テキストの多様性をとらえる分類指標の体系化の試み（2）

小磯 花絵[†]　田中 弥生[‡]　小木曽 智信[†]　近藤 明日子[†]
[†]国立国語研究所　[‡]神奈川大学

koiso@ninjal.ac.jp　yait@jcom.home.ne.jp
togiso@ninjal.ac.jp　kondo@ninjal.ac.jp

1　はじめに

書き言葉の多様性は，新聞・書籍・インターネットといった媒体の違いや主題の違いでは捉えきれない広がりがあり，媒体・主題に加えた新たな指標の体系化が求められている。この種の類型化・体系化の試みは，文体研究や理論研究の中では古くから行われているものの，提案された観点や指標によって，多種多様な書き言葉が具体的にどのように，またどの程度妥当に分類できるのかといったことを実証的に評価した取り組みは，日本語の研究を見る限りあまり行われていない。

そこで小磯ほか（2011）では，人がテキストを読んだ際に感じる印象を表わす表現を調査し，実際のテキストに対して評定実験を行った上で，分類指標を探索的に体系化することを試みた。具体的には,『現代日本語書き言葉均衡コーパス』（以下BCCWJ）に含まれる40のテキストを対象に，予備調査の結果得られた20の評定尺度に基づき，3名の被験者にテキストに対する印象について5段階で評定してもらった。因子分析の結果,「スタイル」「文構成の明晰性」「抑揚・リズム性」という三つの因子が抽出された。

そこで本研究では，サンプルを300に増やして印象評定を再度実施した上で，いかなる言語的特徴が因子「スタイル」「文構成の明晰性」「抑揚・リズム性」の印象形成に関わるかを明らかにすることを試みる。

2　評定実験

2.1　評定尺度

小磯ほか（2011）の結果を受け，因子「スタイル」に関わる評定尺度として「改まった－くだけた」と「硬い－軟らかい」を，因子「文構成の明晰性」に関わる評定尺度として「簡潔な－冗長な」と「整然とした－雑然とした」を，因子「抑揚・リズム性」に関わる評定尺度として「テンポのよい－テンポの悪い」「めりはりのある－単調な」をそれぞれ選択した。

2.2　評定サンプル

テキストサンプルとして，BCCWJのうち自動解析結果を人手修正した精度の高い「短単位」「長単位」情報（小椋ほか2011）が付されたコアと呼ばれるデータセットから300サンプルを選んだ。内訳は，新聞（一般記事・特集・コラムなど含む）124サンプル，書籍（小説以外）55サンプル，雑誌40サンプル，行政白書27サンプル，Yahoo!ブログ54サンプルである。小説には複数の人物の会話文が多く含まれている可能性が高く，テキストから受ける印象を一意に決めづらいことが多いため対象外とした。同様の理由で引用の多いサンプルも対象外とした。

各サンプルのサイズは約300文字（300文字を越えて最初に現れる文末まで）とした。文字数で区切っているため，必ずしも内容的にまとまった単位にはなっておらず，話題の途中で始まったり，あるいは途中で終わったりしているものもある。

2.3　手続き

実験には3名（男性1名，女性2名）の被験者が参加した。被験者には，今回分析対象とする上記評定尺度を含む計13の評定尺度に基づき5段階で評定してもらった。サンプルは被験者毎にランダムに配置し，評定尺度も適宜左右を反転させた。本番に先立ち練習問題として10サンプルを評定してもらった。被験者には，テキストの内容に対する印象ではなくテキストの表現や文体から受ける印象に従って評定してもらうよう，複数の具体例を挙げて指示した。

2.4　結果

小磯ほか（2011）で得られた因子が今回の評定実験の結果についても抽出できるかを確認するために，同様の方法で因子分析（最尤法，バリマックス回転）を行った。各尺度の因子負荷量を表1に示す。表から，因子1では「整然とした－雑然とした」「簡潔な－冗長な」が，因子2では「テンポのよい－テンポの悪い」「めりはりのある－単調な」が，因子3では「改まった－くだけた」「硬い－柔らかい」が，それぞれ高い因子負荷量を示しており，小磯ほか（2011）の結果と同じく,「文構成の明晰性」「抑揚・リズム性」「スタイル」の三つの因子が抽出されたことが確認できた。

