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―

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
独立行政法人　国立国語研究所

1. はじめに

国立国語研究所では現在,『現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese; 以下 BCCWJ と記す)』の構築を進めている[1]。BCCWJ は全体で1億語を超す規模を持ち，図１に示す３つのサブコーパス(SC)から構成される。構築期間は2006～2010 年度であり，現在，サンプリング・電子化・著作権処理・形態論情報付与などの作業が進められている。

図１ BCCWJ の構成

３つのサブコーパスのうち，「生産実態（出版）サブコーパス」に含まれる「書籍」のサンプルは，現在，取得目標値の半数にあたる約 6,600 サンプルが取得できている。

本コーパスの目標の一つは，現代日本語書き言葉の多様な姿をとらえることである。そこで本稿では，サンプリングの過程で観察されたサンプルの多様性を報告するとともに，将来コーパスを利用する際，これらの多様性をどのような観点から分析すべきかについての観測を述べる。以下，2章でBCCWJの設計とサンプリングの概略[2]を述べ，3章で取得した書籍サンプルの多様性を分析する。

2. 生産実態サブコーパスの設計とサンプリングの方針

2.1 生産実態サブコーパスの設計

生産実態サブコーパスは，書き言葉が生産される局面に着目して母集団を定義するものである。対象は，2001～2005 年に出版された全ての書籍，雑誌，新聞である。統計的な言語調査を行うために必要なサンプルサイズとして 1,000 万語を想定し，母集団から 1,000 万語分の「固定長サンプル[3]」を抽出することにした。各媒体から取得するサンプル数は，各母集団の総文字数を推計し，その比率を割り当てる。５年間に出版された各媒体の総文字数を推計したところ，書籍が約 485 億字，雑誌が約 105 億字，新聞が約 64 億字という結果を得た（丸山・秋元 2007)。ここから，書籍 74%，雑誌 16%，新聞 10%，という構成比率を定めた。固定長サンプル 1,000 万語（1,700 万字と推定）を得るための必要数として，書籍 12,604 サンプル，雑誌 2,730 サンプル，新聞 1,700 サンプルという数を算出した。固定長サンプルと同時に「可変長サンプル[4]」も抽出するため，生産実態サブコーパス全体の規模は，約 3,500 万語の見積もりになる。

2.2 NDC による層別

生産実態サブコーパスの「書籍」部分の母集団は，国立国会図書館の蔵書のうち，対象の 5 年間に発行されたすべての書籍から漫画や写真集などを除いた 317,117 冊である。この母集団を，出版年および「日本十進分類法（NDC）」によって層別することにした。国立国会図書館で付与された NDC の１次区分の 10 分類（0. 総記，1. 哲学，2. 歴史，3. 社会科学，4. 自然科学，5. 技術工学，6. 産業，7. 芸術，8. 言語，9. 文学）に「分類なし」を加えた 11 分類と，出版年５年とで，合計 55 層に層別した。各層に含まれる推計総文字数の比例割当により，NDC ごとに抽出するサンプル数を算出した。合計 12,604 サンプルを取得する際の NDC 別のサンプル数と割合を，図２に示す。また，ここから取得される書籍の例を表１に示す。

3. 書籍サンプルの多様性の分析

3.1 多様性をとらえる観点

生産実態サブコーパスの「書籍」の層別には，上記の通り，NDC が用いられている。NDC はそもそもは図書館の資料を分類するための指標であり，書籍の主題や内

[1] 詳細は，http://www2.kokken.go.jp/kotonoha/ ，http://www.tokuteicorpus.jp/を参照。
[2] 詳細は，丸山・秋元(2007)，丸山ほか(2007)を参照。
[3] 1 サンプル 1,000 文字に固定したサンプル。丸山ほか(2007)を参照。
[4] 章や節などの言語的なまとまりを取得するサンプル。可変長１サンプルの平均長は，書籍 3,900 文字，雑誌 3,000 字，新聞 1,000 文字と試算している。よって可変長サンプルのみの合計は約 2,500 万語の見積もり。

