コーパスを用いたテキスト分類指標の検討
—BCCWJの文書構造情報分析を中心に—

間淵洋子†　柏野和佳子　山口昌也　高田智和（国語研）
†mabuchi@ninjal.ac.jp

1.はじめに

国立国語研究所では，2006年～2010年を開発期間として，『現代日本語書き言葉均衡コーパス("Balanced Corpus of Contemporary Written Japanese" 以下，“BCCWJ”と略す)』の構築を進めている。開発最終年度となる2010年度を目前に，データ作成は収束に向かっており，現在は，データ利用に際して有益な研究用付加情報や，既存情報の活用方法について検討を始めている。

本稿では，BCCWJに格納予定の雑誌データを資料として，文書構造情報を数量的に調査し，それらが文書のどのような特徴を示すかを分析する。その上で，文書構造情報が，ジャンル・読者層・文章スタイル等を弁別する際の分類指標として活用できる可能性を示す。

2.研究の背景

BCCWJには，書籍，新聞，雑誌，政府刊行白書，教科書，Webデータなど，多種多様なメディア・ジャンルのテキストが格納されている。これらの多様なテキストを含むコーパスを利用する上で必要となるコンテクスト情報については，NDC等の主題による分野分類を補完する分類の枠組や手法を検討している[1]。

BCCWJのテキストを格納する形式（電子化フォーマット）の設計においても，文・文章を対象とした言語学的分析，辞書記述等を目的とした用例取得，あるいは，教材等の選択を目的とした文書検索などの用途に資するデータの作成を目指し，個々の語・文等の文書要素が文書内でどのような役割として機能しているかを表わす，コンテクスト情報の付与を試みた。見出し，キャプション，注記といった記事構造要素の情報（以下，「文書構造情報」）がそれである[2]。

従来の国語学的な文章・文体研究においては，文長，文字種，品詞，文末表現などの形態論情報を元にした指標による分析が広く行われてきたが[3]，近年は，Webからのデータマイニングを目的とする研究の発展により，レイアウト情報を用いて特定の文書を判別し抽出する手法などの提案もなされている[4]。

本研究では，雑誌にみられる多様な文書構造の型が，文書の機能や対象読者層を考慮した方略的なものであり，その構造類型により分類される文書には特徴的な文体があると仮定する。その特徴的な文体を分析することを最終的な目標として，まずは，構造類型に分類するための指標として，BCCWJの文書構造情報を用いることが可能かを検討する。

3.資料および調査概要

3.1.サンプル概要

今回資料として用いたデータの概要を示す。

BCCWJは，(1)生産実態サブコーパス，(2)流通実態サブコーパス，(3)非母集団サブコーパスの性質の異なる3つのサブコーパスからなる。また，分析目的の違いを考慮し，長さを異にする以下２種類のサンプルを用意している[5]。

固定長サンプル　サンプル長を1000字に固定して取得するサンプル

可変長サンプル　記事，章などの論理的な構造を持つ文章の一まとまりを取得するサンプル

今回分析対象資料としたのは，生産実態サブコーパスに格納予定の雑誌可変長サンプルデータの一部である。2001年～2005年に発行された雑誌よりサンプリングを実施し，文書構造情報付与が終了したものから，可変長サンプルデータとして雑誌の1記事全体を取得し得た1027文書を抽出した[1]。

雑誌を調査対象としたのは，以下の理由による。

1) 他のメディアのデータに比して，多様な文書スタイルを有するため，文書構造の特徴と言語表現の関連性を観察するのに適している。
2) 利用目的や対象読者が想定しやすいため，文書構造の特徴の背景分析に適している。

3.2.付加情報概要

分析用に取得した1027文書には，以下の研究用付加情報が付与されている。

文書構造情報　見出し，著者情報，記事概要，注記，キャプション，引用・発話など

書誌情報　雑誌名，巻号，ジャンル（サンプリング用の層別情報）など

[1] 可変長サンプルは上限文字数を設け，１万字を超える記事の場合，記事内の章・節等一部の構造要素を記事の代用として取得する。記事全体を格納しない文書は，構造の全体像が把握できず，文書構造の特徴分析がし難いため，今回の分析には用いない。