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

表 1: 因子分析の結果－因子負荷量－

	因子 1	因子 2	因子 3
改まった－くだけた	0.287	-0.348	**0.831**
硬い－柔らかい	0.411	-0.288	**0.759**
簡潔な－冗長な	**0.942**	0.101	0.311
整然とした－雑然とした	**0.606**	0.423	0.296
テンポのよい－テンポの悪い		**0.760**	-0.286
めりはりのある－単調な	0.269	**0.794**	-0.235
寄与率	26.7%	26.7%	26.5%

3 各因子に関わる言語的特徴

本節では，前節で改めて確認された三つの因子「スタイル」「文構成の明晰性」「抑揚・リズム性」の印象形成に，それぞれいかなる言語的特徴が関わるかを，評定に用いた 300 の言語サンプルを用いて検討する。

3.1 言語特徴量

先行研究などに基づき，文章の特徴を捉える上で関連が深いと想定される次の言語特徴量を選び，テキストサンプルごとに各特徴量を算出した。

単語に関わる特徴量： （1）名詞（普通名詞・固有名詞・代名詞を含む全ての名詞）率,（2）機能語（助詞・助動詞）率,（3）相類（形容詞・副詞・形状詞・連体詞）率,（4）和語率,（5）外来語率

文に関わる特徴量： （6）文終了率,（7）節終了率,（8）体言終了率,（9）文タイプ異なり率,（10）敬体率,（11）終助詞率,（12）文長,（13）文長変動率

（1）から（5）は単語に関わる特徴である。単語としては短単位（小椋ほか 2011）を用いた。（1）から（3）の品詞に基づく特徴については，記号を除く語の総数を母数に比率を算出した。（4）と（5）の語種率については，体言（固有名を除く）に限定した上で比率を算出した。漢語率は和語率と高い相関を示したため除外した。

（6）から（13）は文に関わる特徴である。BCCWJ で一文と認定されたものの中には，記事や節の見出し，項目列挙なども含まれるが，それらは除外した。分析には，見出しや項目等を除いた上で 1 サンプルに占める文数が 5 以上の 297 サンプルを用いた。

（6）から（8）は，特に文末に着目した文の種類に関わる特徴である。文が文末表現を伴い終了するもの，節で終了するもの（例：「出ないよりはいいんだけど。」），句で終了するもの（例：「まずはフーセンから。」），体言で終了するもの（例：「右手の山門は月見寺で知られる本行寺。」），その他（感動詞など）に分類し，それぞれの比率を算出した。句で終了するもの及びその他については，数が少なかったため分析から除いた。

（9）の文タイプ異なり率は，1 サンプルに含まれる上記文の種類のタイプをトークンで除した値とした。その際，文末表現で終了する文については，時制（過去・非過去），アスペクト・疑問・命令・推量・伝聞・「のだ」の有無の組合せで下位分類した。これらは，文末表現のバリエーションが因子「抑揚・リズム性」と関わりうるという小磯ほか（2011）の観察をふまえて導入したものである。

敬体か常体か，終助詞（疑問の終助詞を除く）を伴うか否かは，スタイルに強く関わりうる要因と考えられるため，文末表現のバリエーションの一部として組み込むのではなく,（10）,（11）として独立させた。(10) については，文末表現を伴い終了する文に限定し，常体と敬体に分類した上で，敬体の比率を求めた。

（12）の文長は，文に含まれる文字数（記号を除く）として,（13）の文長変動率は，一つのサンプルに含まれる文長の分散として求めた。

3.2 分析

2.4 節で求めた三つの因子「スタイル」「文構成の明晰性」「抑揚・リズム性」に，それぞれいかなる言語特徴量が関わるかを明らかにするため，因子ごとに，個々のサンプルの因子得点を従属変数，3.1 節に記した 13 の言語特徴量を独立変数として線形重回帰分析を行った。その上で AIC に基づくモデル選択（変数増減法）で最適なモデルを求めた。なお独立変数（言語特徴量）については，全て標準化得点に変換してから分析に用いた。