容，形式に基づいて，1 次区分で 10 種，2 次区分でさらに 10 種，3 次区分でさらに 10 種と，階層的で詳細な分類が施されている。さらに，国立国会図書館では分類の統一性を図るために分類基準が明文化されており，書籍を客観的に分類する上で有用な情報である。

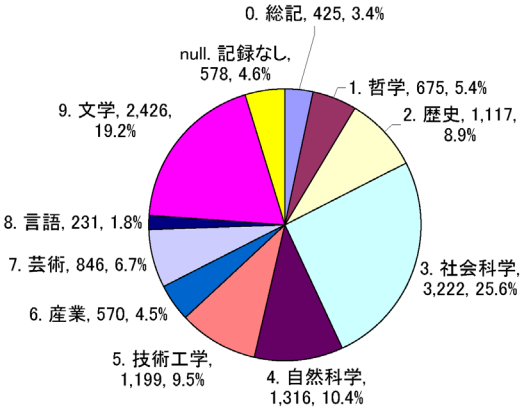


図 2　抽出するサンプル数(NDC 別)とその割合　(%)

表 1　取得する書籍の例　(NDC 別)

NDC	著者	出版年	書名	出版社
007	柴森純著	2004	文系プログラマー奮戦記 同人ゲームとソフトハウスのトンデモ世界	工学社
070	石澤靖治著	2001	大震災とメディア	文藝春秋
134	ヘーゲル[著];長谷川宏訳	2003	歴史哲学講義 下	岩波書店
198	梅原猛著	2004	法然の苦しみ 上	小学館
210	岩田明著	2004	消えたシュメール王朝と古代日本の謎	学習研究社
290	内田芳明著	2001	風景の発見	朝日新聞社
304	梅棹忠夫[ほか]著	2001	逐走日本の原点	新潮社
369	全国動物看護事業協会編	2004	動物看護実務相談Q&A	中央法規出版
451	嶋村克，山内豊太郎著	2002	天気の不思議がわかる本	廣済堂出版
499	宮川博仁[著];ヘルス・システム研究所編	2004	薬と病気	ヘルス・システム研究所
547	水澤純一著	2005	情報通信ネットワーク入門	培風館
537	御川幹夫著	2002	トヨタ成長のカギ 創業期の人間関係	近代文芸社
610	神器俊子，松村和則編	2002	食・農・からだの社会学	新曜社
673	大久保一彦著	2002	誰も言わなかった！飲食店成功の秘密	フォレスト出版
720	石本正著	2001	絵をかくよろこび	新潮社
783	鈴木春洋著	2003	甲子園に賭ける	新潟日報事業社
816	清水義範著	2004	大人のための文章教室	講談社
817	遠藤織枝[他]著	2004	戦時中の話しことば ラジオドラマ台本から	ひつじ書房
913	司馬遼太郎著	2004	坂の上の雲 6	文藝春秋
933	ダン・ブラウン[著];越前敏弥訳	2005	ダ・ヴィンチ・コード	角川書店

NDC の層別にサンプリングすることにより，生産実態サブコーパスの「書籍」部分に集積される書き言葉の多様性は，これまでにない程度で確保されていると言ってよい。従来，小説のテキストデータなどがコーパスとして用いられてきたが，書き言葉の実態を反映するために NDC を基準として設計されたコーパスは存在しなかったからである。

しかし，NDC による分類は，主題や内容によってトップダウン的に一意に分類されているため，書き言葉の多様性を多角的な視点からとらえるには限界がある。実際に集積しつつある書籍のサンプルには，確かに NDC だけではとらえきれない多様性が存在している。書き言葉の多様性をとらえるための観点としては，例えば，次のようなものが考えられる。