文書構造情報については，サンプル文書にXMLタグとして格納されている情報を，以下の６要素にまとめて分析に用いた。分類要素名と説明の後に，統合されるXML要素名を括弧内に示す。

見出し　記事見出しや下位構造要素（章節）の見出し，それに付随する要素など（title，titleBlock，orphanedTitle）

記事情報　記事についての情報に相当する要素。著者情報，目次，記事概要，注記など（authorsData，contents，abstract，profile，noteBody）

図表関連　本文補足・参照の役割を担う図表，それらに付随する図表タイトルやキャプションなど（figure，caption，rejectedBlock）

発話　「地の文」に相当しない発話表現（speech）

引用　「地の文」に相当しない引用表現（citation）

主本文　上記を除く文字列要素。「地の文」(上記以外のsentence要素)

なお，BCCWJのサンプリング層別情報として付与されるジャンル分類は，メディア・リサーチ・センター発行『雑誌新聞総かたろぐ』2004年版における区分である，総合，教育・学芸，政治・経済・商業，産業，工業，厚生・医療の6分類であるが，これはNDCの主題分類に近い。本稿での分析においては，利用目的との関連により文書を分類したいため，利用目的を反映した分類区分として社団法人日本雑誌協会発行の『マガジンデータ2009（2008年版）』を用いた。第1区分の，総合，ライフデザイン，ライフカルチャー，ビジネス，情報，趣味専門，子供誌の7区分を元に，一部の特徴的な文書に対応するため，部分的に第2区分を用いて表1に示す12区分とした[2]。

表1：雑誌分類カテゴリ一覧

分析カテゴリ(括弧内は略記号)	文書数
1. 総合 (G)	215
2. オピニオン (O)	60
3.ライフカルチャー (C)	53
4 . ライフスタイル (L)	112
5. ファッション (F)	88
6. ビジネス (B)	65
7. 情報 (I)	54
8. 趣味 (H)	209
9. 専門 (T)	44
10. スポーツ (S)	80
11. 文芸 (N)	33
12. 子供誌 (c)	14
総計	1027

分類の結果，子供誌に所属する文書数は14と極めて少ないため，今回の分析対象サンプルから外し，参照値として示す。

3.3.調査概要

2節に述べた通り，本研究では方略による文書構造類型があると仮定する。今回分析対象とする雑誌の構造類型としては，以下の5種を想定する。これらの類型は，版面（レイアウト）の異なりとして認識されるだけでなく，文章全体のスタイル（文体），文章の機能，対象とする読者層などに大きく影響するものであり，3.2節で示した利用目的に着目したカテゴリとの相関が期待できる。

a)文章型　主な構成要素は主本文要素であり，文の連続・積み重ねによって説明・解説・論説するテキスト。階層性を持たないフラットなものと，論文のように章節構造を有するものがある。

b)図版型　主な構造要素は写真や図表であり，それに対する補助解説文章を主とする。図版による視覚的効果をねらったテキスト。

c)文章・図版型　文章を主体にしつつ，写真や図表を多用することにより，読み手の理解を促し，興味を惹き付けるねらいを持つテキスト。

d)発話型　発話の書き起こし文を主体とする。対談・座談，インタビューなどに基づき，発話内容を伝達することを目的として書かれたテキスト。

これらの雑誌類型のうち，発話型記事に関しては，後述する構造要素比率のうち，発話率が極めて強い相関を持つことを確認したため[6]，本稿では分析を割愛する。調査項目には，これらの雑誌類型に見られるレイアウト特徴を表わしうる指標として，以下4項目を選択した。計測方法と共に示す。

