書籍テキストへの文体情報付与の試み
—『現代日本語書き言葉均衡コーパス』の収録書籍を対象に—

柏野 和佳子*（国立国語研究所 言語資源研究系）
立花 幸子（国立国語研究所 コーパス開発センター）
保田 祥（国立国語研究所 コーパス開発センター）
飯田 龍（東京工業大学 大学院情報理工学研究科）
丸山 岳彦（国立国語研究所 言語資源研究系）
奥村 学（東京工業大学 精密工学研究所）
佐藤 理史（名古屋大学 大学院工学研究科）
徳永 健伸（東京工業大学 大学院情報理工学研究科）
大塚 裕子（はこだて未来大学 メタ学習センター）
佐渡島 紗織（早稲田大学 留学センター）
椿本 弥生（はこだて未来大学 メタ学習センター）
沼田 寛（はこだて未来大学 メタ学習センター）

Annotation of Writing Styles of the Book Samples in the Balanced Corpus of Contemporary Written Japanese

Wakako Kashino (Dept. Corpus Studies, NINJAL)
Sachiko Tachibana (Center for Corpus Development, NINJAL)
Sachi Yasuda (Center for Corpus Development, NINJAL)
Ryu Iida (Dept. Computer Science Graduate School of Information Science and Engineering, Tokyo Instiute of Technology)
Takehiko Maruyama (Dept. Corpus Studies, NINJAL)
Manabu Okumura (Precision and Intelligence Laboratory, Tokyo Institute of Technology)
Satoshi Sato (Graduate School of Engineering, Nagoya University)
Takenobu Tokunaga (Dept. Computer Science Graduate School of Information Science and Engineering, Tokyo Instiute of Technology)
Hiroko Otsuka (Center for Meta-Learning, Future University Hakodate)
Saori Sadoshima (Center for International Education, Waseda University)
Mio Tsubakimoto (Center for Meta-Learning, Future University Hakodate)
Hiroshi Numata (Center for Meta-Learning, Future University Hakodate)

１．はじめに

『現代日本語書き言葉均衡コーパス』(BCCWJ)1の図書館サブコーパスには，書籍テキストが 10,551 サンプル収録されている。大規模な書籍コーパスをより有効に活用し，テキスト研究を進めるためには，種々の書籍テキストをさまざまな観点から分類できることが望ましい(EAGLES1996)。また，BCCWJ に収録されるテキストの文体を計量的に考察する試みはすでにいくつか行われている（小磯ほか 2008，間淵ほか 2010，小磯ほか 2011）。国立国語研究所の共同研究プロジェクト「テキストの多様性を捉える分類指標の策定」において，書籍テキストを所望の目的で分類するために，書籍テキストの多種多様な形式，内容，表現に関わる特徴を捉えるための分類指標の設計と検証とを行っている（柏野・奥村 2012，保田ほか 2012）。

本稿では，はじめにアノテーション作業の概要を述べる。そして，現時点でのアノテー

* waka@ninjal.ac.jp

1 詳細は http://www.tokuteicorpus.jp/。

ションの途中経過報告として，すでに分類指標を付与した 3,494 テキストの分類結果の内訳と，そこから得られた典型例及び，文体の特徴を支える言語的特徴について述べる。そして，本作業を通し，図書館サブコーパスにどのような文体をもつテキストがどのように分布して収録されているのかを把握することができることを示す。

２．アノテーション作業

２.１　分類指標の設計

BCCWJ に収録されている書籍サンプルには, NDC（日本十進分類法）によるジャンルや, C コード（日本図書コード）による販売対象，発売形態，また，著者情報，形態論情報などが付与されており，それらを利用して，半自動的に種々の観点から分類することは可能である。しかしながら，EAGLES(1996)がコーパスへ付与することが望ましいと挙げる，(A)対象読者に想定される読解レベル（難易度），(B)テキストの作成意図，(C)さまざまな文体情報の 3 種に関する情報は C コード以外には与えられておらず，それらの観点によるテキストの分類や抽出は困難である。そこで，(A)を補う「専門度」，(B)を補う「客観度」，(C)を補う「硬度」「くだけ度」「語りかけ性度」という，あわせて 5 つの分類指標を新たに設計した。EAGLES(1996)でコーパスに備えることが望ましいと議論されている「文体情報」とは，形式性，親疎性，口語性に関わる文体情報だと言える。よって，その形式性，親疎性を問うものとして「硬度」と「くだけ度」の指標を，口語性を問うものとして「語りかけ性度」という指標を設けた。

(a) 専門度

テキストの専門度を想定読者のスケールで測ることとし, 次の 5 段階の選択肢を設けた。

1 専門家向き
　読む前提に高度な専門知識が必要なもの
　それを仕事にしているような人向きのもの
2 やや専門的な一般向き
　読む前提に多少の専門知識が必要なもの
3 一般向き
　特に専門的な内容ではないもの
専門的な内容であっても，読む前提に専門的知識を特に必要とせず，一般向きに書かれているもの
4 中高生向き
　中高生向きに書かれているもの
　専門性の有無にかかわらず，中高生でも読めそうなもの
5 小学生・幼児向き
　明らかに小学生や幼児向きとして書かれているもの

このときの「専門性」はテキストを理解する上での「高度な知識の必要性」の有無と考える。たとえば，パソコン関係では，技術者や研究者向きであれば「専門家向き」であり，やや高度な知識を必要とするパソコンを趣味にしている人向きであれば「やや専門的な一般向き」であり，家庭向きであれば「一般向き」と考える。

(b) 客観度

テキストの書き手の意図を捉えるための指標を検討した。「論説，随筆，報告文，紀行文，手順書・・・」といったような体系的な分類案を作成し，それに基づいた指標の付与が理想であるが，指標の設計やその判断にかかる負荷が大きいことが予測されるため，今回の試行作業ではそのようは指標は設けなかった。代わりに，書き手の態度には「客観的」か「主観的」かの区別があると考え，その判断付与を行うこととした。次の 4 段階の選択肢を設けた。

