『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

国立国語研究所内部報告書（LR-CCG-08-01）

『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

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

平成21年3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

国立国語研究所内部報告書（LR-CCG-08-01）

『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

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

平成21年３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

『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）』は，1976 年から 2005 年までの 30 年間に書かれた現代日本語の書き言葉を収録する，1 億語規模のコーパスである。このコーパスの設計にあたり，我々はいくつかの調査を行ない，コーパスデザインのための基礎資料として用いてきた。これらの調査結果，およびそこから設計されたコーパスデザインの詳細については，2007 年度および 2008 年度に刊行した報告書（丸山・秋元, 2007, 2008）の中で述べてきた。

本報告書で述べるのは，上記の設計に基づいてサンプリングを実施するための，理論的背景と作業方針の詳細である。報告書は，第 I 部と第 II 部から構成される。第 I 部では，サンプリングの理論的背景と BCCWJ で実施するサンプリングの基本方針，そして実際の印刷紙面から書き言葉をサンプリングするための基本的な考え方を提示する。第 II 部では，第 I 部で示した方針に基づいてサンプリング作業を進めていくにあたり，どのような事例をどのように処理しているか，その個別事例と判断基準を挙げていく。

先に刊行してきた報告書，そして本報告書により，コーパスデザインの基盤となる調査結果とそこから得られる構成比率などの詳細な設計図，そしてサンプリングの理論と手順が示されることになり，コーパスの設計からサンプルの取得に至る一連の作業手順が明らかになる。

コーパスの設計にかかる基礎調査とその集計は，丸山岳彦，秋元祐哉，山崎誠，前川喜久雄らが中心となって担当した。実際の書き言葉を対象としたサンプリングの実施手順や基準の作成については，柏野和佳子，稲益佐知子，田中弥生，秋元祐哉，佐野大樹，大矢内夢子らが中心となって担当した。実際のサンプリング作業は，安部達雄，市原乃奈，遠藤直子，久古直，佐藤真奈美，志賀里美，田口久美子，立花幸子，趙恩英，長門美帆子，服部紀子，三浦智子，保田祥らが，これを助けた。

『現代日本語書き言葉均衡コーパス』のサンプリング作業では，以下の各機関・各社よりご協力をいただいています。記して感謝申し上げます。

国立国会図書館，日本図書館協会，立川市中央図書館，東京都立中央図書館，
東京都立多摩図書館，東京都立日比谷図書館，八王子市図書館，
横浜市中央図書館，一橋大学附属図書館，自治大学図書室，
株式会社 学習研究社，株式会社 小学館，ヤフー株式会社　　（順不同）

第I部

BCCWJ構築におけるサンプリングの方針と基準

第1章　理論的背景

丸山岳彦

均衡コーパス（Balanced Corpus）とは，ある言語（の部分集合）の特徴や性質を知るために，その言語（の部分集合）の多様性をできるだけ忠実に反映するようにバランスよくサンプルを収集して構築される言語資源である。この点において，均衡コーパスを構築する作業とは，標本調査を行なうために標本を抽出する作業と基本的に軌を一にする。

本章では，本報告書の前提として，書き言葉を対象としたサンプリングにおける理論的背景について述べる。また，書き言葉を対象として標本調査を実施するためには，どのような点を考えておかなければならないかについて示す。

1.1　標本調査とは何か

標本調査（Sample Survey）とは，ある集団の一部を取り出して調査し，その結果から集団全体の特徴や性質を推定しようとする調査方法である。国民の意識や社会の動向を調査するために行なわれる社会調査や，工業製品の品質管理，薬品の疫学調査のために行なわれる品質検査など，日常のさまざまな場面において統計情報を得るために実施されている。これに対して，集団の全てについて調査を行なう方法を**全数調査**または**悉皆調査（Complete Survey）**と呼ぶ。ある時点における国民の人口・性別・年齢・就業などを調べる国勢調査は，全数調査に分類される。

全数調査は，調査対象となる集合が大きくなるほど，時間・予算・手間などのコストが増えることから，その実施は現実的には不可能になる。そこで対象の一部を標本として取得し，その調査結果から調査対象の全体を推定するという，標本調査の方法が採られることになる。このような考え方は，**推測統計学（Inferential Statistics）**に属する。

このことを，言語調査に置き換えて考えてみよう。例として，以下の 2 つの調査を考える。

- 『源氏物語』で用いられている語彙の調査
- 現代日本語の書き言葉で用いられている語彙の調査

前者であれば,『源氏物語』全体の語数が調査可能な規模である以上，全数調査が可能である。事実，古典作品の用語・用例索引として,『万葉集語彙索引』『源氏物語語彙用例総索引』などの語彙索引が作られており，文学作品の計量的な分析に役立てられている。

一方，後者を全数調査で調べるためには,「現代日本語の書き言葉」の全てを収集し，そこに現れた語を全て数え上げなければならない。「現代日本語の書き言葉」の総体が捉えどころのない以上，このような全数調査は言うまでもなく不可能である。この場合，調査対象となる「現代日本語の書き言葉」を定義し，そこから標本を抽出し，その調査結果をもとに調査対象全体の特徴や性質を推定する，標本調査が選択されることになる。

1.2　標本調査の方法

適切な標本調査を実施するためには，以下の手順を踏む必要がある。

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

以下では，言語調査の中でも「現代日本語の書き言葉調査」を例として，各手順をどのように考えればよいかについて述べる。

1.2.1　母集団の定義

標本調査の実施にとってまず必要とされるのは，調査の対象となる**母集団（Population）**を定義することである。母集団が定義されるためには，その前提として，母集団に含まれる要素が有限個の集合で，それらが数量的に把握できるものであることが求められる。さらに，後の標本抽出のために，母集団を構成する要素はすべて明示的にリスト化されなければならない。

ここで，現代日本語の書き言葉を調査することを考えよう。この場合，母集団を定義するためには，そもそも何を書き言葉と見なすのかをまず考えなければならない。また，その書き言葉を数量的に把握し，明示的な形でリスト化しなければならない。

一口に「現代日本語の書き言葉」と言っても，その外延は実に漠然としか捉えることができず，それ自身を母集団として定義することはできない。これに対して，例えば,「書籍」「新聞」「インターネット上の文書」などの媒体（メディア）による限定を加えることにより，母集団として見なし得る調査対象が得られる。さらに，例えば,「2001 年から 2005 年までに日本国内で発行された書籍に含まれる文章全体」「2003 年に発行された朝日新聞に含まれる新聞記事の全体」といった具合に限定を加えていくことにより，母集団を数量的に定義できる形に近づいていくことになる。

1.2.2　抽出枠，抽出方法の決定

次に，**抽出枠（Frame）**と**抽出方法**を決定する。標本調査で用いられる抽出枠の設定には，母集団の構成要素をすべて一律に扱い，その全体から標本を抽出する「**単純抽出**」と，母集団の構成要素を相互排他的な複数の**層（Stratum）**に分割し（層別し），各層から標本を抽出する「**層化抽出**」とに大別される。また，抽出方法には,「**系統抽出法**」「**無作為抽出法**」「**2 段抽出法**」などがあり，調査の目的や母集団の状態などに応じて選択されることになる。

さらに層化抽出の場合，各層から抽出する標本のサイズをどのように決めるかという問題がある。母集団を構成する各層のサイズに比例して各層から取得する標本サイズを決める方法を

「**比例割当**」と呼ぶ。また，ある変数によって母集団の各層ごとに標準偏差を計算し，その分散が大きい層から標本を多く抽出する方法を「**最適割当**」と呼ぶ。

1964 年にアメリカで公開された Brown Corpus や 1990 年代にイギリスで作られた British National Corpus などの代表的な均衡コーパスでは，書き言葉を “Informative Prose” と “Imaginative Prose” とに大別し，その下位に複数のジャンルを設定して層別を行なっている。Brown Corpus の場合，15 のジャンルに対して実際の出版量に応じて重み付けがされ，その重みに応じて各ジャンルから収集するサンプルの数が決められている。これは，比例割当による層化抽出に相当する。特に均衡コーパスの場合，ジャンル間における言語的特徴の分析に用いられることが多いと想定されるため，適切な層別および層化抽出が行なわれていることが望ましい。

1.2.3 抽出単位，標本サイズ，標本数の決定

次に，**抽出単位**，**標本サイズ**および**標本数**を決定する。個人の意見を問う世論調査であれば個人が抽出単位となり，世帯の収入を問う調査であれば世帯が抽出単位となる。また，それらの抽出単位を合計どれだけ収集するか，その標本サイズを決定し，その標本サイズを確保するだけの標本数を算出する。適切な標本サイズは，調査の目標精度および母集団の分散（**母分散**）に依存して決まる。

書き言葉が調査対象の場合であっても，調査目的に応じて妥当な抽出単位が決まることになる。例えば，書き言葉の文字を調査対象とする場合，最も理想的な抽出単位は「文字」である。同様に，語彙調査にとって最も理想的な抽出単位は「語」である。しかしながら,「文字」や「語」を抽出単位とした均衡コーパスの構築は，現実的ではない。むしろ，一定量の文章（テキスト）を抽出単位として構築されることが通常である。例えば，1959 年にイギリスで作られた SEU コーパスでは，一律 5,000 語の抽出単位が合計 100 万語分収集されている。また，Brown Corpus では一律 2,000 語の抽出単位が合計 100 万語分収集されている。

標本数をどれだけ確保すれば母集団の状態を過不足なく調査できるかという点については，特に言語調査の場合（中でも文法研究，コロケーション研究などを調査・研究の目的とする場合），極めて難しい問題を含む。仮に母集団が有限個の集合であったとしても，そこに含まれる言語現象（あるいは，調査対象である所与の言語における言語現象）の種類が予測できない以上，どれだけの標本サイズがあれば当該の調査・研究に十全かを予測すること自体が困難であるためである。

1.2.4 母集団のリスト化

以上の項目が決まれば，実際の標本抽出を実施するための準備として，母集団を構成する要素を記載した母集団リストを作成する。前述のように，母集団は有限個の集合として数量的に把握できるものでなければならない。既存のリスト，例えば住民基本台帳や選挙人名簿，事業所名簿のような既存の名簿を母集団リストとして使える場合もあるが，独自に母集団リストを作成しなければならない場合も多い。また，無作為抽出法を採用する場合，母集団リストに含まれる抽出単位に通し番号を付し，それらを乱数を用いてあらかじめランダマイズしておく必要がある。

書き言葉を調査対象として標本抽出を実施する場合，母集団をどのようにリスト化するか，という技術的な問題をまず解決しなければならない。個人や世帯，事業所などは抽出単位が明確でリスト化しやすいが，書き言葉を抽出単位ごとにリスト化することは極めて困難であり，何らかの方法によって擬似的に実現せざるを得ない。

1.2.5　標本抽出

実際に標本を抽出する段階では，一定の基準と手順にのっとって，均質的な標本の抽出が実施されなければならない。このためには，調査者や調査期間の間でぶれが生じないように，抽出の手続きが明示的に定められている必要がある。

書き言葉から標本を抽出する作業では，実際に書かれている言葉のうち，どの部分をどのような順序で抽出するか，という原理が問われることになる。この原理を取り決めた上で，抽出単位となる範囲を取り出し，コーパスに収録することになる。このことを概念的に述べ直すと，書き言葉から標本を抽出するためには，**多様な体裁や構造を伴って実現している書き言葉を 1 次元の文字列（1 個以上の文字の連鎖）として把握し，そこから一定範囲の抽出単位を取り出す作業**が必要になる，と言うことができる。

以上が，書き言葉を対象としたサンプリングにとっての，理論的背景である。実際の作業を始めるにあたっては，上記に示した枠組みにしたがって，さまざまな項目を定義したり基準を取り決めたりしておく必要がある。そこで次章では，BCCWJ で実施しているサンプリングの基本方針，および諸項目の定義について述べる。

第2章　BCCWJにおけるサンプリングの基本方針

丸山岳彦・秋元祐哉

前章で示した標本調査の理論的な枠組みに基づいて，以下では，BCCWJ の構築にあたり我々が実施しているサンプリング作業の基本方針を述べる。なお，以下で述べる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，実際の作業手順などについては，丸山・秋元 (2007,2008) でも示してある。

2.1　BCCWJ の内部構成

まず初めに，BCCWJ の内部構成について確認しておく。BCCWJ の内部構成を，図 2.1 に示す。

出版サブコーパス　（生産実態）	図書館サブコーパス　（流通実態）
書籍、雑誌、新聞	書籍
約3,500万語　2001年－2005年	約3,000万語　1986年－2005年
固定長サンプル ＋ 可変長サンプル	固定長サンプル ＋ 可変長サンプル
特定目的サブコーパス（非母集団）	
白書、国会会議録、Web文書（Yahoo! 知恵袋）、ベストセラー、教科書など	
約3,500万語　1976年－2005年	
（固定長サンプル ＋ ）可変長サンプル	

図 2.1: BCCWJ の内部構成

各サブコーパスの概要を，以下に述べる。

出版サブコーパス：　出版サブコーパスは，書き言葉の生産力という側面に着目するサブコーパスである。2001 年から 2005 年の間に国内で出版された全ての**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生産に関するありさまを捉えることを目的とする。

図書館サブコーパス：　図書館サブコーパスは，書き言葉の流通・流布の実態という側面に着目するサブコーパスである。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。

特定目的サブコーパス：　特定目的サブコーパスは，生産・流通という側面からは捉えきれない，あるいは，出版サブコーパス・図書館サブコーパス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サブコーパスである。**白書**，**国会会議録**，**Web 文書（Yahoo!知恵袋）**，**ベストセラー**，**教科書**などを対象として，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が，最長で 1976 年から 2005 年までの 30 年間とする。

これら 3 つのサブコーパスは,「**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。これは，それぞれ以下の 2 つの方針を満たすための設計である。

- 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。

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。

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，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。

可変長サンプルは，3 つのサブコーパス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サブコーパス，すなわち，出版サブコーパス，図書館サブコーパス，および，特定目的サブコーパスの一部（白書など）に対して提供される。

さて，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，出版サブコーパス・図書館サブコーパスの 2 つである。これらの部分については，母集団の定義，抽出枠・抽出方法の決定，母集団のリスト化などの諸手順が，コーパスデザインの段階で厳密に設計される必要がある。以下では，これらのコーパスデザインの中身，およびサンプリングを実施する上での基本方針について述べていくことにする。

2.2 サンプリングの基本方針

BCCWJ で実施しているサンプリングは，図 2.2 に示す基本方針に基づく。

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，言語処理研究など，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，1976 年から 2005 年の間に発行された刊行物を主たる対象とする。

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。出版サブコーパスの固定長サンプルを 1,000 万語分取得することを前提として，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 2.2: BCCWJ におけるサンプリングの基本方針

2.2.1 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，言語処理研究など，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。

2.2.2 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。その上で，BCCWJ で主たる調査対象とするのは,「1976 年から 2005 年の間に発行された刊行物」とした。

また，調査対象の範囲を，以下のように定めた。まず出版サブコーパス・図書館サブコーパスで調査対象とするのは,「書籍」「雑誌」「新聞」という 3 種類のメディアに含まれる書き言葉とした。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媒体（刊行物）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 3 つのメディアを調査対象として定めた。

出版サブコーパスでは，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・新聞という制限を付した。また，図書館サブコーパスでは，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点において，1986 年から 2005 年までに出版された書籍という制限を付した。

さらに，特定目的サブコーパスでは，上記以外のメディアに含まれる書き言葉を収録するという目的から，白書，国会会議録，Web 文書（Yahoo!知恵袋），ベストセラー，教科書などに含まれるさまざまな書き言葉を収録することとした。対象期間は，メディアによってまちまちではあるが，最長で 1976 年から 2005 年までの 30 年間とした。

以下では，出版サブコーパス・図書館サブコーパスが対象とする「書籍」「雑誌」「新聞」という 3 種類のメディアに限って，そのサンプリングの方針を記述する。

2.2.3　母集団

書籍・雑誌・新聞の母集団は，文字数により定義する。これは，書き言葉を構成する最も基本的な要素が文字であり，文字の量によって言語の量的な構造を把握するという見方に立つものである。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）。

書籍（出版サブコーパス）：　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 ページ以下の書籍等を除く。「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。

雑誌（出版サブコーパス）：　2001 年から 2005 年の間に，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定期刊行物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

新聞（出版サブコーパス）：　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。

書籍（図書館サブコーパス）： 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 ページ以下の書籍等を除く。出版サブコーパス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。

2.2.4 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」

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「日本十進分類法（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，出版サブコーパスでは，2001 年から 2005 年までの 5 年間によって 5 層に，図書館サブコーパスでは，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: 「日本十進分類法（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	

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

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・サブコーパス	抽出枠 (1)	抽出枠 (2)	合計層数
書籍（出版サブコーパス）	11 分類	5 分類	55 層
雑誌（出版サブコーパス）	6 分類	5 分類	30 層
新聞（出版サブコーパス）	3 分類	5 分類	15 層
書籍（図書館サブコーパス）	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，出版サブコーパス・図書館サブコーパスの別に，表 2.5，2.6 に示す。

2.2.5　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，無作為に選ばれたページの中に印刷されている文字の中から 1,000 文字を無作為に指定し，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，そこから無作為に 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）。

表 2.5: 推計総文字数の分布（出版サブコーパス）

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%
雑誌	1. 総合	7,421,447,806	11.34%
	2. 教育	877,875,592	1.34%
	3. 政治	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	180,964,513	0.28%
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%
合計		65,471,677,100	100%

表 2.6: 推計総文字数の分布（図書館サブコーパス）

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%

2.2.6　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。母集団の中から無作為に指定された 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ものである。

標本サイズ（コーパスサイズ）は，出版サブコーパス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。

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，出版サブコーパスとして抽出する 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，それぞれ取得されることになる。

さらに，図書館サブコーパスから抽出するサンプル数は，出版サブコーパス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。

出版サブコーパスと図書館サブコーパスから抽出されるサンプル数を，表 2.7，2.8 に示す。

2.2.7　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取得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

[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取得される。

表 2.7: サンプル構成比（出版サブコーパス）

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	1.34%	228
	3. 政治	0.70%	119
	4. 産業	0.17%	29
	5. 工業	2.24%	381
	6. 厚生	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（図書館サブコーパス）

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

分をどのような順序で取得していけばよいか，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。

そこで，続く第 3 章では，サンプリングを実施する上での判断基準として,「印刷紙面からサンプルとして抽出すべき対象を認定する基準」について，詳しく述べる。

第3章　書き言葉の階層的な構造とサンプル範囲の認定基準

丸山岳彦

書籍・雑誌・新聞などの印刷紙面から均質的な手続きによりサンプルを取り出すためには，書き言葉の多様な紙面構成の中からサンプルとして取り出す部分とそうでない部分とを選別する判断基準を取り決めておくことが必要となる。そこで本章では，最も多様な体裁を持つ書籍を対象として，(1) 書籍の構造をどのように捉えるか，(2) そこからどのような基準にもとづいてサンプルの範囲となる部分を認定するか，という 2 点について，詳しく述べる。

3.1　書籍の構成要素とサンプリング

書籍は，印刷物である以上，紙面の上に文字が印刷されていることによって成立している書き言葉である。印刷紙面上に現れる文字は，レイアウトやサイズ，紙面構成上の役割などから,「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」などのさまざまな要素に分類することができる。これらを，**書籍の構成要素**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よう。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，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，印刷紙面上に明示的に表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,「後書き」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 種類のサンプルを取得するという作業は，概念的に言えば，**印刷紙面上に印刷してあるすべての文字を 1 次元上に配置して，そこから当該の範囲を取得していく作業である**と言える。作業者は，印刷紙面の上に現れるあらゆる要素を把握し，その紙面構成を見抜き，そこから一定の順序によって，1 次元の文字の連鎖を取得しなければならない。そこで必要となるのが，書籍の構成要素のうち，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，という判断基準である。このためには，そもそも書籍の構造をどのように捉えるか，どのような基準に基づいて抽出する部分とそうでない部分を判断するか，などについての取り決めが必要となる。

以下では，書籍の構造を階層的に把握することによって，これらの基準を取り決めていく方法について述べる。

3.2　書籍の構造の階層的な把握

3.2.1　書籍の構造を構成する要素の定義

ここでは，書籍の構造を，図 3.1 に示すような諸要素から構成されるものと見る。

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

図 3.1: 書籍の構造

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等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・訳者等の名，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等を表示した部分。

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　書籍の構造の階層的な把握

上記で定義した書籍の構成要素を，書き言葉をサンプリングするという観点から，以下の 7 段階の階層によって捉えることにする。階層が深くなるにしたがって，書籍の構成要素が徐々に排除され，サンプルに収録するための範囲が絞り込まれていくことになる。

第 0 層 物理的実体： 書籍の全体。書籍の物理的な実体そのもの。

手に取って見ることができる書籍の実体。

第 1 層 原紙面： 紙面上に印刷されたすべての内容。

印刷された紙面の集合。第 0 層の物理的実体のうち，表紙のほか，本のケース，カバーなどを除いた部分。また，綴じ込まれたポスターや葉書，添付の CD-ROM など，前付・冊本体・後付には組み込まれない付録の要素も排除する。

第 2 層 実質的な内容： 伝達される主たる内容。

その書籍が伝達する主たる内容に関わる部分。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，広告などは主たる内容以外の要素なのでここで排除し，残った部分を第 2 層とする。

第 3 層 印刷された文字： 伝達される主たる内容のうち，文字を主体とする部分。

第 2 層の印刷された実質的な内容のうち，フィギュア本体を排除して残った部分。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： 現代日本語として表される部分。

第 3 層の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。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，外国語や古典語などは，ここで排除する。

第 5 層 カウント対象文字： サンプルを構成する対象となる文字。

第 4 層の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。固定長サンプル・可変長サンプルを構成する 1,000 字・最大 1 万字としてカウントされるのは，これらの文字種による。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，ここで排除する。

第 6 層 カウント対象要素： サンプルに含まれる文字数のカウント対象となる要素。

第 5 層のカウント対象文字のうち，固定長サンプル・可変長サンプルを構成する 1,000 字・最大 1 万字としてカウントされ得る要素の集合。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，ここで排除する。

以上を図示すると，図 3.2 のようになる。

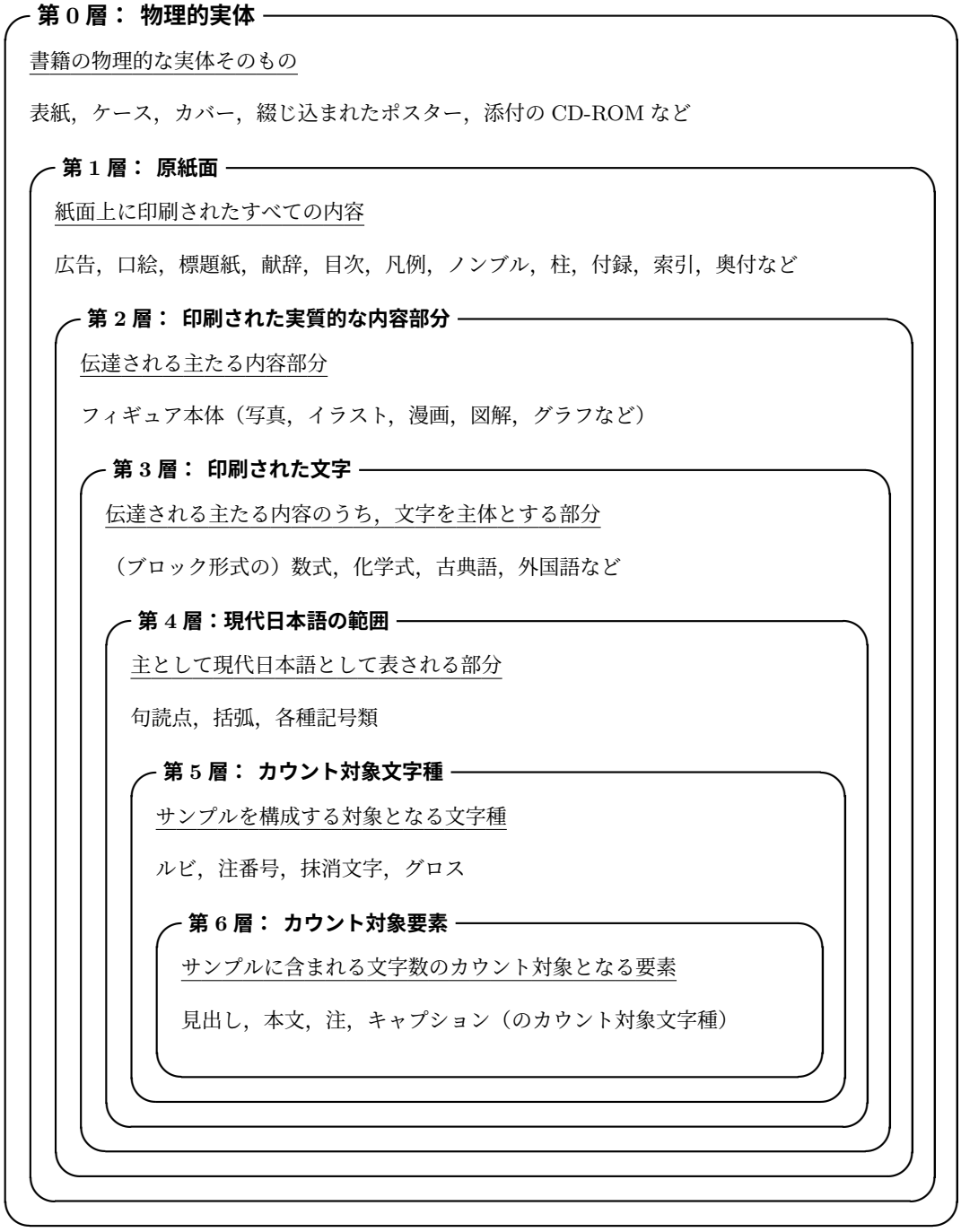


図 3.2: 書籍を構成する各要素の階層

書籍の構造をこのような 7 段階の階層によって捉えることにより，サンプリングの際に排除する要素，およびサンプルの範囲に残る要素を区別するための準備が整った。

3.3　サンプルとして取得する書き言葉の条件

3.3.1　紙面構成要素の排除原則

前節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み得る要素を絞り込んでいく原則について示す。

サンプリングの手順として，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，サンプル範囲に含める対象要素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，書籍の構成要素が徐々に削ぎ落とされ，最後に残った要素がサンプルの範囲に含まれる要素となるわけである。その原則を，以下のように定める。

紙面構成要素の排除原則：

- 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。

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。

逆の視点から言えば，第 4 層以上の要素として残った部分が，サンプルの範囲に含まれる資格を備えるということになる。さらに，第 4 層に含まれる要素のうち，句読点や記号類などの文字，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として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と諸要素の配置を確認し，上記の排除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。

3.3.2　注意を要する事例

書籍の構造を階層的に捉えた上で，紙面を構成する要素の排除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の範囲を定めることができた。しかしながら，実際にこの原則にのっとって作業を進めていく上で，注意を要する事例がいくつかある。その最たる例が，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の扱いである。

まず要点のみを述べれば，注意すべき点は，「**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**」ということである。そして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。

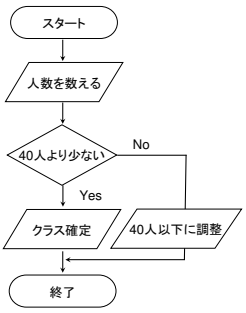


図 3.3: 分岐型フローチャートの例

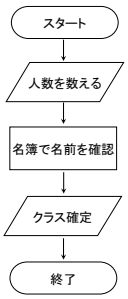


図 3.4: 直線型フローチャートの例

結論から言えば，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ことになる。ここで判断基準となるのは,「紙面構成要素の排除原則」ではなく，むしろ「**そこ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。

	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

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 3 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 次元上に展開できるか否か）のあり方によって，サンプリングの対象とするかどうかを決めているというわけである。

以上に見たように，印刷紙面上にある諸要素を排除する「紙面構成要素の排除原則」を設けたとしても，実際の紙面からサンプルを取得するためには，書き言葉のサンプリングの原理に基づいた上で，個別的な事例に対処していくことが必要になる。そのような作業上の必要性から,「紙面構成要素の排除原則」を柔軟に運用するための基準の策定と，具体事例の処理方法を類型化して整理することを行なっている。原則を運用するための基準については第 4 章で，具体事例の処理方法の類型化と整理については第 II 部で，それぞれ示すことにする。

以上，本章では，(1) 書籍の構成要素を 7 段階の階層によって捉え，(2) その上でサンプル範囲から排除される要素の範囲をどう定めるか，という 2 点について述べた。また，実際のサンプリングを進めていくためには，排除原則とは別に，一方向に読み進められるかという判断基準によってサンプル範囲に含めるか否かを判断しなければならない事例があることを述べた。

第4章　排除原則の運用 — 排除基準と選択基準，運用基準

稲益佐知子・丸山岳彦

第I部では，BCCWJにおけるサンプリングの基本的な方針と，サンプル範囲を絞り込んでいくための「紙面構成要素の排除原則」について述べてきた。続く第II部では，実際のサンプリング作業において，どのような要素がどのような判断基準によってサンプル範囲と認定されるか（あるいはされないか）を，具体的な事例を通して報告していくことにする。ここでは，第II部で事例を示していく前提として,「紙面構成要素の排除原則」が具体的にどのように適用されるか，また，サンプル範囲を認定する上でどのような運用上の規則が必要になるかについて述べる。

以下では，第I部の第3章で提示した「紙面構成要素の排除原則」，つまり「紙面に存在する文字列からどの部分を排除するか」という条件をさらに分解し，**排除基準**として再整理する。そして，排除基準とは逆の視点，すなわち,「紙面に存在する文字列からどの部分を選択するか」という視点から導かれる**選択基準**を示す。さらに，個別の事例ごとに排除基準と選択基準を適用し分けるために設けている**運用基準**を示し，実際の紙面構成に応じて排除基準と選択基準を柔軟に適用し分けていることを示す。

4.1　排除基準

「紙面構成要素の排除原則」は，書籍の構造を7段階の階層によって把握した上で，サンプリングの対象とならない要素を徐々に排除していくというものであった。以下に再掲する。

紙面構成要素の排除原則：

- 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。

この原則を，実際のサンプリング作業で用いるための判断基準として細かく分類・再整理すると，図4.1に掲げる**［排除基準1］**から**［排除基準5］**を得る。

[排除基準 1] 冊本体の構成上，実質的な内容とは見なせない要素（表紙，標題紙，目次，広告など）は，サンプル範囲から排除する。

[排除基準 2] 紙面上に印刷された要素のうち，実質的な内容とは見なせない要素（柱，ノンブルなど）は，サンプル範囲から排除する。

[排除基準 3] 言語表現を主体としない要素（写真，イラスト，図解など）は，サンプル範囲から排除する。

[排除基準 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，サンプル範囲から排除する。

[排除基準 5] 非現代日本語（外国語，古典語，数式など）は，サンプル範囲から排除する。

図 4.1:「排除基準」の一覧

これらの排除基準を順に適用すると，サンプリングの対象から外れる要素が徐々に排除されていき，残った文字列がサンプルに収録される文字列ということになる。

4.2　選択基準

4.2.1　選択基準の一覧

以下で示すのは，排除基準とは逆の視点，すなわち,「紙面に存在する文字列から何を選択するか」という視点から，サンプル範囲を見定めるための基準である。これを，**選択基準**と呼ぶ。「紙面構成要素の排除原則」によれば，第 4 層以降の要素は無条件にサンプル範囲に含まれることになるが，ここでは逆の視点から，作業者がどのような要素を積極的にサンプルの対象として認めていけばよいかについて述べる。

選択基準としては，図 4.2 に示す **[選択基準 1]** から **[選択基準 4]** を挙げる。**[選択基準 1] [選択基準 2]** によって規定される「章節構造」がサンプルを取得する上での基本的な枠組みであり，**[選択基準 3] [選択基準 4]** はそこに付加されるものである。

