンは，さっぱりした味が最高という。
7-9	ポリビビムクッス	江原道江原道地方のめん料理。汁なしの麦の混ぜめん。素朴な味が好まれる。オオムギ原料のめんは，水気を切ってタレをかける。
7-10	マックッス	江原道地方のめん料理。つなぎを使わず，ソバ粉だけの手打ちめん。乾めんを用いてもよい。将製のスープをかける。
7-11	ムルネンミョン	雑穀粉で作る押出めんに，牛肉・鶏肉醤しの冷たいスープをかける。ネンミョンの一般的な食べ方で，日本でも好まれる。

図 4.29: 列見出しを備えない 2 列 その 1

であっても,「ラベル」と「アイテム」両方を排除対象とする，という基準を設けている。「ラベル＋アイテム」形式の実を担う「アイテム」部分を排除するのであれば,「ラベル」だけを読ませる意味はもはやないと考えるためである。この考え方を,「列見出しを備えない 2 列」の場合にも適用する。

次の 4.2.2 で詳述するが，非現代日本語は「第 3 層」に属するものであり，ブロック形式であれば排除対象となる。よって，右列が非現代日本語であれば，右列は一種のブロック形式であるため排除指定をする。この時,「ラベル＋アイテム」と同様に，右列を排除する場合は，同時にその左列も排除する。つまり，結果的に「列見出しを備えない 2 列」全体を排除対象とする。

そこで，作業の効率化と単純化を図るために,「列見出しを備えない 2 列」で右列が非現代日本語であれば，収録すべき文字列を含まない「表」であると判断して全体を排除対象することとしている。図 4.30 に,「ラベル＋アイテム」の「アイテム」が英語の場合も,「列見出しを備えない 2 列」の右列が英語の場合も，結果的に同じように全体が排除対象となるイメージ図を示す。

「列見出しを備えない 2 列」:右列イラスト・記号　「ラベル＋アイテム」の「アイテム」がイラストであれば非現代日本語の場合と同じく「アイテム」が排除対象（イラストは先述のと

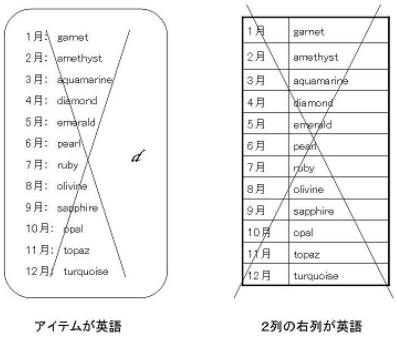


図 4.30: 列見出しを備えない 2 列：右列英語

おり「第 2 層」に属する）であるため,「ラベル」ごと排除対象となる。これに対し,「アイテム」が記号であれば「ラベル」ごと排除対象となるという基準は設けていない（記号はカウント対象文字種ではないが，サンプル構成要素である)。しかし，実際には，イラストと記号との境は曖昧であり，明確な判別はしがたい。このため,「列見出しを備えない 2 列」の右列がイラスト，記号，いずれの場合も，全体を「表」と捉え，排除対象と考えることとした。その典型の 1 つには，交通標識や地図記号などを並べて，図示するようなものがある。また，笑顔などのイラストや★などの記号を並べて，何らかの評価を示すようなものがある。図 4.31 に右列★で評価を表す例を示す。

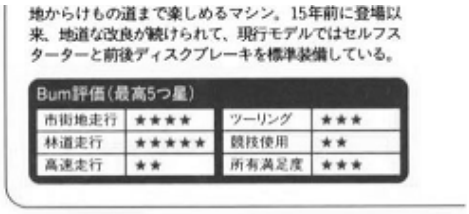


図 4.31: 列見出しを備えない 2 列：右列記号

「列見出しを備えない 2 列」:右列数値　次に，右列が数値であるものを考える。「ラベル＋アイテム」の「アイテム」が数値であれば「ラベル」ごと排除対象となるという基準は設けていない（数値を主に表す「数字」はカウント対象文字種であり，かつ，サンプル構成要素である)。しかしながら，イラストや記号で何らかの評価を表すものと，それらを「3」や「5」といった数値で表す場合は，連続して捉えられるものであると考える。また，例えば，アンケート集計結果における，右側の数値が「○人」「○%」といったものは，円グラフなどのグラフで表されるようなものが「列見出しを備えない 2 列」として表現されたものと考えることができる。グラフは，先に述べた通り,「フィギュア」の類型であり,「第 2 層」に属する排除対象

である。

以上の二つの視点から，右列が数値である場合は「表」と捉え，全体を排除対象とする。

なお,「列見出しを備えない 2 列」の右列に現れやすい数値には，他に，為替のレート表示や，材料の分量表示などがある。それらは，イラストや記号の延長でも，グラフの延長でも捉えられないタイプではある。しかし，イラストや記号の延長で考える場合やグラフの延長で考える場合において，右列が数値であれば排除対象となる類型がある，という事実をもって，作業の効率化と単純化を図るために，右列が数値である場合をひとくくりにし，全体を「表」と認めてよい場合の条件の 1 つとして定めている[1] 。

例えば，図 4.32 は，列見出しを備えない 2 列の右列が数値であることによって「表」と認定した例である。

■選挙費用収支報告

中沢勝子　山梨県玉穂町議選（1999.10.3）

収入総額	100,000
内訳　寄付	0
自己資金	100,000
その他	0
支出総額	76,106
内訳　人件費	0
家屋費	0
通信費	0
交通費	0
印刷費	9,187
広告費	18,900
文具費	2,299
食糧費	31,020
休泊費	0
雑費	14,700
収支差引	23,894
法定選挙費用	1,410,300

131

図 4.32: 列見出しを備えない 2 列：右列数値

「実質一方向に読み進められるもの」　表に関わる記述の最後に,「表」と認定しない「一方向に読み進められるもの」の例を補足するものとして，実質そうであるものについて説明する。例えば，図 4.33 は 6 列あるように見える。しかしながら，0,1,2 … n と，一方向に読み進められるものであり，図 4.26 の「3 列表」と異なり「表組みである」という意識が極めて薄い。むしろ，図 4.27 同様「一方向に読み進められるもの」と言えよう。よってこのようなものも「表」とは認定せず，サンプリング対象とする。

「表組みである」という点に関して，もう 1 例を挙げておく。

[1] ただし，例えば電話番号はいわゆる数値ではないため，電話番号が右列にあるものは，右列が数値という類型には当てはめず，サンプリング対象となる文字列として認めている。「数字」に関しては 3.2.2 で述べたページ単位での回避がもうひとつの例外となる。

日本十進分類法（NDC）の 1 次区分

「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」

図 4.33: 実質一方向で読み進められる文字列のもの

図 4.34 であるが，一見 2 列表であるが，行が分割されている。このような場合は「表組みである」という意識が強いと見て，サンプリング対象外とする。

国語	現代文
数学	数 I
	数 II
	数 III

図 4.34: 2 列であっても行が分割される表

複合的なフィギュア

最後に，複数のフィギュアが上位のキャプションによって統括されている場合について述べる。図 4.35 のような場合は，最上位のキャプション「fig. 人口の推移」のみをサンプリング対象とする。下位のキャプション（「表 1-1...」「図 1-1...」）については，サンプリング対象としない。これは，全体を複合的な図表と捉えるという見方に基づく。

以上で述べてきた各種の認定は，原則的なものである。実際には，前後のレイアウトとも照らし合わせ，サンプル内で統一がとれるよう総合的に判断している。例えば，同一ページに複数の表が提示され，うち 1 つが排除対象とならない体裁の場合には，それらの表を統一して扱うべきか否かを（複数名で）検討し，統一して排除対象とする場合がある。ただし,「行列見出しを備えた表」のように，一方向に読むこと自体が困難なものについて，統一の視点からサンプリング対象にすることはない。

4.2.2　第 3 層の要素

第 3 層に配置されるのは，現代日本語として表されない部分，すなわち数式，化学式，コンピュータ言語，古典語，外国語などがブロック形式で地の文とは切り離された形で出現した場合である。これらの要素は現代日本語の範囲からは外れるものと見なし，サンプリングの対象から排除する。そこで，以下の原則を定める。