1 とても客観的

2 どちらかといえば客観的

3 どちらかといえば主観的

4 とても主観的

ここで「客観的」とは，主に，事実，観察，論証などが述べてあるもの。誰が読んでも納得できる妥当性が高いもの。「主観的」とは，主に，経験や感想などが述べてあるもの。妥当性は筆者の自由と定義する。なお，これはノンフィクションと判断をしたテキストについてのみ付与する。

(c) 硬度

テキストの文体の形式性，親疎性を捉えるために「硬いか軟らかいか」を判断することとした。「硬い」とは，かしこまっている感じ，堅苦しい感じであり，「軟らかい」とは，かしこまっていない感じ，親しみやすい感じである。次の 4 段階の選択肢を設けた。

1 とても硬い

2 どちらかといえば硬い

3 どちらかといえば軟らかい

4 とても軟らかい

(d) くだけ度

テキストの文体の形式性，親疎性を捉えるためのもう一つの指標として，さらに「くだけているか」を問うこととした。「くだけているか」の逆には「改まっているか」を想定するが，「改まっているか」の度合いは問いにくいと考えた。よって，ここでは「くだけている」度合いを問う，次の 3 段階の選択肢を設けた。

1 とてもくだけている

2 どちらかといえばくだけている

3 くだけていない（＝改まっている）

(e) 語りかけ性度

テキストの文体の口語性を問うものとして「語りかけ性度」という指標を設けた。口語性の高いテキストを「語りかけ性がある」ものと捉えることとした。たとえば，「あなた」や「みなさん」などの呼びかけ表現や，「でしょう」「ではないでしょうか」といった問いかけや相づちを求めるような文末表現など，読み手に直接的に語りかけているような表現があるものを，「語りかけ性がある」ものと考える。それら「語りかけ性」の度合を問う選択肢として次の 3 段階のものを設けた。

1 とても語りかけ性がある

2 どちらかといえば語りかけ性がある

3 特に語りかけ性はない

２.２　アノテーション作業の概要

作業対象と内容は次のとおりである。また，アノテーション作業は，次のとおり二段階で進めている。

[作業対象と内容]

- 対象テキスト：BCCWJに収録されている図書館サブコーパス（10,551サンプル）の書籍テキスト。
- 1テキストの範囲と長さ：コーパス収録テキストの分類指標とするため，その一部を字数を揃えて抽出することはせず，1サンプル全体を範囲とする。1テキストの平均はおよそ3,000語。
- 作業ファイル：サンプルを取得した書籍の紙面コピーの電子化ファイルを参照する。

●作業量：１セット約400～500の書籍テキストに対する指標付与を延べ約10日で行う。
●内容：
①形式による判定を行う。構造的に単純なテキストタイプ（例：章節構造）であれば細分類の対象とする[2]。
②細分類をする。「専門度，客観度，硬度，くだけ度，語りかけ性度」の分類指標を付与する。

[作業の一段階目]
●目的:人手付与の作業上の問題点の検討,典型例の抽出,分類指標の検証及び基準の検討。
●態勢：判断のゆれを検証するために同一サンプルを作業者3人で判定。
●判断：付与すべき指標の種類についてごく簡単な説明があるのみ。
●付与済テキスト数：3,324テキストを判定し，細分類を付与したのは2,672テキスト。

[作業の二段階目（途中）]
●目的：全10,551サンプルへの付与。
●態勢：1サンプル1人以上の判定と機械判定の相互参照。
●判断：判断事例付きマニュアルを参照。
●付与済テキスト数：本稿作成時，細分類の付与済みは，現時点3,494テキスト。

３．アノテーション作業結果

３.１　分類指標の付与結果

細分類付与済みの 3,494 テキストについて，その付与結果の内訳を表 1 に示す。なお，客観度はノンフィクションのみ対象としているため，現時点の付与済みテキスト数は 2,314 である。

表 1 分類指標の付与結果 (3,494 テキスト)

専門度	テキスト数・(%)	客観度	テキスト数・(%)	硬度	テキスト数・(%)	くだけ度	テキスト数・(%)	語りかけ性度	テキスト数・(%)
専門家向き	91(3%)	とても客観的	427(18%)	とても硬い	207(6%)	とてもくだけている	186(5%)	とても語りかけ性がある	326(9%)
やや専門的な一般向き	345(10%)	どちらかといえば客観的	904(39%)	どちらかといえば硬い	1135(32%)	どちらかといえばくだけている	1020(29%)	どちらかといえば語りかけ性がある	572(16%)
一般向き	2685(77%)	どちらかといえば主観的	623(27%)	どちらかといえば軟らかい	1834(52%)	くだけていない	2288(65%)	特に語りかけ性はない	2596(74%)
中高生向き	195(6%)	とても主観的	360(16%)	とても軟らかい	318(9%)				
小学生・幼児向き	178(5%)								