(1) 内容・主題:(たとえば，NDC 分類の 1～3 次区分)
(2) 種類 [5]:小説(物語)，手紙，日記，論説文，紀行文，ルポルタージュ，韻文，翻訳，戯曲(シナリオ)，マニュアル，ガイドブック，辞書，事典
(3) 形式:座談，対談，インタビュー，パネル討論，講演集，会話形式，往復書簡形式，リレー執筆形式，Q&A 形式，投稿形式，　辞書・事典形式
(4) 場面設定:時代(現代，江戸時代，平安時代，未来)，場所(日本国内，国外，仮想世界)
(5) 著者の属性 [6]:年代，性別，出身地
(6) 対象読者の属性:年代，性別，好み
(7) 視点:人称，人間以外
(8) 硬軟:難解，堅い，平易，くだけている
(9) 論理構成・紙面構成:章節，キャプション，注記，コラム，引用，ブロック割り構成，図説，カタログ的構成
(10) 文体:口語文，文語文，候文，和漢混淆文，条文
(11) 文末・調子:デスマス調，デアル調，ゴザイマス調，体言止め，語りかけ口調，演説調
(12) 文長:長短
(13) 修辞・比喩:種類，使い方
(14) オノマトペ:種類，使い方
(15) 語彙:語彙の選択，特に位相の異なる語彙の選択(古語，俗語，幼児語，方言など)，語種の選択
(16) 表記:文字種の選択(漢字，カタカナ，ひらがな)，表外漢字の使用，仮名遣い(現代仮名遣い，歴史的仮名遣い)，ローマ字や外国語の使用
(17) 記号類:句読点，記号類の使い方
(18) ルビ・注記:使用量(多少)，使用目的(読み，原語，別の言い換え語，注釈，参考文献)

これらの観点は，サンプリングの過程で経験的に得られたものであり，用語の吟味，体系的な整理が必要であることは言うまでもない。ただし，多様性をとらえる観点・指標を総合的に体系化する作業は，コーパスを有効に利用するために必要不可欠なものであり，今後の大きな課題の一つであることは間違いない。

3.2 多様なサンプル例

以下，書籍サンプルの具体例を示し，多様性のあり方について論じる。( )内の 3 桁の数字は NDC の 3 次区分である。3 次区分の分類名を下線を引いて併記する。

まず，形式に特徴のある 3 例を示す。例 1 は Q&A 形式

[5] 一部 NDC の分類名と重なるが，内容よりも形式的な面からの観点になるものと考え，ここでは「内容・主題」とは分け「種類」と仮に別分類にして挙げる。
[6] サンプルからは判断できないことが多く，簡単にはわからないものであるが，観点の一つに成り得るものであろう。(6)も同様。

である。このような Q&A 形式は，NDC 全般にわたりよく見られる。例 2 は，会話形式の例である。実際の対話とは異なり，発話ごとに著者が異なるわけではない。例 3 は，講義のあまった時間に学生に書かせたものを集めたものであるらしい。段落ごとに実際の著者が違う。しかし，編者がいて，著者明記もないため，「リレー執筆形式」とは言えない。このような形式のサンプルは現時点ほかにはなく，まだうまく類型化できていない。なお，文は，方言もまじり，口語的である。

例 1：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『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（336：経済の経営管理）

Q3-7 マネジメントのための環境会計

マネジメントのための環境会計にはどんなものがありますか？それぞれの特徴を教えて下さい。

A

■内部環境会計の意義

環境会計は，その目的により，外部報告目的の環境会計と内部管理目的の環境会計とに分類されています。わが国では環境省のガイドラインも推進力となって，多数の企業が環境会計を外部に公表するようになってきた一方，企業の意思決定に役立つ内部管理目的の環境会計の研究も進められています。

例 2：野々村花衣『感性ちゃんと頭脳君の対話』文芸社（304：社会科学の論文集.評論集.講演集）

感性 そういうことか。分かったわ。つまり，「肌の表面に何を塗っても，その物質がバリアゾーンを通過して有棘細胞層や基底細胞層にまで到達するわけがない」ってことなのね？