[選択基準 1] 「本文」を同定し，これをサンプリング対象として選択する。

[選択基準 2] 「本文」の内容を意味的にも形式的にも統括するものを「見出し」と同定し，これをサンプリング対象として選択する。

附記 1 「本文」を統括する見出しのことを「**章節見出し**」と呼ぶ。

附記 2 「章節見出し」と「本文」の組み合わせからなるかたまりを「**章節構造**」と呼び，サンプル範囲を考える上での基本的な枠組みとする。

[選択基準 3] 「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して「注」を認め，これをサンプリング対象として選択する。

[選択基準 4] 「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して「キャプション」を認め，これをサンプリング対象として選択する。

図 4.2: 「選択基準」の一覧

以下では，選択基準にもとづく選択を行なう際に注意を要する点として，(1) キャプションの認定，(2) 本文の認定，という 2 点について述べておく。

4.2.2 「キャプション」の認定について

[選択基準 4] でサンプリング対象として選択される「キャプション」に関して，章節構造の中における位置づけを整理しておく。3.2.1 節で示したように，フィギュアとは以下のように定義される。

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。

また，3.3.2 節で述べたように，一方向に読み進められない分岐型のフローチャートや行列見出しを備える表なども，フィギュア本体に相当するものとして扱う。

このうちフィギュア本体については，**[排除基準 3]**（＝言語表現を主体としない），または，**[排除基準 4]**（＝図式化されていて，一方向に読み進められない）が適用され，サンプリング対象から排除される。一方，キャプションはそれ自体が一方向に読み進められる言語表現であるため，サンプルの範囲から排除される理由はない。

キャプションは，直接的には，サンプルから排除されるフィギュア本体について解説を加えるものである。これに似た類型として，章節構造に含まれる本文部分に対して注釈を加える「注」がある。前者がサンプルの対象外要素に解説を加えるものであるのに対して，後者はサンプルの対象要素に注釈を加えるものであるという点において，両者は異なっている。

問題は，サンプルの対象外要素に対して解説を加えるキャプションをサンプルの範囲に含めてよいかという点であるが，キャプションが本文と同一紙面上に存在する文字列である以上，間接的には「章節構造」に含まれる意味内容と関連を持つ要素と認めてよいであろう。キャプションと注は，解説や注釈を加える対象がサンプルの範囲内に入るか否かという点では異なるが，いずれも「章節構造」に包まれる意味内容を言語的に補足するものとして機能している点では同等である。この点において，注とキャプションはいずれも，サンプルの範囲に含まれる要素として認めることにする。

4.2.3　「本文」の認定について

［選択基準 1］では本文の同定について触れているが，この点について追記しておく。3.2.1 節で示したように，本文とは以下のように定義される。

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。

しかしながら,「主になっている部分」とは何か，という定義を考え始めると，これは相対的にしか決まらない問題であり，明確な答えを出すことができない。小説や論文などにおいては，どこが本文であるのかはある意味で自明であるが，現実の書き言葉は，そのようなケースばかりではない。実際のサンプリングで必要とされるのは，本文そのものの定義というよりもむしろ，紙面に存在する文字列のうち，文字列どうしの関係や紙面の体裁などから判断して，作業者がどの部分を「本文」と認定すればよいか，という基準である。

以下では,「カタログ」のような紙面構成から本文となる部分を同定する，という問題を考えてみよう。なお，カタログのような構成は，本文の同定が難しい最大の事例である。

カタログは，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という紙面構成の特徴を持つ。ここで，それぞれの写真・イラストを解説する文字列は，キャプションではなく，本文として認定することにしている。これは，当該の文字列が，当該の紙面において，**［選択基準 2］**で示した「章節構造」を構成する要素であると考えられるからである[1]。

実際のサンプリング作業において，カタログ的な紙面に接した場合，まず写真・イラストに対する解説が「本文」に相当する体裁と量を備えているかを確認し，次にそれらの文字列がどの「見出し」と対応しているかを確かめ，最後にそれらの組み合わせがサンプルを構成する「章節構造」足り得るかを確認する。この確認が取れれば，**［選択基準 1］［選択基準 2］**を適用し，そこに「章節構造」を認定する。以降，当該の文字列は**［選択基準 1］**が適用される「本文」として扱われ，当該の見出しは**［選択基準 2］**が適用される「章節見出し」として扱われることになる。

また，先に「章節見出し」を認定するという手順もあり得る。紙面構成上のフォントやレイアウトなどから，ある見出しが「章節見出し」として機能していると判断できたら，次にその見出しに対応する解説の文字列が「本文」に相当する体裁と量を備えているかを確認する。そ

[1] ここでの「本文」認定は，あくまでもサンプリング上の手続きである。当該の文字列が XML 形式でエンコーディングされる際，そこにどのようなタグが付与されるかについては，ここでは不問とする。

れらが「本文」足り得ると認定できれば，両者の組み合わせを「章節構造」と認め，**[選択基準 1]［選択基準 2］**を適用するわけである。

以上のように，写真やイラストに対する解説の文章は，キャプションとして認定されるだけでなく，本文として認定される場合もあるわけである。このように，本文の認定は実際の紙面構成に応じて相対的に決まるものであり，作業者は個別の事例の中で判断を行なっていくことになる。

4.3 運用基準

4.3.1 運用基準の一覧

以上までで,「紙面に存在する文字列からどの部分を排除するか」という条件を整理した排除基準，および「紙面に存在する文字列からどの部分を選択するか」という条件を整理した選択基準を示した。しかしながら，サンプリングの実作業においては，これらの基準の運用の仕方や適用の順序について，さらに取り決めておくことが必要となる。そこで，図 4.3 に示すような**運用基準**を用意している。

[運用基準 1] サンプリング対象とするか否かの判断が難しい場合は，紙面に存在する同一類型の文字列とあわせて，サンプリング対象とするか否かを決定してよい。

[運用基準 2] ［選択基準 2］で「章節見出し」と認定された文字列については,［排除基準 2］［排除基準 3］［排除基準 4］［排除基準 5］を適用せず，当該の文字列をサンプリング対象とする。

[運用基準 3] ［排除基準 3］［排除基準 4］でサンプリング対象外となる文字列であっても，それらが［選択基準 1］［選択基準 2］により「章節構造」を構成すると認定される場合は，サンプリング対象とする。

[運用基準 4] ［排除基準 3］［排除基準 4］でサンプリング対象外となる文字列であっても,［選択基準 3］または［選択基準 4］に該当する場合は，サンプリング対象としてよい。

附記 1 ［運用基準 4］の適用にあたっては，極力,［運用基準 1］と併用する。

図 4.3:「運用基準」の一覧

以下では，それぞれの運用基準の背景にあるサンプリングの実作業における方針，および適用時に注意している点などについて述べる。

4.3.2 排除対象の不均衡とその解消

排除基準に忠実に従おうとすることによって，同一紙面上にある各要素の扱いに不自然な違いが生まれてしまうことがある。形式的には同じような体裁を持つ 2 つの要素が，排除基準を

適用することによって，片方はサンプルの範囲から排除され，もう片方はサンプルの範囲から排除されずに残る，というような場合である。このような不均衡は，印刷紙面に現れた文字列に対する均質的なサンプリングという点では，好ましくない。

そこで，**［運用基準 1］**を設けることで，この問題に対処する。すなわち，排除基準に照らせば排除される対象の要素であっても，同じ紙面に存在する同じような類型の文字列と統一的な扱いをすることが妥当であると判断できれば，無理に排除対象とはしない，という基準である。これは，排除基準の過度な適用によって，同じような類型の文字列のうちある部分が排除され別の部分が残っているという不均衡を回避するためのものである。

そもそも，サンプリングを実施する上での第一義的な判断基準は，その文字列がどのような形式として実現されているかという点にある。例えば，形式的に「見出し」になっている文字列と，それに続く「本文」として判断できる文字列があったと仮定しよう。ここで，そこで言及されている内容をもっとも端的に表しているのが，本文の中のある 1 文だったとする。仮にそうであったとしても，サンプリングの段階でその 1 文を「見出し」と認定することは無論しない。あくまでも，形式的に「見出し」として把握できる文字列を，本文を統括する「章節見出し」として認定することになる。

別の例として，キャプションの認定の例がある。例えば「地図の中の地名」のように，フィギュア本体の中にレイアウト的に入り込んでいる文字列は，キャプションとは認めず，フィギュア本体とともにサンプリング対象外とする。つまり，それがどのような形式で実現されているかによって，当該の文字列をどのように扱うかが決まるわけである。

しかしながら，その文字列の形式のみから排除基準・選択基準の適用を判断していくことによって，先に述べたような，同じような形式の文字列の間での扱いに不均衡が生じることがある。そのような不均衡をなくすために準備されているのが**［運用基準 1］**である。この基準によって，本来の排除基準による排除対象の範囲をほぼ崩すことなく，印刷紙面上における要素をバランスよく，かつ柔軟に取得することができる。

4.3.3　章節見出しの優位性

次に挙げる**［運用基準 2］**は,「章節見出し」の優位性から導かれる基準である。

先の**［選択基準 2］**で述べたように，サンプルの範囲を考える上での基本的な枠組みは「章節構造」であり，その上で「章節見出し」の認定は極めて重要である。この点において,「章節見出し」として機能していると思われる文字列がサンプル範囲の排除対象となることは，極力避けなければならない。そこで，以下のような場合が問題となる。

1. 「章節見出し」が外国語で表記されている場合
2. 「柱」が「章節見出し」の機能を担っている場合
3. 「吹き出し」が「章節見出し」の機能を担っている場合

特にムックやガイドブック，カタログ類の書籍では，章節見出しが英語で表記される場合が多い。ところが，**［排除基準 5］**によれば，非現代日本語で書かれた要素はサンプル範囲から排除される対象であり，結果，章節見出しが取得できない状態が生じる。そこで**［運用基準 2］**

を適用し，この問題を解決する。すなわち，英語で表記された部分であっても，それが章節見出しとして機能していると判断されれば，**［排除基準 5］**の適用を取り下げ，当該の文字列を章節見出しとして取得してよいと考えるわけである。

また,「柱」は**［排除基準 2］**によってサンプル範囲から排除される要素であるが，その紙面上，章節見出しとなり得る文字列が「柱」にしか示されていないと判断された時点で，**［運用基準 2］**を適用する。これにより，本来であれば排除対象である「柱」にある文字列を「章節見出し」として取り出すことができる。

さらに，やはりムックやガイドブック，カタログ類の書籍において，いわゆる「吹き出し」の形で章節見出しが表されることがある。本来,「吹き出し」はイラストの一部として見なされるため，**［排除基準 3］**によって排除される要素であるが，その紙面上，章節見出しとなり得る文字列が「吹き出し」の中にしかないと判断された場合，**［運用基準 2］**を適用して当該の文字列を「章節見出し」として取り出すのである。

さらに同様のことが,「表」の見出しにも言える。表の見出しが表の中に入り込んでいる場合は，フィギュア本体と一体化したものと見なし，**［排除基準 4］**によって一旦は排除する。ところが，その見出しが章節見出しとして機能していると判断できる場合は，**［運用基準 2］**を適用して当該の文字列を改めて取り出し,「章節見出し」として認定する。

以上のように，章節見出しとして判断された文字列に対しては，例えそれが排除基準の適用対象であったとしても，**［運用基準 2］**を適用することにより，サンプリングの対象として残すことができる。これは，章節構造を作る上で重要な要素である章節見出しを確保するために準備された運用基準である。

ただし，**［排除基準 1］**によって排除された要素（表紙，標題紙，目次，広告など）に対しては，**［運用基準 2］**を適用することはしない。これらの要素は，サンプリングを行なう対象である印刷紙面の上にそもそも存在しないものと考えるからである。

例えば，薄い単行本の中篇小説からサンプル範囲を取り出そうとしたところ，本文全体が 1 万字以内に収まり，本文がまるごとサンプル範囲として認定されたとする。さらに，その本文全体を統括する章節見出しが，標題紙にしか記載されていなかったとする。この場合，標題紙に記載された小説のタイトルを章節見出しとして取り出せば章節構造を得ることができるが，これは行なわない。なぜなら，標題紙は**［排除基準 1］**によって排除される要素であり，その書籍が持つ実質的な内容の外側に位置づけられるものと解釈するからである。この場合は，全体を統括する章節見出しの存在しないまとまりとして処理せざるを得ない。

4.3.4 フィギュア本体に含まれる章節構造

写真，イラスト，図解などのフィギュア本体は，**［排除基準 3］**によってサンプルの範囲から排除される要素であるが，これらの中にある文字列が，印刷紙面上,「章節構造」を構成していることがある。ここでは，図解の中に章節構造が含まれる例を取り上げる。

特に子供向けの図鑑や，コンピュータの解説書・マニュアル類で典型的に見られるように，ページ全体に写真やイラストが多数配置され，その中に差し込まれた図解によって解説が付されている場合がある（典型的には，イラストが引き出し線によって注記を加える解説部分と結

ばれるようなもの)。このような場合，引き出し線によって結ばれた解説部分の中に，章節見出しや本文が記述されていることが多い。

このような場合，章節構造を取り出すために**［運用基準 3］**を適用し，フィギュア本体の中にある文字列を，サンプルの範囲に含めることにする。すなわち，当該の文字列を**［排除基準 3］**の適用対象から外し，**［選択基準 1］**および**［選択基準 2］**の適用対象とするわけである。

これと同様のことが，**［排除基準 4］**によってサンプルの範囲から排除される要素（分岐型のフローチャートなど）についても言える。すなわち，そのような要素の中にある文字列が「章節構造」を構成していると判断される場合は，**［排除基準 4］**を適用せず，**［運用基準 3］**を適用する。すなわち，**［選択基準 1］**および**［選択基準 2］**の適用対象として，それらの文字列をサンプルの範囲に含めることにする。

4.3.5　フィギュア本体に含まれる「注」「キャプション」

さらに，フィギュア本体の中にある文字列を「注」や「キャプション」と見なすことができる場合がある。例えば,「吹き出し」に該当する形式のものが当該の紙面において「注」や「キャプション」として機能している場合である。

本来,「吹き出し」は**［排除基準 3］**によって排除される要素であるが，その紙面の中で「注」や「キャプション」として機能していると判断された場合，**［運用基準 4］**を適用し，当該の文字列を「注」や「キャプション」として認定することにする。

同様に，分岐型のフローチャートなどの中にある文字列が「注」や「キャプション」として機能していると判断できる場合は，**［排除基準 4］**を適用せず，**［運用基準 4］**を適用して，それらの文字列をサンプルの範囲に含めることにする。

ただし，**［運用基準 4］**を過度に適用すると，フィギュア本体に含まれる文字列をすべてサンプル範囲に残すことができるようになってしまい，本来の排除基準の意図が損なわれることになる。そこで，**［運用基準 4］**は極力**［運用基準 1］**と併用することにより，過度に適用されることのないようにする。

4.4　排除基準，選択基準，運用基準の整理

本章では，(1) サンプル範囲内から要素を排除するための「排除基準」，(2) 紙面上に存在する文字列をサンプリングの対象として認めるための「選択基準」，そして (3) 両者を運用する上での基準である「運用基準」について示した。実際のサンプリング作業では，このような形で，サンプリングの原則を柔軟に解釈し，かつ事例ごとに運用していくことが求められる。

以下では，排除基準によって排除される要素の一覧を表 4.1 に，選択基準によって選択される要素の一覧を表 4.2 に，運用基準によって，排除基準が適用されない場合の一覧を表 4.3 に，それぞれまとめておく。

表 4.1: 排除基準によってサンプル範囲から排除される要素

	排除される要素
[排除基準 1]	表紙，標題紙，目次，広告など
[排除基準 2]	柱，ノンブルなど
[排除基準 3]	写真，イラスト，図解など
[排除基準 4]	分岐型フローチャート，行列見出しを備える表など
[排除基準 5]	外国語，古典語，数式など

表 4.2: 選択基準によってサンプル範囲内の要素として選択される要素

	選択される要素
[選択基準 1]	本文
[選択基準 2]	章節見出し
[選択基準 3]	注
[選択基準 4]	キャプション

表 4.3: 運用基準が適用される組み合わせ

	[選択基準 1] 本文	**[選択基準 2]** 章節見出し	**[選択基準 3]** 注	**[選択基準 4]** キャプション
[排除基準 2] 柱，ノンブルなど		**[運用基準 2]**		
[排除基準 3] 写真，イラスト，図解など		**[運用基準 2]**		
	[運用基準 3]	**[運用基準 3]**		
			[運用基準 4]	**[運用基準 4]**
[排除基準 4] 分岐型フローチャート，行列見出しを備えた表など		**[運用基準 2]**		
	[運用基準 3]	**[運用基準 3]**		
			[運用基準 4]	**[運用基準 4]**
[排除基準 5] 外国語，古典語，数式など		**[運用基準 2]**		

※ 運用基準が適用される場合，排除基準が適用されず，その要素はサンプル範囲から排除されない。

第Ⅱ部

収録テキストの抽出

第１章　収録するテキストの抽出基準とその手順

柏野和佳子・稲益佐知子・田中弥生・秋元祐哉

第Ⅰ部を受け，この第Ⅱ部では，書籍の場合を例にして，実際の紙面からどのように収録テキストを抽出しているかを，サンプル作成の作業段階を追って，具体例とともに詳述する。また，抽出過程の段階ごとに生じる作業上の問題を明らかにする。

1.1　サンプル作成の作業段階

第Ⅰ部に示した基準にのっとって，サンプル作成の作業は，おおよそ次のように，段階的に収録テキストを絞り込んでいく手順で行う。詳細は第2章以下に順次述べる。

作業段階 1: サンプル抽出基準点を取得するページの指定
作業段階 2: サンプル範囲の指定
作業段階 3: 収録対象外要素の排除指定
作業段階 4: 収録対象要素の確定と入力順の指定

1.2　サンプル紙面の作成

サンプル作成の実作業は，サンプル紙面の作成である。書籍の印刷紙面をコピーしたものに，コーパスに収録するテキストとして抽出する部分の指定と，電子テキストとして入力する際の順番を書き込むというものである。第Ⅱ部では，作業内容を説明するために，その書き込んだ紙面例を多く引用している。よって，ここであらかじめ紙面上に書き込むものの概要と，紙面上の書き込みが意味するもののあらましを説明する。

1.3　紙面上に書き込まれる内容

（1）「サンプル抽出基準点」　・・・作業段階 1

「サンプル抽出基準点」として取得する文字に ● 印をつける。また，「サンプル抽出基準点」のある紙面であることが分かるよう，当該のサンプル紙面の右端に太線を引く。

（2）「可変長サンプル」と「固定長サンプル」の範囲　・・・作業段階 2

固定長サンプルと可変長サンプルは，同一のサンプル抽出基準点から取得する。両者の関わりには，次の３つの類型が存在する。

- 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

以上の類型にあわせ，紙面上にはサンプル範囲を次のように指示する。

・included の場合　　可変長サンプル範囲のみ「　」（かぎ括弧）の印をつける。
・overflow の場合　　可変長サンプル範囲に「　」の印をつけ，終了の　」　には，固定長範囲が可変長範囲を超えて続くことを意味する →（矢印）を書き加え，さらに，固定長サンプルの終了範囲の　」の印をつける。
・separated の場合　　可変長サンプル範囲，固定長サンプルの範囲，それぞれに「　」の印をつける。そして，可変長サンプル終了範囲の 」 には，固定長範囲が可変長範囲の先にあることを意味する →（矢印）を書き加える。

また，いずれの場合も，紙面に範囲外となる部分を含む場合は，その部分に大きく × 印をつける。

（３）収録対象外要素の排除指定　・・・作業段階 3

収録対象外要素として排除指定する部分に × 印をつける[1]。その時，非現代日本語文字列の排除要素（4.2 節参照）であれば，サンプリング作業の後，電子化テキストが構造化される際に適切なタグに置き換わるまでの代用タグの入力を指示する「*d*」の文字を × 印に加えて付与する。

（４）収録対象要素の確定と入力順の指定　・・・作業段階 4

収録対象要素部分に，文字列を電子テキストとして収録する際の入力順が分かるよう，適宜，連番を付与する。また，最終番号を△で囲む。

以上をまとめると，次の通りである。

１．●印	サンプル抽出基準点
２．紙面右端の太線	サンプル抽出基準点のある紙面
３．「　と　」	サンプル範囲の始まりと終わり
４．　」に →	可変長の範囲を超えて，あるいはその先に，固定長の範囲あり
５．「　」の外の×印	サンプル範囲外
６．×印のみ	非現代日本語以外の排除要素
７．×印と *d*	非現代日本語の排除要素
８．番号と△	テキスト収録のための入力順，△は最終番号

[1] ノンブルや柱など，サンプル内に繰り返し出てくる排除要素には，初出時のみ指示することもある。

1.4　サンプル紙面の例

はじめに，固定長サンプルのすべてが可変長サンプルに含まれる included のサンプル例を用いて，そこに書き込まれている指示の持つ意味を示す。図 1-1 は「８章」が可変長サンプル範囲であり，その中に固定長サンプル範囲を含むという例である。

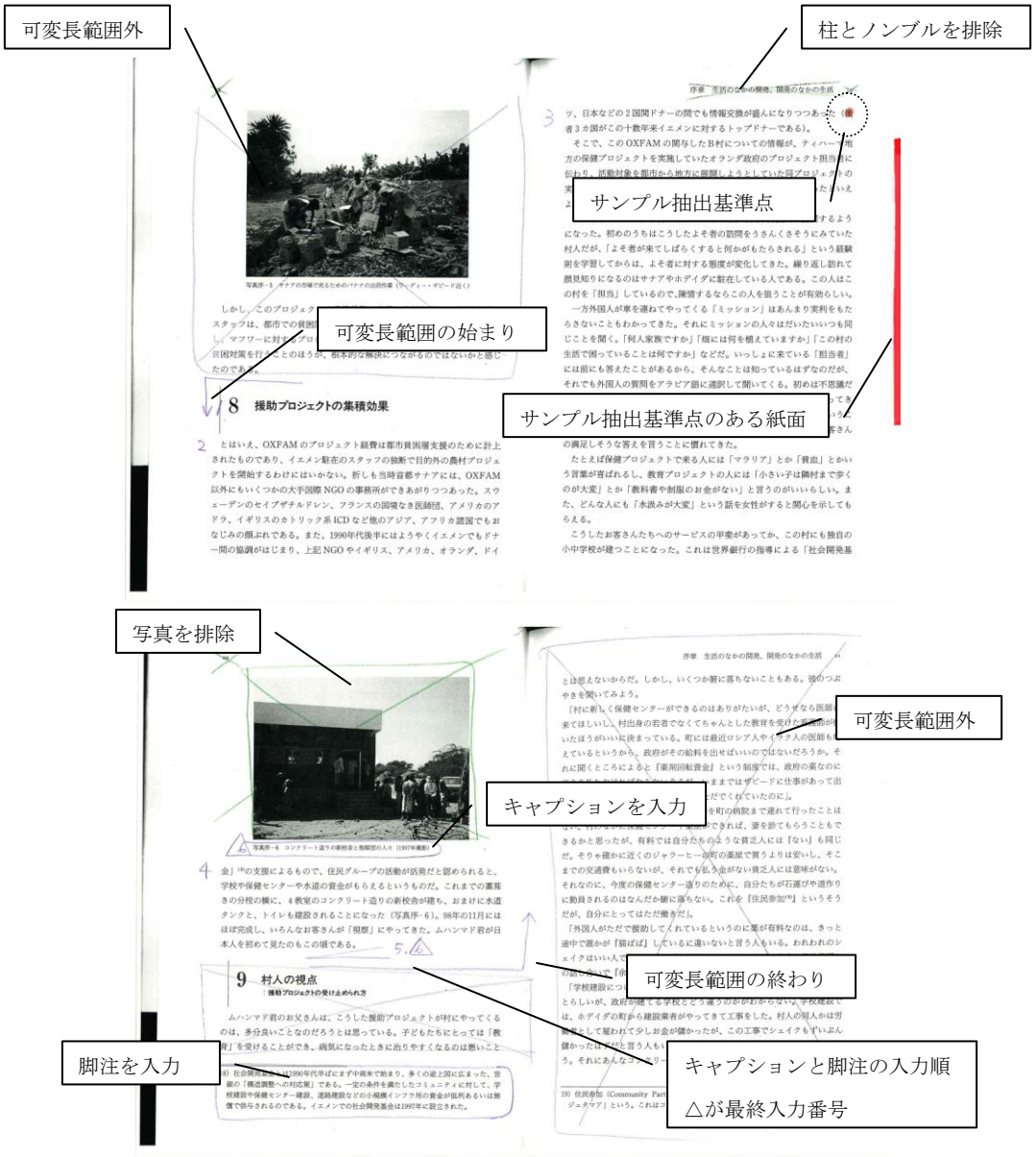


図 1-1 included のサンプル例

次に，固定長サンプル範囲が可変長サンプル範囲を超えて続く overflow のサンプル例を用いて，そこに書き込まれている指示の持つ意味を示す。図 1-2 から図 1-3 でひと続きのサンプルである。このサンプルは「Q3-7」が可変長サンプルの範囲である。固定長サンプルの範囲は，その終わりの方から，次の「Q3-8」の中ごろのあたりまでである[2]。

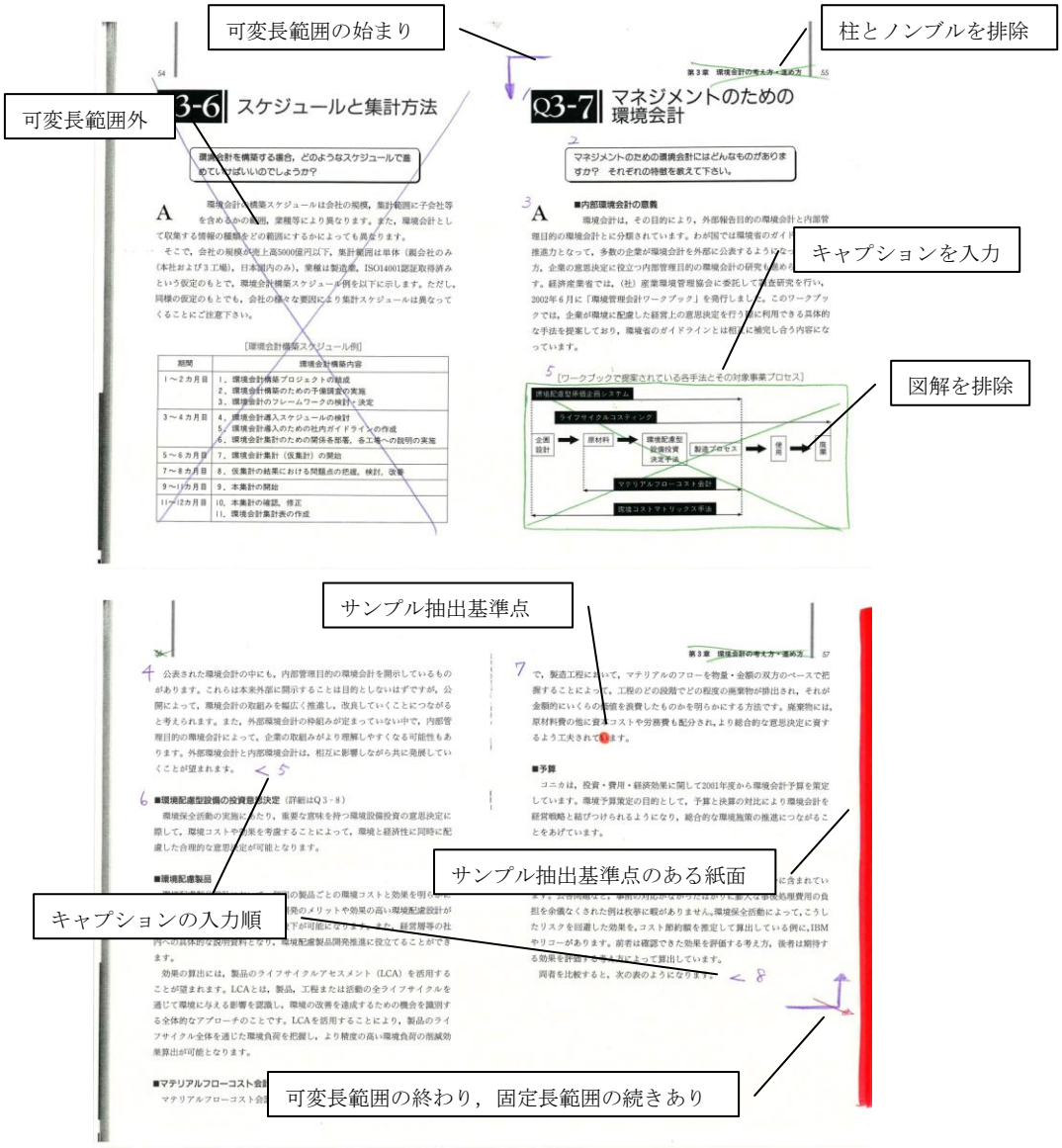


図 1-2 overflow のサンプル例

2　Q＆A形式という類型の扱いについては，3.3.1 節にて改めて説明する。

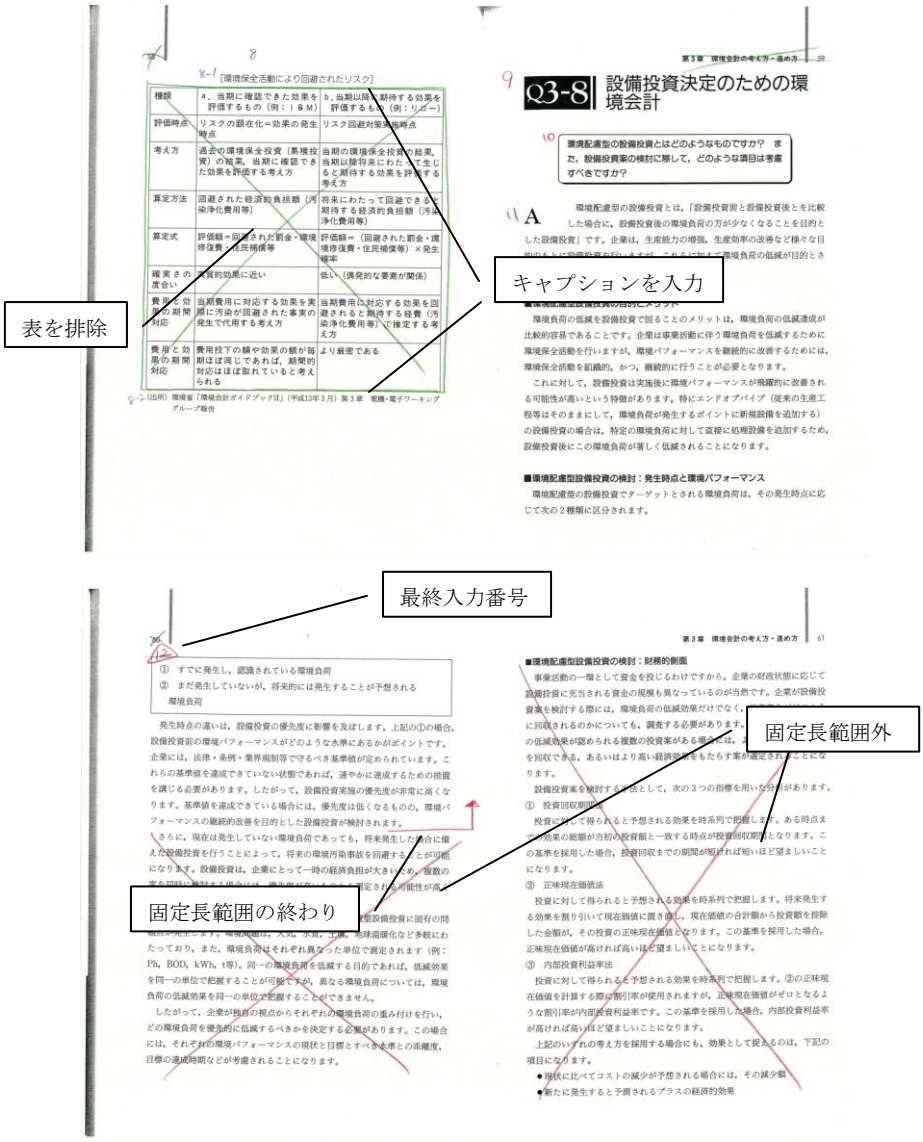


図 1-3 overflow のサンプル例（つづき）

以上の，図 1-1～図 1-3 で示した例を含め，第Ⅱ部において紙面例を引いた書籍の出典情報は，第Ⅱ部の末にまとめて示す。なお，ほとんどの例が実際のサンプリング作業でサンプル作成を行った紙面を引いているが，より分かりやすいものを明示するため，サンプリング対象外のページや，サンプリング対象外の書籍より例を引いたものもある[3]。