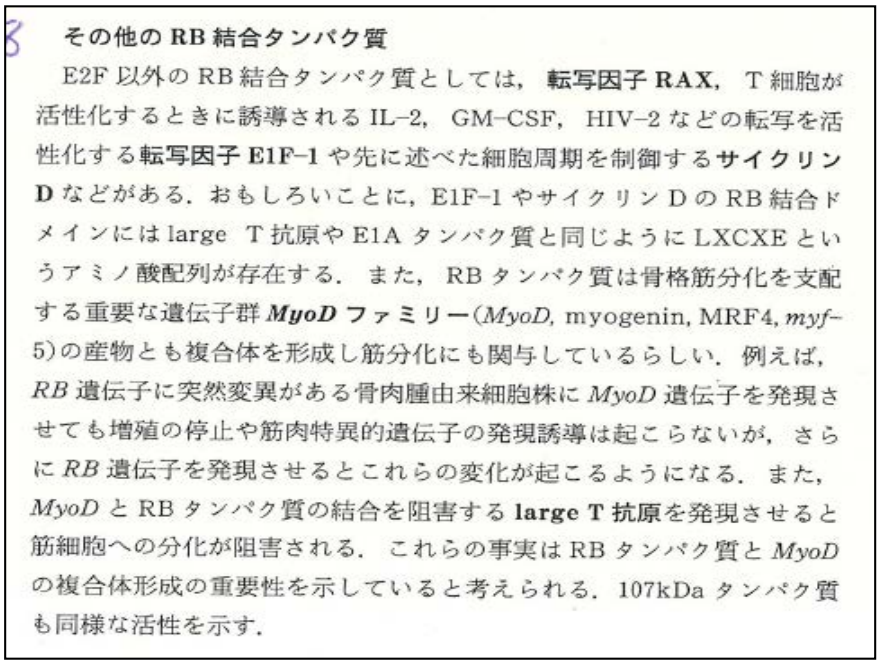
専門度は「一般向き」が多い。客観度も真ん中あたりが多いが，そのうち，「どちらかといえば客観的」の方が多い。硬度はさらに真ん中あたりが多いが，そのうち，「どちらかといえば軟らかい」の方が多い。くだけ度は約 3 分の 1 近く「どちらかといえばくだけている」である。語りかけ性度は「とても」と「どちらかといえば」をあわせて約 4 分の 1 である。

[2] 対象外とした形式が特徴的なテキスト（例：対談，Q&A 形式，図解，用語解説）については，一定量が分類されてから細分類を検討する予定でいる。

３.２　分類の典型例

ここでは, [作業の二段階目]で使用したマニュアルに掲載した典型例（サンプルの出典は, BCCWJのサンプルIDと書名とで記す）を示す。

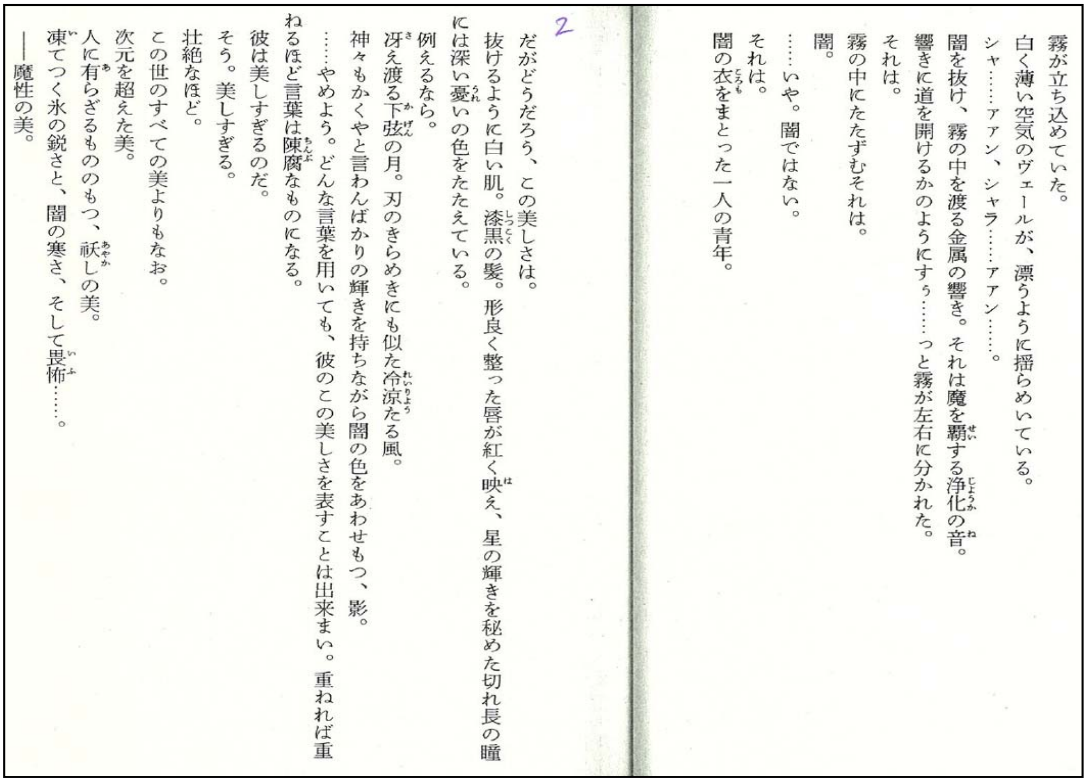
(1)専門度：1 専門家向き　　（LBi4_00021『がんと遺伝子』）



その他の RB 結合タンパク質

E2F 以外の RB 結合タンパク質としては，**転写因子 RAX**，T 細胞が活性化するときに誘導される IL-2，GM-CSF，HIV-2 などの転写を活性化する**転写因子 E1F-1** や先に述べた細胞周期を制御する**サイクリン D** などがある．おもしろいことに，E1F-1 やサイクリン D の RB 結合ドメインには large T 抗原や E1A タンパク質と同じように LXCXE というアミノ酸配列が存在する．また，RB タンパク質は骨格筋分化を支配する重要な遺伝子群 ***MyoD* ファミリー**(*MyoD*, myogenin, MRF4, *myf*-5)の産物とも複合体を形成し筋分化にも関与しているらしい．例えば，*RB* 遺伝子に突然変異がある骨肉腫由来細胞株に *MyoD* 遺伝子を発現させても増殖の停止や筋肉特異的遺伝子の発現誘導は起こらないが，さらに *RB* 遺伝子を発現させるとこれらの変化が起こるようになる．また，*MyoD* と RB タンパク質の結合を阻害する **large T 抗原**を発現させると筋細胞への分化が阻害される．これらの事実は RB タンパク質と *MyoD* の複合体形成の重要性を示していると考えられる．107kDa タンパク質も同様な活性を示す．