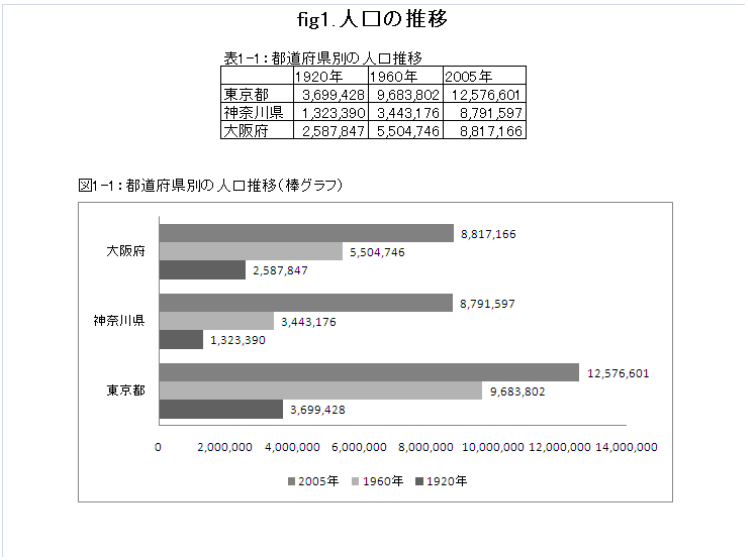


図 4.35: 複合的な図表

排除原則 5： 第 3 層の要素，すなわち，非現代日本語（数式，化学式，コンピュータ言語，古典語，外国語など）は，サンプル範囲から排除する。

なお，ブロック形式ではなく，本文中にインライン形式で数式や外国語が入り込んでいる場合は，本文の一部を構成する要素と見なし，この第 3 層で除外する対象には含めない。図 4.36 〜図 4.38 に，非現代日本語の表現がブロック形式で現れている例を示す。

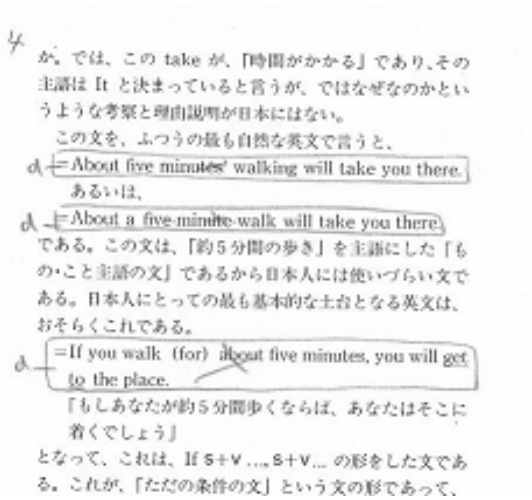

が。では、この take が、「時間がかかる」であり、その主語は It と決まっていると言うが、ではなぜなのかというような考察と理由説明が日本にはない。
この文を、ふつうの最も自然な英文で言うと、
a. =About five minutes' walking will take you there.
あるいは、
b. =About a five-minute walk will take you there.
である。この文は、「約5分間の歩き」を主語にした「もの・こと主語の文」であるから日本人には使いづらい文である。日本人にとっての最も基本的な土台となる英文は、おそらくこれである。
a. =If you walk (for) about five minutes, you will get to the place.
「もしあなたが約5分間歩くならば、あなたはそこに着くでしょう」
となって、これは、If S+V ..., S+V... の形をした文である。これが、「ただの条件の文」という文の形であって、

図 4.36: ブロック形式の英語

例えば，出典の明示がない場合に，古典語か否かを判断することが難しいことがある。ま

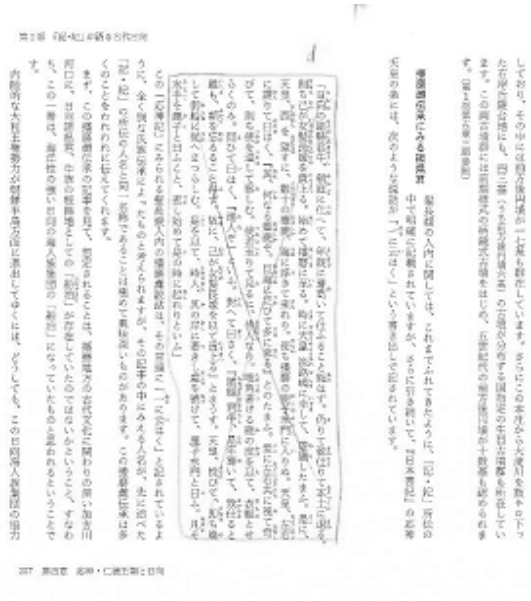


図 4.37: ブロック形式の古典語

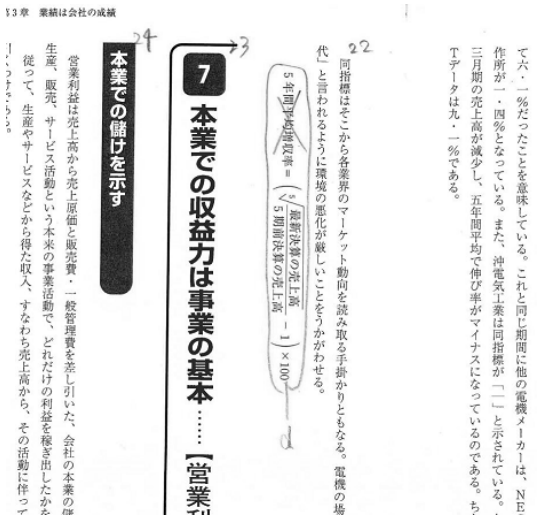


図 4.38: ブロック形式の数式

た，ほとんどが非現代日本語であるブロックにおいて，現代日本語が一部分混じるような場合，それが（ ）内であれば，従要素と見なし主体は非現代日本語であると判断してよい，といった細かな作業基準の検討が必要になる。例えば，英語に（ ）が付く例を図 4.39 に，古典語に（ ）が付く例を図 4.40 に示す。

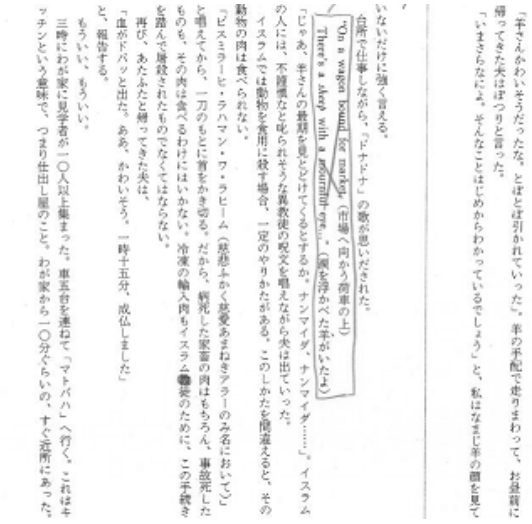


図 4.39: ブロック形式の英語に（ ）が付くもの

4.3 サンプル構成要素の確定と入力順の指定

以上までで示した排除要素の指定によって，第 4 層「現代日本語の範囲」以上の要素が残ったことになる。これらはサンプルとして抽出する対象要素であり，おおむね次のいずれかに分類される。

- 見出し
- キャプション
- 本文
- 注（脚注・後注）

さて，次に必要となるのは，これらの要素を電子テキストとして収録するための入力順を指示する作業である。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確認することである。

以下，4 種類のサンプル対象要素について，最終的にサンプリング対象要素を確定させ，テキスト収録の入力順を認定する際に留意する点を述べる。

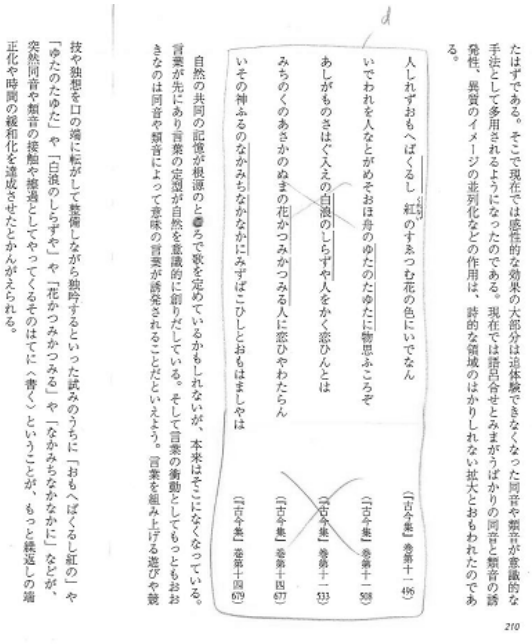


図 4.40: ブロック形式の古典語に（ ）が付くもの

4.3.1 「見出し」

可変長サンプル範囲を考える際には文章構造の把握が必要であり，そのために「見出し」の認定は欠かせないものである。それに加え，収録後のテキストの構造化の際にも見出しは重要な意味を持つ。BCCWJ において，収録した後のテキストの構造化の際には，章節構造を明示させるため,「本文」を統括する「見出し」の認定が重要になるのである。よって，サンプル作成の最終段階において,「見出し」を再確認し，その「見出し」を収録テキストの頭に配置するよう，入力順の指示を工夫する必要がある。