3.3 結果

モデル選択の結果，三つの因子「スタイル」「文構成の明晰性」「抑揚・リズム性」ごとに，それぞれ表 2〜4 に示す言語特徴量が変数として選択された。

因子「スタイル」の結果から見て行く（表 2）。推定値が負の特徴量はスタイルの高さ（硬い・改まった）と，正の特徴量はスタイルの低さ（軟らかい・くだけた）と関連することを意味する。結果から，名詞率の高さと文の長さはスタイルの高さに，外来語率・和語率・敬体率・終助詞率・体言終了率の高さはスタイルの低さに，それぞれ影響することが分かる。

因子「文構成の明晰性」では（表 3），推定値が負の特徴量は明晰性の高さ（簡潔な・整然とした）と，正の特徴量は明晰性の低さ（冗長な・雑然とした）と関連することを意味する。結果から，体言終了率・文終了率の高さと文の長さは文構成の明晰性の高さに，相類率と文長変動率の高さは文構成の明晰性の低さに，それぞれ影響することが分かる。

因子「抑揚・リズム性」では（表 4），推定値が負の特徴量は抑揚・リズム性の高さ（テンポのよい・めりはりのある）と，正の特徴量は抑揚・リズム性の低さ（テンポの悪い・単調な）と関連することを意味する。結果から，体言終了率・文タイプ異なり率・外来

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

表 2: 因子「スタイル」の結果

パラメタ	推定値	標準誤差	t 値	p 値
（切片）	0.009346	0.031319	0.298	0.76
外来語率	0.297008	0.034431	8.626	$< .001$
和語率	0.244418	0.046330	5.276	$< .001$
敬体率	0.154759	0.035223	4.394	$< .001$
終助詞率	0.103362	0.036136	2.860	$< .01$
体言終了率	0.094046	0.036750	2.559	$< .05$
文長	−0.086522	0.037811	−2.288	$< .05$
名詞率	−0.183417	0.063711	−2.879	$< .01$

表 3: 因子「文構成の明晰性」の結果

パラメタ	推定値	標準誤差	t 値	p 値
（切片）	−0.01046	0.05028	−0.208	0.84
相類率	0.16279	0.05550	2.933	$< .01$
文長変動率	0.21701	0.07910	2.744	$< .01$
体言終了率	−0.41279	0.13682	−3.017	$< .01$
文長	−0.25995	0.08534	−3.046	$< .01$
文終了率	−0.46069	0.14266	−3.229	$< .01$

語率の高さは抑揚・リズム性の高さに，敬体率・文長変動率・名詞率・相類率・終助詞率の高さと文の長さは抑揚・リズム性の低さに，それぞれ影響することが分かる。言語特徴量と因子との関係をまとめて表 5 に示す。

3.4 考察

名詞率： 佐野・丸山（2008）は，文章の複雑さの指標として Halliday（1985）により提案された語彙密度を対象に，BCCWJ に含まれる白書と書籍（文学）の比較を行い，書籍よりも白書の方が語彙密度が高い傾向にあること，そのことが Halliday（1990）の「綿密に計画された，あるいはよりフォーマルな文章ほど語彙密度が高い」という主張からの予測と一致することを指摘している。佐野・丸山は Halliday の定義に従い，「述語をもつ節に含まれる内容語の割合」で語彙密度を定義しているが，仮に本研究における機能語率の逆数が「内容語の占める割合」として語彙密度に概略相当すると考えるならば，フォーマルな，あるいはより綿密に計画された文章の場合，機能語率は下がることになる。残念ながら機能語率自体はいずれの因子についても選択されなかったが，Halliday の語彙密度は，複雑な文章ほど動詞群の名詞化により機能語に対する内容語の比率が高くなることに由来する指標であり，品詞として見た場合に名詞率と強く関係する可能性がある。このことは，名詞率が高くなるほどスタイルの高い（改まった・硬い）印象を受けるという本研究の結果と整合的である。また名詞率が高くなる要因の一つは「原子力災害対策特別措置法」のような複合語が多く見られることである。この種の複合語は専門性の高い文章に多く見られることを考えると，スタイルの高さにつながることは十分に考えられる。

名詞率の高さが抑揚・リズム性の低下を招くという点も興味深い。以下は名詞率の高いサンプルである。

【例 1】独立行政法人森林総合研究所においては、中期計画に基づき、森林のもつ多面的機能に関する研究、地球温暖化対策に関する研究、木質資源の有効利用に関する研究等、森林・林業・木材産業に関する総合的な試験研究を実施した。（ID：OW6X_00007）