(1)構造要素比率　3.2節に示した6構造要素の文書内での比率。記事全体の構造や主要構成要素を把握するための指標。記事内の文字列を各構造要素に分類し，全体の文字数に対する比率として示す。

(2)図表数　一定のテキストに含まれる図表要素の数。図表要素の多さ（＝図表要素への依存度）を把握するための指標。記事内の図表関連要素数を計測し，4000字（今回調査した雑誌の１可変長サンプルあたりの平均文字数）あたりの図表数として示す。

(3)階層深度　小見出しごとに認定される章節構造（cluster要素）の深さ。階層のないフラットなテキストか，階層性を有するテキストかを把握するための指標。各階層に含まれる文字数を計測し，文字数×階層深度の和を記事全体の文字数で割った，平均の階層数として示す。

[2] 『マガジンデータ2009（2008年版）』に掲載のない雑誌については，適宜類似の雑誌と同じカテゴリを付与した。

(4)章節数　一定のテキストに含まれる 章・節の数。並立して言及される事柄の多さ・細かさ（トピックの多さ）を把握するための指標。上記階層深度における並立項（兄弟関係にあるcluster要素）の数を計測し，4000字あたりの章節数として示す。

また，これら文書構造情報による指標に加えて，従来から文体との相関が認められている以下２項についても調査し，カテゴリの文体的特徴を確認した。

(5)文長　１文あたりの文字数平均。

(6)文字種比率　漢字・平仮名・片仮名・アルファベット・算用数字・その他(記号等)の文書内での比率。

4. 調査結果と分析

4.1. カテゴリの文体的特徴

まず，文長と文字種比率について，各カテゴリにおける統計量を示し，文体的特徴を把握しておく。

文長の計測結果を図1に，文字種比率の平均値を図2に示す。図1では箱の上端が第1四分位数，下端が第3四分位数，黒い水平線が中央値を表す。

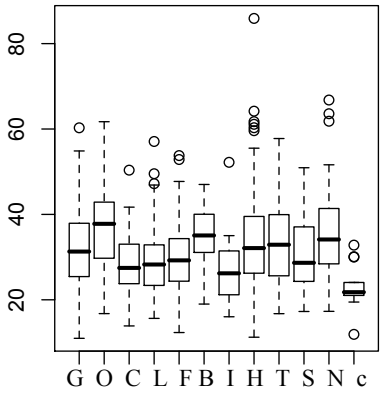

図1：カテゴリ別の文長[3]

図1，2より，オピニオン，ビジネス，専門，文芸で文長が長く，かつ漢字率が高い，情報，ライフカルチャー，ライフスタイル，ファッション，スポーツで文長が短く，かつ片仮名・アルファベット・数字・記号の比率が高いという特徴を見いだすことができる。

これらの特徴は，学術・専門的文章と一般・日常的文章に見られる文体的特徴として，多くの先行研究で言及されてきた結果と一致するものであり，雑誌という同一メディア内においても，このような文体差が認められることを示唆している。

4.2 カテゴリごとの文書構造特徴

次に，文書構造を示す指標として挙げた4項目についての計測結果を，図3〜図6に示す。

4.1節において類似の傾向を見せたビジネス，専門，文芸で主本文が記事の大半を占め(図3)，図表数が極めて少なくサンプル間でのばらつきも少ない(図4)という特徴が見て取れる。総合，オピニオンで発話比率が高い(図3)のは，これらの雑誌にインタビュー・座談書き起こし形式の記事が多く含まれることに起因する。これらの特殊な形式を除くと，概ね，オピニオン，ビジネス，専門，文芸はa)文章型の構造を持つ文書であると判断できる。ただし，階層数において文芸とそれ以外とで差異が見られ(図5)，章節数においてビジネス，オピニオンと専門に差異が見られる(図6)。ビジネス，オピニオンはトピックについてじっくり文を重ねて述べる形式を取り，専門では事柄を多様に細かく述べる形式を取るという機能に基づく方略の差異と解釈することができる。