例えば，図 4.41 と図 4.42 は，同じサンプルの別紙面の画像である。図 4.42 が見出しのあるページであり，図 4.41 はその次の見開きページの左上に示されていた図 4.42 の部分の拡大である。この「柱」や，目次のタイトル表示などを参考にし，図 4.42 では,「見出し」部分の入力順の指示をしている。

図 4.41: 見出しの入力順指示に留意するものの「柱」部分

図 4.42: 見出しの入力順指示に留意するものの「見出し」部分

また，構造化における,「見出し」認定の必要性の高さにより，通常は収録対象外となる非現代日本語であっても，それが「見出し」相当と認められれば，その部分を収録対象とする（柏野ほか (2009) を参照)。

例えば，図 4.43 に示すように,「見出し」が非現代日本語（英語や古典語）であれば，それをそのまま「見出し」として入力するよう指示する。また，図 4.44 では，テレビのイラストの中の「7」という章番号に当たる文字を入力するよう指示しているが，このように,「見出し」の文字列がイラストの中に入っている場合は，その文字列を取り出して入力するよう指示する。

4.3.2 「本文」

テキスト収録という観点において,「見出し」同様に,「本文」の確定と入力順の指示においても，論理構造の把握が重要である。「見出し」の認定の際には，まとまったテキスト部分を統括するものを探すが,「本文」の確定には，逆に,「見出し」として認定したものが統括する範囲を再確認することになる。

例えば，図 4.45 に示すようなガイドブックのような紙面は，大小さまざまなレベルの「見出し＋本文」のまとまりが複数存在する。それらまとまりが分かるよう，入力順を指示する必要がある。

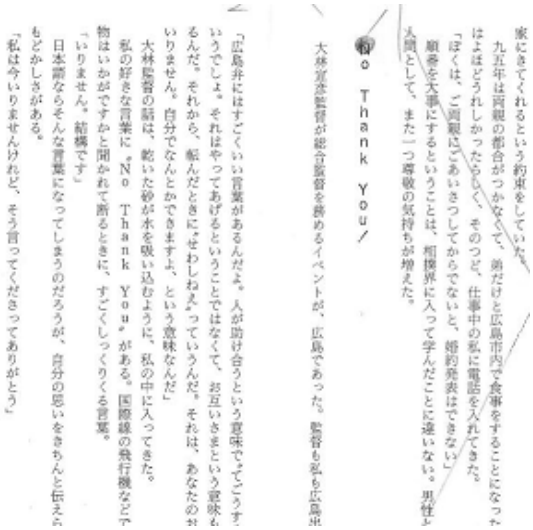


図 4.43: 「見出し」が英語

図 4.44: 「見出し」の文字列をイラストの中から取り出すもの

図 4.45: 入力順指示に留意するもの

入力順の指示で留意するものの例として，ほかに,「コラム」がある。その内容や形式に応じて，道なりに入力すべきか，適当な章節末に位置づけて入力すべきかの指示が必要になる。また，章節末の位置を指示する際には，コラムが本文のどの階層構造に位置づけられるものであるかの判断も必要になる。例えば，図 4.46 は，コラムも各節も同じ階層にあると見て，コラムはそのまま道なりに入力することを指示した例である。

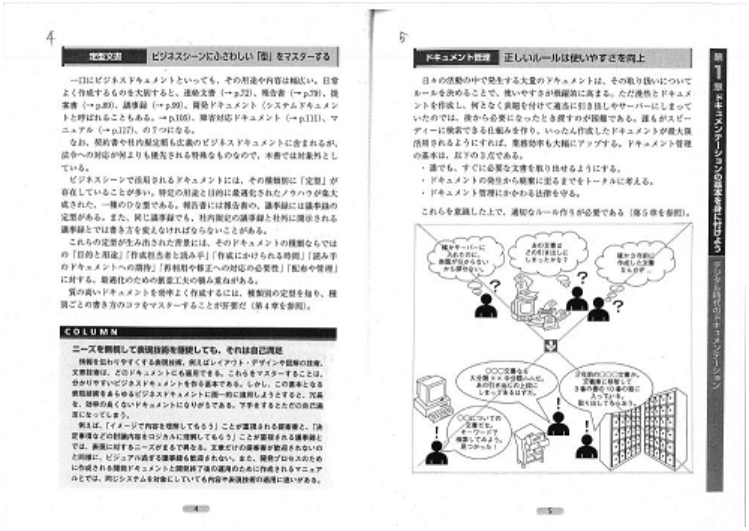


図 4.46: 「コラム」を道なりに入力するもの

一方，次の図 4.47 は，コラムが「I 章 2 節 (1)」の本文途中に挿入されているものである。挿入箇所では道なりには入力しがたいため,「I 章」「2 節」「(1)」のうちいずれかの章節末での

入力指示が必要である。この例では，内容，及び他の章節にある同様の「コラム」との形式の比較などにより，この書籍においてコラムは「節」の階層に位置づけられるものと判断し,「2 節」末で入力するよう，指示をしたものである。

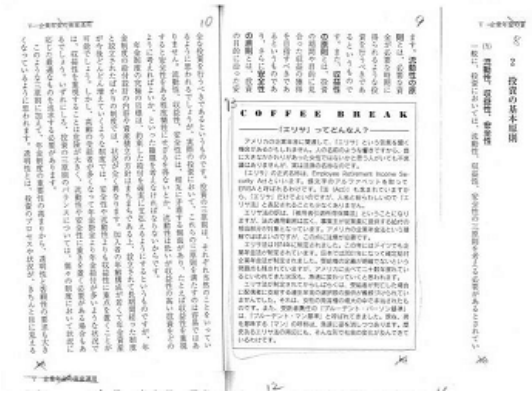


図 4.47: 「コラム」を章節末に入力するもの

4.3.3 「キャプション」

柏野ほか (2009) で述べたように,「キャプション」について「章節構造」が包含する意味内容を言語的に補足するものと認め，サンプリング対象とする。写真に伴う「キャプション」の典型例を図 4.48 に，表に伴う「キャプション」の典型例を図 4.49 に示す。

「キャプション」の入力順は,「フィギュア」の種類に関わらず，他の「本文」などとあわせて道なりに入力するか，あるいは,「本文」などのまとまりを一通り入力し終えた後にまとめて入力するか，いずれか適当と判断する方を指示する。

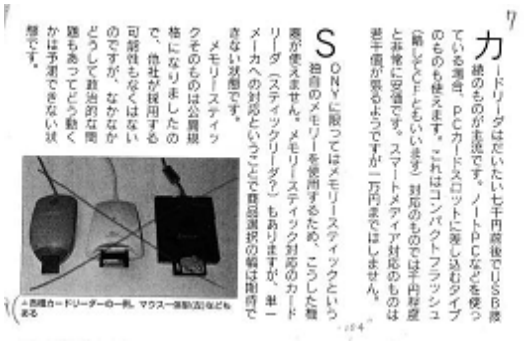


図 4.48: 写真の「キャプション」

「カタログ」のような紙面の場合，写真・イラストと，それらに対する解説が大量に配置される構成を取る。ここで，写真やイラストを解説する文字列，すなわち「キャプション」に相

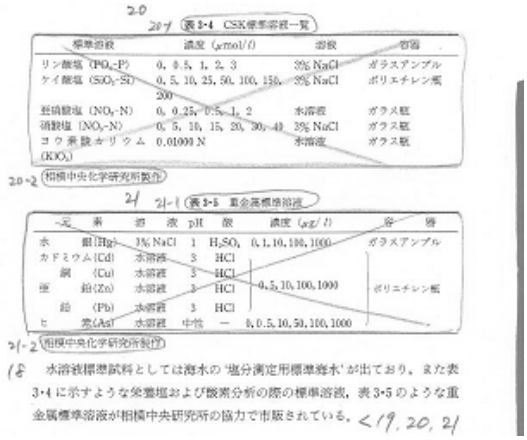


図 4.49: 図表の「キャプション」

当する文字列は，章節構造を構成する要素と考え,「本文」として認定する。例えば，次の図 4.50 のようなものである。また，その次に示す，図 4.51，図 4.52 も，写真やイラストの「キャプション」に相当する文字列が「本文」として認定される例である。

図 4.50: カタログ様の紙面で写真に伴う「キャプション」相当文字列が「本文」であるもの

4.3.4 「注」

柏野ほか (2009) で述べたように,「注」について,「章節構造」が包含する意味内容を言語的に補足するものと認め，サンプリング対象とする。注には，基本的には同一ページにある「脚注」と呼ばれるものと，巻末や章節末にある「後注」と呼ばれるものとがある。いずれも，注