[3] その場合，図 1-1～図 1-3 に示したような書き込みはない。

第２章　サンプル抽出基準点を取得するページの指定

柏野和佳子・稲益佐知子・田中弥生・秋元祐哉

第Ｉ部で述べた通り，ページの無作為抽出によってサンプル抽出基準点を取得するページの順位が決まっている。その順位に沿い，そのページからサンプル抽出基準点の取得が可能か否かを判断する。その判断は，三段階の手順を踏む。第一段階で，そのページがサンプル抽出基準点を取得可能である文字列を有するページであるかを判断する。第二段階で，そのページが「書籍」のどの位置にあるものかを判断する。書籍の実質的本体である「冊本体」に位置するページであれば，この時点でサンプル抽出基準点の取得ページとして確定する。それより前の「前付」か，後の「後付」に位置するページであれば，次の第三段階に進む。第三段階では，「前付」か「後付」に位置する場合，そのページを含む文章類型が，文章量のある類型であれば，そのページをサンプル抽出基準点の取得ページとして確定する。

第一段階，あるいは，第三段階において，サンプル抽出基準点の取得が不可と判断された場合は，最初に戻り，ページの無作為抽出の指示する次の順位のページについて，第一から同じ判断を繰り返す。以下，詳述する。

2.1 サンプル抽出基準点の取得が可能か否かの確認

サンプル抽出基準点の取得が不可能なページであれば，そのページを回避する必要がある。例えば，白紙や，文字列がまったくないページは即座に回避する。全面広告のページは，第Ｉ部で述べた通り，書籍の主たる内容ではないため，回避する。また，第 4 章にて述べるサンプリング対象外要素である「図解」や「グラフ」などの類型のみで構成されるページであれば，そのページも回避する。回避した場合は，次の順位で同様の判断を行う。

2.2 冊内での位置の確認

サンプル抽出基準点の取得が可能であることを確認できたら，次は，サンプル抽出基準点の取得ページが，書籍の実質的本体である「冊本体」に位置するページであるか，それより前の「前付」か，後の「後付」に位置するページであるかを判断する。「冊本体」であると判断されれば，第 3 章の「可変長サンプル範囲の確定」作業へと進む。「前付」「後付」と判断された場合には，次の 2.3 節に述べる確認を行い，収録対象になり得ると判断されれば，第 3 章の「可変長サンプル範囲の確定」作業へと進むことになる。

2.3 「前付」「後付」の場合に必要な確認

ここでは，文章量のあるものであるか否かの判断が必要となる。原則として，以下に記述する通り，文章類型で判断している。

2.3.1 「前付」「後付」のうち収録対象とするもの

「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典型例は，「前書き」と「後書き」である。図 2-1 と図 2-2 とに，「前書き」と「後書き」の例を一つずつ挙げる。

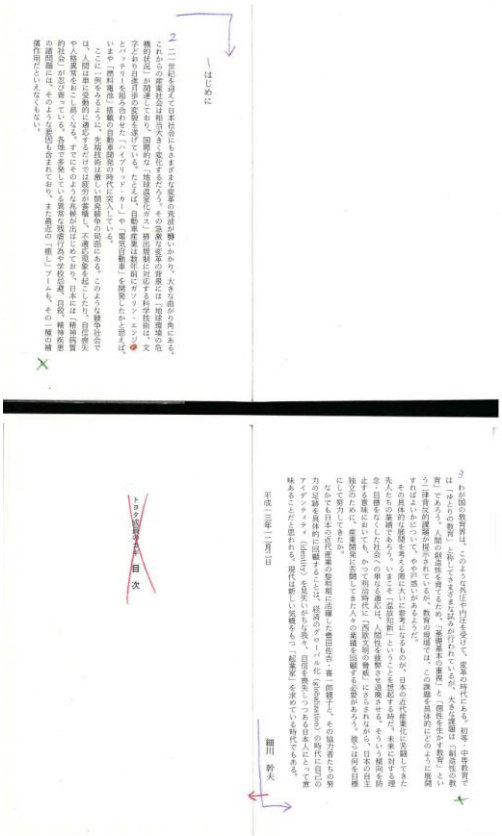


図 2-1 「前書き」

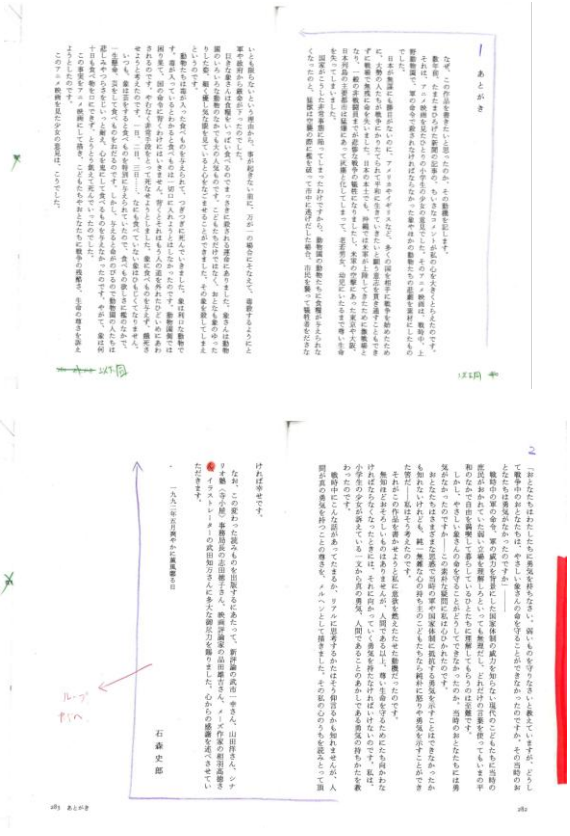


図 2-2 「後書き」

2.3.2 「前付」「後付」のうち収録対象としないもの

第 I 部 3.2.1 節で述べた通り，「前書き」と「後書き」のほかに，「前付」に位置するものとしては「口絵，標題紙，献辞，目次，凡例」などがあり，「後付」に位置するものとしては「付録，索引，奥付，広告」などがある。それらのうち，「広告」や，図表のみからなるまとまりである場合の「付録」は，先の 2.1 節で述べたサンプリング抽出基準点の取得ページの判断時において，すでに収録の対象から外されている。よって，ここで収録対象としない判断をするのは，「口絵，標題紙，献辞，目次，凡例，索引，奥付」である。それらは，原則的には，文章量の少ない類型として考え，サンプル抽出基準点を取得するページとはしない。例えば，次に示す図 2-3 から図 2-11 は，原則通り，対象外とするような例である。

図 2-3　「口絵」（イラスト）

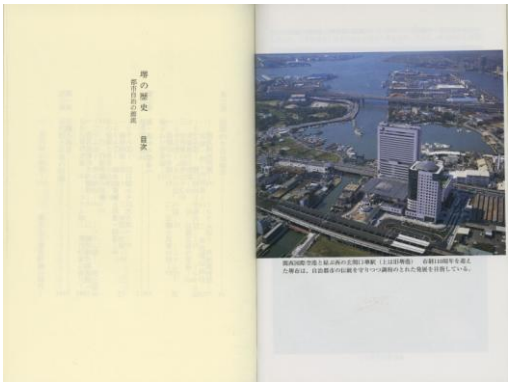


図 2-4　「口絵」（写真）

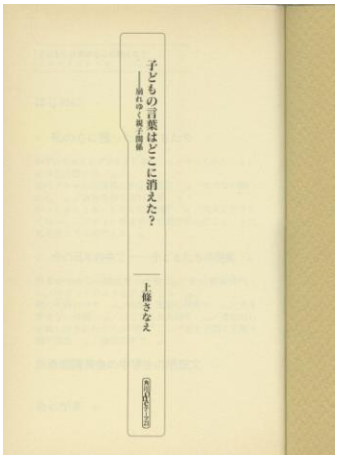


図 2-5　「標題紙」

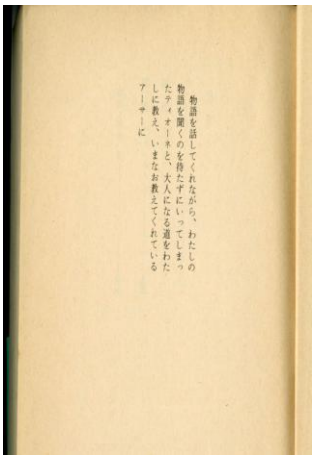


図 2-6　「献辞」

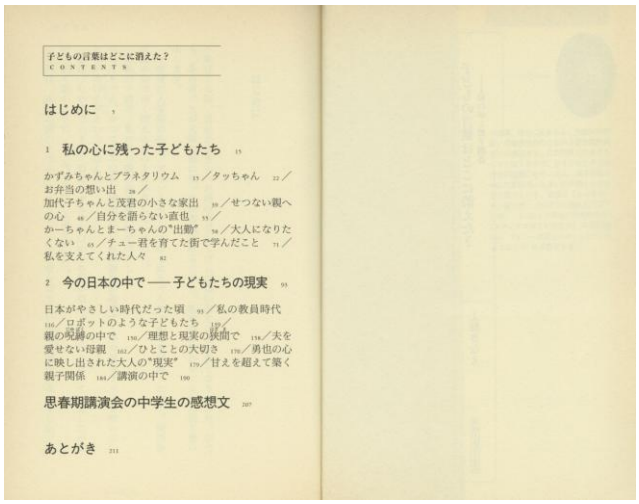


図 2-7　「目次」

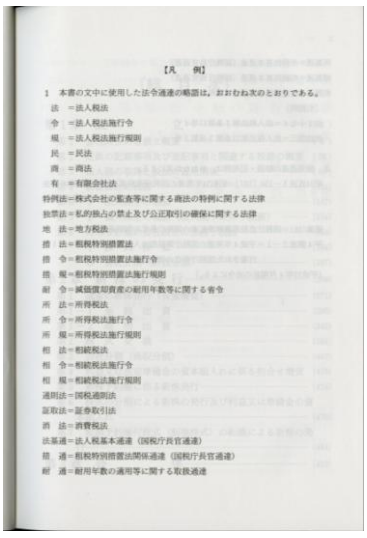


図 2-8　「凡例」

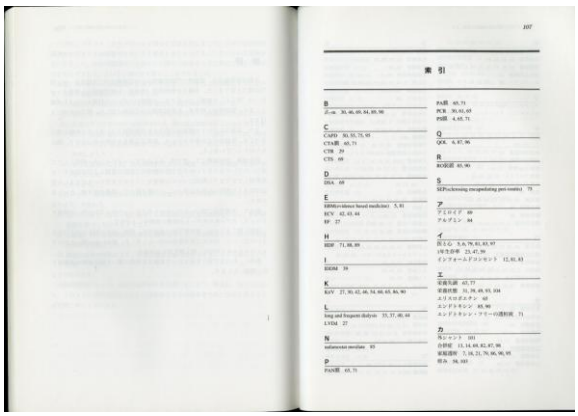


図 2-9 「索引」

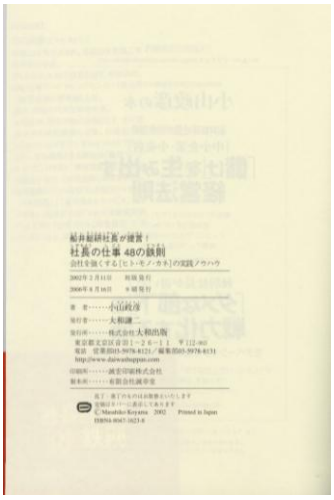


図 2-10 「奥付」

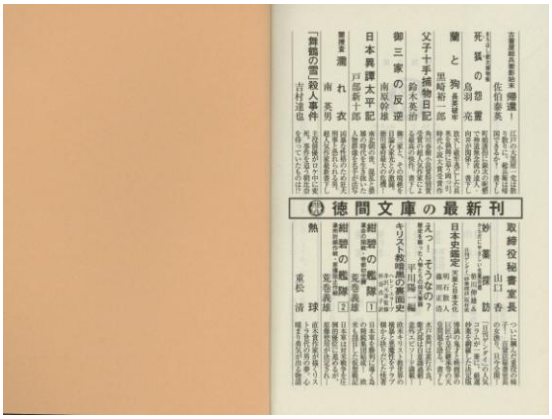


図 2-11 「広告」

2.4 サンプル抽出基準点の取得ページの指定に関わる問題点

サンプル抽出基準点の取得ページの指定で問題になる点は大きく 2 点ある。

まず，「前付」「冊本体」「後付」を分割しがたい場合があることである。「前付」の判断には，最初に「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。例えば，「目次」より前に書籍のタイトルを有するページがあれば，それは「前付」に位置する「標題紙」と見なしてサンプリングの対象ページにしない。しかし，「目次」と「冊本体」の間に書籍のタイトルを有するページがあれば，それは「冊本体」の「中扉」と見なして，サンプリングの対象ページとする。この冊本体の「中扉」がある場合は，「前付」の境界を，「目次」からこの「中扉」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，「中扉」より前に存在するものがあれば，それも「前付」になる。例えば，図 2-12 に示したパターンＡ，Ｂ，Ｃのように捉えるのである。ところが，「目次」がない，「中扉」がないなど，この認定手順を適用できない体裁を持つ書籍も少なくはなく，その

際には個別判断が必要になる。

「後付」の判断には，「前付」以上に個別判断が必要になる。「奥付」は「後付」の典型と言えるため，「奥付」とそれ以降は「後付」と言えるが，「奥付」より前において，「冊本体」との「後付」との境を決められる典型はない。

2 点目の問題は，文章量の多少を決めがたい点にある。原則は，2.3.2 節で述べた通り，「口絵，標題紙，献辞，目次，凡例，索引，奥付」などは文章量の少ない文章類型として捉えている。しかし，それらと似ているもので文章量が多いと考えられる場合がある。また，図表のみからなるまとまりである場合以外の「付録」も，文章量が多い場合がある。それら文章量が多いものはサンプリング対象となるかを個別に判断する必要がある。

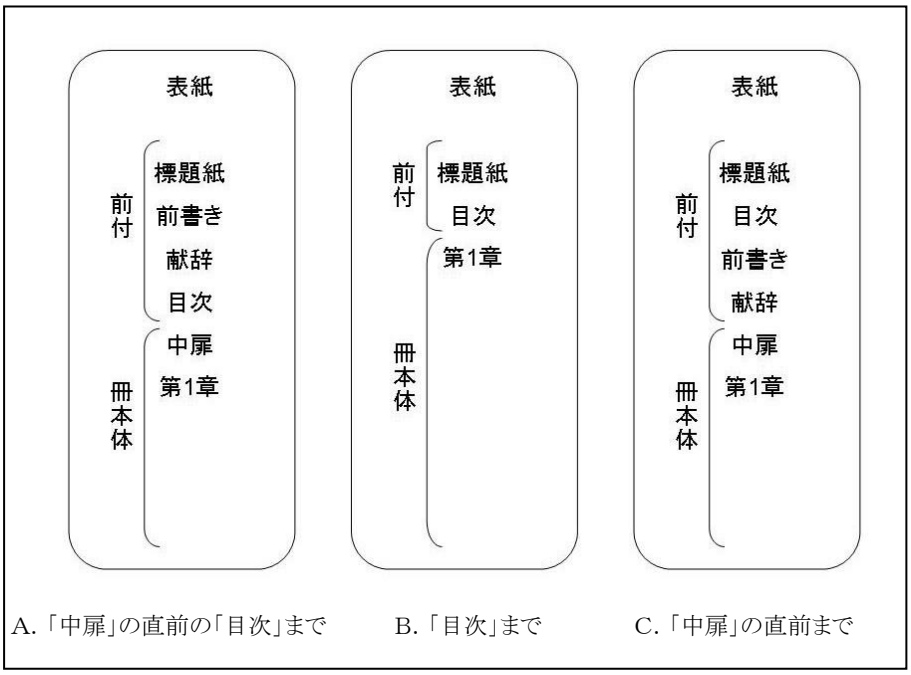


図 2-12 「前付」の捉え方のパターン例

第３章　可変長サンプル範囲の指定

柏野和佳子・稲益佐知子・田中弥生・秋元祐哉

続く作業は，可変長サンプルと固定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，第Ⅱ部では取り上げない。

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，論理的なまとまりを取得しようとするものである。以下，可変長サンプルの範囲と構造の認定基準について詳述する。

3.1　「理想範囲」と「完結構造」

可変長サンプルの範囲は，「理想範囲」と「完結構造」という二点から捉えて把握する。ここで，「理想範囲」とは，同一著者が書いた論理的な文書構造の全体を便宜的に呼ぶものである。また，「完結構造」とは，文書を構成する各階層，例えば，全体，章，節といった各階層のまとまりを呼ぶものである。

「理想範囲」は，同一著者が書いた単行本であれば，その「冊本体」，全文である。「前付」「後付」にはその範囲は及ばない。逆に，「前付」「後付」にサンプル抽出基準点が定まる場合は，「理想範囲」は，その位置において完結した構造を取得できる範囲である。また，分担執筆の著作（3.3.1 節）や，同一著者による複数作品の著作（3.3.3 節）のような場合は，その「理想範囲」は，その「冊本体」の中に分割される，同一の著者が書いた一つの論理的な文書構造の全体である。該当する文書構造全体が上限 1 万字におさまる場合は，その全文を収録対象とする。このような場合を，「理想範囲」が「完全」で，「完結構造」が「完結」の可変長サンプルであると呼ぶ。

しかしながら，多くの場合は，「冊本体」の全文が上限 1 万字を超えるため，その全文を取得することは少ない。「前付」「後付」であっても，その全文が上限 1 万字を超えることがある。そのような場合は，同一著者の書いた文章構造の中で，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，章や節相当と見なせる，論理的なまとまりを取得する。このような場合を，「理想範囲」は「不完全」で，「完結構造」は「完結」である可変長サンプルであると呼ぶ。

なお，サンプル抽出基準点を章題の文字列より取得する場合に，該当章が 1 万字を超えるような場合がある。このような場合は，章題に加え，至近の「一つ下の階層構造」を取得する。本来取得したかった「完結構造」（該当章）を部分的にしか取得できていないため，これを「理想範囲」は「不完全」で「完結構造」は「一部完結」と呼ぶ。

さらに，1 万字におさまる範囲内に，適当な論理的なまとまりを認定できない場合が少なからずある。このような場合は，サンプル抽出基準点を含む，含まないに関わらず，冒頭より機械的に 1 万字を可変長範囲とする。このような場合を，「理想範囲」が「不完全」なことに加え，完結構造も「不完結」である可変長サンプルであると呼ぶ。

以上述べた通り，可変長サンプルの「理想範囲」と「完結構造」の組合せパターンは，「完全・完結」，「不完全・完結」「不完全・一部完結」「不完全・不完結」の 4 つのうちのいずれかになる。例えば，「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，次の通りである。

① 第 3 章に著者名の記名があり，第 3 章全体が 1 万字以内である。：理想範囲「完全」，完結構造「完結」
② 第 3 章より上の範囲が同一著者の同一著作である。1 万字以内におさまる最大が第 3 章全体である。：理想範囲「不完全」，完結構造「完結」
③ 第 3 章より上の範囲が同一著者の同一著作である。第 3 章は 1 万字を超える。サンプル抽出基準点が第３章の章題に当たる。1 万字以内におさまる最大が第 3 章の一つ下の構造第 1 節である。：理想範囲「不完全」，完結構造「一部完結」
④ 第 3 章より上の範囲が同一著者の同一著作である。第 3 章は 1 万字を超える。その下には階層構造がない。冒頭から 1 万字が範囲である。：理想範囲「不完全」，完結構造「不完結」

以上を図示すると，次の図 3-1 の通りである。

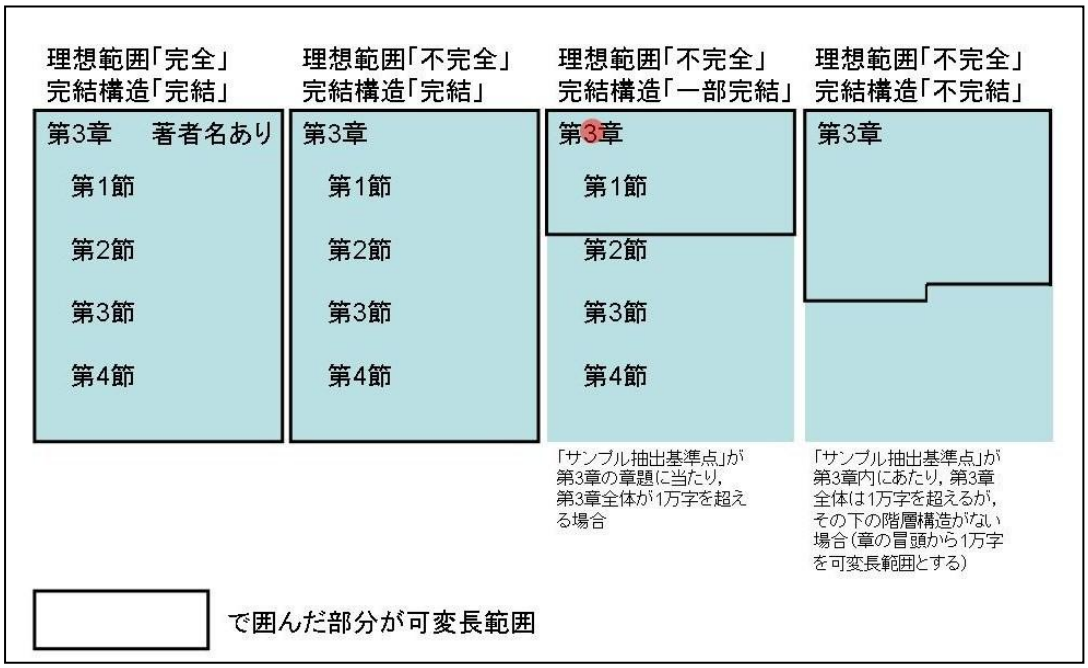


図 3-1 可変長サンプル範囲 4 パターンの例

3.2　可変長サンプル例

先に述べたように，可変長サンプルの「理想範囲」と「完結構造」の組合せパターンは，「完全・完結」，「不完全・完結」「不完全・一部完結」「不完全・不完結」の 4 つのうちの

いずれかになる。「不完全・不完結」となる例はここでは引かないが，それ以外のものを順次例示する。

図 3-2 は，「立松晃」の節の末に著者表示があることにより，「立松晃」の節が「理想範囲」である。その全体が 1 万字以内におさまるため，「完全・完結」の例である。

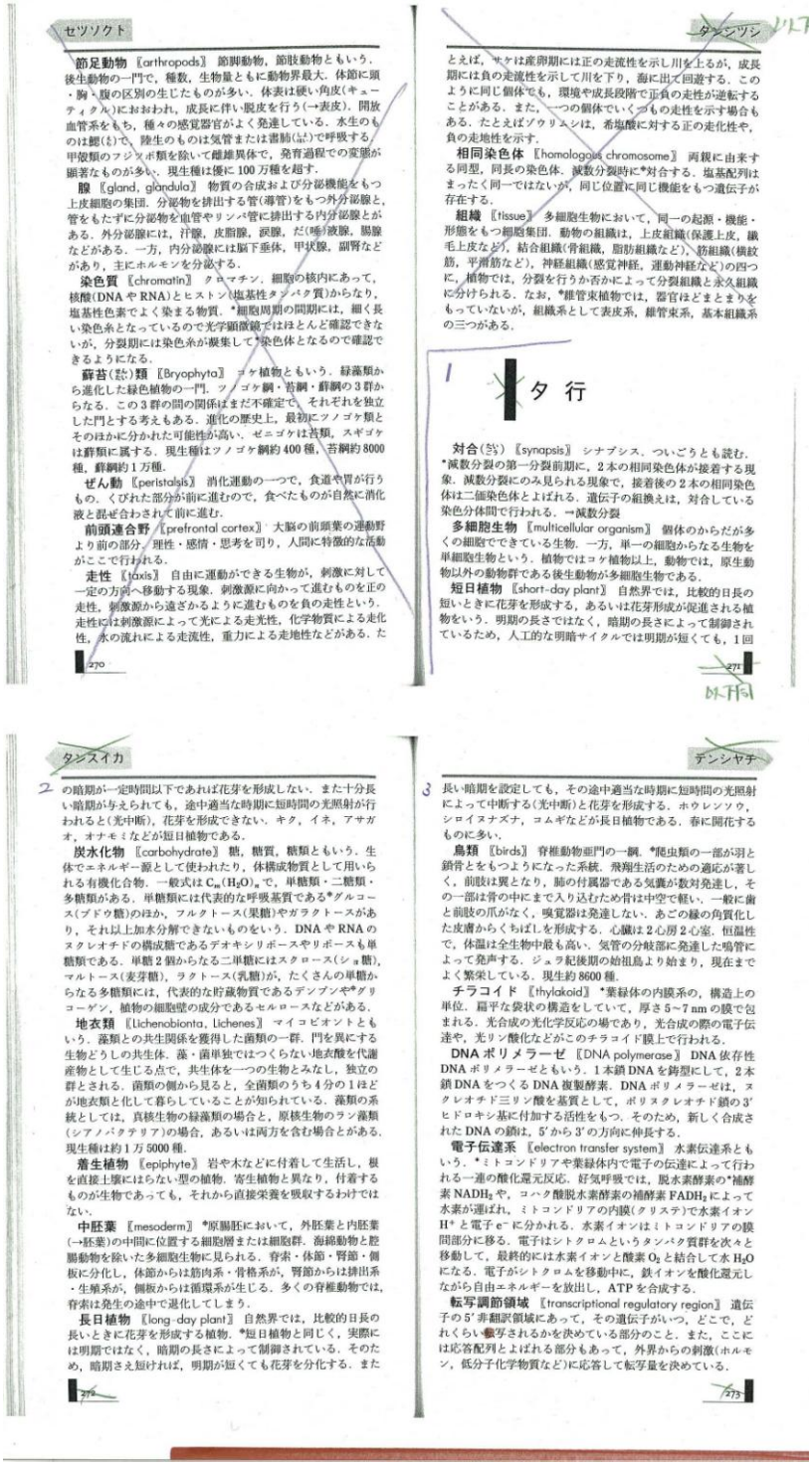
図 3-2 理想範囲「完全」，完結構造「完結」

上限 1 万字制約により，「理想範囲」が「不完全」となる場合は非常に多い。そこで，実際にどのようなサンプルが取得されているのか，「不完全・完結」の例を 4 例示す。また，最後に，「不完全・一部完結」の例を引く。

図 3-3 は，「冊本体」はもとより，「3 章」全体でも 1 万字を超えているため，その下の「3.1 節」が可変長範囲となったものである（紙幅の都合により冒頭ページのみ表示）。

図 3-3 理想範囲「不完全」，完結構造「完結」　その 1

図 3-4 は，「冊本体」では 1 万字を超えてしまうため，その下の「タ行」という階層が可変長範囲となったものである（紙幅の都合により冒頭ページのみ表示）。


セツソクト

節足動物〔arthropods〕 節脚動物，節肢動物ともいう．後生動物の一門で，種数，生物量ともに動物界最大．体節に頭・胸・腹の区別の生じたものが多い．体表は硬い角皮（キューティクル）におおわれ，成長に伴い脱皮を行う（→表皮）．開放血管系をもち，種々の感覚器官がよく発達している．水生のものは鰓（えら）で，陸生のものは気管または書肺（しょはい）で呼吸する．甲殻類のフジツボ類を除いて雌雄異体で，発育過程での変態が顕著なものが多い．現生種は優に 100 万種を超す．

腺〔gland, glandula〕 物質の合成および分泌機能をもつ上皮細胞の集団．分泌物を排出する管（導管）をもつ外分泌腺と，管をもたずに分泌物を血管やリンパ管に排出する内分泌腺とがある．外分泌腺には，汗腺，皮脂腺，涙腺，だ（唾）液腺，腸腺などがある．一方，内分泌腺には脳下垂体，甲状腺，副腎などがあり，主にホルモンを分泌する．

染色質〔chromatin〕 クロマチン．細胞の核内にあって，核酸（DNA や RNA）とヒストン（塩基性タンパク質）からなり，塩基性色素でよく染まる物質．*細胞周期の間期には，細く長い染色糸となっているので光学顕微鏡ではほとんど確認できないが，分裂期には染色糸が凝集して*染色体となるので確認できるようになる．

蘚苔（せんたい）類〔Bryophyta〕 コケ植物ともいう．緑藻類から進化した緑色植物の一門．ツノゴケ綱・苔綱・蘚綱の 3 群からなる．この 3 群の間の関係はまだ不確定で，それぞれを独立した門とする考えもある．進化の歴史上，最初にツノゴケ類とそのほかに分かれた可能性が高い．ゼニゴケは苔類，スギゴケは蘚類に属する．現生種はツノゴケ綱約 400 種，苔綱約 8000 種，蘚綱約 1 万種．

ぜん動〔peristalsis〕 消化運動の一つで，食道や胃が行うもの．くびれた部分が前に進むので，食べたものが自然に消化液と混ぜ合わされて前に進む．

前頭連合野〔prefrontal cortex〕 大脳の前頭葉の運動野より前の部分．理性・感情・思考を司り，人間に特徴的な活動がここで行われる．

走性〔taxis〕 自由に運動ができる生物が，刺激に対して一定の方向へ移動する現象．刺激源に向かって進むものを正の走性，刺激源から遠ざかるように進むものを負の走性という．走性には刺激源によって光による走光性，化学物質による走化性，水の流れによる走流性，重力による走地性などがある．た

270

タンシツシ

とえば，サケは産卵期には正の走流性を示し川を上るが，成長期には負の走流性を示して川を下り，海に出て回遊する．このように同じ個体でも，環境や成長段階で正負の走性が逆転することがある．また，一つの個体でいくつもの走性を示す場合もある．たとえばゾウリムシは，希塩酸に対する正の走化性や，負の走地性を示す．

相同染色体〔homologous chromosome〕 両親に由来する同型，同長の染色体．減数分裂時に*対合する．塩基配列はまったく同一ではないが，同じ位置に同じ機能をもつ遺伝子が存在する．

組織〔tissue〕 多細胞生物において，同一の起源・機能・形態をもつ細胞集団．動物の組織は，上皮組織（保護上皮，繊毛上皮など），結合組織（骨組織，脂肪組織など），筋組織（横紋筋，平滑筋など），神経組織（感覚神経，運動神経など）の四つに，植物では，分裂を行うか否かによって分裂組織と永久組織に分けられる．なお，*維管束植物では，器官ほどまとまりをもっていないが，組織系として表皮系，維管束系，基本組織系の三つがある．

タ行

対合（ついごう）〔synapsis〕 シナプシス．ついごうとも読む．*減数分裂の第一分裂前期に，2 本の相同染色体が接着する現象．減数分裂にのみ見られる現象で，接着後の 2 本の相同染色体は二価染色体とよばれる．遺伝子の組換えは，対合している染色分体間で行われる．→減数分裂

多細胞生物〔multicellular organism〕 個体のからだが多くの細胞でできている生物．一方，単一の細胞からなる生物を単細胞生物という．植物ではコケ植物以上，動物では，原生動物以外の動物群である後生動物が多細胞生物である．