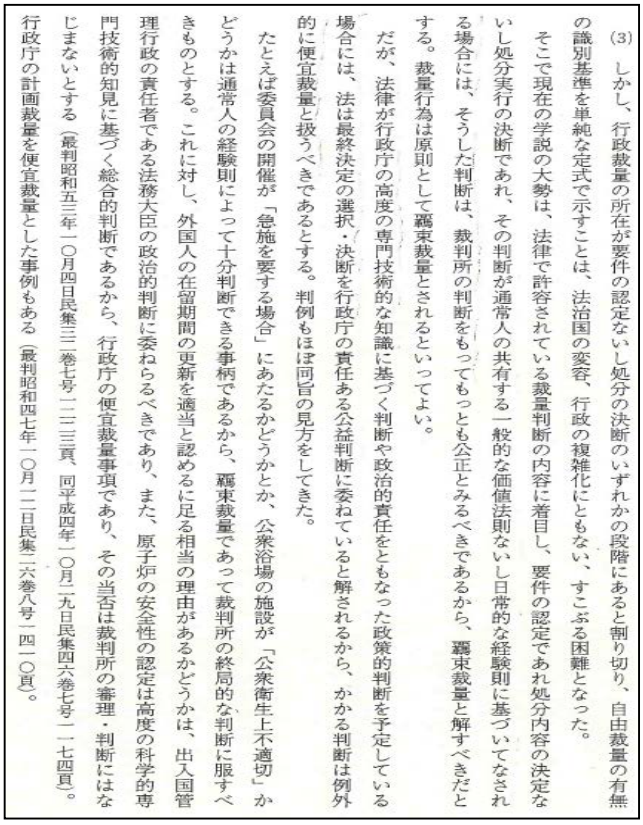
(2)専門度：4 中高生向き　　（LBf9_00090『超魔炎獄変』）



霧が立ち込めていた。
白く薄い空気のヴェールが、漂うように揺らめいている。
シャ……ァァン、シャラ……ァァン……。
闇を抜け、霧の中を渡る金属の響き。それは魔を覇する浄化の音。
響きに道を開けるかのようにすぅ……っと霧が左右に分かれた。
それは。
霧の中にたたずむそれは。
闇。
……いや。闇ではない。
それは。
闇の衣をまとった一人の青年。

だがどうだろう、この美しさは。
抜けるように白い肌。漆黒の髪。形良く整った唇が紅く映え、星の輝きを秘めた切れ長の瞳には深い憂いの色をたたえている。
例えるなら。
冴え渡る下弦の月。刃のきらめきにも似た冷涼たる風。
神々もかくやと言わんばかりの輝きを持ちながら闇の色をあわせもつ、影。
……やめよう。どんな言葉を用いても、彼のこの美しさを表すことは出来まい。重ねれば重ねるほど言葉は陳腐なものになる。
彼は美しすぎるのだ。
そう。美しすぎる。
壮絶なほど。
この世のすべての美よりもなお。
次元を超えた美。
人に有らざるもののもつ、祇しの美。
凍てつく氷の鋭さと、闇の寒さ、そして畏怖……。
――魔性の美。

(3)客観度：1 とても客観的
（LBo3_00158『行政法要論』）



(3) しかし、行政裁量の所在が要件の認定ないし処分の決断のいずれかの段階にあると割り切り、自由裁量の有無の識別基準を単純な定式で示すことは、法治国の変容、行政の複雑化にともない、すこぶる困難となった。

そこで現在の学説の大勢は、法律で許容されている裁量判断の内容に着目し、要件の認定であれ処分内容の決定ないし処分実行の決断であれ、その判断が通常人の共有する一般的な価値法則ないし日常的な経験則に基づいてなされる場合には、そうした判断は、裁判所の判断をもってもっとも公正とみるべきであるから、羈束裁量と解すべきだとする。裁量行為は原則として羈束裁量とされるといってよい。

だが、法律が行政庁の高度の専門技術的な知識に基づく判断や政治的責任をともなった政策的判断を予定している場合には、法は最終決定の選択・決断を行政庁の責任ある公益判断に委ねていると解されるから、かかる判断は例外的に便宜裁量と扱うべきであるとする。判例もほぼ同旨の見方をしてきた。

たとえば委員会の開催が「急施を要する場合」にあたるかどうかか、公衆浴場の施設が「公衆衛生上不適切」かどうかは通常人の経験則によって十分判断できる事柄であるから、羈束裁量であって裁判所の終局的な判断に服すべきものとする。これに対し、外国人の在留期間の更新を適当と認めるに足る相当の理由があるかどうかは、出入国管理行政の責任者である法務大臣の政治的判断に委ねらるべきであり、また、原子炉の安全性の認定は高度の科学的専門技術的知見に基づく総合的判断であるから、行政庁の便宜裁量事項であり、その当否は裁判所の審理・判断にはなじまないとする（最判昭和五三年一〇月四日民集三二巻七号一三三頁、同平成四年一〇月二九日民集四六巻七号一一七四頁）。行政庁の計画裁量を便宜裁量とした事例もある（最判昭和四七年一〇月二日民集二六巻八号一四一〇頁）。