図 4.51: 写真に伴う「キャプション」相当文字列が「本文」であるもの

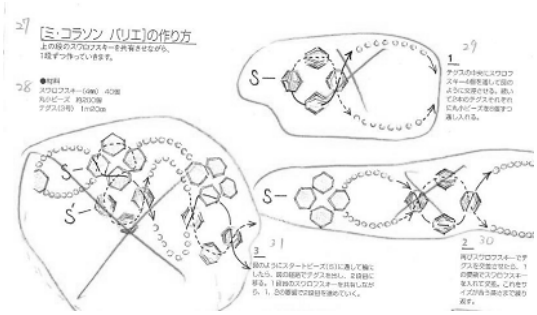
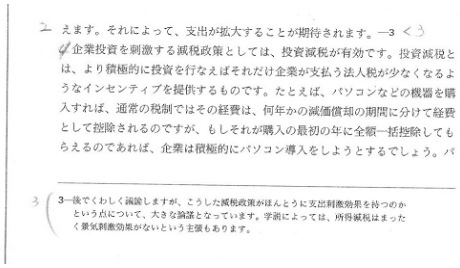


図 4.52: イラストに伴う「キャプション」相当文字列が「本文」であるもの

マーカーのある形式段落の最後で入力するよう，指示することとしている。例えば，次の図 4.53 では「脚注 3」を注マーカー「—3」のある形式段落末で入力するよう指示している。

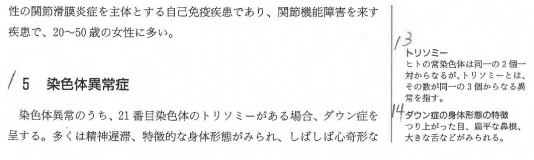

えます。それによって、支出が拡大することが期待されます。—3
企業投資を刺激する減税政策としては、投資減税が有効です。投資減税とは、より積極的に投資を行なえばそれだけ企業が支払う法人税が少なくなるようなインセンティブを提供するものです。たとえば、パソコンなどの機器を購入すれば、通常の税制ではその経費は、何年かの減価償却の期間に分けて経費として控除されるのですが、もしそれが購入の最初の年に全額一括控除してもらえるのであれば、企業は積極的にパソコン導入をしようとするでしょう。パ

3 3—後でくわしく論議しますが、こうした減税政策がほんとうに支出刺激効果を持つのかという点について、大きな論議となっています。学派によっては、所得減税はまったく景気刺激効果がないという主張もあります。

図 4.53: 脚注：マーカーあり

一方，注マーカーがない場合もある。その場合は，太字，下線，フォント差などから，あるいは，形式的な手がかりがなくても，語句の対応が容易に分かる場合には，対応のとれる形式段落の最後で入力するよう指示する。

例えば，図 4.54 は，注マーカーのない傍注の例である。語句の対応から，傍注を形式段落の最後に入力するよう，指示しているものである。


性の関節滑膜炎症を主体とする自己免疫疾患であり、関節機能障害を来す疾患で、20〜50 歳の女性に多い。

5 染色体異常症

染色体異常のうち、21 番目染色体のトリソミーがある場合、ダウン症を呈する。多くは精神遅滞、特徴的な身体形態がみられ、しばしば心奇形な

13 トリソミー
ヒトの常染色体は同一の 2 個一対からなるが、トリソミーとは、その数が同一の 3 個からなる異常を指す。

14 ダウン症の身体形態の特徴
つり上がった目、扁平な鼻根、大きな舌などがみられる。

図 4.54: 脚注：マーカーなし

対応が取れない時は，それが脚注の場合は，章節末にまとめて入力するか，可変長サンプルの最後でまとめて入力するよう，指示する。それが後注の場合は，可変長サンプルの範囲内に後注が存在する場合は道なりに入力するよう指示するが，可変長サンプルの範囲外にあれば，範囲外のものとして，収録対象とはしない。

第5章　固定長サンプルの抽出

本章の概要：　本章では，書籍から固定長サンプルを抽出する一連の手続きとその基準について示す。「言語的な構造のまとまり」を抽出する可変長サンプルとは異なり，固定長サンプルはサンプル抽出基準点から一律「1,000 文字」を抽出するものである。この 1,000 文字を抽出する際の判断基準や問題点などについて，具体例を挙げながら示す。

5.1　固定長サンプルを抽出する原理

はじめに，書籍から固定長サンプルを抽出する原理について述べる。第 1.3 で提示した固定長サンプルの定義を，一部再掲しておく。

> **固定長サンプル**：「固定長サンプル」は，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，その文字を始点として 1,000 文字目までの範囲を抽出するサンプルである。母集団（＝推計された総文字数）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。

固定長サンプルを抽出する際に前提となるのは，繰り返し述べてきている通り，**母集団を構成する文字が 1 次元に配列されている**ということである。すなわち，出版 SC では 1 文字目から 48,539,925,351 文字目までが，図書館 SC では 1 文字目から 47,877,656,072 文字目までが，それぞれ 1 列に配置されていると仮定する。その中からランダムに指定された 1 文字をサンプル抽出基準点として，そこから 1,000 文字目まで一続きの文字列を取り出すことが，固定長サンプルを抽出する原理であると言える。

実際の印刷紙面から「1,000 文字」の範囲を取り出す際に問題となるのは，**紙面上に構成された文字をどのような順序で抽出していくか**，そして**そのうちどの文字を 1,000 文字としてカウントする対象とするか**，という 2 点である。前者については，基本的には，可変長サンプルの抽出の中で示した「入力順」（4.3 節）による順序を用いる。すなわち，入力順として指定された連番通りに文字が 1 列に配置されていると考えるのである。一方，後者については，1,000 文字としてカウントする文字を特定する必要があるが，これはすでに「書籍の階層的な成立に関わる要素」で挙げた「第 6 層」で定義されている。すなわち,「現代日本語の範囲」のうち,「仮名，漢字，数字，アルファベット」で表記された「見出し，本文，注，キャプション」

である。これらの要素を，サンプル抽出基準点から入力順に 1,000 文字分を取得すれば，固定長サンプルが得られることになる。

なお，固定長サンプルの取得においては，可変長サンプルで問題にした紙面構成上の役割の区別，あるいは「理想範囲」「完結構造」という把握の仕方は不要である。抽出した 1,000 文字の中に見出しやキャプションなど異なる役割を持つ要素が含まれていても，第 6 層に所属する要素である限り，それらが固定長サンプルの中で区別されることはない。あるいは，1,000 文字の中に異なる著者が執筆した複数の理想範囲が含まれていても，それぞれの著者によってサンプル範囲が区切られることはない。あくまでも 1 次元に並んだ文字列によって構成される母集団から 1,000 文字分を切り出した範囲が固定長サンプルの本質であり，紙面構成上の役割の違いや著者の違いによって，その中に含まれる文字の扱いが影響を受けるわけではない。

固定長サンプルを抽出する原理は，以上に述べたとおりである。以下では，固定長サンプルを構成する 1,000 文字としてカウントされる文字の種類について，その詳細を示す。

5.2　固定長サンプルを構成する文字種

5.2.1　カウント対象とする文字の定義

固定長サンプルを構成する文字は，上述の通り,「仮名，漢字，数字，アルファベット」で表記された「見出し，本文，注，キャプション」に相当する要素である。このうち，カウント対象にする文字の種類について，さらに詳細に示す。

固定長サンプルの 1,000 文字としてカウント対象にする文字を，原則的に，以下のように定義する。

固定長サンプルの 1,000 文字としてカウントする文字

- 原則的に，現代日本語の文章に含まれる語を構成する音に対応付けられた文字を，固定長サンプルの 1,000 文字としてカウントする。

その上で，カウント対象とする文字の種類を分類すると，以下のようになる。

- **カウント対象にする文字**

  a. 仮名文字（平仮名・片仮名・変体仮名）

  b. 漢字（簡体字・繁体字も含む）

  c. 準漢字（例：踊り字（「々」「ゝ」）

  d. 長音記号（「ー」）

  e. 数字（アラビア数字・ローマ数字）

  f. アルファベット（ローマ字・ギリシャ文字）

また，これらの文字が丸で囲まれている場合や（例：㊙，㊤，①，㉑），四角や☆など，丸以外の記号類で装飾されている場合も，カウント対象に含める。

逆に，固定長サンプル・可変長サンプルを構成する要素として入力はされるものの，1,000文字としてカウントしない文字には，以下のようなものがある。

- **カウント対象にしない文字**

  a. 句読点類（「,」「。」「,」「.」「…」「・」「：」「；」）
  b. 疑問符，感嘆符（「？」「！」...）
  c. 括弧類（「(」「「」「{」「＜」「《」「【」「〔」...）
  d. 線記号類（「-」「‐」「—」「～」...）
  e. 矢印類（「→」「↑」「⇒」「⇔」...）
  f. 算術記号類（「＋」「－」「×」「÷」「＝」「±」「≠」「>」「∴」「１／２」「１.２」...）
  g. 通貨・単位記号類（「£」「＄」「￥」「%」「‰」 ...）
  h. 音符類（「♪」...）
  i. 絵文字
  j. その他記号類（「○」「▲」「□」「◎」「※」「＃」「&」「☆」...）