短日植物〔short-day plant〕 自然界では，比較的日長の短いときに花芽を形成する，あるいは花芽形成が促進される植物をいう．明期の長さではなく，暗期の長さによって制御されているため，人工的な明暗サイクルでは明期が短くても，1 回

271

タンシュカ

の暗期が一定時間以下であれば花芽を形成しない．また十分長い暗期が与えられても，途中適当な時期に短時間の光照射が行われると（光中断），花芽を形成できない．キク，イネ，アサガオ，オナモミなどが短日植物である．

炭水化物〔carbohydrate〕 糖，糖質，糖類ともいう．生体でエネルギー源として使われたり，体構成物質として用いられる有機化合物．一般式は $C_m(H_2O)_n$ で，単糖類・二糖類・多糖類がある．単糖類には代表的な呼吸基質である*グルコース（ブドウ糖）のほか，フルクトース（果糖）やガラクトースがあり，それ以上加水分解できないものをいう．DNA や RNA のヌクレオチドの構成糖であるデオキシリボースやリボースも単糖類である．単糖 2 個からなる二単糖にはスクロース（シュ糖），マルトース（麦芽糖），ラクトース（乳糖）が，たくさんの単糖からなる多糖類には，代表的な貯蔵物質であるデンプンや*グリコーゲン，植物の細胞壁の成分であるセルロースなどがある．

地衣類〔Lichenobionta, Lichenes〕 マイコビオントともいう．藻類との共生関係を獲得した菌類の一群．門を異にする生物どうしの共生体．藻・菌単独ではつくられない地衣酸を代謝産物として生じる点で，共生体を一つの生物とみなし，独立の群とされる．菌類の側から見ると，全菌類のうち 4 分の 1 ほどが地衣類と化して暮らしていることが知られている．藻類の系統としては，真核生物の緑藻類の場合と，原核生物のラン藻類（シアノバクテリア）の場合，あるいは両方を含む場合とがある．現生種は約 1 万 5000 種．

着生植物〔epiphyte〕 岩や木などに付着して生活し，根を直接土壌にはらない型の植物．寄生植物と異なり，付着するものが生物であっても，それから直接栄養を吸収するわけではない．

中胚葉〔mesoderm〕 *原腸胚において，外胚葉と内胚葉（→胚葉）の中間に位置する細胞層または細胞群．海綿動物と腔腸動物を除いた多細胞生物に見られる．脊索・体節・腎節・側板に分化し，体節からは筋肉系・骨格系が，腎節からは排出系・生殖系が，側板からは循環系が生じる．多くの脊椎動物では，脊索は発生の途中で退化してしまう．

長日植物〔long-day plant〕 自然界では，比較的日長の長いときに花芽を形成する植物．*短日植物と同じく，実際には明期ではなく，暗期の長さによって制御されている．そのため，暗期さえ短ければ，明期が短くても花芽を分化する．また

272

デンシヤテ

長い暗期を設定しても，その途中適当な時期に短時間の光照射によって中断する（光中断）と花芽を形成する．ホウレンソウ，シロイヌナズナ，コムギなどが長日植物である．春に開花するものに多い．

鳥類〔birds〕 脊椎動物亜門の一綱．*爬虫類の一部が羽と鎖骨とをもつようになった系統．飛翔生活のための適応が著しく，前肢は翼となり，肺の付属器である気嚢が数対発達し，その一部は骨の中にまで入り込むため骨は中空で軽い．一般に歯と前肢の爪がなく，嗅覚器は発達しない．あごの縁の角質化した皮膚からくちばし（嘴）を形成する．心臓は 2 心房 2 心室．恒温性で，体温は全生物中最も高い．気管の分岐部に発達した鳴管によって発声する．ジュラ紀後期の始祖鳥より始まり，現在までよく繁栄している．現生約 8600 種．

チラコイド〔thylakoid〕 *葉緑体の内膜系の，構造上の単位．扁平な袋状の構造をしていて，厚さ 5～7 nm の膜で包まれる．光合成の光化学反応の場であり，光合成の際の電子伝達や，光リン酸化などがこのチラコイド膜上で行われる．

DNA ポリメラーゼ〔DNA polymerase〕 DNA 依存性 DNA ポリメラーゼともいう．1 本鎖 DNA を鋳型にして，2 本鎖 DNA をつくる DNA 複製酵素．DNA ポリメラーゼは，ヌクレオチド三リン酸を基質として，ポリヌクレオチド鎖の 3′ ヒドロキシ基に付加する活性をもつ．そのため，新しく合成された DNA の鎖は，5′ から 3′ の方向に伸長する．

電子伝達系〔electron transfer system〕 水素伝達系ともいう．*ミトコンドリアや葉緑体内で電子の伝達によって行われる一連の酸化還元反応．好気呼吸では，脱水素酵素の*補酵素 $NADH_2$ や，コハク酸脱水素酵素の補酵素 $FADH_2$ によって水素が運ばれ，ミトコンドリアの内膜（クリステ）で水素イオン H^+ と電子 e^- に分かれる．水素イオンはミトコンドリアの膜間部分に移る．電子はシトクロムというタンパク質群を次々と移動して，最終的には水素イオンと酸素 O_2 と結合して水 H_2O になる．電子がシトクロムを移動中に，鉄イオンを酸化還元しながら自由エネルギーを放出し，ATP を合成する．

転写調節領域〔transcriptional regulatory region〕 遺伝子の 5′ 非翻訳領域にあって，その遺伝子がいつ，どこで，どれくらい転写されるかを決めている部分のこと．また，ここには応答配列とよばれる部分もあって，外界からの刺激（ホルモン，低分子化学物質など）に応答して転写量を決めている．

273

図 3-4 理想範囲「不完全」，完結構造「完結」 その２

続く，図 3-5 と図 3-6 も，一つ上の階層が 1 万字を超えてしまうため，その下の階層が可変長範囲となったものであるが，特に文字数の少ない範囲となったものである。ＢＣＣＷＪでは，可変長サンプルの文字数の上限 (1 万字) は設けているが，下限は設けていない。よって，文字数の少ない可変長サンプルであっても不問としている。図 3-5 の可変長範囲は「メイショウジンライ」の部分だけであり，図 3-6 の可変長範囲は「ままかり」の部分だけである。これらはいずれも可変長範囲だけでは 1,000 字の固定長範囲が取得できないため，第１章で述べた，固定長範囲が可変長範囲を超えて先に延びる，「overflow」タイプのサンプルとなる。

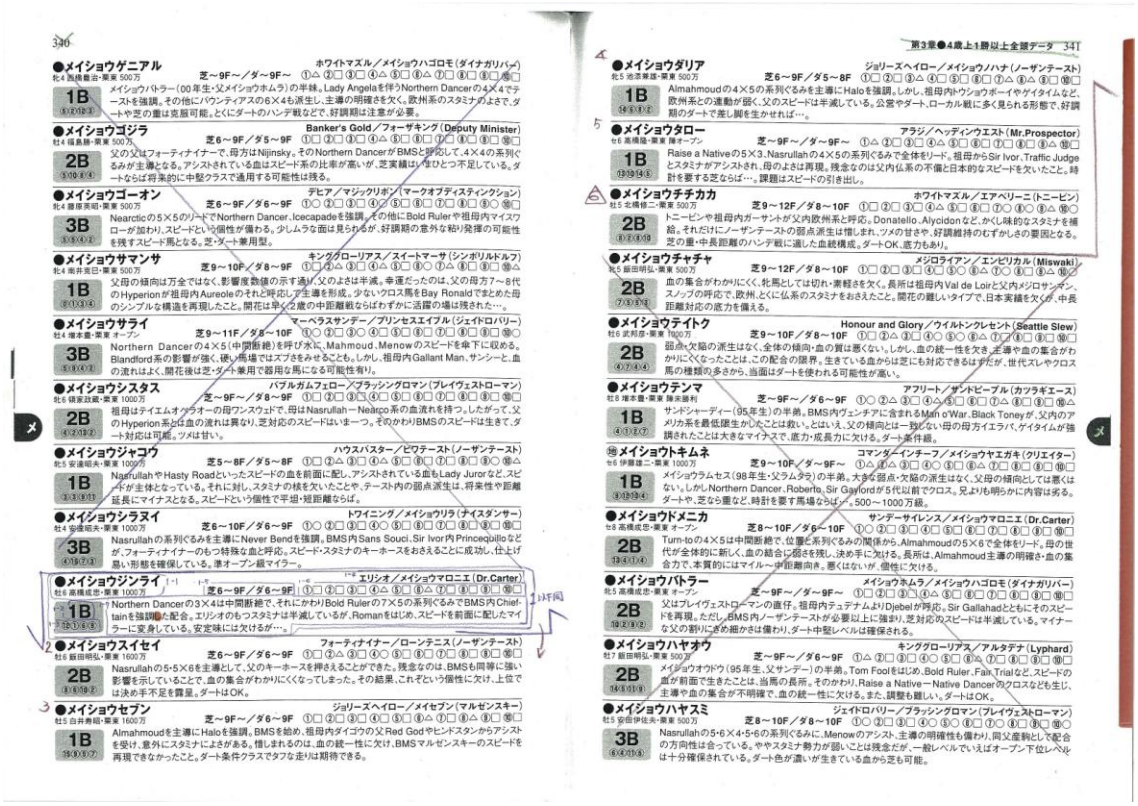


図 3-5 理想範囲「不完全」，完結構造「完結」 その３

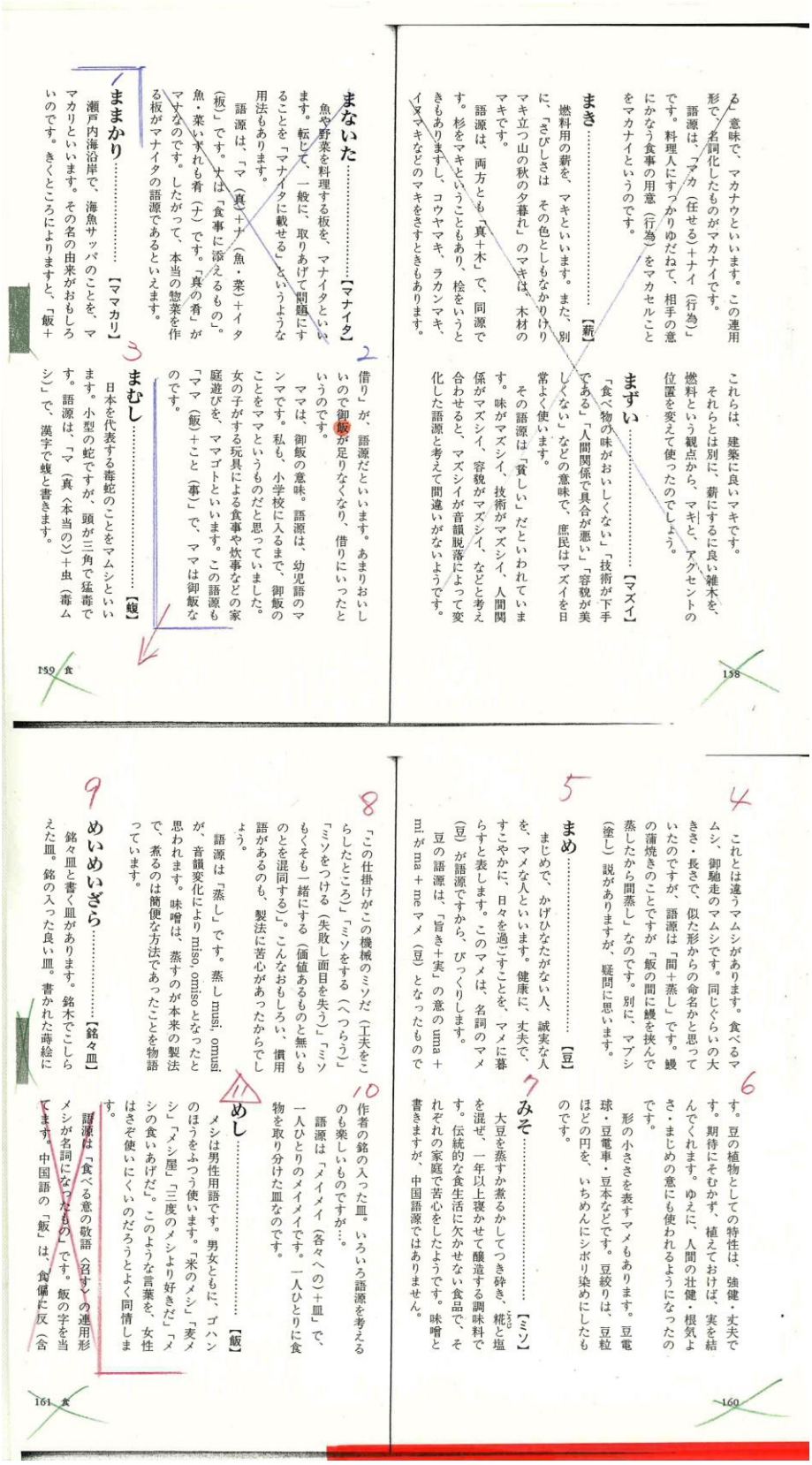


図 3-6 理想範囲「不完全」，完結構造「完結」 その４

最後に，サンプル抽出基準点を章題の文字列より取得する場合に，該当章が 1 万字を超えたため至近の「一つ下の階層構造」を取得し，理想範囲「不完全」，完結構造「一部完結」となった例を示す（図 3-7）。[1]

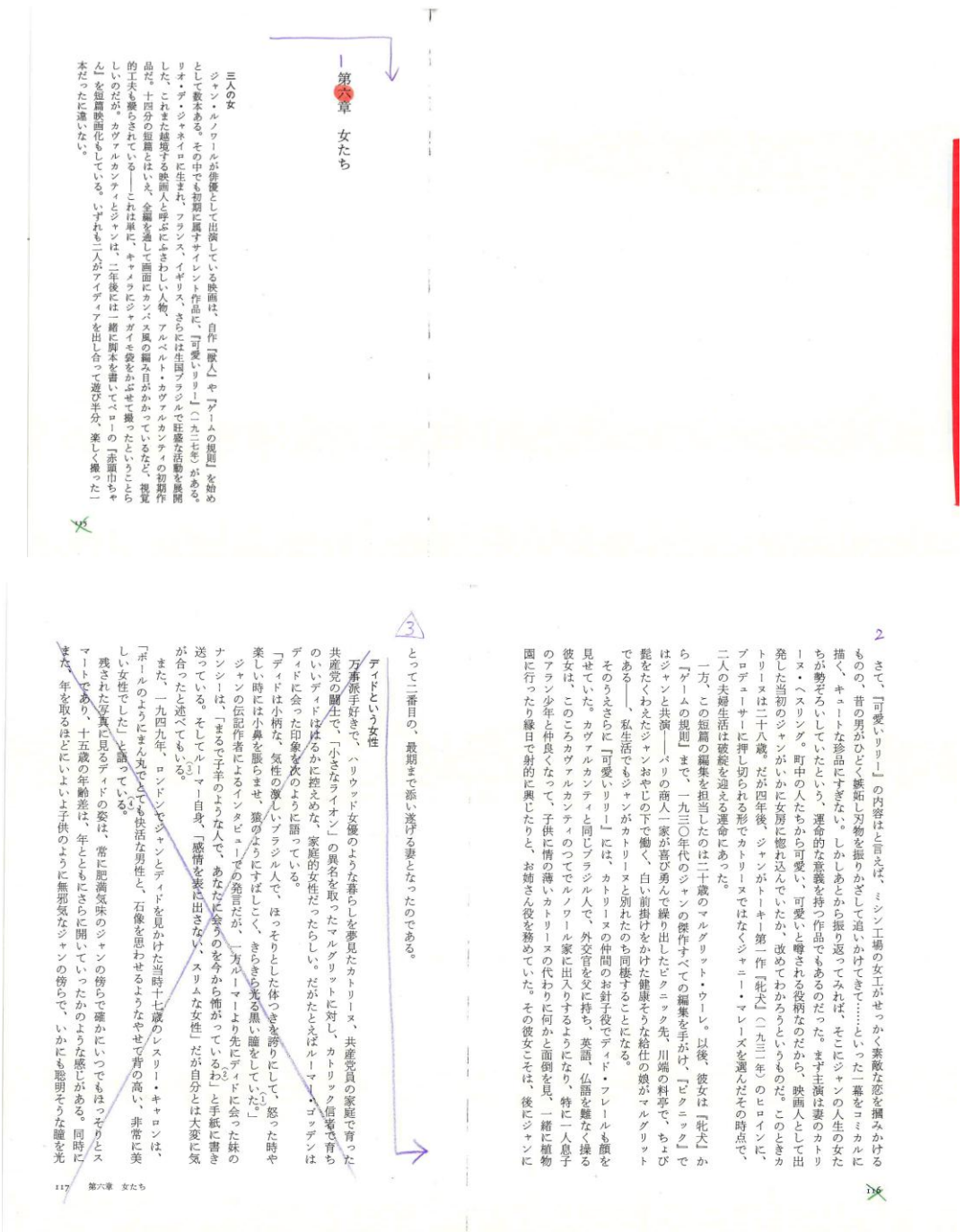


図 3-7 理想範囲「不完全」，完結構造「一部完結」

1 このサンプルの場合，「一つ下の階層構造」は，小見出しで区切られる「三人の女」である。区切りの認定については，3.4 節で解説する。

3.3 「理想範囲」の捉え方

3.3.1 著者とその「理想範囲」の認定

同一著者が書いた論理的構造の範囲を把握するために，著者とその「理想範囲」の認定が重要である。例えば，共著の場合，章ごとに，紙面，目次，奥付等に分担明記があれば，各章それぞれの著者による「理想範囲」と捉える。もし，分担明記がなければ，共著による全体が一つの「理想範囲」と捉えることになる。図 3-8 は，分担明記があるため，「三頭山」という一項目を「理想範囲」とする例である。

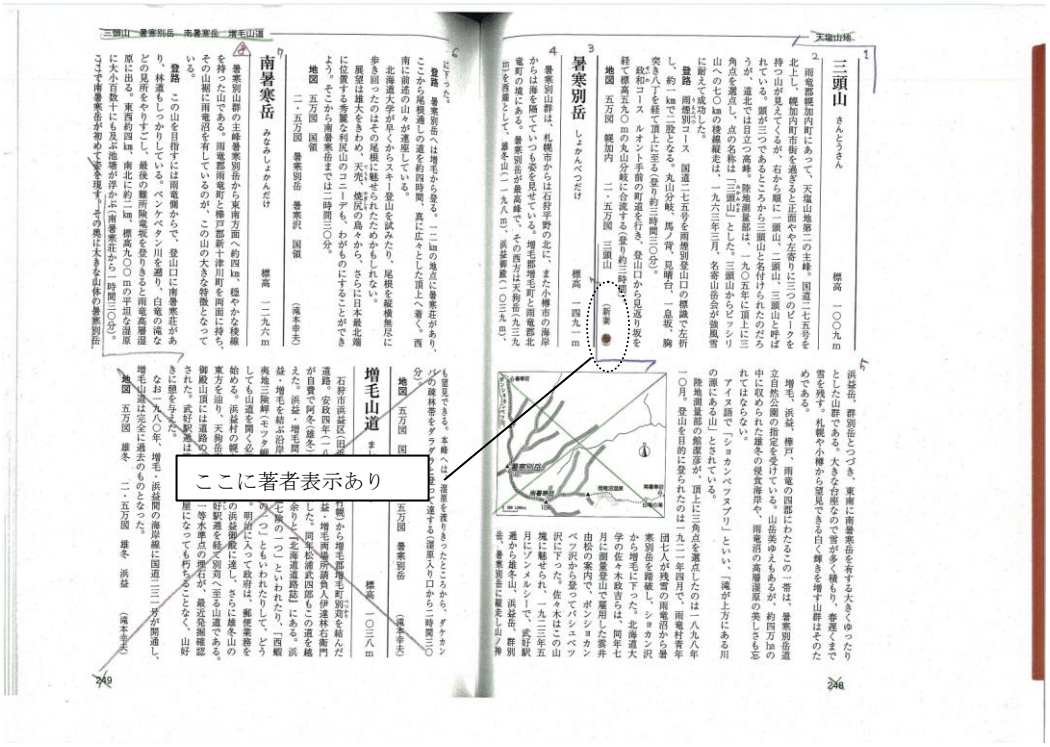


図 3-8 分担執筆で紙面に著者表示あり

また，対談，座談，インタビューなどの場合は，いずれも原則は，発話者一人ひとりを単独著者として「理想範囲」を分割せずに，まとまり全体を共著の「理想範囲」であるものとして扱う類型である。図 3-9 に対談，図 3-10 に座談，図 3-11 にインタビューの例を示す。

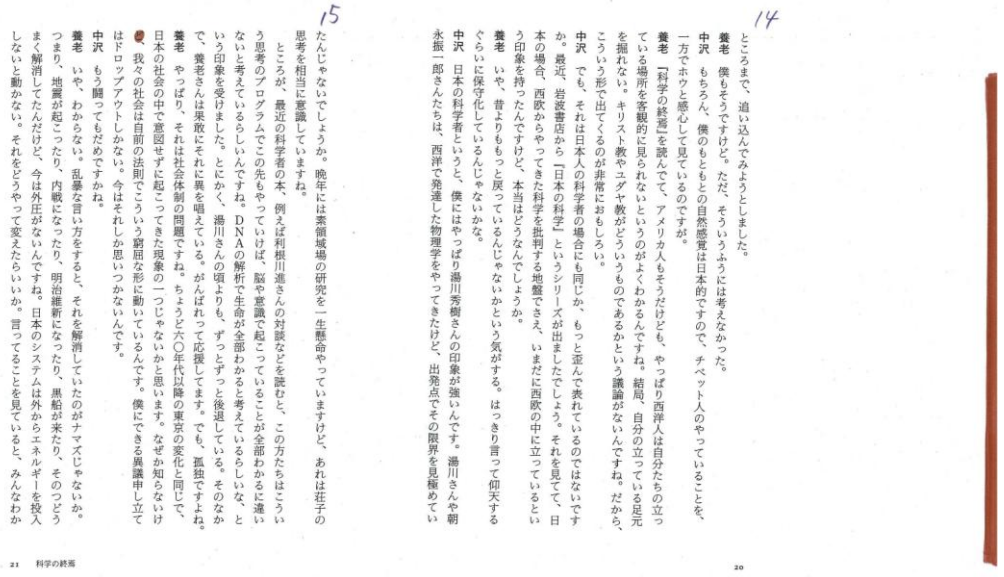


図 3-9 対談

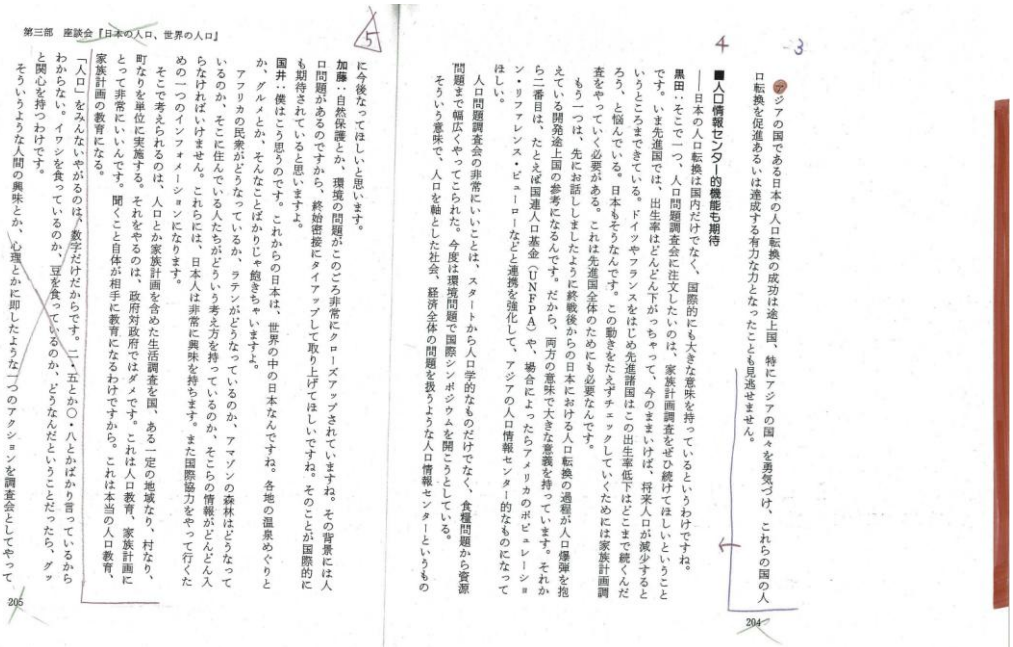


図 3-10 座談

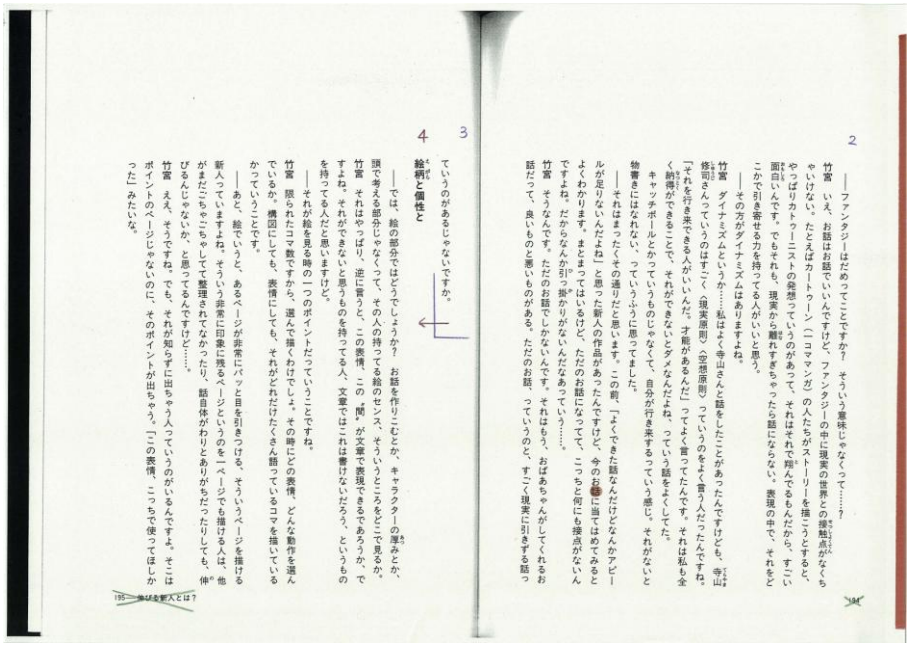


図 3-11 インタビュー

次の図 3-12 も対談の例ではあるが，この例は文章の終わりに，対談後の対談者のみの単独記述があるものである。そのような場合も文章類型を優先させ，終わりの部分も含めて対談全体を一つの「理想範囲」としている。単独記述の部分だけを単独著者の一つの「理想範囲」として認めるという考え方はしていない。

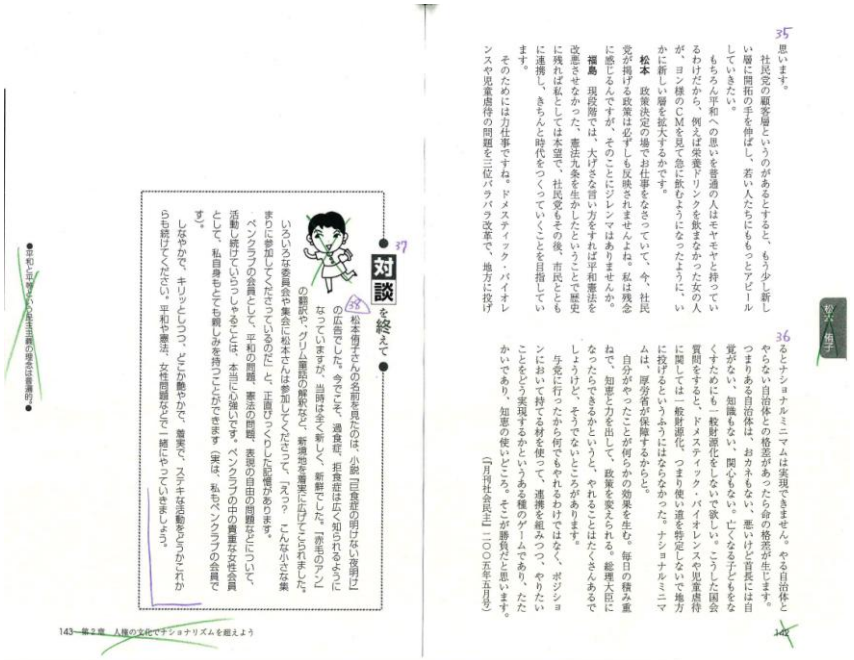
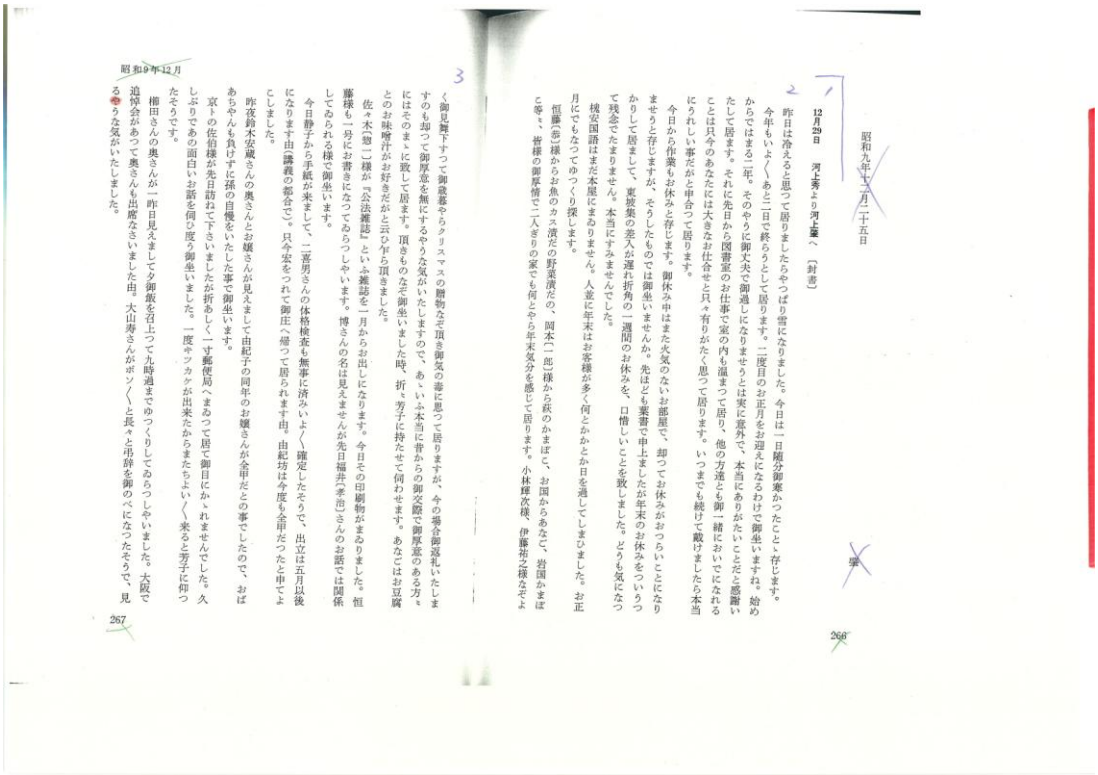


図 3-12 対談（終わりに対談者の単独著述部分あり）

次に，「本文」の前に「あらすじ」が示されているような例を考える。場合によっては，「あらすじ」は「本文」と別著者である可能性も否定できない。しかしながら，「あらすじ」という内容の性質上，「本文」に付随するものと考える。よって，「あらすじ」と「本文」という類型は，全体が「本文」の著者の一つの「理想範囲」として捉えるものである。例えば，次の図 3-13 がその例である。