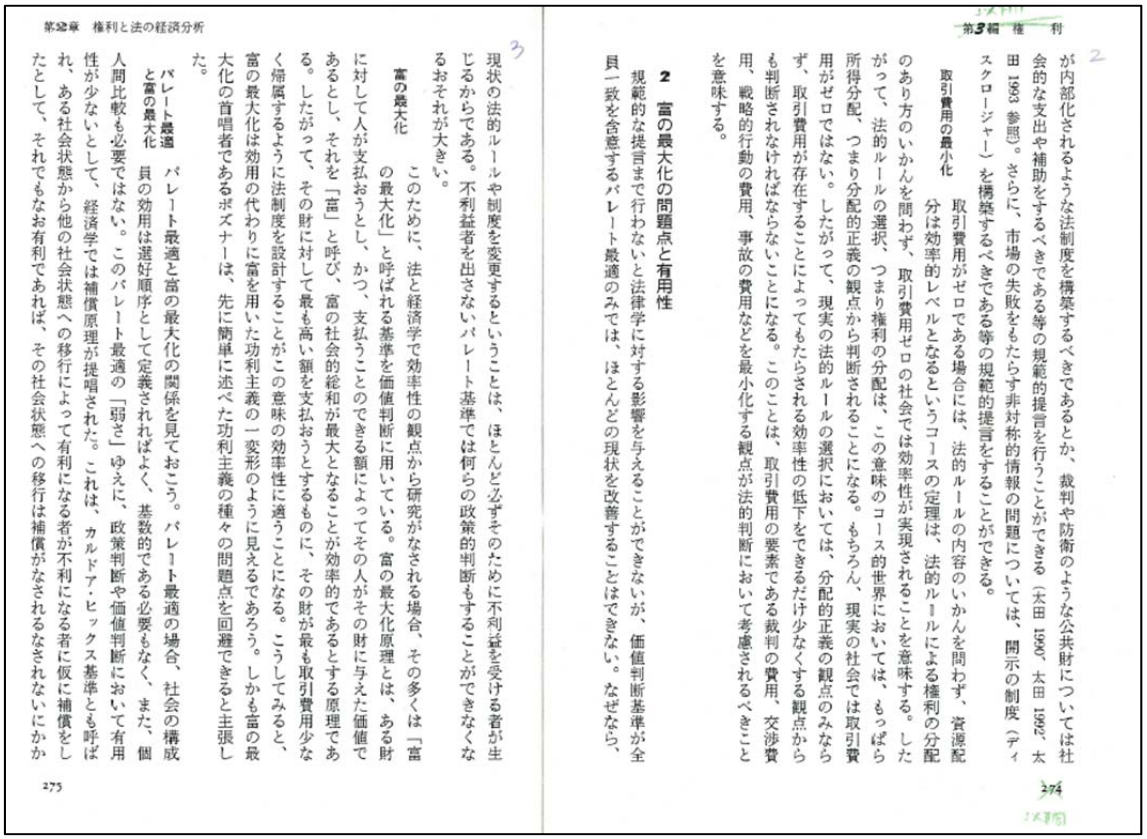
(4)客観度：4 とても主観的
（LBo3_00132『教師をめざす若者たち』）

人形は、私と木下さんを大胆にさせてくれました。人形を持った私たちは子供たちのなかに自然に入っていけました。教壇から降り、人形を片手にして、子供たちと触れ合ったのです。子供たちは人形に触れ、私も人形を通して子供たちの顔や体に触れていきました。私と子供たちの間に、大きな橋が架かったのです。言葉が互いに通じ合わないからこそ、見えてくるものがたくさんありました。言葉は、嘘をつくこともできるし、自分を隠すこともできます。しかし、心は嘘をつけないことを実感したのです。

どんなに上手な言葉を使っても、思っていないことを発すれば、子供に伝わらない。どんなに下手な言葉でも、心から伝えたいという愛情があれば、伝わるものであるということを信じることができました。この実感は日本でも通じる「教育の原則」であると思いました。

二日目、子供たちと綿花摘みを一緒にしました。敦煌の子供たちの手は「仕事をしている手」でした。その表情を持った手が「一緒に綿花を摘もうね」と私の手を引いてくれたとき、何かが私に伝わってきました。小さなその手から、人間としての強さが伝わってきました。それは、この子供たちが、それまでに培ってきた力だと感じました。

(5)硬度：1 とても硬い（LBi3_00033『現代法社会学入門』）



第3章 権利

が内部化されるような法制度を構築するべきであるとか、裁判や防衛のような公共財については社会的な支出や補助をするべきである等の規範的提言を行うことができる（太田 1993、太田 1995、太田 1983 参照）。さらに、市場の失敗をもたらす非対称的情報の問題については、開示の制度（ディスクロージャー）を構築するべきである等の規範的提言をすることができる。

取引費用の最小化　取引費用がゼロである場合には、法的ルールの内容のいかんを問わず、資源配分は効率的レベルとなるというコースの定理は、法的ルールによる権利の分配のあり方のいかんを問わず、取引費用ゼロの社会では効率性が実現されることを意味する。したがって、法的ルールの選択、つまり権利の分配は、この意味のコース的世界においては、もっぱら所得分配、つまり分配的正義の観点から判断されることになる。もちろん、現実の社会では取引費用がゼロではない。したがって、現実の法的ルールの選択においては、分配的正義の観点のみならず、取引費用が存在することによってもたらされる効率性の低下をできるだけ少なくする観点からも判断されなければならないことになる。このことは、取引費用の要素である裁判の費用、交渉費用、戦略的行動の費用、事故の費用などを最小化する観点が法的判断において考慮されるべきことを意味する。

2 富の最大化の問題点と有用性

規範的な提言まで行わないと法律学に対する影響を与えることができないが、価値判断基準が全員一致を含意するパレート最適のみでは、ほとんどの現状を改善することはできない。なぜなら、