実際に音声を発話する場合，名詞の途中で「間（ま）」が置かれることはあまりなく，節や文節の境界に置かれる。この種の「間」を置くタイミングが抑揚・リズム構成の一翼を担うと考えるならば，複合語が多く文節が長くなると抑揚・リズムの阻害につながることも考えられる。実際には音を発しなくても何らかの影響を与える可能性は十分にある。また単純に名詞が多い（機能語が少ない）ことに伴う「見ための黒さ（濃淡のなさ）」が影響している可能性もある。

相類率： 相の類には主に修飾の働きをなす形容詞・副詞・形状詞・連体詞がまとめられており，この率が高いほど冗長性が増すという傾向と整合的である。形容詞・副詞は，ニュースよりも日常談話に，専門性の高い学会講演よりも個人の体験談などを話すスピーチにより多く見られる傾向にあることが指摘されており（国語研究所 1955; 小椋 2005），スタイルとの関係が見られることも予想されたが，今回の分析では両者の関係は見られなかった。

和語率・外来語率： 樺島（1963）は，名詞率と漢語率は正の相関にあることを指摘している。漢語率と和語

表 4: 因子「抑揚・リズム性」の結果

パラメタ	推定値	標準誤差	t 値	p 値
（切片）	−0.004882	0.039674	−0.123	0.90
文長	0.336928	0.073470	4.586	$< .001$
敬体率	0.193029	0.045888	4.207	$< .001$
文長変動率	0.215776	0.063089	3.420	$< .001$
名詞率	0.247408	0.072582	3.409	$< .001$
相類率	0.161162	0.061197	2.633	$< .01$
終助詞率	0.117244	0.053203	2.204	$< .05$
体言終了率	−0.253817	0.115243	−2.202	$< .05$
文タイプ異なり率	−0.136362	0.049907	−2.732	$< .01$
外来語率	−0.122188	0.043202	−2.828	$< .01$

表 5: 言語特徴量と因子との関係（言語特徴量の値が大きくなる場合の各因子の傾向の概要）

言語特徴量	スタイル	明晰性	抑揚リズム
名詞率	高スタイル	—	抑揚なし
機能語率	—	—	—
相類率	—	非明晰	抑揚なし
和語率	低スタイル	—	—
外来語率	低スタイル	—	抑揚あり
文終了率	—	明晰	—
節終了率	—	—	—
体言終了率	低スタイル	明晰	抑揚あり
文タイプ異なり率	—	—	抑揚あり
敬体率	低スタイル	—	抑揚なし
終助詞率	低スタイル	—	抑揚なし
文長	高スタイル	明晰	抑揚なし
文長変動率	—	非明晰	抑揚なし

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

率は負の相関があることから，名詞率と和語率は負の相関を示すことになる。実際，今回のデータでも −0.70 という高い負の相関を示している。このことは，和語率の高さは上述の名詞率とは逆の傾向，つまり「軟らかく，くだけた」印象を与えることにつながる。また野元（1959）は話し言葉では漢語よりも和語が用いられる傾向にあることを指摘しており，小磯ほか (2009) でも，名詞率の低さだけでは説明できない和語率の高さが話し言葉や書籍 (文学) などに見られることを指摘している。この点もスタイルとの関係に影響を与えると考えられる。外来語率の高さが和語と同様にスタイルの低さに影響するという点については，結果として硬い印象を与える漢語率の低下につながること，片仮名で書かれるため見た目として軟らかい印象を与える可能性があることなどが考えられる。

文・節・体言終了率： 文末表現をいわば省いた体言止めが相対的にスタイルの低下を招くことは十分予想できることだが，一方でそれが煩雑さではなく簡潔さを導く点は興味深い。また体言止めの使用は抑揚・リズム性にも関与する。次項の文タイプ異なり率とも関わることだが，体言で文を終えることそのものが直接的に抑揚・リズムを産み出すのではなく，文末の変化がその効果を産み出している可能性がある。