図 3-13 あらすじと本文

さて，以上までの例とは逆に，あるまとまりが一つの「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，往と復の組合せで共著による一つの「理想範囲」とは捉えず，一つの書簡を一著者による個別の「理想範囲」と考える。図 3-14 がその例である。



昭和九年十二月二十五日

12月29日　河上秀より河上肇へ〔封書〕

昨日は冷えると思つて居りましたらやつぱり雪になりました。今日は一日随分御寒かつたことゝ存じます。

今年もいよ〳〵あと二日で終らうとして居ります。二度目のお正月をお迎えになるわけで御坐いますね。始めからではまる二年。そのやうに御丈夫で御過しになりませうと実に意外で、本当にありがたいことだと感謝いたして居ます。それに先日から図書室のお仕事で室の内も温まつて居り、他の方達とも御一緒においでになれることは只今のあなたには大きなお仕合せと只々有りがたく思つて居ります。いつまでも続けて戴けましたら本当にうれしい事だがと申合つて居ります。

今日から作業もお休みと存じます。御休み中はまた大気のないお部屋で、却つてお休みがおつらいことになりませうと存じますが、そうしたものでは御坐いませんか。先ほども薬書で申上ましたが年末のお休みをついつかりして居まして、東坡集の差入が遅れ折角の一週間のお休みを、口惜しいことを致しました。どうも気になつて残念でたまりません。本当にすみませんでした。

槐安国語はまだ本屋にまゐりません。人並に年末はお客様が多く何とかかとか日を過してしまいました。お正月にでもなつてゆつくり探します。

恒藤（恭）様からお魚のカス漬だの野菜漬だの、岡本（一郎）様から萩のかまぼこ、お国からあなご、岩国かまぼこ等々、皆様の御厚情で二人ぎりの家でも何とやら年末気分を感じて居ります。小林躋造様、伊藤祐之様などよ

昭和九年十二月

く御見舞下すつて御歳暮からクリスマスの贈物など頂き御気の毒に思つて居りますが、今の場合御返礼いたしますのも却つて御厚意を無にするやうな気がいたしますので、あゝいふ本当に昔からの御交際で御厚意のある方々にはそのまゝに致して居ます。頂きものなぞ御坐いました時、折々芳子に持たせて伺わせます。あなたはお豆腐とのお味噌汁がお好きだがと云ひ乍ら頂きました。

佐々木（惣一）様が『公法雑誌』といふ雑誌を一月からお出しになります。今日その印刷物がまゐりました。恒藤様も一号にお書きになつてゐらつしやいます。博さんの名は見えませんが先日福井（孝治）さんのお話では関係してゐられる様で御坐います。

今日静子から手紙が来まして、二喜男さんの体格検査も無事に済みいよ〳〵確定したそうで、出立は五月以後になります由（講義の都合で）。只今宏をつれて御庄へ帰つて居られます由。由紀坊は今度も全治だつたと申てよこしました。

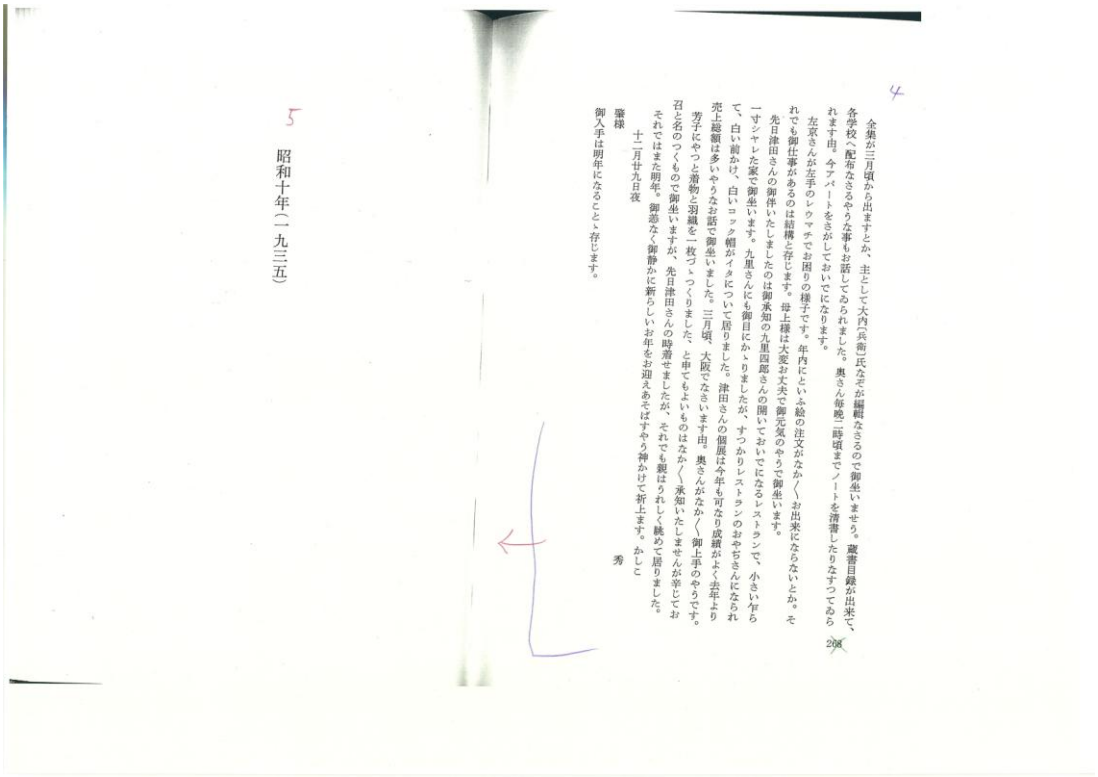
昨夜鈴木安蔵さんの奥さんとお嬢さんが見えまして由紀子の同年のお嬢さんが全甲だとの事でしたので、おばあちやんも負けずに孫の自慢をいたした事で御坐います。

京一の佐伯様が先日詰ねて下さいましたが折あしく一寸郵便局へまゐつて居て御目にかゝれませんでした。久しぶりであの面白いお話を伺つて度う御坐いました。一度ヤツカが出来たからまたよい〳〵来ると芳子に仰つたそうです。

櫛田さんの奥さんが一昨日見えまして夕御飯を召上つて九時過までゆつくりしてゐらつしやいました。大阪で追悼会があつて奥さんも出席なさいました由。大山寿さんがポツ〳〵と長々と弔辞を御のべになつたそうで、見るような気がいたしました。

267

268



全集が三月頃から出ますとか、主として大内（兵衛）氏などが編輯なさるので御坐いませう。蔵書目録が出来て、各学校へ配布なさるやう事もお話してゐられました。奥さん毎晩二時頃までノートを清書したりなすつてゐられます由。今アパートをさがしておいでになります。

左京さんが左手のレウマチでお困りの様子です。年内にといふ絵の注文がなか〳〵お出来にならないとか。それでも御仕事があるのは結構と存じます。母上様は大変お丈夫で御元気のやうで御坐います。

先日津田さんの御伴いたしましたのは御承知の九里四郎さんの開いておいでになるレストランで、小さい乍らすシヤレた家で御坐います。九里さんにも御目にかゝりましたが、すつかりレストランのおやぢさんになられて、白い前かけ、白いコック帽がイタについて居りました。津田さんの個展は今年も可なり成績がよく去年より売上高額は多いやうなお話で御坐いました。三月頃、大阪でなさいます由。奥さんがなか〳〵御上手のやうです。

芳子にやつと着物と羽織を一枚づゝつくりました、と申てもよいものはなか〳〵承知いたしませんが幸じてお召と名のつくもので御坐いますが、先日津田さんの時着せましたが、それでも親はれしく眺めて居りました。それではまた明年。御恙なく御静かに新らしいお年をお迎えあそばすやう神かけて祈上ます。かしこ

十二月廿九日夜　秀

肇様

御入手は明年になること、存じます。

268

5

昭和十年（一九三五）

図 3-14 往復書簡

最後に，Q＆A形式の場合を考える。Q＆A形式の典型は，Aの著者がQを引用しているものと考え，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の組合せを「理想範囲」として区切ることもしない。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，Q＆A集の場合，Aの著者が複数いて，各Q＆Aに分担明記がある場合は，一つのQ＆Aが一つの「理想範囲」になる。例えば，次の図 3-15 のような場合である。

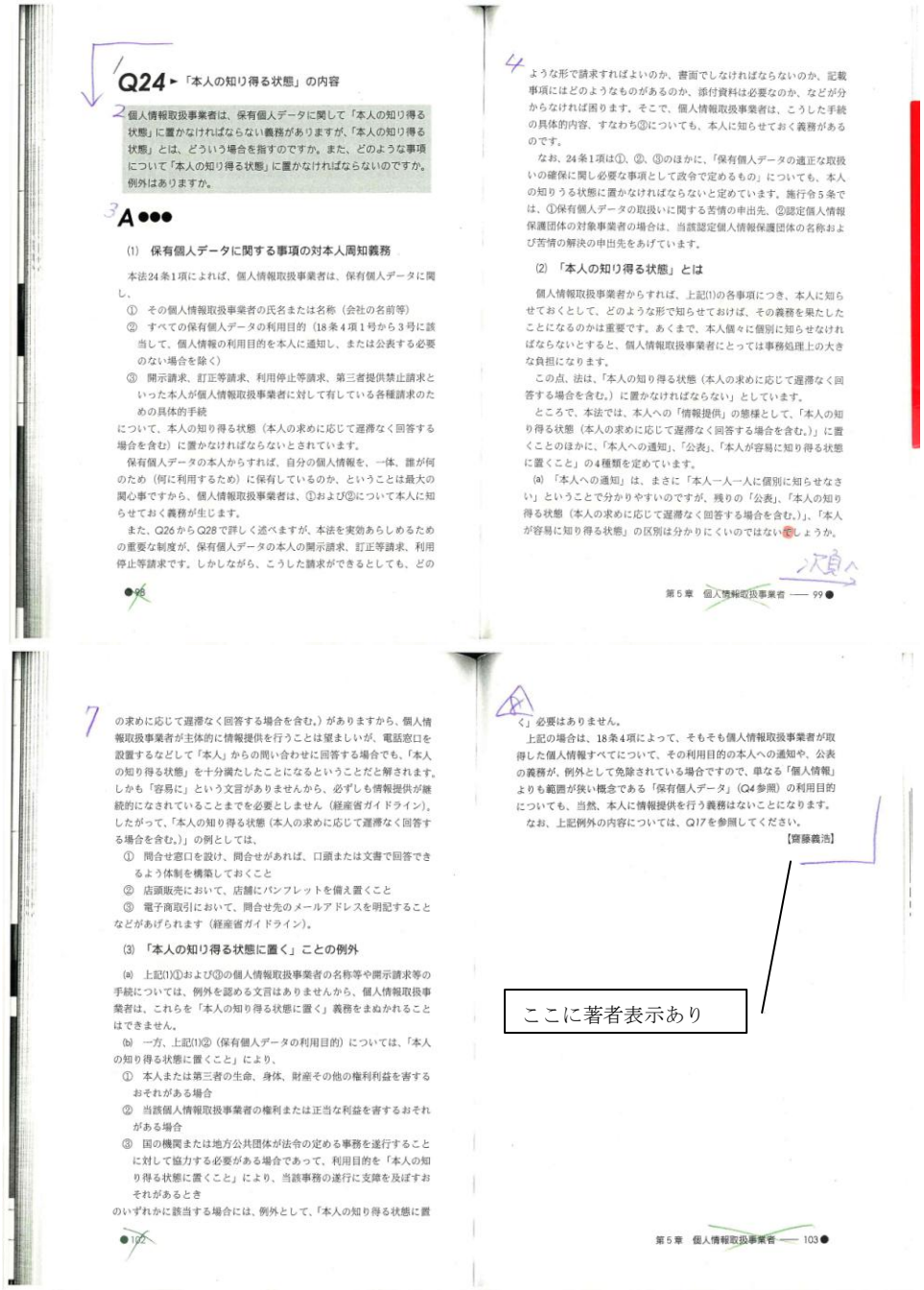


図 3-15　Q＆A（分担執筆の著者表示あり）

3.3.2　著者とその「理想範囲」の認定に関わる問題

共著である場合，あるいは，共著と考えられるような場合に，単独の著者別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，原著者一人に対し，訳者が各章で異なる場合は，各章がそれぞれ原著者と訳者の共著による「理想範囲」である。

また，著者が一見して判断しにくい場合も問題である。例えば，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。例えば，図 3-16 は分担明記のある料理人を著者と認定できる例である。このような場合は，料理人の明記されている一つの料理手順の部分が「理想範囲」と捉えられる（「夏梅」氏を，「肉じゃが」の節の著者と認め，「肉じゃが」全体を「理想範囲」とする）。ただし，このサンプルの場合，サンプル抽出基準点が章題に当たっているため，分担明記の有無に関わらず，冊本体全文「理想範囲」になる。1 万字におさまる最大が「1 章」であるため，ここでは「1 章」が可変長範囲になっている。

図 3-16 料理人の著者認定（分担執筆の明記あり）によらないもの

3.3.3　作品集等の場合の「理想範囲」

同一著者が書いた単行本であれば，その「冊本体」全体を一作品と見なし「理想範囲」とする原則を 3.1 節で述べたが，個人全集のような作品集の場合には，その原則を適用せずに，一作品を一つの「理想範囲」であると考える。

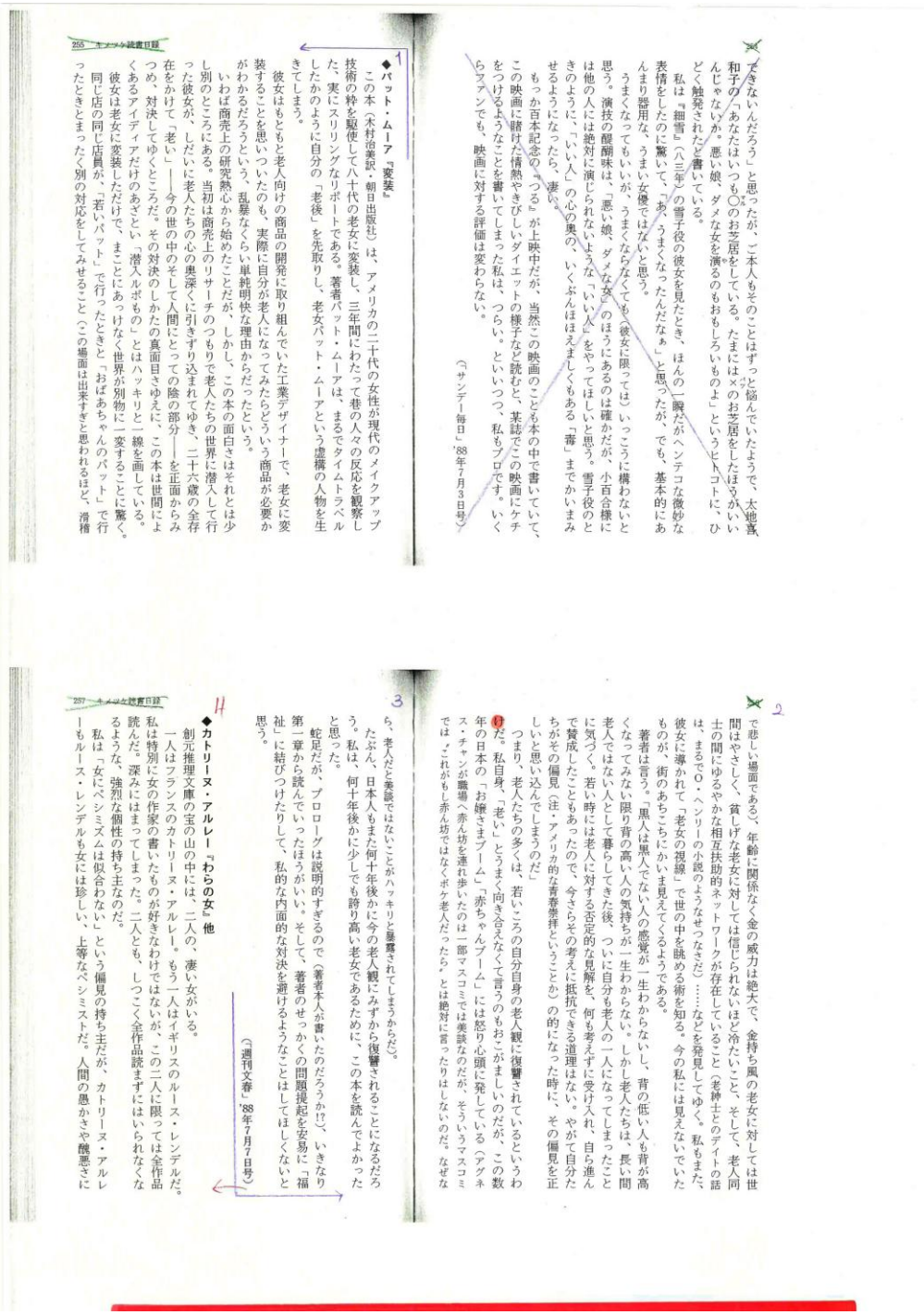


図 3-17　一作品が一つの「理想範囲」（初出表示あり）

一作品を一つの「理想範囲」と考える理由は，作品集におさめられた作品は，別時期に，別媒体にて公表された作品の再録であることが多く，一冊の単行本におさめられているとしても，それらの作品は寄せ集められたものと考えられるためである。よって，まずは書籍のタイトル等に「集」とある場合は，作品単位で「理想範囲」を認定する。また，タイトル等に「集」がない場合でも，各作品が再録であることが初出表示等により明確であれば，一作品を一つの「理想範囲」としている。典型は，個人全集であるが，短篇小説集や，エッセー集等の場合も一作品を一つの理想範囲と認めることが多くある。例えば，図 3-17 のような例である。

なお，そのような作品集については，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，一作品で一つの「理想範囲」であると考える。ただし，サンプル抽出基準点が章立ての文字列の一つである場合は，原則に立ち戻り，その「冊本体」を「理想範囲」とする。

3.4 「完結構造」の捉え方

可変長サンプルの範囲指定における問題として，「完結構造」の把握，すなわち，章節構造の把握そのものの難しさがある。

例えば，図 3-18 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当たった場合である。この「結論」を含む「冊本体」の全体を取ろうとすると，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がたい。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，章節構造の把握には，内容にも踏み込んだ判断がしばしば必要になる。

図 3-18 物理的に 4 章の下位に位置づけられている「結論」

加えて，章節構造の把握のためには，見出しや区切り記号の認定が必要であるが，そこにも問題が多く存在する。見出しや区切記号りについては，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，イラスト，線，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。

以下，図 3-19 はイラストで区切る例，図 3-20 は空行で区切る例である。なお，見出しについては第５章で述べる。

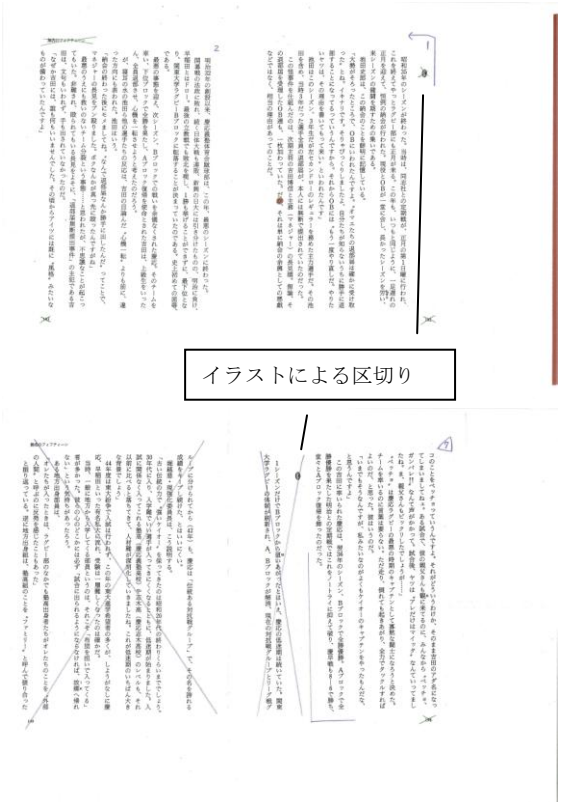


図 3-19 イラストによる区切り

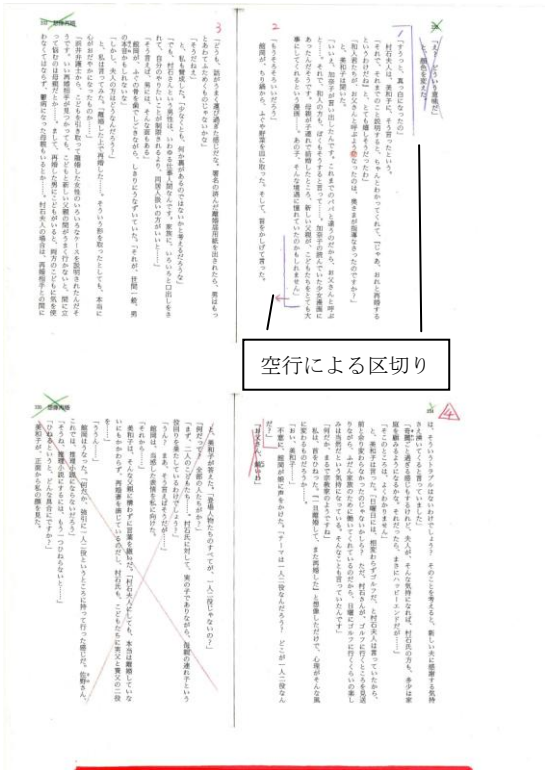


図 3-20 空行による区切り

第４章　対象外要素の排除指定

柏野和佳子・稲益佐知子・田中弥生・秋元祐哉

ここでは，第Ⅰ部第４章で述べた［排除基準 3］～［排除基準 5］の適用によって排除指定をする対象のうち，次の２点について説明する。

①言語表現を主体としない，あるいは，文字列が図式化されている「フィギュア」［排除基準 3］［排除基準 4］

②現代日本語を主体としないブロック形式部分［排除基準 5］

4.1　「フィギュア」

第Ⅰ部で述べた通り，「フィギュア」は「フィギュア本体」とそれに付帯する「キャプション」とに分かれる（以降，「フィギュア本体」のみを「フィギュア」と呼び，「キャプション」は「キャプション」で呼び分ける）。「キャプション」はサンプリング対象であるため，ここでは言及せず，次の第５章で説明する。

「フィギュア」は，［排除基準 3］［排除基準 4］によって，排除されるものである。そのおおよその類型は，次の通りである（類型別の詳細は 4.1.1 節以降に後述する）。

○類型 1
- ・写真
- ▪写し込み

○類型 2
- ・イラスト
- ▪漫画

○類型 3
- ・図解
- ▪グラフ

○類型 4
- ・分岐型フローチャート
- ▪表[1]

類型 1 の写真や，類型 2 のイラストにおいて，そこに一切文字がない場合は，収録可能な文字列が存在しないため，当然，排除対象である。その判断に迷うことは，まず，ない。よって，そのようなものはここではこれ以上言及しない。また，類型 1～4 を通し，「フィギュア」とともにある文字列が「フィギュア」の一部であるのか，そうではなく，［選択基準 4］でサンプリング対象とする「キャプション」であるのか，という判断もまた，しばし

[1] 第Ⅱ部では，サンプリング対象外とするものだけを「表」と呼ぶ。すなわち，第Ⅰ部図 3-5 で示した行列見出しを備えたようなものだけを「表」と呼び，第Ⅰ部図 3-6 のようなものは「表」とは呼ばない。

ば問題になるが，この問題についてもこれ以上は言及しない。

ここでは，「フィギュア」に該当する場合，しない場合の判断に伴う問題について取り上げる。類型 1〜3 においては，文字列を含む「フィギュア」が，「フィギュア」が主体であるものであるかを判断する必要がある。「フィギュア」が主体である場合の類型化が一つの課題であり，当該部分がそれら類型に該当するものであるかの判断が問題になる。類型 4 においては，「文字列が図式化されているか」の判断が問題になる。

以下，類型別に，事例を挙げて上記問題について説明する。

4.1.1 類型 1: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。

図 4-1 写真の一部に文字列（看板の文字）

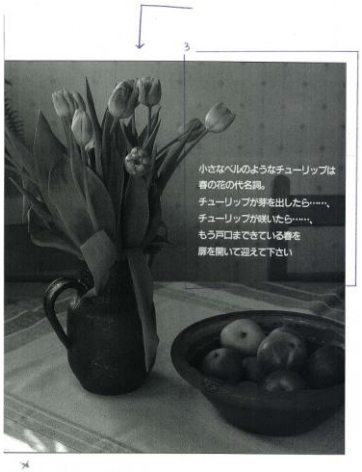


図 4-2 写真の上に文字列

4.1.2　類型 1:写し込み

画像として写し込んだものを，「写し込み」と呼び，写真の下位類型として考える。よって，一部に文字列を含んでいるもの，文字列のみであるものも，この類型で捉えるものは，文字列を含めてすべてを「フィギュア」として排除対象とする。

例えば，DVD などのパッケージ，書籍の表紙等の画像，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のまま貼りつけたようなものも，この延長で考える。

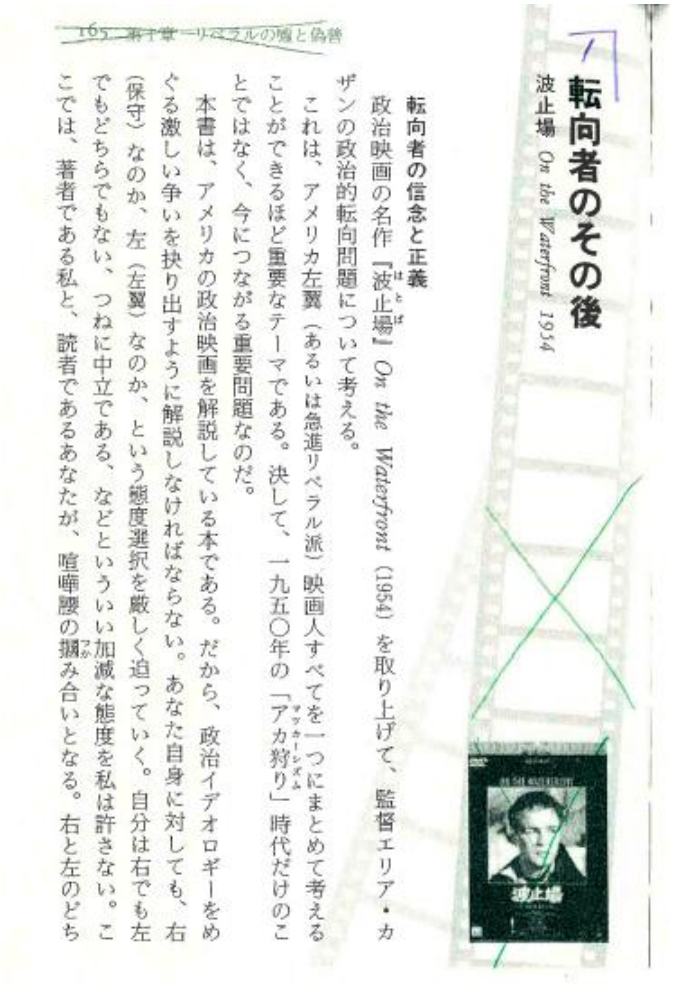


図 4-3 パッケージ

図 4-4 書籍の表紙

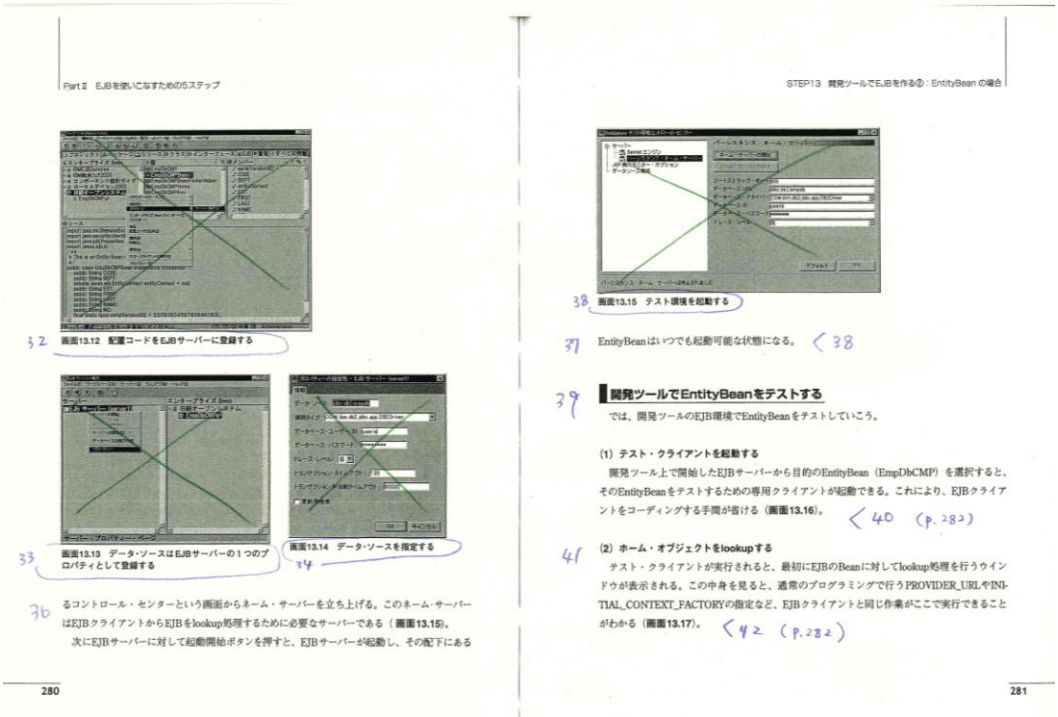


図 4-5 コンピュータのキャプチャ画面

また，文書形式を示す目的の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ものも，この「写し込み」の類型で捉える。例えば，婚姻届，確定申告の書類などである。図 4-6 は，明細書の書式が挿入されている例である。また，紙面をそのまま切り取って提示するようなものも，「写し込み」である。例えば，新聞記事（図 4-7）や週刊誌の記事などである。