5.2.2 カウント対象とする文字の判断基準

文字種の別による判断基準

ある文字をカウント対象とするかしないかの判断については，その文字が具体的な文脈の中で担う意味・用法はなるべく考慮せず，先に定めた文字種から判断することを原則とする。例えば，以下のような文脈で用いられる文字は,「現代日本語の文章に含まれる語を構成する音に対応付けられた文字」とは言い難いが，文字種という観点から，カウント対象とする。

- 箇条書きにおける項目にラベルとして用いられる「１.」「２.」「３.」「い.」「ろ.」「は.」,「前：」（「前川」という発言者の略記）など。この場合,「１」「２」「３」「い」「ろ」「は」「前」をそれぞれサンプリング対象とする。
- 「０４２－５４０－４３００」など，独立した言語表現（の大半）が数字のみから構成されている場合など。この場合,「０４２５４０４５５０」それぞれをカウント対象とする。

逆に，以下に挙げるような文字は「現代日本語の文章に含まれる語」の一部を表記していると見られるが，記号であると見なし，カウント対象とはしない。

- 「そうだよね～。」「と～っても」のように,「～」が長音記号として用いられている場合など。この場合,「～」はカウント対象としない。

- 「モーニング娘。」「藤岡弘、」「つんく♂」など，カウント対象でない文字が，語の一部を構成すると思われる場合。この場合,「。」「,」「♂」はカウント対象としない。
- 「Q &A」「果汁 100 %」のように，カウント対象でない文字が読みと強く対応している場合でも,「&」「%」はカウント対象としない。

漢字・外字の種類による判断基準

「漢字」については，JIS 第一水準〜第四水準の全漢字，および，簡体字，繁体字をカウント対象とする。なお，電子テキストとして入力できない漢字には，入力時に「**=**」が充てられるが，カウント対象文字の代用である限りは，これらもカウント対象とする。逆に，絵文字などにも「**=**」が充てられるが，カウント対象文字の代用でない限りは，これらはカウント対象としない。

ローマ数字のカウント方法

ローマ数字の場合，1 から 10 までの数「I」「Ⅱ」「Ⅲ」「Ⅳ / Ⅲ」「V」「Ⅵ」「Ⅶ」「Ⅷ」「Ⅸ」「X」），あるいは一定数を表わす数（「L」「C」「D」「M」）は，すべて 1 文字分としてカウントする。11 以上の数字（「Ⅺ」「Ⅻ」...）は，それを構成する 2 文字以上を開いた形でカウントする。例えば,「Ⅺ」は 2 文字,「Ⅻ」は 3 文字としてカウントする。

外国語の表現

カウント対象となる文字種で構成される外国語の表現は，1 文字ずつをカウント対象とする。例えば，行中（インライン）に「Ｉ　ＬＯＶＥ　ＹＯＵ♡」と書かれていた場合，8 文字としてカウントする。

一方，カウント対象とならない文字種で構成される外国語の表現は，一律，カウント対象とはしない。キリル文字，アラビア文字，ハングルなどの文字は，それが用いられる文脈にかかわらず，一律カウント対象から除外する。

組み文字

組み文字については，入力対象となり得る文字種が組み合わさってできているものについては，その文字数分をカウント対象とする。「ﾄﾙ」ならば 2 文字,「ﾘｯﾄﾙ」ならば 4 文字分をカウントする。

アスキーアート

行中（インライン）にアスキーアートが出現した場合，そのアスキーアートを構成する文字の種類によってカウント対象とするか否かを判断する。例えば,「ありがと（＾＾）」という例の「(＾＾)」は記号のみであるためカウント対象には含めず，全体を 4 文字としてカウントする。一方,「ゴメン m（__ __）m」の場合は「m」をカウント対象文字として数え，全体を 5 文字としてカウントする。

なお，ブロック単位のアスキーアートはフィギュアを構成するものと見なし，丸ごとカウント対象外とする。

5.3 可変長サンプルと固定長サンプルの相互関係

サンプル抽出基準点を指定した後，可変長サンプルと固定長サンプルの 2 つを同時に取得する。この際，サンプル抽出基準点の位置により，両者の関係性には以下のようなパタンが存在する。

included： 固定長サンプルのすべてが可変長サンプルに含まれる形

overflow： 可変長サンプルの最後から固定長サンプルが飛び出す形

separated： 可変長サンプルと固定長サンプルとが一切重なっていない形

「included」は，固定長サンプル，すなわちサンプル抽出基準点から 1000 文字目までの範囲が可変長サンプルの内部に包含される場合である[1]。「overflow」は，サンプル抽出基準点は可変長サンプル内部に包含されるが，終端部が可変長サンプルから超過するパタンである。

「separated」は，3.4.2 で述べた「冒頭 1 万字」の場合において，まれに生じるパタンである。この場合，可変長サンプルの内部にサンプル抽出基準点が包含されないことになるが，サンプル抽出基準点を可変長サンプル内に移動させるのではなく，別途固定長サンプルを取得する。

これら 3 つの関係を図示すると，図 5.1 のようになる。

[1] 2.7 で述べたように，固定長サンプルはサンプル抽出基準点が含まれる文の文頭，およびサンプル抽出基準点から数えて 1,000 文字目が含まれる文末までが合わせて抽出される。

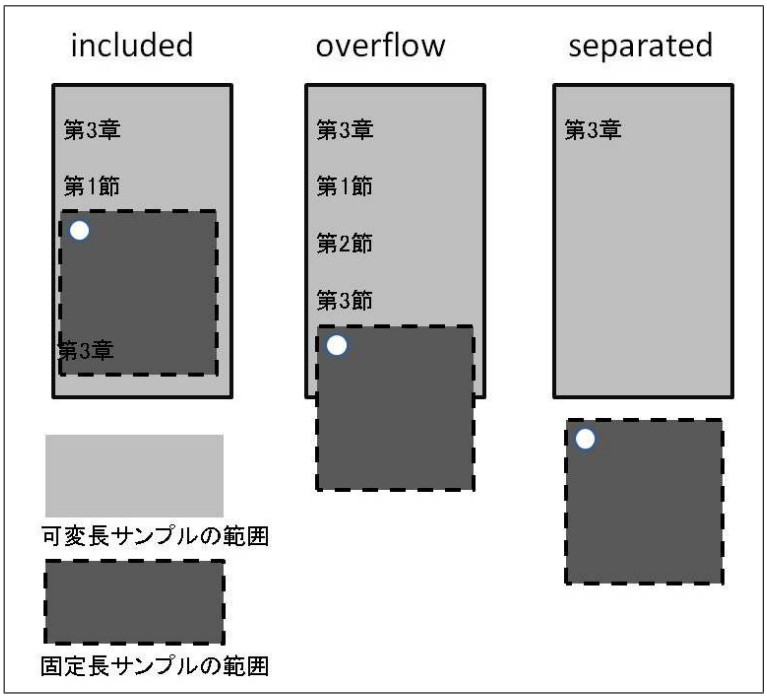


図 5.1: 可変長サンプルと固定長サンプルの相互関係のパタン

第III部

雑誌・新聞における
サンプリングの原理と運用

第6章　雑誌におけるサンプリング

本章の概要：　第II部では，書籍を対象として，書き言葉の構造をどのように把握するか，その中からどの部分をサンプリングの対象とするか，という点について述べてきた。書き言葉の構造を捉えるための見方は，雑誌からサンプルを取得する場合においても有効である。しかしながら，雑誌というメディアが持つ特有の問題により，書籍の場合とは異なる問題点が生じることがある。ある記事が「広告」であるかをどのように判別するか，ある紙面の理想範囲をどのように定めるか，といった問題である。