一方，ファッション，ライフスタイル，ライフカルチャーは，いずれも図表キャプションの占める割合が高く(図3)，図表数が多い(図4)。したがって，b)図版型のテキストであると判断できる。これらの雑誌は主に，文章を読ませることよりも，図版によって視覚的に訴えることを方略として作られた雑誌であることが，計測結果にはっきりと表れている。

なお，4.1節の文長，文字比率調査においてファッション等と同様の傾向を見せたスポーツは，主本文の比率が高く，図表の比率や図表数がb)図版型のテキストほど多くない(図3,図4)というように，異なる文章類型であることが分かる。文書構造要素の計測結果においてスポーツと類似する傾向が見られる情報，趣味も合わせて，c)文章・図版型のテキストと考えられる。文章による解説的な要素と図表を用いて明快に説明する要素を複合することによって，分かりやすさや情報量の豊富さを求めるタイプの雑誌と解釈できる。

総合については，機能と対応付けられる他のカテゴリと異なり，様々な性質の記事が含まれることが特徴であるため，構造特徴を抽出することは難しい。個々の記事を対象に，他ジャンルとの類似度を計る等の分析が必要となる。

5.おわりに

本稿では，現在構築中のBCCWJ雑誌データを用いて，文書構造情報の計量調査を行った。文書構造類型や読者層と大きく関連する，雑誌ジャンル情報を元に分析を行った結果，構造要素比率や図表数，階層数などの情報が，一部の雑誌ジャンルにおいて構造の特徴を示しうることを指摘した。

今回の報告では，分析指標として利用可能な情報を

[3] 図1のx軸は左から，1.総合，2.オピニオン，3.ライフカルチャー，4.ライフスタイル，5.ファッション，6.ビジネス，7.情報，8.趣味，9.専門，10,スポーツ，11.文芸（，12.子供誌；参考値）の分布。以降の図4〜6も同様。

探索する足掛かりとして，雑誌ジャンルごとの計測結果の概観にとどまったが，調査結果を元に，今後，各種パタン認識手法等を試行するなど，分類手法の検討を進める予定である。

資料：分析カテゴリに含まれる雑誌タイトル（一部抜粋）

総合(AERA,Yomiuri Weekly, 週刊新潮, 女性セブン)/オピニオン(現代, Voice, 新潮45, 論座)/Lカルチャー(おしゃれ工房, オレンジページ,すくすく子育て,住まい100選)/Lスタイル(an・an, BRUTUS, Tarzan, クロワッサン, サライ, 家庭画報)/ファッション(Domani, JJ, POPEYE)/ビジネス(エコノミスト，経済界)/情報(DIME, Hanako,TVガイド,Weeklyぴあ)/趣味(YOMIURI PC, アサヒカメラ, きものサロン, 愛犬の友, 囲碁, 趣味の園芸,旅の手帖)/専門(農耕と園藝,メディカル朝日,小一教育技術)/スポーツ(大相撲, 月刊SKI JOURNAL)/文芸(オール読物，文芸ポスト)

付記：本研究は，文部科学省科学研究費補助金特定領域研究「日本語コーパス」による補助を得た。

参考文献

[1]柏野和佳子他(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)
[2]山口昌也他 (2008)『『現代日本語書き言葉均衡コーパス』における電子化フォーマット ver.2.0』(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)
[3]樺島忠夫，寿岳章子(1965)『文体の科学』綜芸舎
[4]田島幸恵，奥村学(2005)「Web上の料理レシピの抽出とその利用」『言語処理学会第11回年次大会発表論文集』
[5]丸山岳彦他(2005)「代表性を有する書き言葉コーパスのサンプリング手法について」『言語処理学会第12回年次大会発表論文集』
[6]山口昌也他(2009) 「『現代日本語書き言葉均衡コーパス』における電子化フォーマットとその応用」『特定領域研究「日本語コーパス」平成20年度公開ワークショップ予稿集』