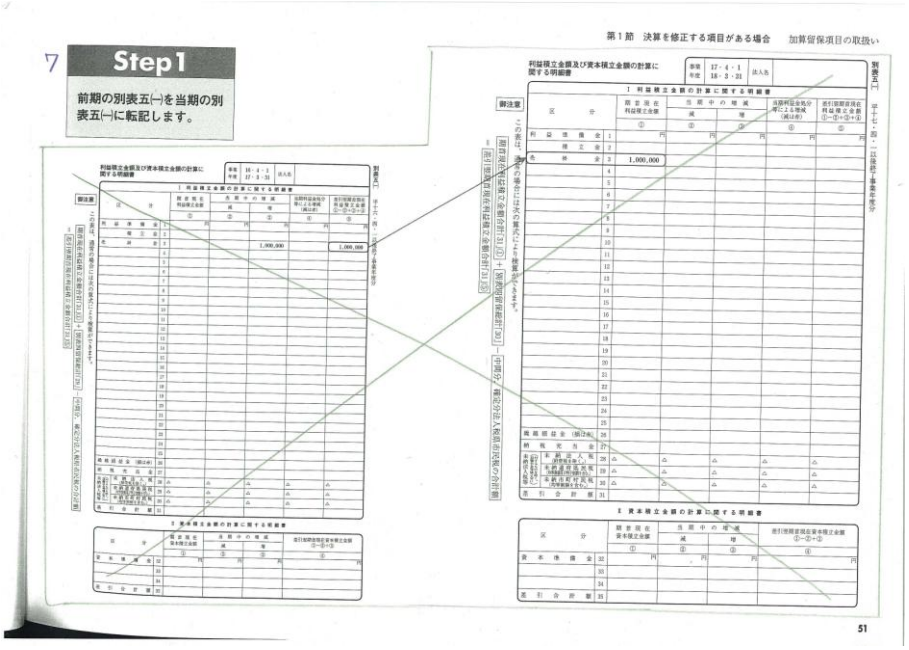


図 4-6 文書形式を示す書式

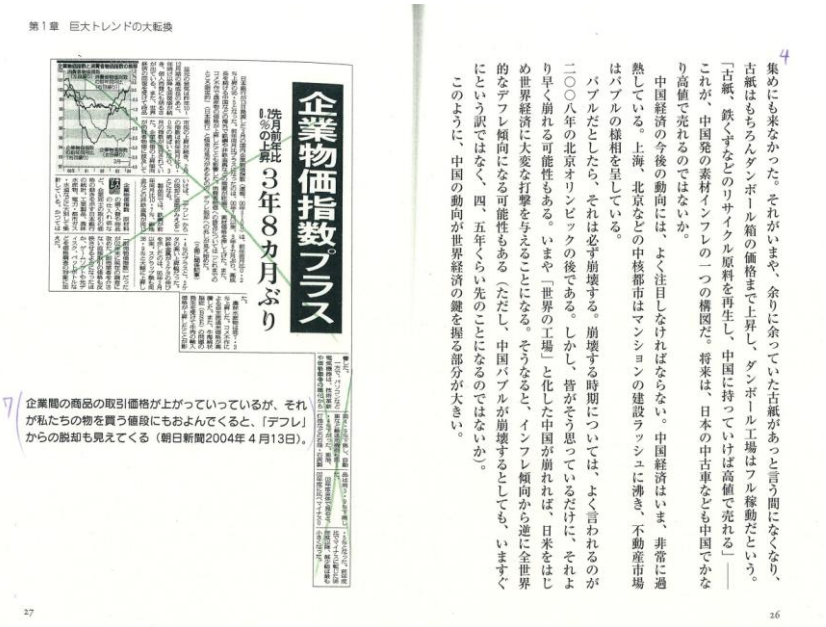


図 4-7 新聞記事

4.1.3 類型 2: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。イラストに文字列が含まれる場合は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，必ずしも表現の主体が文字列であるとも認めがたい。漫画は少なくとも，視覚表現と言語表現の併存によって成り立っているため，文字列のみを抽出し列挙したところで，十分な言語表現であるとは言いがたい。ＢＣＣＷＪでは，漫画をそのようなものと考え，文字列ごと排除対象とする。なお，そもそも本１冊丸ごと漫画であるような場合は，第Ⅰ部 2.2.3 節で述べた通り，言語表現が主体ではないという理由から母集団の定義の時点で除かれている。

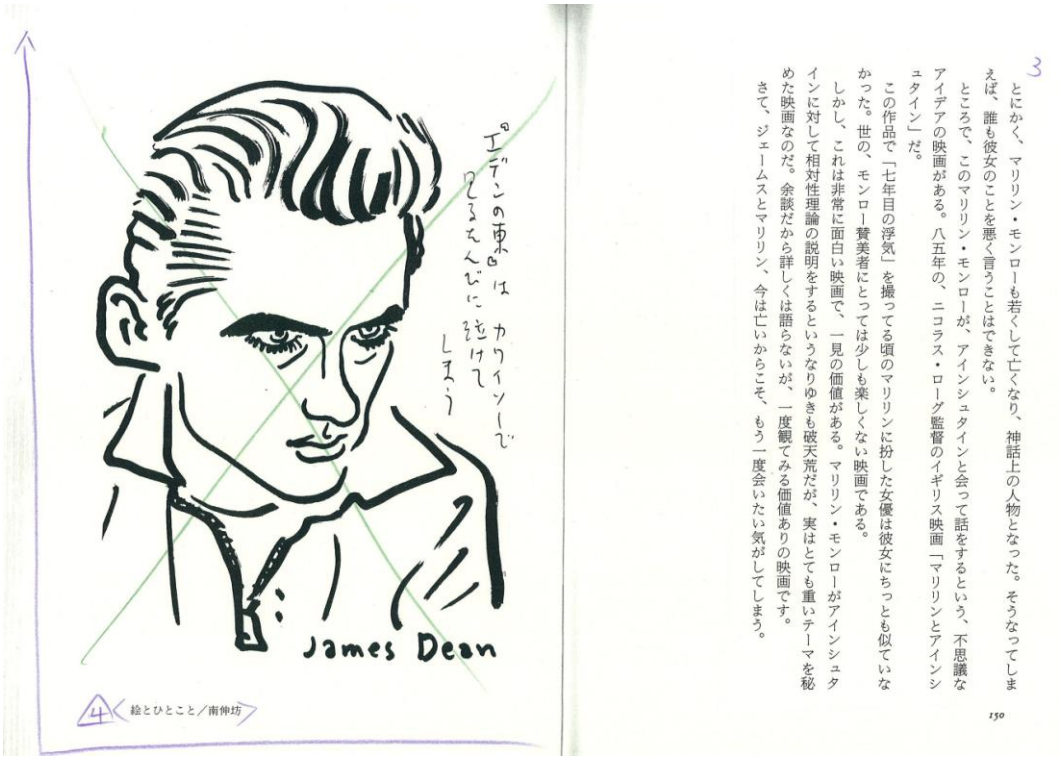


図 4-8 イラスト（中に文字列有り）

4.1.4　類型 3: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，第Ⅰ部第 4 章で述べた［選択基準 2］の章節構造を持つ本文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。

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。

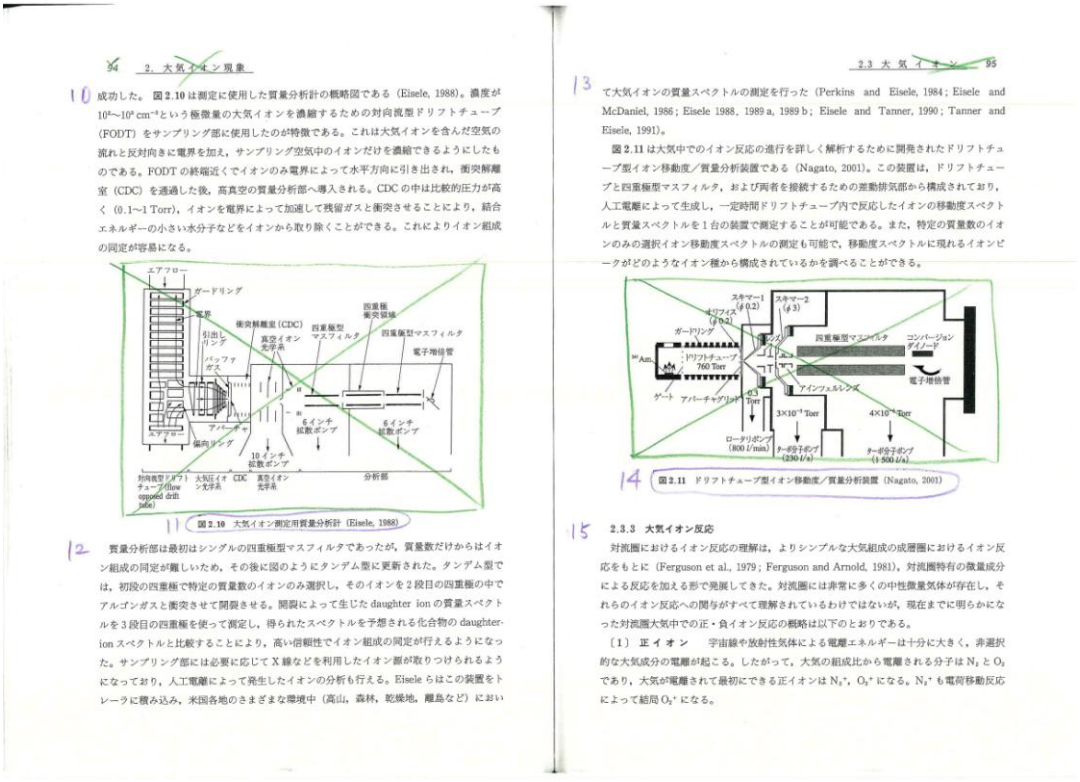


図 4-9 引き出し線付き図解 その 1

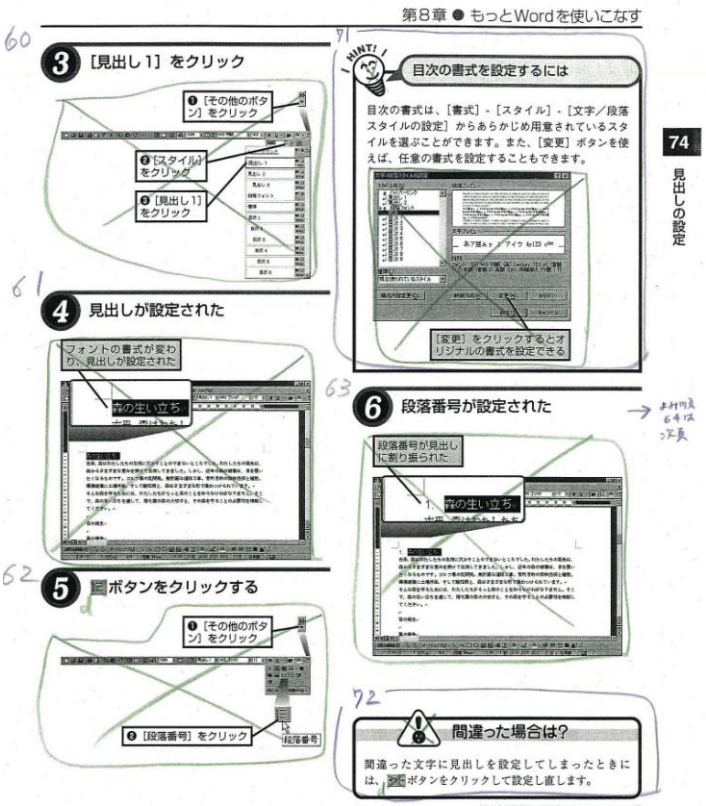


図 4-10 引き出し線付き図解その 2

図 4-11 引き出し線付き文字列部分は本文

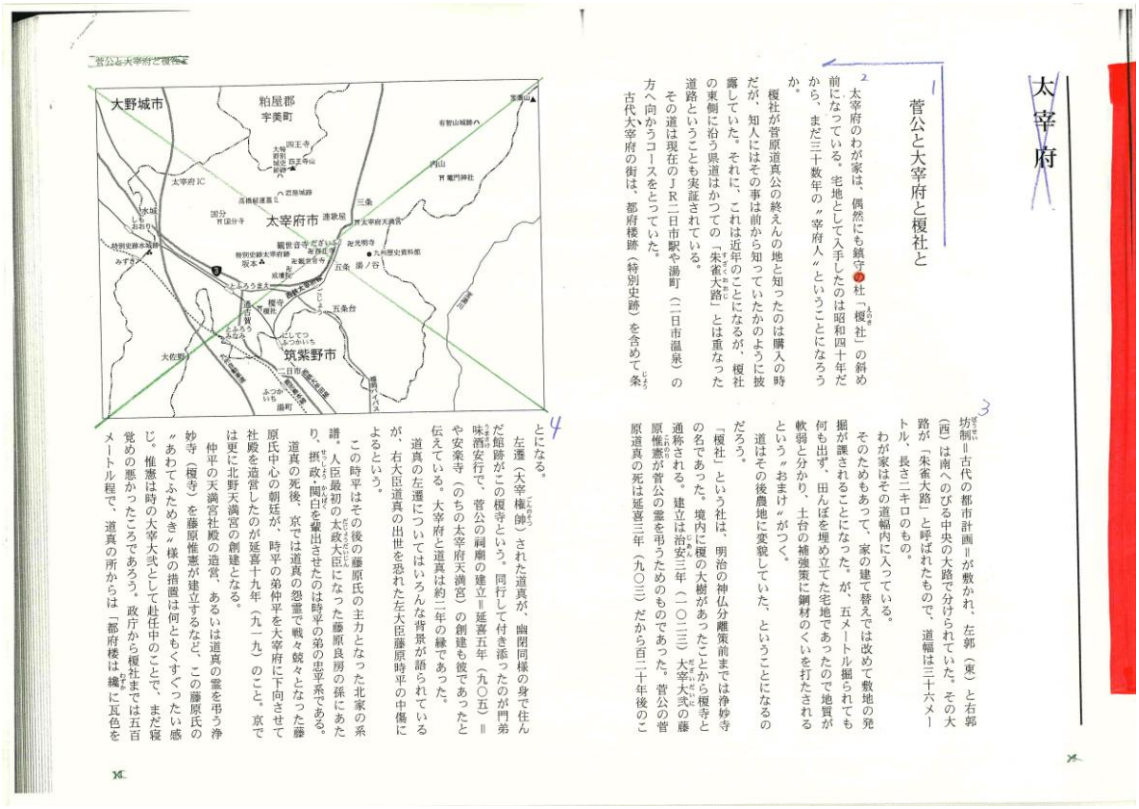
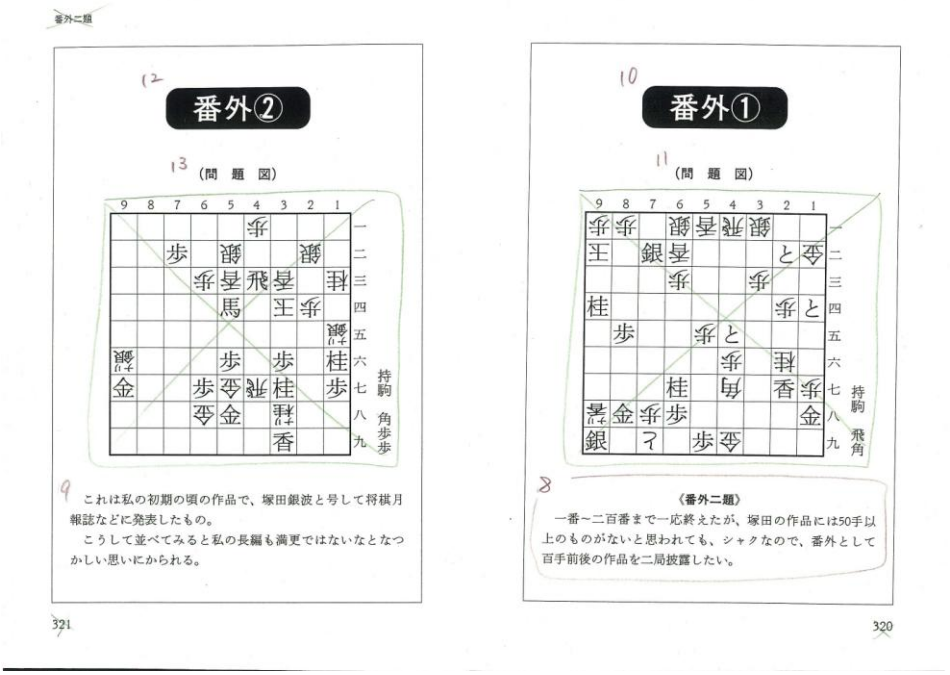


図 4-12 地図

図 4-13 スポーツポジション図



番外二題

番外②

（問 題 図）

持駒 角歩歩

これは私の初期の頃の作品で、塚田銀波と号して将棋月報誌などに発表したもの。

こうして並べてみると私の長編も満更ではないなとなつかしい思いにかられる。

321

番外①

（問 題 図）

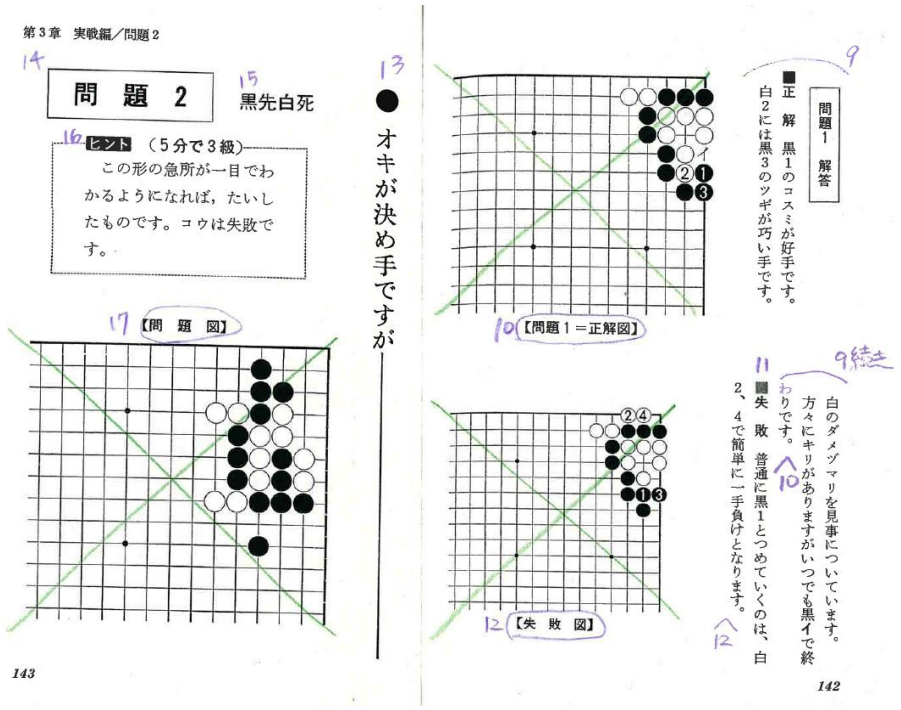
持駒 飛角

《番外二題》

一番～二百番まで一応終えたが、塚田の作品には50手以上のものがないと思われても、シャクなので、番外として百手前後の作品を二局披露したい。

320

図 4-14 棋譜



第3章 実戦編／問題2

問 題 2

黒先白死

ヒント （5分で3級）

この形の急所が一目でわかるようになれば，たいしたものです。コウは失敗です。

【問 題 図】

143

● オキが決め手ですが——

【問題1＝正解図】

問題1 解答

■正 解 黒1のコスミが好手です。白2には黒3のツギが巧い手です。

【失 敗 図】

■失 敗 普通に黒1とつめていくのは、白2、4で簡単に一手負けとなります。

白のダメヅマリを見事についています。方々にキリがありますがいつでも黒イで終わりです。

142

図 4-15 碁譜

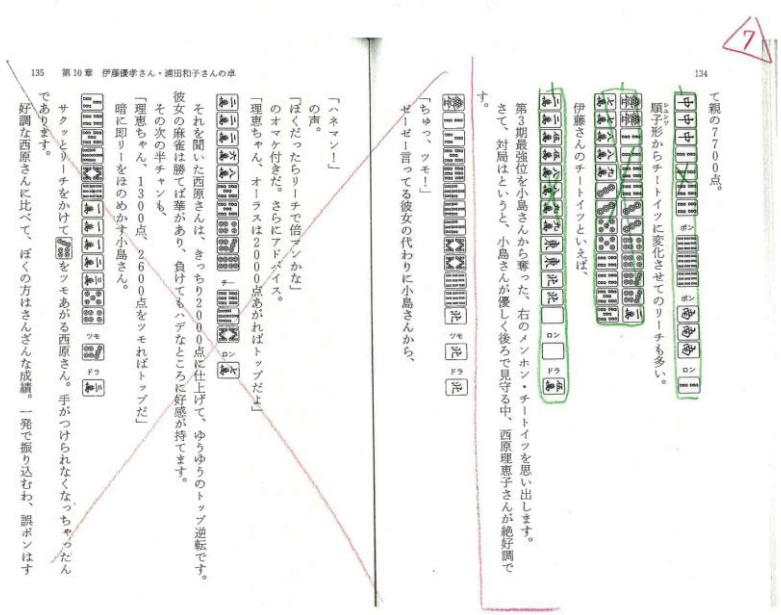


図 4-16 牌図

4.1.5　類型 3: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。

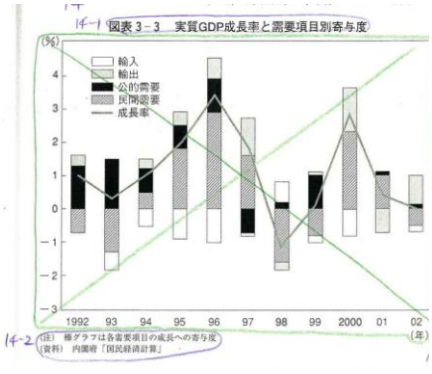


図 4-17 棒グラフ

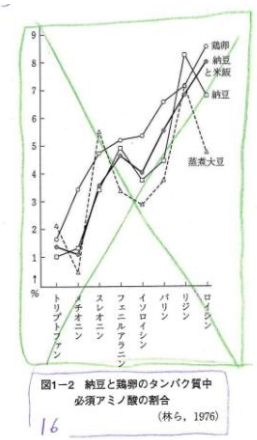


図 4-18 折れ線グラフ

図 4-19 円グラフ

4.1.6　類型 4:分岐型フローチャート

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」の類型の一つと考える。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。

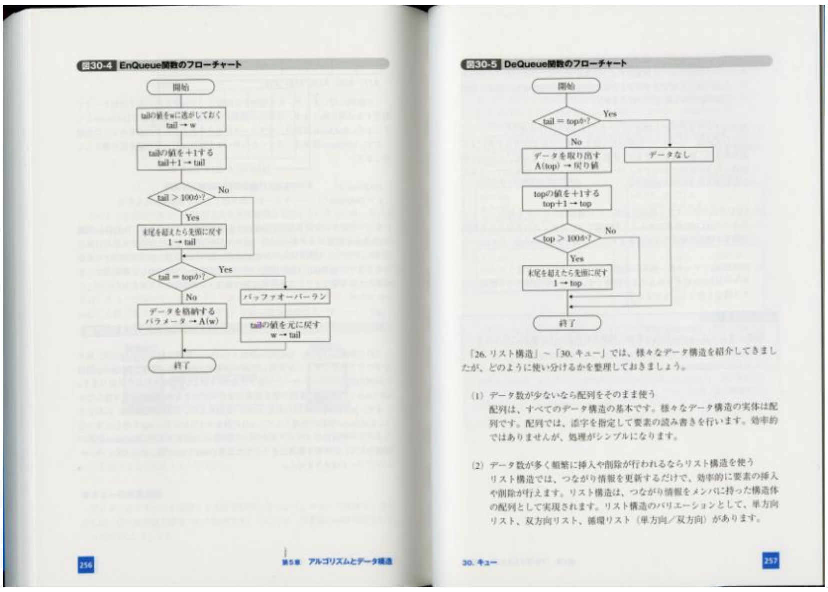


図 4-20 分岐型フローチャートその 1

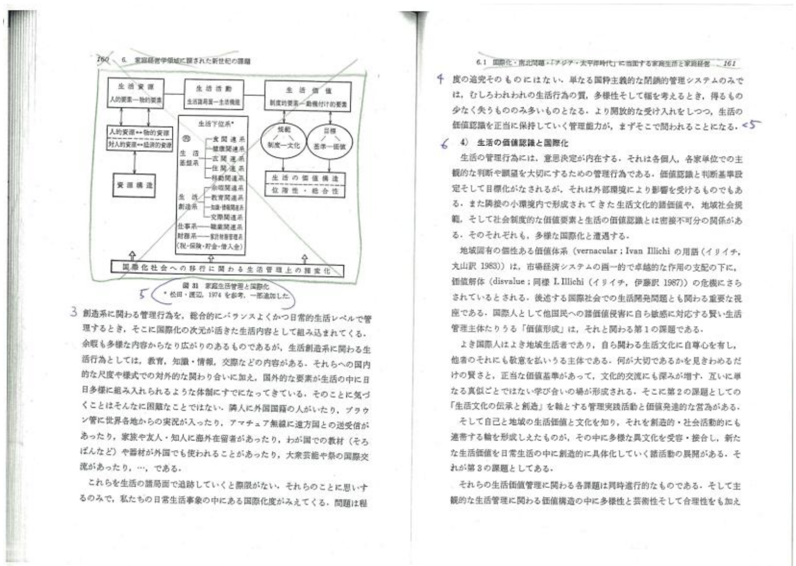
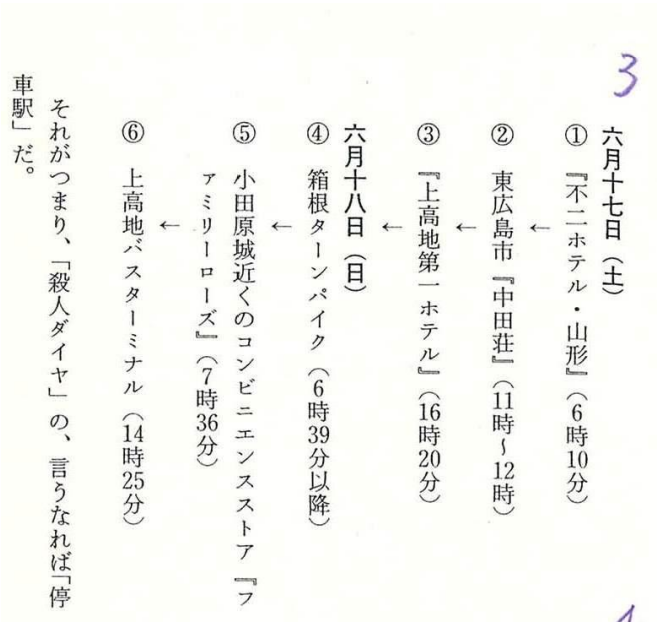


図 4-21 分岐型フローチャート その 2



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。常識的には、通貨はある程度変動させたほうがよいといえますが、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。

経済統合へ進むプロセス

第1段階
自由貿易地域
関税をゼロにしたり、貿易に関する規制をなくす。NAFTAはこれに当たる。

第2段階
関税同盟
自由貿易地域をつくった国が同じ関税を域外の国に適用。つまり域外に対して一致団結すること。

第3段階
共通市場
物だけでなく、人やお金も自由に移動する市場をつくる。

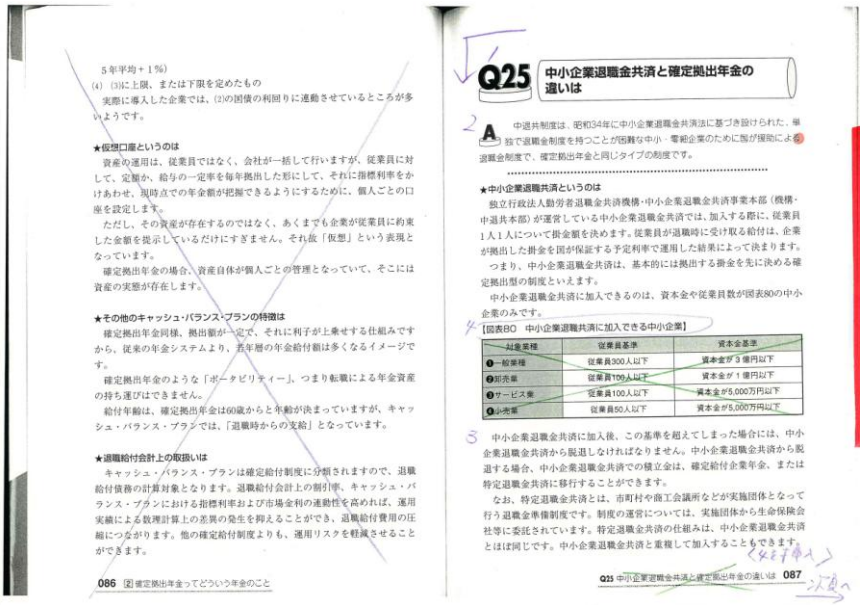
第4段階
経済統合
薬や車などの商品の認可基準を統一する。マーケットが広がるメリットがある。

第5段階
完全な市場統合
通貨統合をしたり、共通の政府をつくるなどそれぞれの国が同一の市場になる。EUは通貨統合が実現した段階。

図 4-23 直線型フローチャート その２

4.1.7 類型 4:表

第 I 部 3.3.2 節でも述べた通り，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。さらに言えば，より典型的な表は罫線が引かれている。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」の類型の一つと考える。



5年平均＋1%)
(4) (3)に上限、または下限を定めたもの
実際に導入した企業では、(2)の国債の利回りに連動させているところが多いようです。

★仮想口座というのは
資産の運用は、従業員ではなく、会社が一括して行いますが、従業員に対して、定期か、給与の一定率を毎年拠出した形にして、それに指標利率をかけあわせ、現時点での年金額が把握できるようにするために、個人ごとの口座を設定します。
ただし、その資産が存在するのではなく、あくまでも企業が従業員に約束した金額を提示しているだけにすぎません。それが「仮想」という表現となっています。
確定拠出年金の場合、資産自体が個人ごとの管理となっていて、そこには資産の実態が存在します。

★その他のキャッシュ・バランス・プランの特徴は
確定拠出年金同様、拠出額が一定で、それに利子が上乗せする仕組みですから、従来の年金システムより、若年層の年金給付額は多くなるイメージです。
確定拠出年金のような「ポータビリティー」、つまり転職による年金資産の持ち運びはできません。
給付年齢は、確定拠出年金は60歳からと年齢が決まっていますが、キャッシュ・バランス・プランでは、「退職時からの支給」となっています。

★退職給付会計上の取扱いは
キャッシュ・バランス・プランは確定給付制度に分類されますので、退職給付債務の計算対象となります。退職給付会計上の割引率、キャッシュ・バランス・プランにおける指標利率および市場金利の連動性を高めれば、運用実績による数理計算上の差異の発生を抑えることができ、退職給付費用の圧縮につながります。他の確定給付制度よりも、運用リスクを軽減させることができます。

086 図 確定拠出年金ってどういう年金のこと

Q25 中小企業退職金共済と確定拠出年金の違いは

A 中退共制度は、昭和34年に中小企業退職金共済法に基づき設けられた、単独で退職金制度を持つことが困難な中小・零細企業のために国が援助による退職金制度で、確定拠出年金と同じタイプの制度です。

★中小企業退職共済というのは
独立行政法人勤労者退職金共済機構・中小企業退職金共済事業本部（機構・中退共本部）が運営している中小企業退職金共済では、加入する際に、従業員1人1人について掛金額を決めます。従業員が退職時に受け取る給付は、企業が拠出した掛金を国が保証する予定利率で運用した結果によって決まります。
つまり、中小企業退職金共済は、基本的には拠出する掛金を先に決める確定拠出型の制度といえます。
中小企業退職金共済に加入できるのは、資本金や従業員数が図表80の中小企業のみです。

【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、または特定退職金共済に移行することができます。
なお、特定退職金共済とは、市町村や商工会議所などが実施団体となって行う退職金準備制度です。制度の運営については、実施団体から生命保険会社等に委託されています。特定退職金共済の仕組みは、中小企業退職金共済とほぼ同じです。中小企業退職金共済と重複して加入することもできます。

Q25 中小企業退職金共済と確定拠出年金の違いは 087

図 4-24 行列見出しを備えた表

しかしながら，第 I 部 3.3.2 節に，「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」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，積極的にサンプリング対象要素と指示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。よって，本節では，「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。

（１）「列見出しを備えた２列」，もしくは，「3 列以上」〔表と認定〕

「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

表 5-5　登録外国人流入数と登録外国人の出身上位国（1999 年）

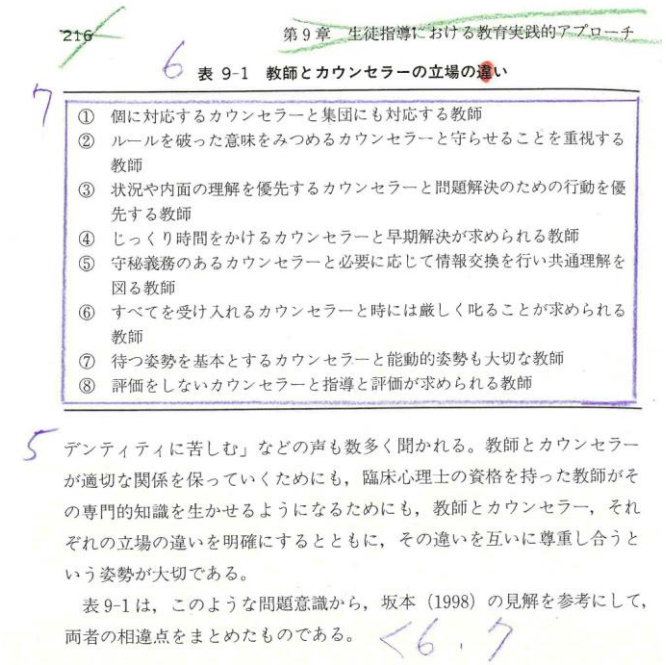
（単位：1000 人）

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	57.8	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	91.4	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