そこで本章では，第II部で示したサンプリングの原理を踏まえた上で，雑誌のサンプリングにおいて問題となる点とその判断基準について示すことにする。

6.1　雑誌の特徴と紙面構成

ここでは雑誌という定期刊行物が持つ一般的な特徴として，以下の3点を挙げておく。

- 書籍に比べて速報性が高いこと
- 多種多様な記事[1]が写真やイラストなどとともに複雑に構成されていること
- 「広告」のページが多く含まれること

雑誌の大半は「週刊」や「月刊」のように短い刊行周期を持っており，新聞ほどではないものの，速報性の高い出版メディアであると言える。かつその中には，論説記事，エッセイ，料理のレシピ，映画の情報，洋服の値段，広告など，実に多くの種類の記事が（場合によっては所狭しと）詰め込まれていることが多い。1人の著者が時間をかけてじっくりと1冊の雑誌を編集する，ということはまずない。むしろ，多人数の執筆者による小さな記事が集まって1冊の雑誌が構成されていると言ってよい。

このような雑誌の特性を踏まえた上で，雑誌の印刷紙面から文章をサンプリングしようとする場合，いくつか注意を要する点がある。以下では順に，雑誌に特有の問題とそれに対する判断基準を挙げていく。

[1] ここでは便宜的に，著者の別や目次により他の部分と区別される一定範囲を「記事」とする。

6.2　サンプリングの対象外とする要素の認定

3.3 節で示したように，書き言葉の成立を階層的に把握し，サンプリングの対象を絞り込んでいくという見方は，雑誌のサンプリングの場合にも当てはまる。以下では，その絞り込みの過程において問題となる事例をいくつか記述する。

6.2.1　「付録」の扱い

雑誌を対象にサンプリングを実施する場合,「付録」の扱いを決めなければならない。雑誌の付録には,「別冊付録」や「付録 CD-ROM」など，冊の中に含まれないものもあれば,「綴じ込み付録」のように冊の中に綴じ込まれたものもある。

28 ページの図 3.2 に挙げた書き言葉の階層構造に従えば，付録は第 0 層に位置づけられる要素であり，基本的にサンプリングの対象外となる要素である。別冊付録のように冊の中に含まれない付録は，当然ながら，サンプリングの対象とはならない。

ところが「綴じ込み付録」のように雑誌本体に綴じ込まれている場合，それが付録か否かの判断に迷う場合がある。例えば，冊の中ほどに，その雑誌の判型よりも小さいサイズの判型で綴じ込まれた「記事」がある場合，その部分は冊本体のようにも付録のようにも見える。そこに切り取り線が付いていれば，別冊付録が綴じ込まれていると考えることもできる。

そこで，雑誌の中のある部分が付録であるか否かを，図 6.1 のようなフローにより判断した。

なお，該当部分の範囲に対して，冊本体の通しページとは別のページ番号が与えられている場合は，その範囲は冊本体に含まれていないものと見なし，サンプリングの対象外とする。

6.2.2　「広告」の扱い

次に,「広告」の判別基準について示す。広告は第 1 層に位置づけられる要素であり，サンプリングの対象外となるものである。ところが，広告は先に見た「付録」とは異なり，一般紙面と同様，冊に綴じ込まれている。そのため，その範囲が広告なのか否かを判別する必要がある。そこで，広告をいくつかの類型に分け，以下のように考えることにした。

他社広告

他社広告とは，当該の雑誌の出版社以外の会社が出した広告（のように見える記事）を指す。他社広告であるか否かを判別する手がかりとして,「広告（のように見える記事）が目次に記載されているか否か」という点を重視した。それが目次に記載されていない場合は「広告」と認定し，サンプリングの対象から除外した。

a. 目次または当該部分に「付録」と書いてあるか

→ 書いてあれば,「付録」と認定し，サンプリングの対象外とする。
→ 書いていなければ，b. へ進む。

b. 当該部分に切り取り線が付いているか

→ 付いていれば,「付録」と認定し，サンプリングの対象外とする。
→ 付いていなければ，c. へ進む。

c. 判型は表紙と同じか

→ 同じならば,「付録」とは認定せず，サンプリングの対象とする。
→ 大きい，または小さければ，d. へ進む。

d. 目次または当該の記事の中に「保存版」などの文字列があるか

→ あれば,「付録」と認定し，サンプリングの対象外とする。
→ なければ,「付録」とは認定せず，サンプリングの対象とする。

図 6.1: 雑誌の「付録」を認定するフロー

また，目次に記載がある場合でも，目次そのもの，または当該の記事の中に「PR」「広告」「広告特集」などの文字列がある場合は，その記事を「広告」と認定し，サンプリング対象から除外した。

さらに，上記以外でも，非常に広告らしく見える記事がある。例えば，ある特定の会社のある特定の商品を紹介する記事で，記事の末尾に「お問い合わせ先」としてその会社の連絡先が記載されているような場合である。このような場合，そこにノンブルが振られていなければ「広告」と認定し，サンプリングの対象から除外した。逆に，ノンブルが振られており，かつ目次で他の通常の記事と同じように大きく扱われている場合には「広告」とは認定せず，サンプリングの対象とした。目次での扱いが小さい場合は，当該記事自体の体裁や内容，および目次での扱いが小さい他の記事が広告であるかも併せて参照し,「広告」と認定するか否かを個別に判断した。

自社広告

自社広告とは，当該の雑誌の出版社および関連会社が出した広告（のように見える記事）を指す。これらについても，上記の「他社広告」を判別する基準を用いて「広告」と認定されたものはサンプリング対象外とした。

ただし,「当該の出版社（関連会社）が運営している通信販売の記事」「当該の出版社（関連

会社）が運営しているチケット販売の記事」などについては，広告か否かを内容的に判断することが難しかった。そこで，その記事が主に文章で構成されている場合には，基本的にサンプリング対象とした。

自誌広告

自誌広告とは，その雑誌自身の広告（のように見える記事）を指す。これらについても,「他社広告」を判別する基準を用いて「広告」と認定されたものはサンプリング対象外とした。

ただし,「レポーター募集」「キャッチフレーズ募集」の記事などは，広告か否かを内容的に判断することが難しかった。そこで，その記事が主に文章で構成されている場合には，基本的にサンプリング対象とした。

6.3　理想範囲の認定

6.3.1　「著者」による理想範囲の認定

先にも述べたとおり，雑誌は多人数の執筆者による小さな記事の集合によって構成されていると言える。ここで，3.4 節で述べた「理想範囲」を認定することを考えると，雑誌の場合，著者の違いによって理想範囲を認定できることが多い。

雑誌の記事の冒頭や末尾には,「取材・文 前田洋二」「構成 稲元正子」などのように役割とともに著者が明示されていることが多い。また，著者が目次に記載されていることもしばしばである。「理想範囲」とは「当該文書のうち，同一著者によって同一テーマのもとに書かれた範囲の全体」であるから，このような著者表示があれば，比較的容易に理想範囲が認定できることになる。

一方，著者の認定に迷うケースとして,「談話」風の記事がある。実際には雑誌社の編集部が取材し，後に書き起こして編集したと推測される内容が，いかにも取材を受けた相手（談話の語り手）が文章を書いたような体裁を取っている記事のことである。この場合の著者をどのように認めるか，という問題がある。

この問題については，その記事が「取材を受けた相手が文章を書いた」ような体裁になっており，かつ第三者が書いた文章が他に存在しない場合には，その取材の相手を著者として認めることにした。一方，その記事の主たる書き手が書いた文章（地の文）が存在する場合は，談話の部分は「引用」されたものとして考えることにした。

ただし，実際には談話は多様な形で現れるものであり，それぞれのケースで判断をする必要があった。

6.3.2 「目次」による理想範囲の認定

著者表示がどこにもない場合は，目次に従って理想範囲の認定を行なう。基本的に，雑誌の記事はすべて目次に記載されているものと考える。逆に，目次に記載のない記事は，原則的にサンプリング対象外とする。

その際，理想範囲を認定する判断に迷う事例がいくつかあった。以下，事例の種類ごとに挙げる。

記事の終端に関する問題

目次に記載のある記事の最後部に，一定量の文章が配置されていることがある。この文章が，直前の記事に含まれるか否かの判断によって，目次に記載のある記事の一部としてサンプリングの対象となるか，目次に記載のない記事としてサンプリングの対象外となるかが決まることになる。

この場合，最後部に配置された一定量の文章が「広告」と認められるか否かによって判別する。6.2.2 節で示した基準によって広告であると判断された場合はサンプリングの対象外として扱い，そうでないと判断された場合は基本的にサンプリングの対象とする。なお，直前にある記事の中に当該の個所への言及がある場合は，言及内容を確認の上，サンプリングの対象とする可能性がある。