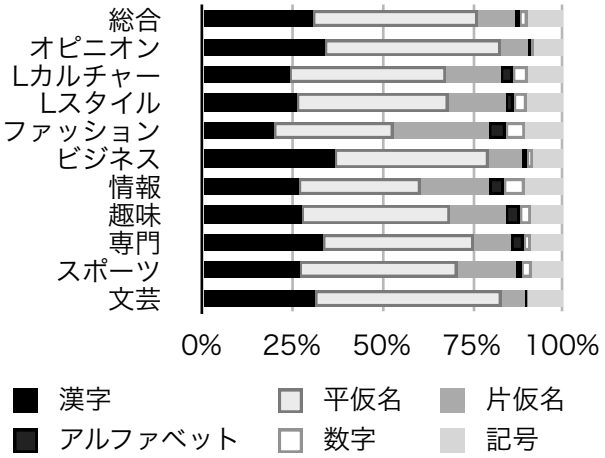

図2：文字種比率

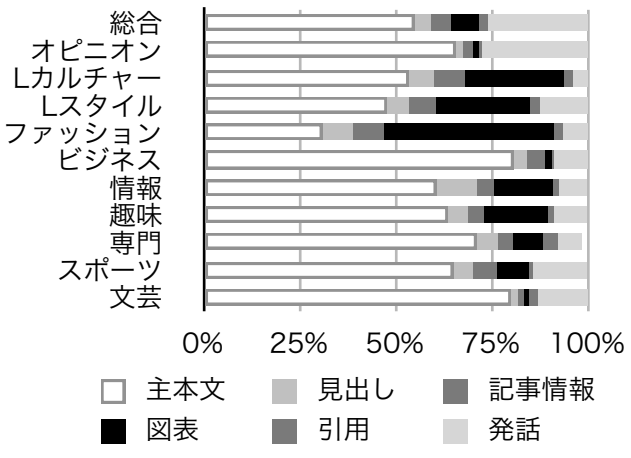

図3：構造要素比率

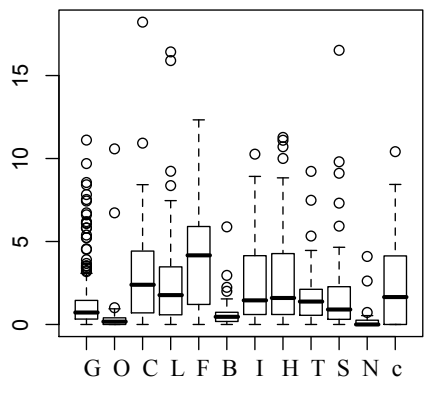

図4：図表数

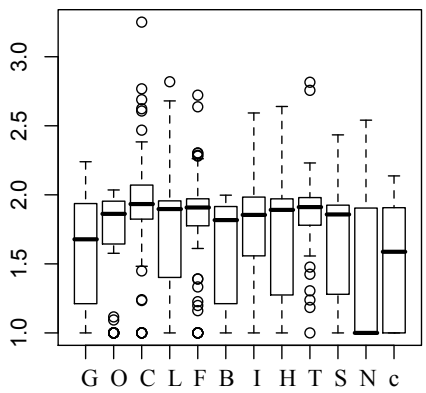

図5：階層深度

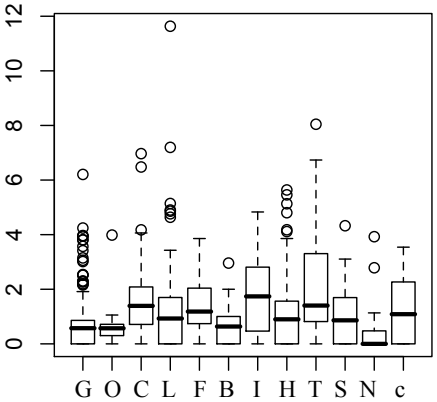

図6：章節数