274

第2章 権利と法の経済分析

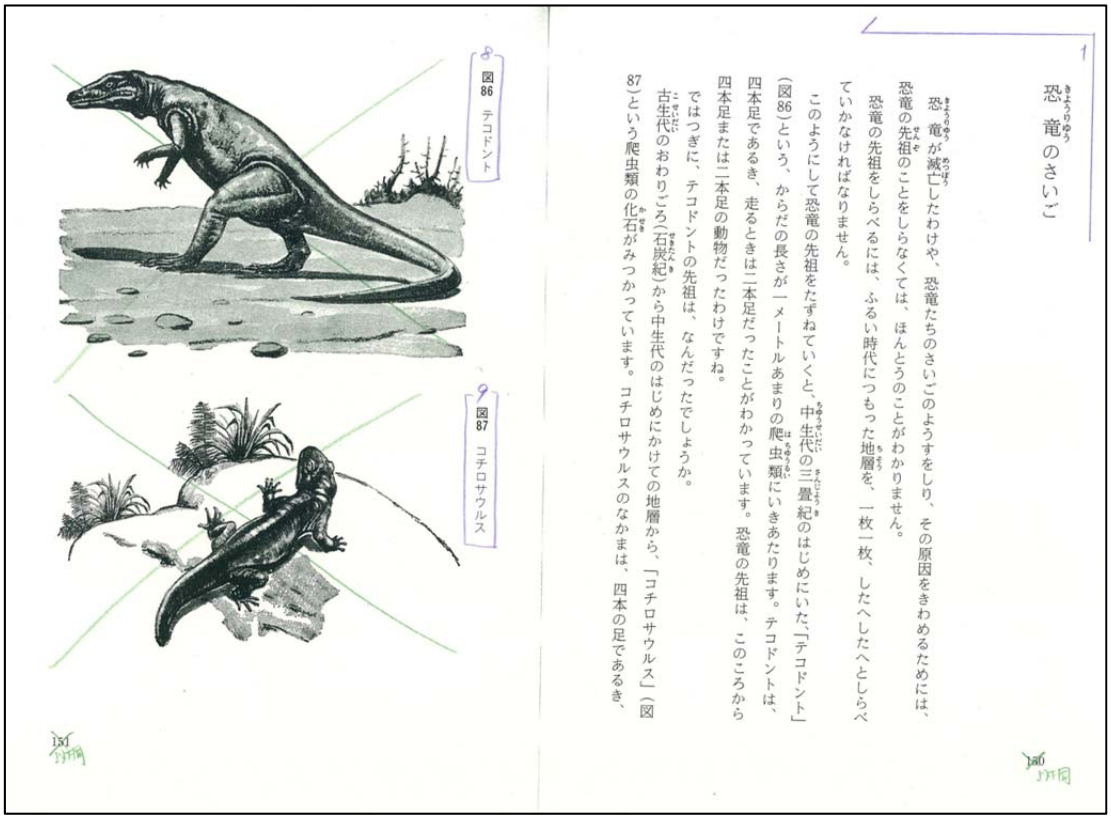
現状の法的ルールや制度を変更するということは、ほとんど必ずそのために不利益を受ける者が生じるからである。不利益者を出さないパレート基準では何らの政策的判断もすることができなくなるおそれが大きい。

富の最大化　このために、法と経済学で効率性の観点から研究がなされる場合、その多くは「富の最大化」と呼ばれる基準を価値判断に用いている。富の最大化原理とは、ある財に対して人が支払おうとし、かつ、支払うことのできる額によってその人がその財に与えた価値であるとし、それを「富」と呼び、富の社会的総和が最大となることが効率的であるとする原理である。したがって、その財に対して最も高い額を支払おうとするものに、その財が最も取引費用少なく帰属するように法制度を設計することがこの意味の効率性に適うことになる。こうしてみると、富の最大化は効用の代わりに富を用いた功利主義の一変形のように見えるであろう。しかも富の最大化の首唱者であるポズナーは、先に簡単に述べた功利主義の種々の問題点を回避できると主張した。

パレート最適と富の最大化　パレート最適と富の最大化の関係を見ておこう。パレート最適の場合、社会の構成員の効用は選好順序として定義されればよく、基数的である必要もなく、また、個人間比較も必要ではない。このパレート最適の「弱さ」ゆえに、政策判断や価値判断において有用性が少ないとして、経済学では補償原理が提唱された。これは、カルドア・ヒックス基準とも呼ばれ、ある社会状態から他の社会状態への移行によって有利になる者が不利になる者に仮に補償をしたとして、それでもなお有利であれば、その社会状態への移行は補償がなされるなされないにかか

273

(6)硬度：4 とても軟らかい（LBa4_00010『恐竜の世界をたずねて』）



恐竜のさいご

恐竜が滅亡したわけや、恐竜たちのさいごのようすをしり、その原因をきわめるためには、恐竜の先祖のことをしらなくては、ほんとうのことがわかりません。

恐竜の先祖をしらべるには、ふるい時代につもった地層を、一枚一枚、したへしたへとしらべていかなければなりません。

このようにして恐竜の先祖をたずねていくと、中生代の三畳紀のはじめにいた、「テコドント」（図86）という、からだの長さが一メートルあまりの爬虫類にいきあたります。テコドントは、四本足であるき、走るときは二本足だったことがわかっています。恐竜の先祖は、このころから四本足または二本足の動物だったわけですね。