頭脳 そうだよ。そんなことは不可能なんだよ。もしもそれが可能だとしたら，肌の防衛網が機能していないことになるから，おそらくそういう人は生きていけないだろうね。

感性 物質のサイズを小さく，細かくしてもダメなの？

頭脳 ダメだよ。無理だね。バリアゾーンが健全な場合には，水の分子一個ですら通さないんだ。

例 3：吉村英夫，シネカブ撰『ほろっと本音キラッと青春 紙上チャット こんな大学生しています』アールズ出版（377：教育の大学・高等・専門教育.学術行政）

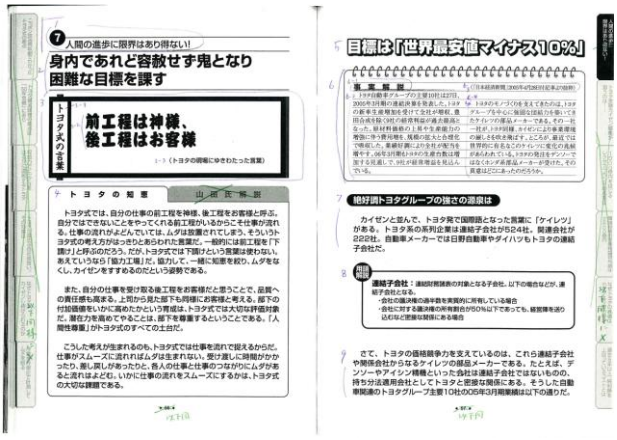
一八歳ってこんなものかなあ。ちょっと予定とはちがう。

なんだか毎日平凡。だけど，毎日平凡に過ごせていることを幸せだと思う。何も特別じゃなくていいと思いながら，毎日を平凡に頑張ってます。

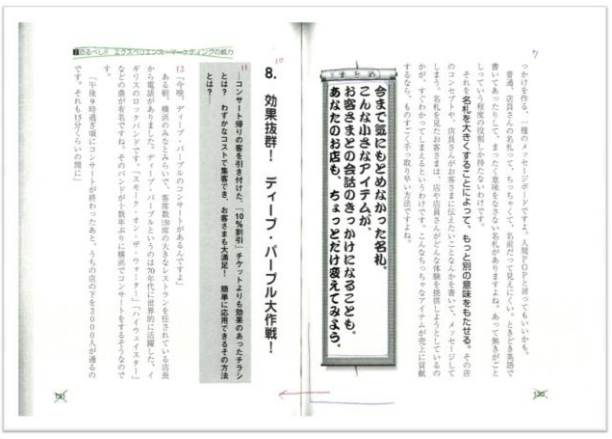
友よ！ おまえらみんなさめすぎや。もっと毎日，感動的に生きろよ。

次に，紙面構成に特徴のある 2 例を示す。例 4，例 5 は，以下に示した紙面のとおり，視覚的な工夫がされている。このようなものを仮に「ブロック割り構成」と呼ぶこととする。ガイドブックや PC 関連のマニュアル，図説・図解の必要なものなど，なかば必然的に視覚的工夫をとるものもあるが，ここではそれら以外にも工夫のあるものの例として，ビジネス書の例を示す。

例 4：山田真哉監修『トヨタだけがなぜ儲かるのか！？財務を「カイゼン」する 12 のルール』宝島社（null）



例 5：藤村正宏『「モノ」を売るな！「体験」を売れ！ 2 時間でわかる！』オーエス出版（673 商業の商業経営．商店）



続いて，例 6 と例 7 に，語彙や表記に特徴のある論説文 2 例を示す。例 6 は，漢字に関する論説文である。表外漢字が数多く出現し，非常に難解に見える。例 7 は，言葉遣いは平易であるが，歴史的仮名遣いであるため，この仮名遣いに馴染みがないとやはり難解に見える。

例 6：白川静『白川静著作集』平凡社（222：アジア史.東洋史の中国）

「柱也，从木盈聲」とあり，「春秋傳曰，丹桓宮楹」と春秋傳の文を引くが，文は莊廿三年の經文である。釋名釋宮室に，「楹，亭

也，亭亭然孤立，旁無所依也」とあり，段注にその文を引いて，「按禮言東楹西楹，非孤立也，自其一言之耳」というも，字義に關しない。盈は盈滿の義であるから，柱というもおそらく太みのある圓柱の意であろう。