記事の入れ子構造

ある 2 つの記事が隣接しているとき，それらが並列的に並んでいる場合を「並列関係」と呼び，片方の記事の内部にもう 1 つの記事が入り込んでいる場合を「主従関係」と呼ぶ。後者は，上位の記事が下位の記事を内包する「入れ子構造」の形である。ここで，サンプル抽出基準点が下位の記事に当たった場合は，下位の記事のみを取得する。一方，サンプル抽出基準点が上位の記事に当たった場合は，下位の記事を含めた上位の記事全体を取得する。

ここで，以下のような場合に，2 つの記事を入れ子構造として扱ってよいのか，あるいは 2 つの独立した記事と扱うべきか，判断に迷うことになる。

- 紙面上では入れ子構造になっている上位の記事と下位の記事が，目次では別立てになっており，並列的な 2 つの記事として扱われているように見える。

この場合，むしろ肝要なのは，紙面上にある文字列によって，2 つの記事の間に上位・下位の関係を作ることが構造上可能かどうかである。

例えば，2 つの記事を統括する見出しがある場合，それらは並列した 2 つの記事としても，入れ子構造の記事としても，認めることができる。この場合は，目次での扱いや記事の内容をもとに，並列関係とするか入れ子構造とするかを判断する。

一方で，2 つの記事を統括する見出しが紙面上にない場合は，原則として 2 つは独立した別の記事であると見なす。

「特集」の扱い

雑誌では「特集」が組まれることが多い。特集とは，同一テーマに基づく複数の記事の集合によって構成される範囲である。この「特集」全体も 1 つの理想範囲と見なされるが，実際には各記事ごとに著者がいる場合が多いため，下位に複数の理想範囲を含む上位の理想範囲という，入れ子構造が生じていることになる。

なお，特集の途中に，特集とは関係のない別の記事が入り込んでいる場合がある。この場合，特集としての統括が崩れているものと見なし，個々の記事を最大の理想範囲とした。ただし，途中に入り込んでいる要素が広告の場合は，この限りではない。

6.4　入力順序の指定

理想範囲が確定したら，サンプル抽出基準点をもとに可変長サンプル・固定長サンプルを抽出するが，ここで文字の入力順序の指定について述べておく。

例えば，文芸雑誌に掲載されている「小説」がサンプリング対象となった場合，書籍における「小説」同様，一方向に読み進めていく通りに入力順序を指定すればよい。一方，特にカタログ状の体裁を持つファッション誌などでは，紙面上のあちこちに文章が散らばっている場合が多く，どこから読み進めていけばよいのか迷うことがある。入力順序の指定は，特に 1,000 字の固定長サンプルにとってその中身を決める重要な問題であり，ある程度の原則があることが望ましい。

そこで，以下の 3 点を原則として，入力順序を定めることにした。

1. 基本的に，紙面の右上から左下へ向けて入力順序を付けていくものとする。
2. 「トピックが近いものは連続する」と考え，内容が近いものは入力順序でも隣接させる。
3. 枠線などで囲まれて取り立てられた記事がある場合，入力順序はその紙面の最後に回すこともある。

第7章　新聞におけるサンプリング

本章では，新聞のサンプリングにおいて問題となる点とその判断基準について示す。

7.1　新聞の特徴と紙面構成

新聞という定期刊行物が持つ一般的な特徴として，以下の 3 点を挙げておく。

- 書籍・雑誌に比べて速報性が高いこと
- 小さな記事が大量に集まって全体が構成されていること
- 記事を弁別する手がかりとなる「目次」が存在しないこと

新聞の大半は「日刊」でかつ朝刊・夕刊が分かれており，日々生産され続けるメディアであると言える。6.1 で，雑誌が持つ一般的な特徴として，速報性の高さ，および記事の多様性を挙げたが，新聞は雑誌以上に速報性を有するメディアであり，また，雑誌同様，多人数の執筆者による小さな記事の集合体であると見なすことができる。なお，雑誌と同様，ここでは便宜的に著者やトピックの切れ目により判別された範囲を「記事」と呼ぶ。

さて，新聞の場合，書籍・雑誌と異なり，記事の範囲を認定する手がかりとなる「目次」というものが存在しない。そのため，理想範囲の認定には「著者」および「トピック」という 2 つの判断基準を用いる。

以下では，新聞の理想範囲を認定する際に「著者」および「トピック」がそれぞれどのように関わるかについて述べる。

7.2　理想範囲の認定

7.2.1　「著者」による理想範囲の認定

新聞の場合，多くの記事には著者の氏名が表示されていない。第 6 章で見た雑誌のように，著者または目次から理想範囲を認定することはできないため，7.2.2 節で後述する「トピック」による範囲の認定が必要になる。

しかし，寄稿や連載，書評などといった一部の記事には，著者が表示されているものがある。また，新聞記者の氏名が記載されている場合もある。これらを「記名記事」と呼ぶ。記名

記事の場合，その著者によって書かれた部分を理想範囲として認定する。ただし，雑誌と同様，「談話」風の記事の場合に，著者をどう判断するかという問題がある。これについては，6.3.1 節を参照のこと。

また，記名記事であっても，著者表示の及ぶ範囲が分かりにくい場合がある。例えばスポーツ欄によく見られる類型として，(a) 概況を示した後に，(b) 解説が入る，というものがある。(b) の末尾に著者表示がある場合，この著者は (b) のみに対する記名とするか，(a) と (b) 全体に対する記名とするかの判断が難しい。

この場合，(b) がその新聞の定常的なコラムであれば (a) の下位の記事に当たると判断し，そうでない場合はレイアウトや前後の記事を確認した上で，個別に判断を行なった。

7.2.2　「トピック」による理想範囲の認定

新聞では，多くの場合,「トピック」という概念で記事の範囲を捉える。トピックとは，そこで述べられているテーマのことを指し，同じトピックでまとまっている範囲を取り出して理想範囲と認定するのである。トピックによる範囲は，基本的には，それを統括する見出しを持つ。その見出し（および，ある場合にはリード内で言及された内容）を判断基準として，理想範囲がどこまでかを特定する。

また,「社説」の欄は，通常，2 つのトピックに分かれていることが多い。この場合，社説は 2 つの理想範囲を含むという格好になる。

「面種」の扱い

「面種」とは，日付などが記載されている柱の部分に「社会」「経済」「国際」「生活」などと記載される文言でまとめられる，各ページの種類を指す。新聞の各ページは，それぞれ「社会面」「経済面」「国際面」「生活面」としてのまとまりを持つと考えることもできる。しかし，これらの面種によるまとまりを，一つのトピックして認めることはしない。理想範囲となる範囲は，あくまでも「面」の下に展開されている個別の部分に収まると考え，その上位にある「面」にまでトピックの範囲を拡張することはしない。

「欄」の扱い

「欄」とは，基本的に，ある見出しによって面全体が統括されている範囲，あるいはそれに近い大きさを持つ範囲を指す。すぐ上で，トピックの範囲は面の下に展開されている個別の部分に収まると述べたが，欄は面のすぐ下の構造として位置付けられる，統括する見出しを持つ単位であるため，理想範囲として認めることができる。

ただし，その新聞で定常的に掲載されている欄は「面」に相当するものと見て，欄全体ではなく欄の下位の構造を理想範囲として取得する。その上で，欄名にサンプル抽出基準点が

当たった場合は，欄全体をサンプルとして取得する。また，定常的な欄の中に「連載小説」や「コラム」が一見入り込んでいるように見える場合があるが，これは別記事として考え，欄の内部には含めない。

統括する見出しがない記事

記事の類型によっては，それを統括する見出しが存在しない場合がある。テレビ欄の番組案内や，訃報記事などがこれに該当する。これらはトピックとしてのまとまりの強さを重視し，統括する見出しはないものの，まとめて 1 つの理想範囲とした。

ページの切れ目について

基本的には，見出しの統括する範囲が及ぶのはページの終端までであり，ページが変わるとそこには新たな記事があると考える。ある記事の末尾に「〜面に続く」「社会面に関連記事」などの文言がある場合でも，その先の記事までを取得することはしなかった。

「著者」の優位性

著者表示の存在は，トピックとしてのまとまりに優先する。直感的には同じトピックでまとめられているように見える複数の文章群であっても，それらの文章がそれぞれ著者表示を持っている場合は，その部分が 1 つの理想範囲を構成することになる。

7.3 「広告」の認定

雑誌の場合と同様に，新聞でも本文部分と広告とが混在している。新聞の広告は「全面広告」のような面単位のもの，紙面下部に現れる定常的なもの，そして紙面の中に置かれる小さな広告枠などさまざまであるが，これらはいずれもサンプリングの対象外となる。