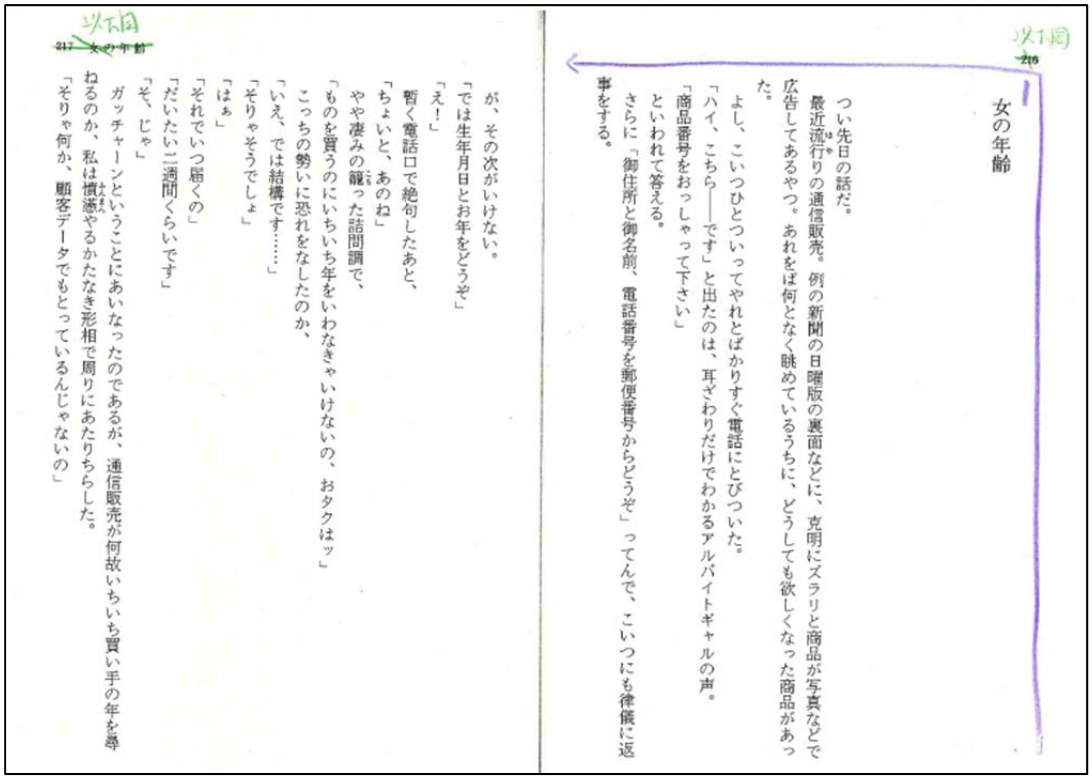
ではつぎに、テコドントの先祖は、なんだったでしょうか。

古生代のおわりごろ（石炭紀）から中生代のはじめにかけての地層から、「コチロサウルス」（図87）という爬虫類の化石がみつかっています。コチロサウルスのなかまは、四本の足であるき、

図86 テコドント

図87 コチロサウルス

(7) 1 とてもくだけている（LBf9_00067『男はオイ！女はハイ…』）



女の年齢

つい先日の話だ。

最近流行りの通信販売。例の新聞の日曜版の裏面などに、克明にズラリと商品が写真などで広告してあるやつ。あれをば何となく眺めているうちに、どうしても欲しくなった商品があった。

よし、こいつひとついってやれとばかりすぐ電話にとびついた。

「ハイ、こちら――です」と出たのは、耳ざわりだけでわかるアルバイトギャルの声。

「商品番号をおっしゃって下さい」

といわれて答える。

さらに「御住所と御名前、電話番号を郵便番号からどうぞ」ってんで、こいつにも律儀に返事をする。

が、その次がいけない。

「では生年月日とお年をどうぞ」

「え！」

暫く電話口で絶句したあと、

「ちょいと、あのね」

やや凄みの籠った詰問調で、

「ものを買うのにいちいち年をいわなきゃいけないの、おタクはッ」

こっちの勢いに恐れをなしたのか、

「いえ、では結構です……」

「そりゃそうでしょ」

「はぁ」

「それでいつ届くの」

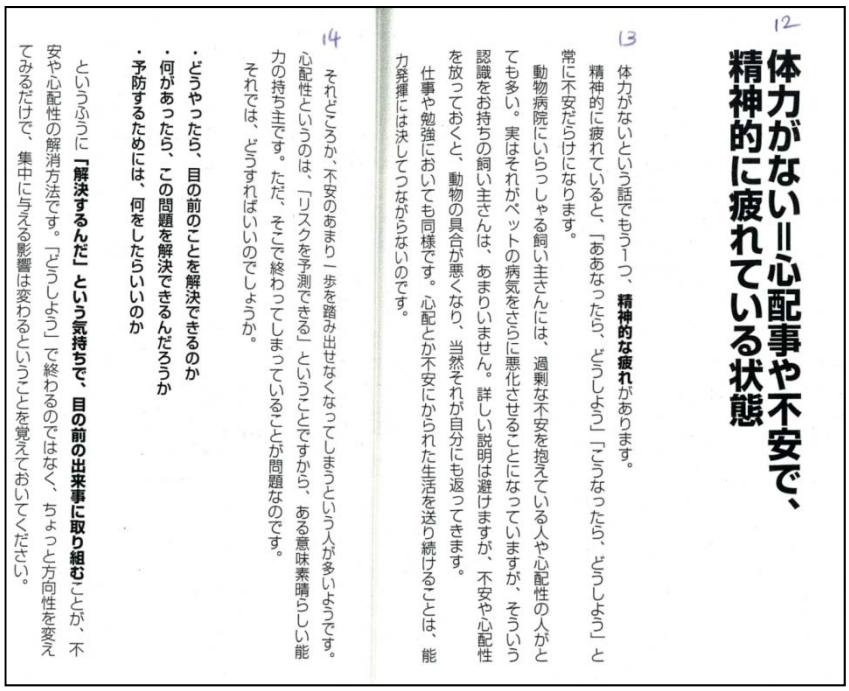
「だいたい二週間くらいです」

「そ、じゃ」

ガッチャーンということにあいなったのであるが、通信販売が何故いちいち買い手の年を尋ねるのか、私は憤懣やるかたなき形相で周りにあたりちらした。

「そりゃ何か、顧客データでもとっているんじゃないの」

(8) 1 とても語りかけ性がある（LBt1_00013『5 分間集中力トレーニング』）



12
体力がない＝心配事や不安で、精神的に疲れている状態

13
体力がないという話でもう一つ、**精神的な疲れ**があります。
精神的に疲れていると、「ああなったら、どうしよう」「こうなったら、どうしよう」と常に不安だらけになります。
動物病院にいらっしゃる飼い主さんには、過剰な不安を抱えている人や心配性の人がとても多い。実はそれがペットの病気をさらに悪化させることになっていますが、そういう認識をお持ちの飼い主さんは、あまりいません。詳しい説明は避けますが、不安や心配性を放っておくと、動物の具合が悪くなり、当然それが自分にも返ってきます。
仕事や勉強においても同様です。心配とか不安にかられた生活を送り続けることは、能力発揮には決してつながらないのです。