（２）「一方向に読み進められるもの」〔表と認定せず〕

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。



216 第9章 生徒指導における教育実践的アプローチ

表 9-1 教師とカウンセラーの立場の違い

①	個に対応するカウンセラーと集団にも対応する教師
②	ルールを破った意味をみつめるカウンセラーと守らせることを重視する教師
③	状況や内面の理解を優先するカウンセラーと問題解決のための行動を優先する教師
④	じっくり時間をかけるカウンセラーと早期解決が求められる教師
⑤	守秘義務のあるカウンセラーと必要に応じて情報交換を行い共通理解を図る教師
⑥	すべてを受け入れるカウンセラーと時には厳しく叱ることが求められる教師
⑦	待つ姿勢を基本とするカウンセラーと能動的姿勢も大切な教師
⑧	評価をしないカウンセラーと指導と評価が求められる教師

デンティティに苦しむ」などの声も数多く聞かれる。教師とカウンセラーが適切な関係を保っていくためにも，臨床心理士の資格を持った教師がその専門的知識を生かせるようになるためにも，教師とカウンセラー，それぞれの立場の違いを明確にするとともに，その違いを互いに尊重し合うという姿勢が大切である。

表 9-1 は，このような問題意識から，坂本（1998）の見解を参考にして，両者の相違点をまとめたものである。

図 4-27 罫線で囲まれた「一方向に読み進められるもの」

（3）「列見出しを備えない 2 列」〔表と認定せず〕

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。

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」

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 や，図 4-30 のようなものである。

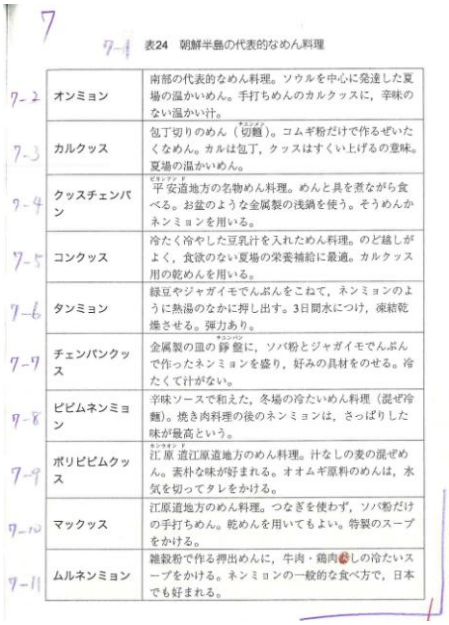


表24　朝鮮半島の代表的なめん料理

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。
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。
クッスチェンパン	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。
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。
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，凍結乾燥させる。弾力あり。
チェンパンクッス	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。
ピビムネンミョン	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。
ポリピビムクッス	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。
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。特製のスープをかける。
ムルネンミョン	雑穀粉で作る押出めんに，牛肉・鶏肉だ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

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

図 4-30 列見出しを備えない２列 その２

（４）「列見出しを備えない２列」：右列非現代日本語〔表と認定〕

「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」を主たる構成要素と認める。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 節で詳述するが，非現代日本語は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 4-31 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。

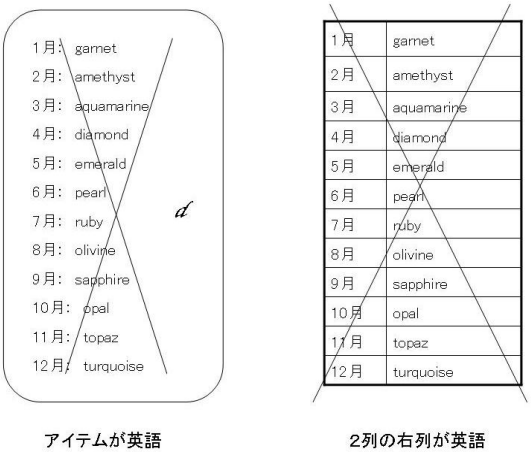


図 4-31 列見出しを備えない２列：右列英語

（５）「列見出しを備えない２列」:右列イラスト・記号〔表と認定〕

「ラベル＋アイテム」の「アイテム」がイラストであれば（４）の非現代日本語の場合と同じく「アイテム」が排除対象であるため，「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。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。その典型の一つには，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 4-32 に右列★で評価を表す例を示す。

図 4-32 列見出しを備えない２列：右列記号

（６）「列見出しを備えない２列」：右列数値〔表と認定〕

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。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，右列の数値が，例えば，アンケート集計結果における，「○人」や「○%」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，4.1.5 節で述べた通り，「フィギュア」の類型であり，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一つとして定めている[1]。

例えば，図 4-33 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

図 4-33 列見出しを備えない２列：右列数値

[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。

（7）「実質一方向に読み進められるもの」〔表と認定せず〕

最後に，（2）で述べた「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 4-34 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る。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 2.1:「日本十進分類法（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	

図 4-34 実質一方向で読み進められる文字列のもの

4.2　現代日本語を主体としないブロック形式

本節では「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部分は排除対象と判断する。その該当条件は三つあり，その一つ目は，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である。非日本語は外国語の場合，非現代語は古典語の場合，非言語は，数式やコンピュータ言語の場合である。該当条件の二つ目は，現代日本語が混在したとしても，非現代日本語が主体と見なせることである。そして三つ目は，インラインではなくブロック形式であることである。

明らかに非現代日本語である要素のみで構成され，ブロック形式であれば，上記の三条件が揃うため，排除対象であることが瞬時に判断可能である。図 4-35～図 4-38 に，ブロック形式においてそれら非現代日本語が現れている，典型的な例を示す。

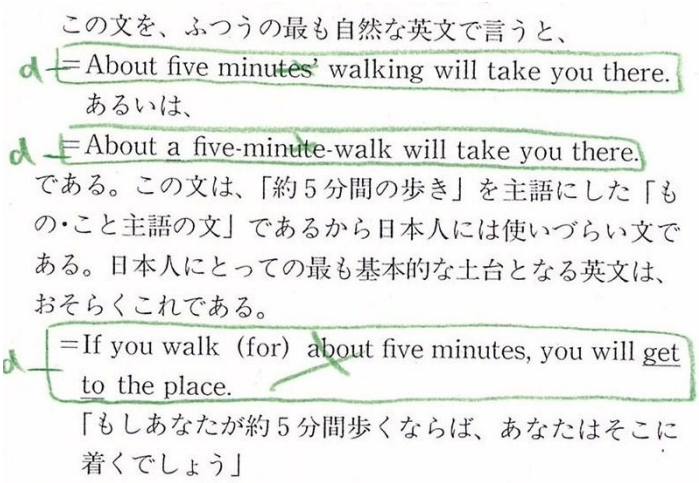

この文を、ふつうの最も自然な英文で言うと、
＝About five minutes' walking will take you there.
あるいは、
＝About a five-minute-walk will take you there.
である。この文は、「約5分間の歩き」を主語にした「もの・こと主語の文」であるから日本人には使いづらい文である。日本人にとっての最も基本的な土台となる英文は、おそらくこれである。
＝If you walk (for) about five minutes, you will get to the place.
「もしあなたが約5分間歩くならば、あなたはそこに着くでしょう」

図 4-35 ブロック形式の英語

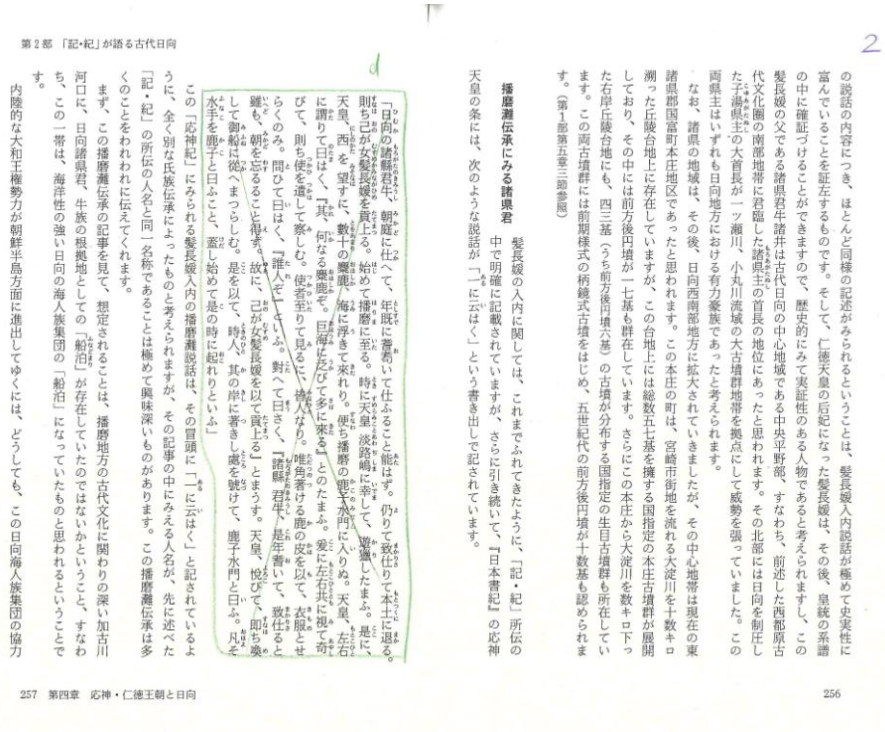


図 4-36 ブロック形式の古典語

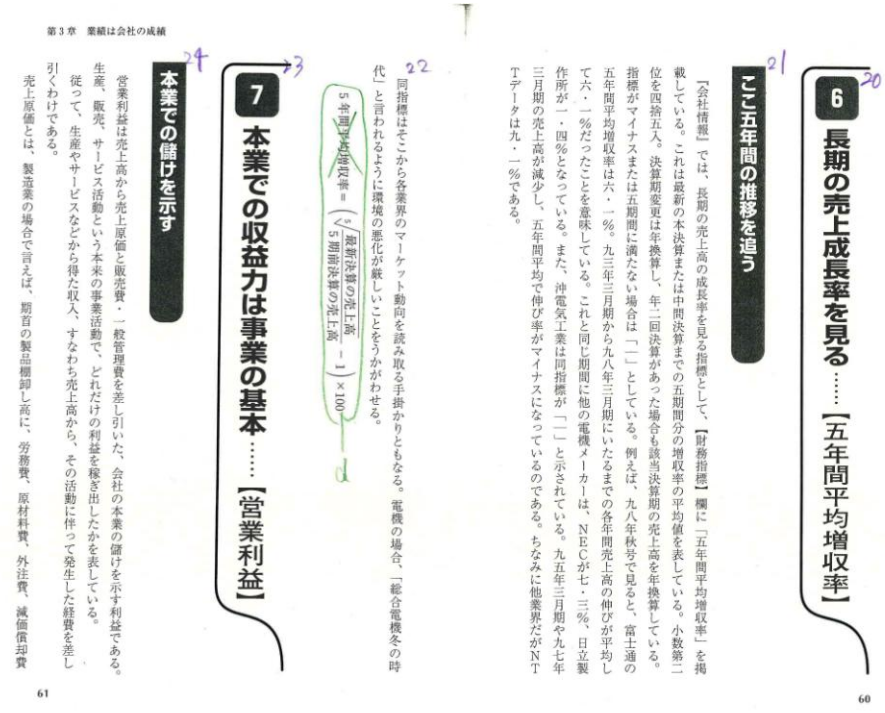
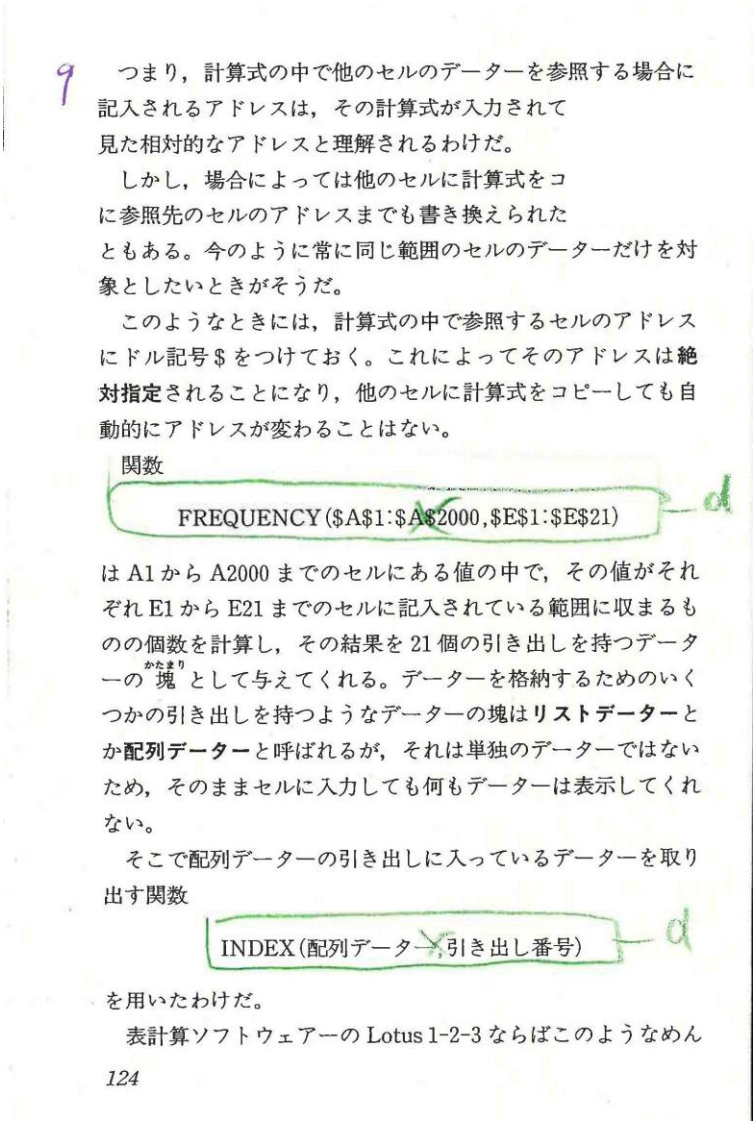


図 4-37 ブロック形式の数式



つまり，計算式の中で他のセルのデーターを参照する場合に記入されるアドレスは，その計算式が入力されて見た相対的なアドレスと理解されるわけだ。

しかし，場合によっては他のセルに計算式をコに参照先のセルのアドレスまでも書き換えられたともある。今のように常に同じ範囲のセルのデーターだけを対象としたいときがそうだ。

このようなときには，計算式の中で参照するセルのアドレスにドル記号＄をつけておく。これによってそのアドレスは**絶対指定**されることになり，他のセルに計算式をコピーしても自動的にアドレスが変わることはない。

関数

FREQUENCY(＄A$1:＄A$2000,＄E$1:＄E$21)

はA1からA2000までのセルにある値の中で，その値がそれぞれE1からE21までのセルに記入されている範囲に収まるものの個数を計算し，その結果を21個の引き出しを持つデーターの塊（かたまり）として与えてくれる。データーを格納するためのいくつかの引き出しを持つようなデーターの塊は**リストデーター**とか**配列データー**と呼ばれるが，それは単独のデーターではないため，そのままセルに入力しても何もデーターは表示してくれない。

そこで配列データーの引き出しに入っているデーターを取り出す関数

INDEX(配列データー,引き出し番号)

を用いたわけだ。

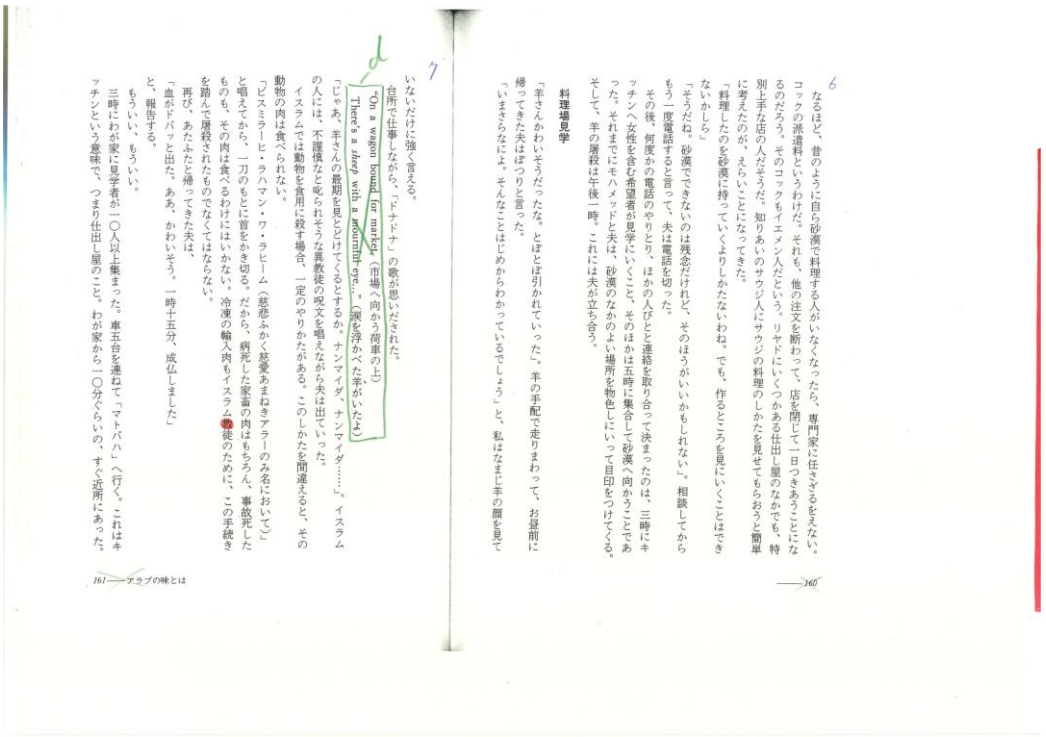
表計算ソフトウェアーのLotus 1-2-3ならばこのようなめん

124

図 4-38 ブロック形式のコンピュータ言語

以上，図 4-35～図 4-38 に，順に，ブロック形式の英語，古典語，数式，コンピュータ言語の部分が排除要素として指定される例を示した。

実作業においては，「非現代日本語か」「非現代日本語が主体か」「ブロック形式か」という条件ごとに，その吟味，判断が必要になる。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。



なるほど、昔のように自ら砂漠で料理する人がいなくなったら、専門家に任さざるをえない。コックの派遣料というわけだ。それも、他の注文を断わって、店を閉じて一日つきあうことになるのだろう。そのコックもイエメン人だという。リヤドにいくつかある仕出し屋のなかでも、特別上手な店の人だそうだ。知りあいのサウジ人にサウジの料理のしかたを見せてもらおうと簡単に考えたのが、えらいことになってきた。

「料理したのを砂漠に持っていくよりしかたないわね。でも、作るところを見にいくことはできないかしら」

「そうだね。砂漠でできないのは残念だけれど、そのほうがいいかもしれない」。相談してからもう一度電話すると言って、夫は電話を切った。

その後、何度かの電話のやりとり、ほかの人びとと連絡を取り合って決まったのは、三時にキッチンへ女性を含む希望者が見学にいくこと、そのほかは五時に集合して砂漠へ向かうことであった。それまでにモハメッドと夫は、砂漠のなかのよい場所を物色しにいって目印をつけてくる。そして、羊の屠殺は午後一時。これには夫が立ち会う。

料理場見学

「羊さんかわいそうだったな。」とぼとぼ引かれていった」。羊の手配で走りまわって、お昼前に帰ってきた夫はぽつりと言った。

「いまさらなによ。そんなことはじめからわかっているでしょう」と、私はなまじ羊の顔を見ていないだけに強く言える。

台所で仕事しながら、「ドナドナ」の歌が思いだされた。

「じゃあ、羊さんの最期を見とどけてくるとするか。ナンマイダ、ナンマイダ……」。イスラムの人には、不謹慎だと叱られそうな異教徒の呪文を唱えながら夫は出ていった。

イスラムでは動物を食用に殺す場合、一定のやりかたがある。このしかたを間違えると、その動物の肉は食べられない。

「ビスミラーヒ・ラハマン・ワ・ラヒーム（慈悲ふかく慈愛あまねきアラーの名において）」と唱えてから、一刀のもとに首をかき切る。だから、病死した家畜の肉はもちろん、事故死したものも、その肉は食べるわけにはいかない。冷凍の輸入肉もイスラム教徒のために、この手続きを踏んで屠殺されたものでなくてはならない。

再び、あたふたと帰ってきた夫は、

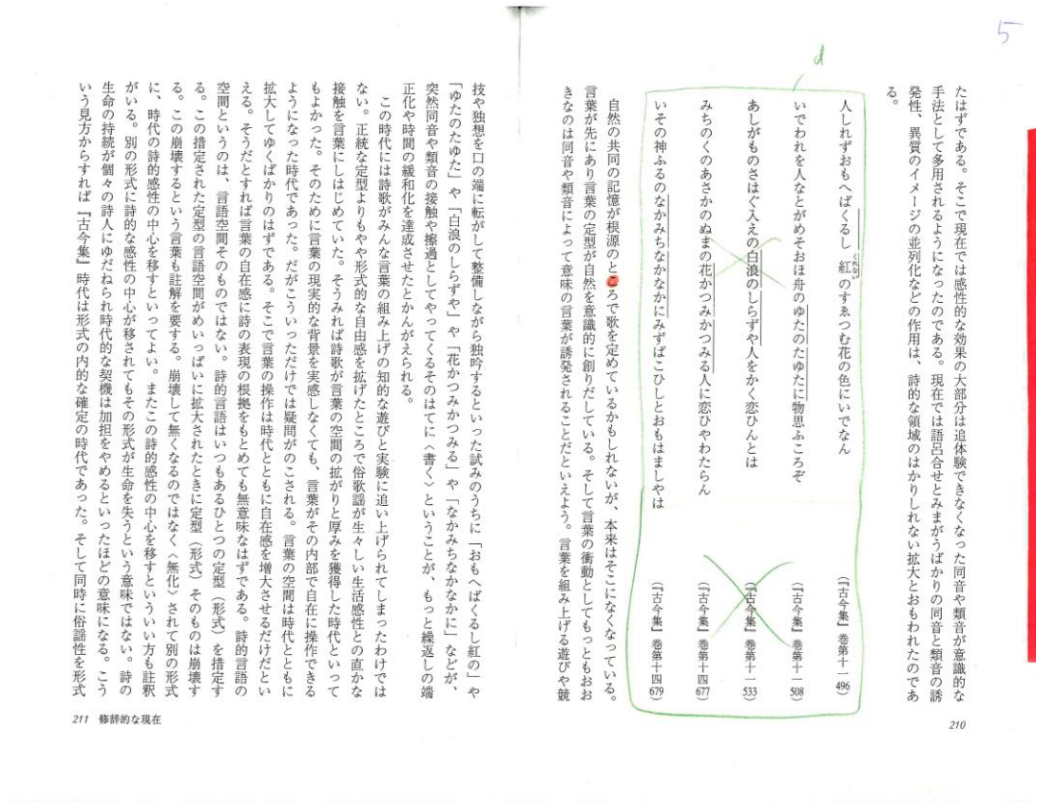
「血がドバッと出た。ああ、かわいそう。一時十五分、成仏しました」

と、報告する。

もういい、もういい。

三時にわが家に見学者が一〇人以上集まった。車五台を連ねて「マトバハ」へ行く。これはキッチンという意味で、つまり仕出し屋のこと。わが家から一〇分ぐらいの、すぐ近所にあった。

図 4-39 ブロック形式の英語に（　）が付くもの



たはずである。そこで現在では感性的な効果の大部分は追体験できなくなった同音や類音が意識的な手法として多用されるようになったのである。現在では語呂合せとみまがうばかりの同音と類音の誘発性、異質のイメージの並列化などの作用は、詩的な領域のはかりしれない拡大とおもわれたのである。

人しれずおもへばくるし（紅）のすそつむ花の色にいでなん　（古今集）巻第十一　496

いでわれを人なとがめそおほ舟のゆたのたゆたに物思ふころぞ　（古今集）巻第十一　508

あしがものさはぐ入えの白浪のしらずや人をかく恋ひんとは　（古今集）巻第十一　533

みちのくのあさかのぬまの花かつみかつみる人に恋ひやわたらん　（古今集）巻第十四　677

いその神ふるのなかみちなかなかにみずぼこひしとおもはましやは　（古今集）巻第十四　679

自然の共同の記憶が根源のところで歌を定めているかもしれないが、本来はそこになくなっている。言葉が先にあり言葉の定型が自然を意識的に創りだしている。そして言葉の衝動としてもっともおおきなのは同音や類音によって意味の言葉が誘発されることだといえよう。言葉を組み上げる遊びや競技や独想を口の端に転がして整備しながら独吟するといった試みのうちに「おもへばくるし紅の」や「ゆたのたゆた」や「白浪のしらずや」や「花かつみかつみる」や「なかみちなかなかに」などが、突然同音や類音の接触や擦過としてやってくるそのはてに〈書く〉ということが、もっと繰返しの端正化や時間の緩和化を達成させたとかんがえられる。

この時代には詩歌がみんな言葉の組み上げの知的な遊びと実験に追い上げられてしまったわけではない。正統な定型よりもやや形式的な自由感を拡げたところで俗歌謡が生々しい生活感性との直かな接触を言葉にしはじめていた。そうみれば詩歌が言葉の空間の拡がりと厚みを獲得した時代といってもよかった。そのために言葉の現実的な背景を実感しなくても、言葉がその内部で自在に操作できるようになった時代であった。だがこういっただけでは疑問がのこされる。言葉の空間は時代とともに拡大してゆくばかりのはずである。そこで言葉の操作は時代とともに自在感を増大させるだけだいえる。そうだとすれば言葉の自在感に詩の表現の根拠をもとめても無意味なはずである。詩的言語の空間というのは、言語空間そのものではない。詩的言語はいつもあるひとつの定型（形式）を措定する。この措定された定型の言語空間がめいっぱいに拡大されたときに定型（形式）そのものは崩壊する。この崩壊するという言葉も註解を要する。崩壊して無くなるのではなく（無化）されて別の形式に、時代の詩的感性の中心を移すといってよい。またこの詩的感性の中心を移すといういい方も註釈がいる。別の形式に詩的な感性の中心が移されてもその形式が生命を失うという意味ではない。詩の生命の持続が個々の詩人にゆだねられ時代的な契機は加担をやめるといったほどの意味になる。こういう見方からすれば『古今集』時代は形式の内的な確定の時代であった。そして同時に俗謡性を形式

図 4-40 ブロック形式の古典語に（　）が付くもの

第５章　サンプリング対象要素の確定と入力順の指定

柏野和佳子・稲益佐知子・田中弥生・秋元祐哉

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を，サンプルとして収録すべき対象部分として確定し，電子テキストとして収録するための入力順を指示することである。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，一次元の文字列として取得できているかを最終確認することである。

これまでの作業段階において対象外要素を排除した結果，収録対象として残る文字列部分は，次のいずれかである。

・見出し[1]
・本文
・キャプション
・注（脚注・後注）

以下，それぞれについて，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ふる際に留意する点を述べる。

5.1　「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 5-1 と図 5-2 は，同じサンプルの別紙面の画像である。図 5-2 が見出しのあるページであり，図 5-1 はその次の見開きページの左上に示されていた図 5-2 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 5-2 では，「見出し」部分の入力順の指示をしている。

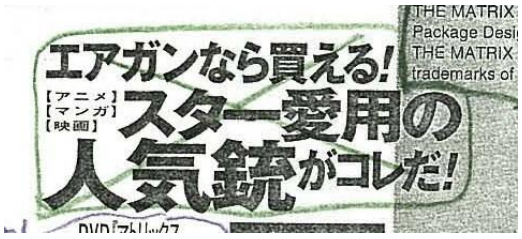


図 5-1 見出しの入力順指示に留意するものの「柱」部分

[1] ここで「見出し」とは，第Ｉ部 4.2 節「選択基準」で述べた，「本文」を統括する見出しである「章節見出し」のことを言う。以下，同じ。

図 5-2 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。このことは，第Ⅰ部 4.3 節で[運用基準 3]として述べた通りである。

例えば，図 5-3 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 5-4 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。さらに，非言語で入力できない，例えばイラストそのものが「見出し」相当である場合は，そのことを表すタグの入力を指示する。

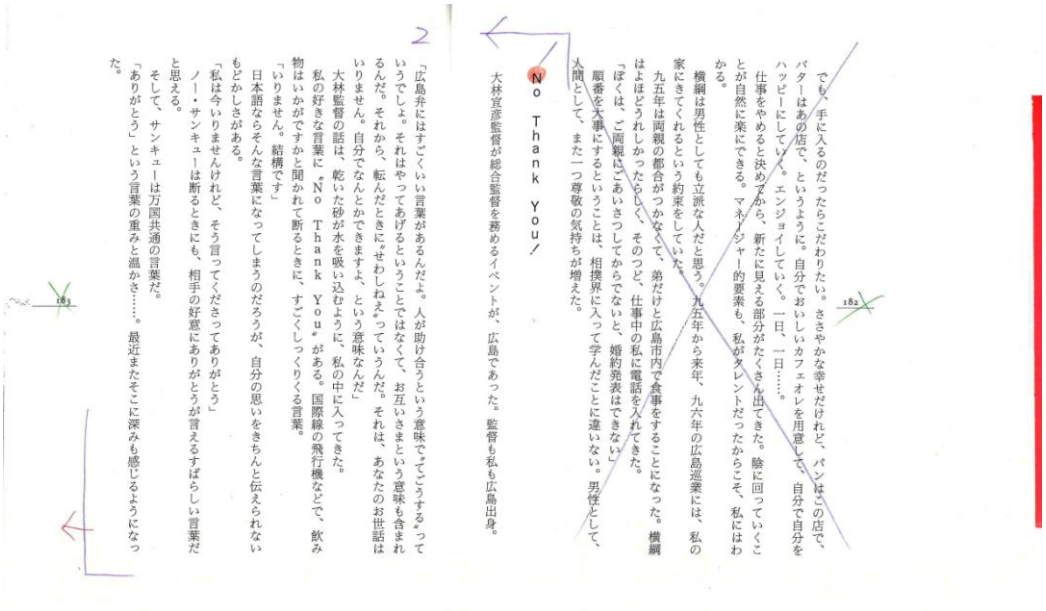


図 5-3 「見出し」が英語

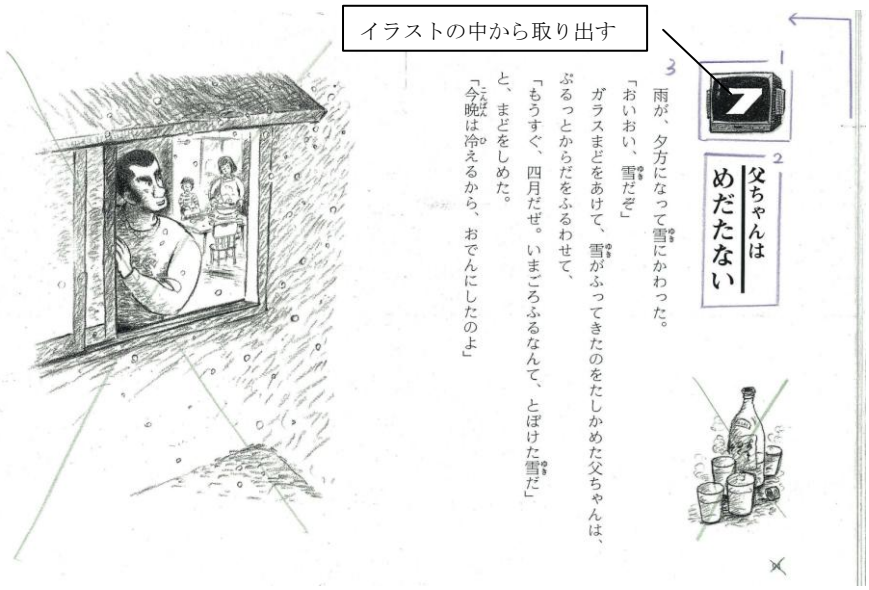


図 5-4 「見出し」の文字列をイラストの中から取り出すもの

5.2 「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 5-5 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 5-5 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 5-6 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。