全面広告の場合，柱に「広告特集」などとある場合や，紙面中に「広告」「AD」の文字列がある場合は，広告と判断する。また，制作者として「〜新聞広告部」と書かれていたら，広告と判断する。

他社広告

商品を紹介する記事の場合，それを広告と見なす指標がなく，またレイアウト的にも他の記事と区別がない場合は，サンプリング対象外とすることはしない。新譜紹介，新刊紹介，新製品紹介などの記事が，これに該当する。

自社広告

「社会人野球日本選手権・主催：毎日新聞社」や「夏の高校野球・主催：朝日新聞社等」など，その新聞を発行する新聞社が主催する催しに関する案内で，かつレイアウト的にも他の記事と区別がないように提示されている場合がある。これらは，広告として機能しているとも思われるが，サンプリングの対象外とすることはしない。

7.4　入力順序の指定

理想範囲が確定したら，サンプル抽出基準点をもとに可変長サンプル・固定長サンプルを抽出する。ここでも，雑誌の場合と同様，文字の入力順序の指定が問題になる。

新聞の場合，文字や写真・図などが紙面全体を埋めているものの，そこに含まれる記事のレイアウトは常に一定でなく，読む順序が定められているわけではない。そこで，雑誌と同様，以下の 3 点を原則として，入力順序を定めることにした。

1. 基本的に，紙面の右上から左下へ向けて入力順序を付けていくものとする。
2. 「トピックが近い記事は連続する」と考え，内容が近い記事は入力順序でも隣接させる。
3. 枠線などで囲まれて取り立てられた記事がある場合，入力順序はその紙面の最後に回すこともある。

おわりに

最後に，日本語を対象としたコーパス言語学において，均衡コーパスの構築という作業がどのような意義を持つか，その中でサンプリングという作業がどのような役割を占めるか，という 2 点について述べておきたい。

均衡コーパスの歴史は，1959 年の SEU，あるいは 1964 年の Brown Corpus にまで遡ることができる。母集団をジャンルに区分し，層化抽出法に基づいてサンプルを抽出してコーパスを構築するという方法は，すでに 50 年も前から実施されているものであり，見方によっては，古典的な方法論であると言える。その一方で，近年では Web 上に存在する膨大なテキストを自動的に収集してコーパスとして用いる方法論（Web as Corpus）が提案されてきている。また，1990 年代以降に出てきた BOE（Bank of English）をはじめとする「モニターコーパス」では，全体のバランスを均衡させるのではなく，新しいテキストを次々に追加することで，コーパス全体を巨大化させる方針が取られている。均衡コーパスとしての設計よりも，テキストの量が重視される傾向が強まっていると言える。

しかしながら，Brown Corpus や BNC をはじめとする英語の均衡コーパスが，これまでに数多くの研究成果をもたらしてきたのは紛れもない事実であり，現代においてもなお，サンプリングに基づく均衡コーパスの意義が失われているとは言えない。ましてや，現代日本語を対象とした均衡コーパスがこれまで存在していなかった以上，それを設計し構築するという作業そのものが大きな意義を持つ。

また，収録語数を順次追加・拡張していく方針のモニターコーパスとは異なり，綿密に設計された均衡コーパスを一度構築すれば，学界内の言語研究者にとって共通のプラットホームができることになる。すなわち，誰の手元にも同じデータが存在し，それをもとに研究者自身が独自の視点で分析を進めるという状況が生じるわけである。このような環境では，主観的（場合によっては恣意的）な言語研究は排除され，誰が実施しても同じ結果が得られる客観的な分析が進められることなる。かつ，これまでほとんど実施されてこなかった研究結果の追試・検証なども可能になる。

また，ジャンルごと，年代ごと，執筆者ごと，などの詳細かつ正確な区別を書誌情報として持つ均衡コーパスでは，そこに観察される言語事象が，社会的な位相の中のどこに位置づけられるのかを記述することができる。サンプルの属性を表わす書誌情報などを利用することで，社会言語学的なテーマを定量的に扱うことができることになる。Web 上のテキストは書誌情報をほとんど持たないため，この点においては均衡コーパスに明らかな優位性がある。

さらに言えば，現代日本語を対象とした均衡コーパスの設計および構築という作業自体がこれまでに行なわれてこなかった以上，それを実践してその有用性を議論することそのものが，今後のコーパス日本語学における重要な課題になるであろう。どのようなテキストを，どれくらい，どのような方法でサンプリングすれば，より「適切な」均衡コーパスが構築できるのか，その方法論自体を検証することが今後求められることになる。

その検証過程において重要な役割を持つのが，サンプリングの設計である。母集団をどのように定義し，どのように層別し，どのような手続きで言語表現を抽出したのか，その設計過程の全てを詳らかにしておくことが，均衡コーパスとしての評価を決定する上で重要な手掛かりになる。その点において，均衡コーパスの構築におけるサンプリングの役割は極めて重要であり，かつ，その内実を記録として留めておくこともまた重要であると言える。

関連文献

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–』, 特定領域研究「日本語コーパス」平成 18 年度研究成果報告書 (JC-D-06-02), 特定領域研究「日本語コーパス」データ班.

柏野和佳子, 丸山岳彦, 秋元祐哉, 稲益佐知子, 佐野大樹, 田中弥生, 山崎誠 (2008). 『『現代日本語書き言葉均衡コーパス』における書籍サンプルの多様性』, 特定領域研究「日本語コーパス」平成 19 年度研究成果報告書 (JC-D-07-02), 特定領域研究「日本語コーパス」データ班.

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–』, 特定領域研究「日本語コーパス」平成 19 年度研究成果報告書 (JC-D-07-01), 特定領域研究「日本語コーパス」データ班.

佐野大樹, 丸山岳彦, 山崎誠, 柏野和佳子, 秋元祐哉, 稲益佐知子, 田中弥生, 大矢内夢子 (2009). 『語彙密度を利用した『現代日本語書き言葉均衡コーパス』テクスト分類の試み』, 特定領域研究「日本語コーパス」平成 20 年度研究成果報告書 (JC-D-08-02), 特定領域研究「日本語コーパス」データ班.

柏野和佳子, 丸山岳彦, 稲益佐知子, 田中弥生, 秋元祐哉, 佐野大樹, 大矢内夢子, 山崎誠 (2009). 『『現代日本語書き言葉均衡コーパス』における収録テキストの抽出手順と事例』, 特定領域研究「日本語コーパス」平成 20 年度研究成果報告書 (JC-D-08-01), 特定領域研究「日本語コーパス」データ班.

丸山岳彦, 山崎誠, 柏野和佳子, 佐野大樹, 秋元祐哉, 稲益佐知子, 田中弥生, 大矢内夢子 (2011). 『『現代日本語書き言葉均衡コーパス』におけるサンプリングの原理と運用』, 特定領域研究「日本語コーパス」平成 22 年度研究成果報告書 (JC-D-10-01), 特定領域研究「日本語コーパス」データ班.

丸山岳彦, 山崎誠, 柏野和佳子, 佐野大樹, 秋元祐哉, 稲益佐知子, 田中弥生, 大矢内夢子 (2011). 『『現代日本語書き言葉均衡コーパス』に含まれるサンプルおよび書誌情報の設計と実装』, 特定領域研究「日本語コーパス」平成 22 年度研究成果報告書 (JC-D-10-02), 特定領域研究「日本語コーパス」データ班.

コーパス開発センター（サンプリングサブグループ）

山崎 誠　（言語資源研究系准教授、コーパス開発センター（兼））
柏野 和佳子　（言語資源研究系准教授、コーパス開発センター（兼））
丸山 岳彦　（言語資源研究系助教、コーパス開発センター（兼））
佐野 大樹　（コーパス開発センタープロジェクト特別研究員）
田中 弥生　（コーパス開発センタープロジェクト特別研究員）
秋元 祐哉　（コーパス開発センタープロジェクト奨励研究員）
大矢内 夢子　（コーパス開発センタープロジェクト奨励研究員）
稲益 佐知子　（派遣社員、マンパワー・ジャパン株式会社）

国立国語研究所内部報告書（LR-CCG-10-01）

『現代日本語書き言葉均衡コーパス』におけるサンプリングの原理と運用

平成23年2月25日
執筆者　丸山岳彦　山崎誠　柏野和佳子　佐野大樹
秋元祐哉　稲益佐知子　田中弥生　大矢内夢子
発行者　大学共同利用機関法人 人間文化研究機構 国立国語研究所
〒190–8561 東京都立川市緑町10番地の2
電話　042（540）4300（代表）

ISBN 978-4-906055-00-5

国立国語研究所