例 7：柳田国男『柳田國男全集 第 30 巻』筑摩書房（380：風俗習慣，民俗学，民族学）

それから又同じ穀物でも，碎けやかけらや粃などのやうに，粒のまゝでは用ゐられぬものが多い。さういふのは粉に挽いて食べるの他は無い。蕎麦も小麦も粉にするのが元の食べ方であつて，或はそれを飯や粥の上にふりかけて食ふこともあるが，多くはその粉だけを食べる算段をして居た。

最後に，例 8 に戯曲の例，例 9 にルポルタージュの例を示す。この「戯曲」や「ルポルタージュ」は NDC9 番台の 3 次区分にある分類名であるが，以下の本は「全集」「論文集」であることが優先されて 0 番台に分類されている。例 8 は，時代設定が江戸時代であり，語彙や語法はその時代を表すように表現されている。例 9 は，一人称の「僕」が語りかけるような口調が特徴的であり，くだけた印象を与えている。

例 8 ：梅原猛『梅原猛著作集 19』小学館（081：叢書，全集，選集の日本語）

（照手姫，男衆や召使いを呼んで指図する。しばらくして，国府の役人登場。）
国府の役人　頼もう，頼もう。国守様が介殿とともにおいでじゃ。
小糸　よくおいで下さいました。
（小糸，長殿と女将に知らせに行く。）

例 9：桝田武宗『「社会の窓」から何が見えるのか』桜桃書房（049：一般論文集．一般講演集の雑著）

確かに，〈カフェ・パタゴニア〉なんて章題をつけて気を惹こうとした僕は，「ミミッコイ」って言われても仕方ないと思います。「セコイ」でも仕方ありません。しかし，多少弁解をさせて頂きますと，これから書こうとしている〈カフェ・パタゴニア〉ての，実言うと並のサ店だなんて思ったら大間違いなんですよ。なんちゃってへへへ。とか言いつつなんとか気を惹こうなんて，姑息だね。アー。ヤダヤダァダナ姿の洗髪って，よく分かんないでしょうけど，深く考えないで下さい。別に意味はないんです。

4. 多様性の分析と文体論研究との接点

本稿の最後に，コーパスに格納された文章（サンプル）の多様性を分析するとはどういう作業であるか，という点について考えてみたい。コーパスが言語研究に用いられる限り，そこに記録された言語表現は言語学的な視点から分析されることになるが，そこに存在する多様性をどのようにとらえるかは，立場によって異なるであろう。たとえば，従来の文章・談話・テクスト研究の中では「文体」や「スタイル」，あるいは「位相」としてとらえられてきた。

このうちの「文体」について，林(1991)は，以下のように述べている(p.32)。

> 文体論：文体に関する理論，理論的追及。文章は，その表出の目的(内容)，時代的制約による記載様式・語彙・語法などの違い，および書き手の個性，個性に基づく言語表現に関する美的理想の違い等から，多様な形態を示し，読み手に違った印象を与える。この違いを類型的に，あるいは個別的にとらえたものを文体とし，これについて論ずる分野を文体論という。

現実に存在している書き言葉の実態を反映するようにサンプルを抽出し，その言語的な特徴や類型を多角的な視点から明らかにしていく作業は，上記の文体論の方法論に近い。NDC を基準として設計された現代日本語書き言葉のコーパスが，どのような多様性を有するかを検討することは，そのまま，現代日本語の書き言葉全体を対象とした文体論的な分析となり得ると考えられる。

5. おわりに

5 年間に出版された書籍より，NDC の層別にランダムに取得した書籍約 6,600 サンプルを概観し，そこから多様性をとらえる観点を抽出した。そして，現在構築中のコーパスがどのような多様性を有するかを分析した結果を報告した。文章に関する研究が言語学的に，あるいは心理学的に進められている一方，図書館や書店においては，図書分類というものが，NDC 以外にも様々に工夫，検討されている。それら従来の議論と，今まさに実現しつつある大規模コーパス分析とをあわせ，文章の多様性について，さらなる分析，議論を進めていきたい。

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

［**参考文献**］

林巨樹(1991)「文体論の領域」『文体論の世界』三省堂.
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).
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.