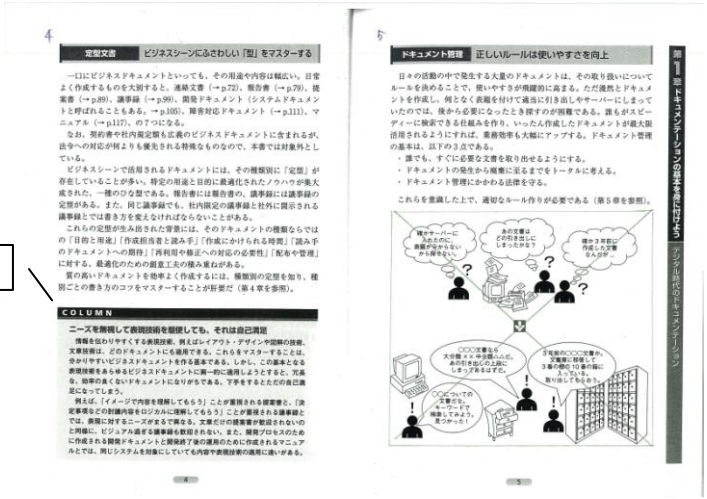


図 5-6 「コラム」を道なりに入力するもの

一方，次の図 5-7 は，コラムが「Ⅴ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，「Ⅴ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」との形式の比較等により，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。

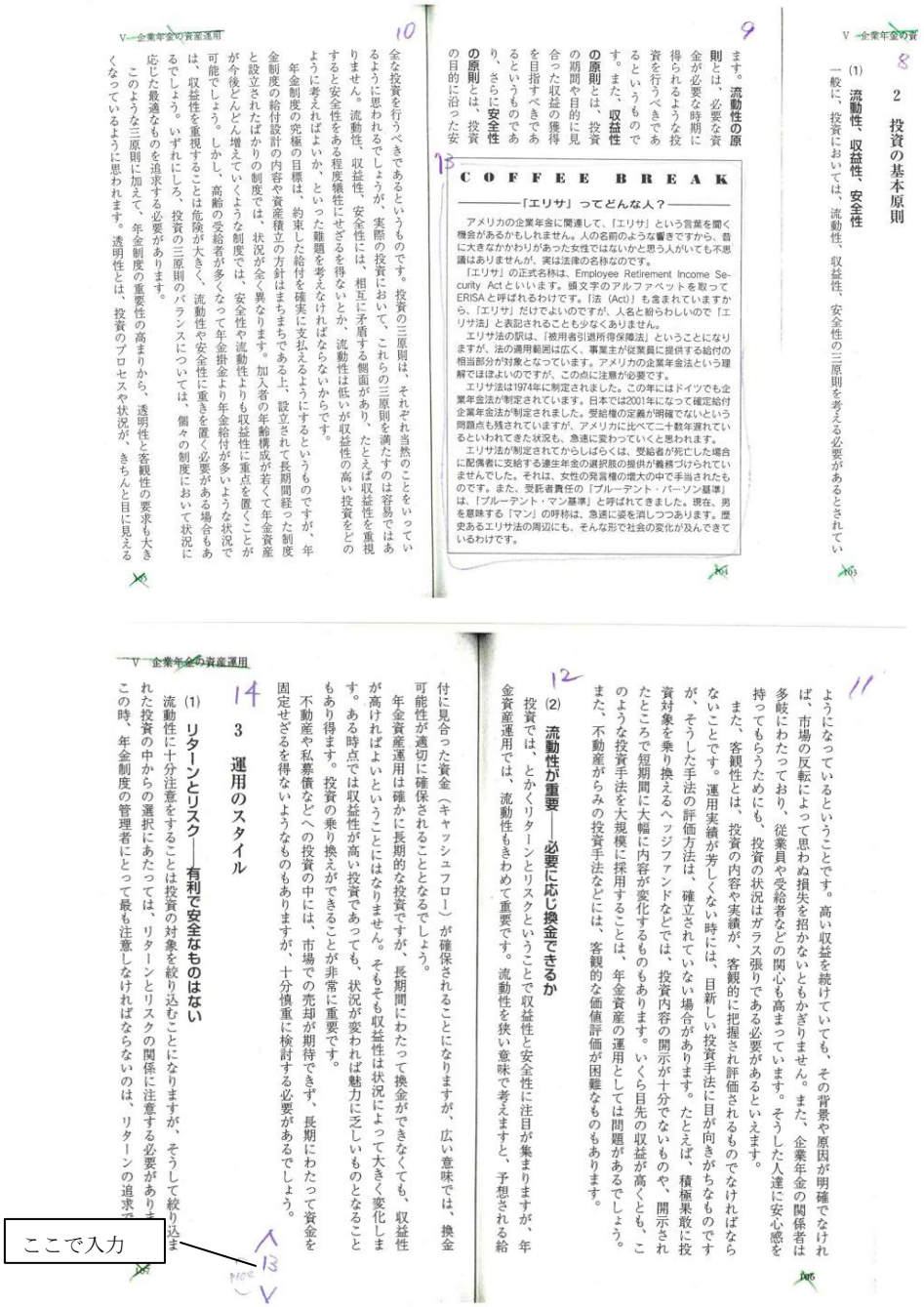
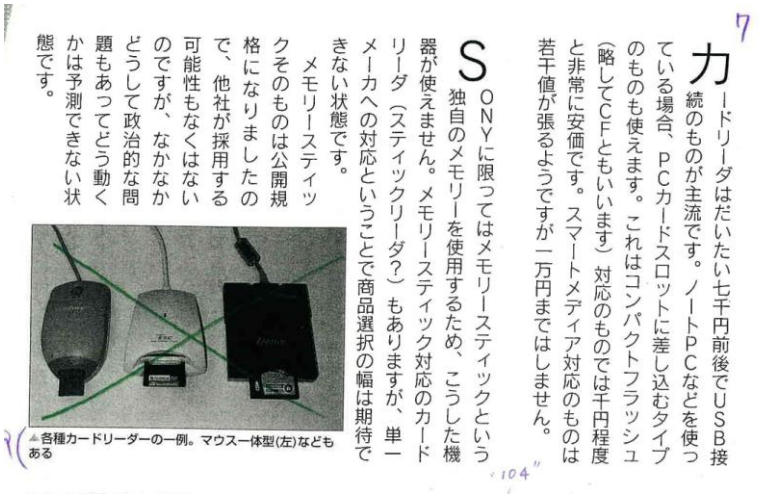


図 5-7 「コラム」を章節末に入力するもの

5.3 「キャプション」

第Ⅰ部第4章で述べたように，「キャプション」は，［選択基準4］により，収録対象である。写真に伴う「キャプション」の典型例を図 5-8 に，表に伴う「キャプション」の典型例を図 5-9 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。


力ードリーダはだいたい七千円前後でUSB接続のものが主流です。ノートPCなどを使っている場合、PCカードスロットに差し込むタイプのものも使えます。これはコンパクトフラッシュ（略してCFともいいます）対応のものでは千円程度と非常に安価です。スマートメディア対応のものは若干値が張るようですが一万円まではしません。

SONYに限ってはメモリースティックという独自のメモリーを使用するため、こうした機器が使えません。メモリースティック対応のカードリーダ（スティックリーダ?）もありますが、単一メーカへの対応ということで商品選択の幅は期待できない状態です。

メモリースティックそのものは公開規格になりましたので、他社が採用する可能性もなくはないのですが、なかなかどうして政治的な問題もあってどう動くかは予測できない状態です。

▲各種カードリーダーの一例。マウス一体型(左)などもある

図 5-8 写真の「キャプション」

表3・4 CSK標準溶液一覧

標準溶液	濃度（μmol/l）	溶液	容器
リン酸塩（PO_4-P）	0, 0.5, 1, 2, 3	3% NaCl	ガラスアンプル
ケイ酸塩（SiO_2-Si）	0, 5, 10, 25, 50, 100, 150, 200	3% NaCl	ポリエチレン瓶
亜硝酸塩（NO_2-N）	0, 0.25, 0.5, 1, 2	水溶液	ガラス瓶
硝酸塩（NO_3-N）	0, 5, 10, 15, 20, 30, 40	3% NaCl	ガラス瓶
ヨウ素酸カリウム（KIO_3）	0.01000 N	水溶液	ガラス瓶

相模中央化学研究所製作

表3・5 重金属標準溶液

元素	溶液	pH	酸	濃度（μg/l）	容器
水銀（Hg）	3% NaCl	1	H_2SO_4	0, 1, 10, 100, 1000	ガラスアンプル
カドミウム（Cd）	水溶液	3	HCl	0, 5, 10, 100, 1000	ポリエチレン瓶
銅（Cu）	水溶液	3	HCl		
亜鉛（Zn）	水溶液	3	HCl		
鉛（Pb）	水溶液	3	HCl		
ヒ素（As）	水溶液	中性	—	0, 0.5, 10, 50, 100, 1000	

相模中央化学研究所製作

水溶液標準試料としては海水の‘塩分測定用標準海水’が出ており，また表3・4に示すような栄養塩および酸素分析の際の標準溶液，表3・5のような重金属標準溶液が相模中央研究所の協力で市販されている．

図 5-9 図表の「キャプション」

なお，第Ⅰ部4.4節で説明した通り，例えば，「カタログ」のような紙面の場合，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」に相当する文字列は，「本文」とし

て認定される。例えば，次の図 5-10 のようなものである。また，その次に示す，図 5-11，図 5-12 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 5-10 カタログのような紙面で写真に伴う「キャプション」相当文字列が「本文」であるもの

図 5-11 写真に伴う「キャプション」相当文字列が「本文」であるもの

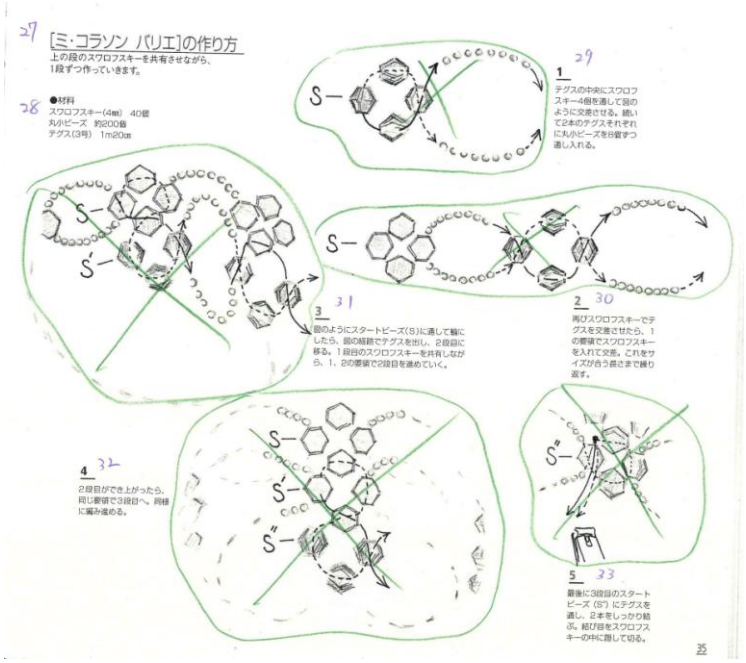


図 5-12 イラストに伴う「キャプション」相当文字列が「本文」であるもの

5.4 「注」

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 5-13 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。

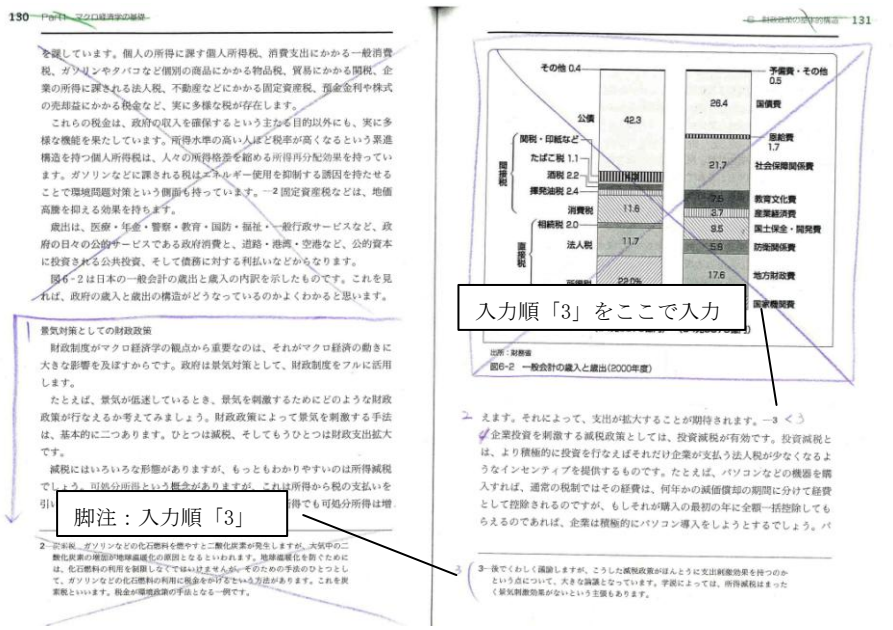


図 5-13 脚注：マーカーあり

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 5-14 は，注マーカーのない脚注の例である。語句の対応から，脚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。

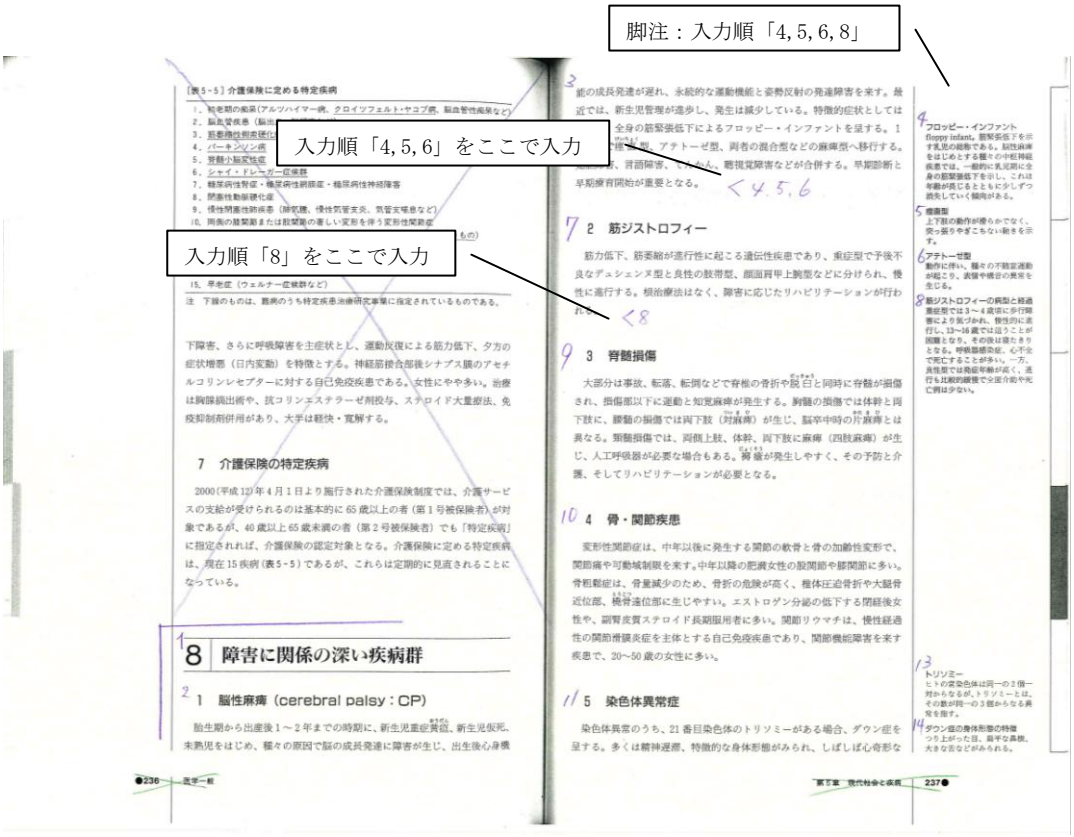


図 5-14 脚注：マーカーなし

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，可変長サンプル外にあれば，範囲外のものとして，収録対象とはしない。

第６章　まとめ

柏野和佳子・稲益佐知子・田中弥生・秋元祐哉

第Ⅱ部では，文章構造に沿って範囲を絞り，サンプリング対象文字列を絞り，最終的にＢＣＣＷＪに収録するテキストを抽出するまでに至る方法を，可能な限りゆれなく実行するための，大小様々な，数多くの基準や手順を示した。そして，実作業に含まれる複雑さや作業負担についても言及した。

我々の基準や手順のポイントは，文章構造に基づく範囲判断と，抽出する文字列の選択にある。

もし，文章構造を気にせずに，ページ単位に無作為抽出することにすれば，サンプル作成の作業効率は上がる。しかし，1 ページを超えるまとまった文章構造を取得したいという要求に応えられなくなるなどの問題が生じる。よって，文章構造に基づく範囲判断を外すことはできない。

また，もし，我々が排除要素として定める，ノンブルや柱，フィギュアの文字列，非現代日本語などを排除するという判断をなくし，ページから文字列すべてを抽出する方法をとったとすれば，飛躍的にテキスト抽出の作業効率は上がるだろう。しかしながら，一方向に読み進められないようなものが混在するテキストがそのまま収録されただけのコーパスは，非常に使いにくい。文字列の種々のレベルの違いに細かく対応した電子テキストの詳細な構造化が必要になるうえ，その後の形態素解析時にも大きな負荷がかかることが予想される。よって，収録テキストの抽出段階で，一方向に読み進められないようなものを排除する必要性は高いのである。

第Ⅱ部では，5 年間で，汎用的な 1 億語規模の大規模コーパスを均質的に構築するという課題実現のための，具体的なサンプル作成，収録テキスト抽出のための作業基準，および手順を報告した。本報告をきっかけとして，様々な観点から議論が深まることを期待したい。

出典一覧

図 1-2：佐藤寛，青山温子|編著(2005)『生活と開発』日本評論社
図 1-3：中央青山監査法人，中央青山 PwC サステナビリティ研究所|編(2003)『環境経営なるほど Q&A 環境先進企業へのヒント』中央経済社
図 2-1：細川幹夫|著(2002)『トヨタ成長のカギ創業期の人間関係』近代文芸社
図 2-2：石森史郎|著(1992)『エレ Once upon a time in…』新評論
図 2-3：竹河聖|著(1990)『後ろのローラさん』集英社
図 2-4：朝尾直弘|ほか著(1999)『堺の歴史都市自治の源流』
図 2-5：上條さなえ|著(2001)『子どもの言葉はどこに消えた？崩れゆく親子関係』角川書店
図 2-6：デイヴィッド・エディングス|著;宇佐川晶子|訳(1988)『予言の守護者』早川書房
図 2-7：上條さなえ|著(2001)『子どもの言葉はどこに消えた？崩れゆく親子関係』角川書店
図 2-8：中野百々造|著(2003)『会社法務と税務設立、増資・減資、合併、組織変更、解散、分割、株式交換、株式移転、企業組織再編等の会社実務』税務研究会出版局
図 2-9：金田浩|著;三村信英|監修(2001)『21 世紀の慢性透析治療法を革命しよう健常者に限りなく近い長期生存を可能にするために』東京医学社
図 2-10：小山政彦|著(2002)『社長の仕事 48 の鉄則船井総研社長が提言！　会社を強くする「ヒト・モノ・カネ」の実践ノウハウ』大和出版
図 2-11：南原幹雄|著(2004)『御三家の反逆』徳間書店
図 3-2：キネマ旬報社|編(1996)『日本映画人名事典』男優篇　下巻キネマ旬報社
図 3-3：田中一|編(2001)『社会情報学』培風館
図 3-4：石浦章一，小林秀明，塚谷裕一|著(2001)『生物の小事典』岩波書店
図 3-5：久米裕|著(2005)『血統クラシックロード』2005　春-秋白夜書房
図 3-6：増井金典|著(2005)『語源を楽しむ知って驚く日常日本語のルーツ』ベストセラーズ
図 3-7：野崎歓|著(2001)『ジャン・ルノワール越境する映画』青土社
図 3-8：日本山岳会|編著(2005)『新日本山岳誌』ナカニシヤ出版
図 3-9：養老孟司|ほか著(1999)『脳が語る科学』青土社
図 3-10：毎日新聞社|編(1989)『地球環境の危機人口環境開発』毎日新聞社
図 3-11：竹宮惠子|著(2001)『竹宮惠子のマンガ教室』筑摩書房
図 3-12：福島みずほ|著(2005)『戦争と憲法危機の時代に政治をあきらめない 』明石書店
図 3-13：柳家花緑，小野幸恵|著;大野伸彦|写真(2003)『花緑の落語江戸ものがたり師匠小さんの想い出とたどる』近代映画社
図 3-14：一海知義|編(1986)『河上肇獄中往復書簡集』上岩波書店
図 3-15：三宅弘|編(2005)『Q&A 個人情報保護法解説』三省堂
図 3-16：(2005)『簡単！おいしい！煮もの上手料理ならおまかせ』世界文化社
図 3-17：中野翠|著(1998)『ムテッポー文学館』文藝春秋
図 3-18：C. キース・コナーズ，ジュリエット・L. ジェット|著;佐々木和義|訳(2004)『ADHD 注意欠陥/多動性障害の子への治療と介入』金子書房
図 3-19：松本正志|著(1986)『ラガーメン日本の大学ラグビーを創った男たちの人間ドラマ』飛鳥新社
図 3-20：佐野洋|著(2000)『一人二役時代連作推理小説』光文社
図 4-1：菅間誠之助|著(1987)『酒つくりの匠たち老杜氏の語る日本の酒』柴田書店
図 4-2：(1994)『春の花の贈り物英国式フラワーアレンジメントリフレッシュ版』同朋舎出版
図 4-3：副島隆彦|著(2004)『ハリウッド映画で読む世界覇権国アメリカ』下講談社
図 4-4：勝尾金弥|著(1999)『伝記児童文学のあゆみ 1891 から 1945 年』ミネルヴァ書房
図 4-5：樋口研究室|著;日経オープンシステム|監修(2001)『基礎からわかるサーバー・サイド Java,Java サーブレット、JSP、JavaBeans、EJB による Web アプリケーション開発』日経 BP 社;日経 BP 出版センター(発売)
図 4-6：山崎郁雄|編著(2002)『契約書式の作成全集』自由国民社
図 4-7：浅井隆|著(2004)『いよいよインフレがやってくる！』第二海援隊
図 4-8：清水義範|著(1993)『映画でボクが勉強したこと 』毎日新聞社

図 4-9：日本大気電気学会|編(2003)『大気電気学概論』コロナ社
図 4-10：田中亘，インプレス書籍編集部|著(1999)『できる Word 2000Windows 版』インプレス;インプレス販売(発売)
図 4-11：小山健治，海野京子|著(2002)『スキャナ+パソコン画像かんたん活用術 Windows XP　このとおりやればすぐできる！！』技術評論社
図 4-12：岬茫洋|著(2002)『九州の苗字を歩く』福岡編梓書院
図 4-13：ワールドサッカーダイジェスト|責任編集(2005)『ヨーロッパサッカー・トゥデイ』2005-2006　シーズン開幕号日本スポーツ企画出版社
図 4-14：塚田正夫|著(1999)『塚田詰将棋代表作』日本将棋連盟
図 4-15：南善己|著(2005)『初級者の詰碁入門』山海堂
図 4-16：西原理恵子，山崎一夫|著(2001)『サクサクさーくる』角川書店
図 4-17：浜田宏一，堀内昭義，内閣府経済社会総合研究所|編(2004)『論争日本の経済危機長期停滞の真因を解明する』日本経済新聞社
図 4-18：渡辺杉夫|著(2002)『納豆原料大豆の選び方から販売戦略まで』農山漁村文化協会
図 4-19：竹内誠|監修(2002)『お江戸の歩き方歴史を体感する、タイムマシン時代の観光ガイド』学習研究社
図 4-20：矢沢久雄|著(2003)『新人 SE のための「基本情報技術者」入門』翔泳社
図 4-21：日本家政学会|編(1989)『家庭生活の経営と管理』朝倉書店
図 4-22：津村秀介|著(1995)『上高地・芦ノ湖殺人事件』光文社
図 4-23：竹中平蔵|著(2000)『竹中教授のみんなの経済学』幻冬舎
図 4-24：原彰宏|著(2005)『「Q&A」確定拠出年金ハンドブック』セルバ出版;創英社(発売)
図 4-25：坂元良暢|著(2003)『限界突破の社長塾心の時代の必勝戦略』郁朋社
図 4-26：西川潤|著(2004)『世界経済入門』岩波書店
図 4-27：田中雄三，森谷寛之|共編(2001)『生徒指導と心の教育』実践編培風館
図 4-28, 4-30, 4-34：(本報告書 第 I 部)
図 4-29：岡田哲|編(2001)『コムギの食文化を知る事典』東京堂出版
図 4-32：(2001)『オフロードバイク&ギアカタログ』2001[エイ]出版社
図 4-33：市川房枝記念会出版部|編(2002)『〞市川房枝政治参画センターで学ぶ 47 人の挑戦』市川房枝記念会出版部
図 4-32：中野昭一，重田定義|編(1992)『図説からだの事典』朝倉書店
図 4-35：副島隆彦|著(1997)『英文法の謎を解く』続筑摩書房
図 4-36：日高正晴|著(2003)『西都原古代文化を探る東アジアの視点から』鉱脈社
図 4-37：日本経済新聞社|編(1999)『日経会社情報徹底活用法』日本経済新聞社
図 4-38：保江邦夫|著(2001)『Excel で学ぶ量子力学量子の世界を覗き見る確率力学入門』講談社
図 4-39：樋口健夫，樋口容視子|著(1986)『住んでみたサウジアラビアアラビア人との愉快なふれあい』サイマル出版会
図 4-40：吉本隆明|著(2005)『戦後詩史論』思潮社
図 5-1, 5-2：(2003)『熱狂！ホビー王』v. 2 ベストセラーズ
図 5-3：相原勇|著(1997)『ありがとう』近代映画社
図 5-4：辻邦|文;夏目尚吾|画(1990)『父ちゃんはナンバーワン！』童心社
図 5-5：発想工房ジースタッフ|編(2004)『北海道　にっぽんの旅』昭文社
図 5-6：永山嘉昭，山崎紅，黒田聡|著(2005)『説得できるドキュメンテーション 200 の鉄則デジタル時代の文書はこう作成・管理する』日経 BP 社;日経 BP 出版センター(発売)
図 5-7：久保知行|著(2004)『わかりやすい企業年金』日本経済新聞社
図 5-8：園田誠|著(2001)『デジタルカメラ 100 の技読んで納得！うまくなる！』技術評論社
図 5-9：西村雅吉|著(1991)『環境化学』裳華房
図 5-10：K-Writer's Club|著(1997)『世界と日本のウィスキー・カタログ』西東社
図 5-11：岩下聆|監修(2001)『手軽にかんたんフィットネスできる、つづける、若がえる』新星出版社
図 5-12：ウタ・オーノ|著(2002)『エレガント・ビーズ・リングウタ・オーノのすてきな世界』小学館
図 5-14：伊藤元重|著(2002)『マクロ経済学』日本評論社
図 5-14：日本社会福祉士養成校協会|監修(2003)『社会福祉士のための基礎知識』3 中央法規出版

おわりに

均衡コーパス（Balanced Corpus）の価値は、それがいかにバランスのよい言語データの集合になっているかによって決まる。その一方で、バランスが取れていることの妥当性を客観的に評価するための尺度は、今のところ存在しない。

そのような中で、均衡コーパスの妥当性を評価するための手がかりとなり得るのは、設計方針と構築手順のあり方であろう。すなわち、そのコーパスがどのように設計され、どのように構築されたか、ということの内実を明らかにしておくことが、コーパスの評価を考える上でまず必要であると思われる。

BCCWJ の持つ最大の特徴は、綿密な調査に基づいて推計した文字数による母集団の厳密な定義とそれに基づく構成比率の算出、そして数多くの基準をもとに均質的な手続きで無作為に抽出されたテキストを収録した、大規模な均衡コーパスであるという点にある。均衡コーパスとしての妥当性を測る尺度の開発は今後の課題としても、コーパスの設計方針およびサンプリングの実施手順がこのような形で明示され、かつ実践されているという点において、従来の均衡コーパスには見られなかった品質が実現されていると考える。

本報告書では、BCCWJ 構築にかかるサンプリングの基本方針、および作業手順について、具体的な事例を挙げながら示した。多様な体裁を持つ書き言葉の印刷紙面から、コーパスに収録するテキストを均質的な手続きで抽出するためには、書き言葉の構造の大局的な把握や、種々の事例に即した数多くの基準と手順の取り決めが必要になる。特に、本報告書で示したような、書き言葉の構造を踏まえたサンプル範囲の認定基準，およびそこから抽出する文字列の選択基準は、質の高いサンプリングを実現する上で、不可欠なものと考える。

関連文献

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). 「書籍の生産実態を反映するサンプリング —NDC ごとに取得したサンプルの多様性の分析—」. 『言語処理学会 第 14 回年次大会 発表論文集』, 939-942. 言語処理学会.

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.

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 (2009). 「『現代日本語書き言葉均衡コーパス』のサンプル収録方法」. 『言語処理学会 第 15 回年次大会 発表論文集』. 言語処理学会.

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.

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.

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 (2007). 「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—」. 『言語処理学会 第 13 回年次大会 発表論文集』. 言語処理学会.

丸山岳彦・柏野和佳子・山崎誠・佐野大樹・秋元祐哉・稲益佐知子・吉田谷幸宏 (2007). 「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概要」. 『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.

丸山岳彦・柏野和佳子・山崎誠・佐野大樹・秋元祐哉・稲益佐知子・田中弥生 (2008). 「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概要 (2) —流通実態サブコーパスの設計—」. 『特定領域「日本語コーパス」平成 19 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.

丸山岳彦・柏野和佳子・山崎誠・佐野大樹・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子 (2008). 「『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概要（3）—

代表性を実現するためのサンプリング手法—」『特定領域「日本語コーパス」平成 20 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.

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.

山崎誠・丸山岳彦・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 (2009). 「現代日本語書き言葉均衡コーパスのサンプル長と言語的特徴 —固定長サンプルと可変長サンプルの質的な違い—」. 『言語処理学会 第 15 回年次大会 発表論文集』. 言語処理学会.

Maruyama, Takehiko, Makoto Yamazaki, and Kikuo Maekawa (2009). "Statistical sampling method used in the Balanced Corpus of Contemporary Written Japanese", *The proceedings of 18th International Congress of Linguists*.

研究開発部門言語資源グループ （サンプリングサブグループ）

山崎 誠　　（研究開発部門グループ長（副））
柏野 和佳子 *　（研究開発部門主任研究員）
丸山 岳彦 *　（研究開発部門研究員）
佐野 大樹　　（研究開発部門特別奨励研究員）
秋元 祐哉 *　（研究開発部門研究補佐員）
稲益 佐知子 *　（研究開発部門研究補佐員）
田中 弥生 *　（研究開発部門研究補佐員）
大矢内 夢子　（研究開発部門研究補佐員）

（* 印は主たる執筆者）

国立国語研究所内部報告書（LR-CCG-08-01）

『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

平成21年3月24日
執 筆 者　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
発 行 者　独立行政法人国立国語研究所
〒190–8561 東京都立川市緑町10番地の2
電 話　042（540）4300（代表）

　（平20–8）

国立国語研究所