14
それどころか、不安のあまり一歩を踏み出せなくなってしまうという人が多いようです。
心配性というのは、「リスクを予測できる」ということですから、ある意味素晴らしい能力の持ち主です。ただ、そこで終わってしまっていることが問題なのです。
それでは、どうすればいいのでしょうか。

・どうやったら、目の前のことを解決できるのか
・何があったら、この問題を解決できるんだろうか
・予防するためには、何をしたらいいのか

というふうに**「解決するんだ」という気持ちで、目の前の出来事に取り組む**ことが、不安や心配性の解消方法です。「どうしよう」で終わるのではなく、ちょっと方向性を変えてみるだけで、集中に与える影響は変わるということを覚えておいてください。

3.3 文体の特徴を支える言語的特徴

5 つの分類指標のうち，硬度，くだけ度，語りかけ度を判定する際に手掛かりとなりそうな言語的特徴を，これまでの作業で得られている典型例の範囲で分析した。判定基準とするにはさらに分析が必要であるが，現段階に捉えられた特徴を以下に述べる。

硬度判定の参考情報を次の表 2 に示す。

表 2 硬度判定の参考情報

	硬い	軟らかい
和語：漢語（平均は 6：2）	5:3	7:1
語彙	難解	平易
	新密度の低い語（学術・専門用語）	接頭辞「御（お・ご）」
	感動詞ほとんどなし	副助詞（～か，たり，や，まで等）
	数詞が多い	
	文語助動詞（べし）	
副詞	いかに，より等のかしこまった語	バリエーションが豊富
	出現頻度は平均より低い	
接続詞	ないし，いかに等のかしこまった語	
	使用率は平均より少し高め	
文末	終助詞はほとんどなし	です・ます
		「である」はほとんどなし
主語・述語	抽象物の主語＋受動態の述語	
その他	疑問・回答が対応	記号が多い
	断定・定義を表す文	

くだけた印象を与えるサンプルに特徴的な点を以下に列挙する。

- 述語省略など，文法的に破格の文がある
- 一人称が主語の文が多い
- 平易な語に加え，俗語がある
- 音変化（拗音化，撥音化など）の語がある
- オノマトペが多い
- 感覚や感情表現が多い
- 回答のない，いいっぱなしの疑問文がある
- 終助詞（ね，よ等）がある
- 「～だ。」で終わる文が平均より多め
- 外来語が多い
- 副詞はバリエーションが豊富で，出現頻度が平均より高い

最後に，語りかけ性度判定の参考情報を表 3 に示す。

表 3 語りかけ性度判定の参考情報

	語りかけ性がある	語りかけ性はない
語種		固有名詞（人名・地名など）が多い
語	「あなた，みなさん」などの呼びかけ	
	「自分」が多い	
文末	「ます」／終助詞「ね」が多い	「た」が多い
	意志推量「だろう，でしょう」などが多い	
	体言化（「～のです，～ということだ」など）が多い	

４．おわりに

BCCWJ に収録する書籍コーパスの有効活用を可能とするための分類指標の人手付与作業の概要と，作業の途中経過を報告した。

人手判定の一方で，表層的な情報を利用した機械判定を試みている。Support Vector Regression によるランキングを行ったところ，機械判定と人手判定の相関は見られた。中でも，専門度，硬度，くだけ度の相関は高かった。現在，機械判定と人手判定のずれを対照させ，人手判定の見直しを行っている。さらに人手判定の基準を明確にすることで，機械判定の精度向上も目指したい。

今年度がプロジェクトの最終年度である。最終成果として，分類指標の明確な基準を示すとともに, BCCWJ の図書館サブコーパスに収録される 10,551 サンプルの全てに分類指標を付与し，コーパスの研究や教育の利用価値を高めることを目指す。

さらに文体的な特徴を支える言語表現の分析を進め，辞書記述への応用を考えている。

謝　辞

本研究は，国立国語研究所の共同研究プロジェクト「テキストの多様性を捉える分類指標の策定」に基づくものです。また，BCCWJ の構築は，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）による補助を得たものです。

文　献

EAGLES. 1996. EAGLES Preliminary recommendation on Text Typology, *EAGLES Document EAG–TCWG–TTYP/P*, Version of Jun 1996. (http://www.ilc.cnr.it/EAGLES/texttyp/texttyp.html)

柏野和佳子，奥村学(2012)「書籍テキストへの分類指標人手付与の試み—『現代日本語書き言葉均衡コーパス』の収録書籍を対象に—」『言語処理学会第 18 回年次大会予稿集』B5-6.

小磯花絵，小木曽智信，小椋秀樹，冨士池優美，宮内佐夜香(2008)「『現代日本語書き言葉均衡コーパス』にもとづくジャンル間の文体差に関わる要因の分析」『社会言語科学会第 22 回研究大会発表論文集』pp.192-195.

小磯花絵，田中弥生，小木曽智信，近藤明日子(2011)「評定実験に基づくテキスト分類尺度の体系化の試み」『現代日本語書き言葉均衡コーパス』完成記念講演会予稿集，pp.47-52.

間淵洋子，柏野和佳子，山口昌也，高田智和(2010)「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—」『言語処理学会第 16 回年次大会予稿集』PA1-11.

保田祥，柏野和佳子，立花幸子，丸山岳彦(2012)「「語り性」を有する書きことばの典型例の分析」『第 1 回コーパス日本語学ワークショップ』予稿集， pp.139-146.