文タイプ異なり率： 次に挙げる例 2 と例 3 は小磯ほか (2011) で取り上げたサンプルである。いずれもスタイルと文構成の明瞭性が高いという意味では共通しているが，例 2 は抑揚・リズム性が低く例 3 は高いという点において異なる。例 2 は一貫して過去形が用いられているのに対し，例 3 は，時制の変化，アスペクトや伝聞形式の使用など，バリエーションに富んでいることが分かる。文タイプ異なり率はこの観察を裏付けるために導入した特徴量であるが，予想した通り抑揚・リズム性との関連が見られた。

【例 2】「森林・林業・木材産業分野の研究・技術開発戦略」及び「林木育種戦略」に基づき、＜省略＞効率的かつ効果的に推進 した。独立行政法人森林総合研究所及び独立行政法人林木育種センターにおいては、＜省略＞研究・技術開発等を 実施した。また、研究・技術開発等の実施に当たっては、＜省略＞評価と見直しを 行った。(1) 試験研究の効果的推進 森林・林業・木材産業分野の研究・技術開発戦略に基づき、試験研究の効果的・効率的推進を 図った。(ID:OW6X_00007)

【例 3】一方、公立高一年の少女（十六）＝殺人予備容疑で逮捕＝は、＜省略＞百円ショップで 購入していた。価格や切れ味などの点で二人の凶器に隔たりがあり、＜省略＞関連があるとみて 調べている。調べでは、＜省略＞購入していたという。(ID:PN3d_00013)

敬体率： 常体と比べて敬体の方が軟らかい印象を与えることはよく指摘されることだが，今回の分析においてもその点が確認された。意外なことに敬体率の高さが抑揚・リズム性の低さと関係している。敬体の持つ軟らかい印象が抑揚・リズム性の低下を招いた可能性はある。

終助詞率： 「ね」「よ」など対人的機能を有する終助詞の使用は，相手との距離感を縮め，結果としてくだけた軟らかい印象を与えることになると考えられる。文末に終助詞が置かれると一定のリズムを生む効果があると考えたが，むしろ逆の結果となった。この種の終助詞の置かれるテキストが限られていたことに影響している可能性もある。

文長・文長変動率： 文の長さは，複文か単文かといった文構成の複雑さだけでなく，文末表現やモダリティ表現、修飾表現の有無・多少にも影響するため一概には言えないが、総じて長い文ほど文構成が複雑で硬い印象を与えることにつながるものと思われる。また長い文に比べて短い文の方が抑揚・リズムを作ることは小磯ほか (2011) の観察とも合うが，文長変動率についてはむしろ抑揚・リズム性を抑制する効果があり，長短の文を織りまぜた文章は抑揚・リズムを産み出すのではないかという予想は外れた。単純に指標の作り方に問題があった可能性もあるため，この点については今後の課題としたい。

4 おわりに

本稿では，テキストの印象評定に基づき抽出された三つの因子「スタイル」「文構成の明晰性」「抑揚・リズム性」に対し，いかなる言語的特徴がその印象形成に関わるかを検討した。今後より詳細な分析・考察を経た上で，個々の言語的特徴からどの程度テキストが分類できるかを検討したい。

参考文献

Halliday(1985) *Spoken and Written Language*. Victoria: Deakin University.

Halliday(1990) Some grammatical problems in scientific English, *Annual Review of Applied Linguistics*, 6, pp.13–37.

小椋 (2005) 「『日本語話し言葉コーパス』の資料性—形態論情報を用いた分析から—」『国語語彙史の研究』24, 259–275, 和泉書院.

小椋ほか（2011）『『現代日本語書き言葉均衡コーパス』形態論情報規定集第 4 版』国立国語研究所内部報告書.

樺島（1963）「漢語をめぐって」『計量国語学』27, 14–19.

小磯ほか（2009）「コーパスに基づく多様なジャンルの文体比較—短単位情報に着目して—」『言語処理学会第 15 回年次大会発表論文集』pp.594–597.

小磯ほか（2011）「テキストの多様性をとらえる分類指標の体系化の試み」『言語処理学会第 17 回年次大会発表論文集』pp.683–686.

国語研究所（1955）『談話語の実態』国立国語研究所報告 8, 秀英出版.

佐野・丸山（2008）「システミック文法に基づく書きことばの複雑さ測定— 日本語大規模コーパスを用いた語彙密度計測—」『言語処理学会第 14 回年次大会予稿集』, pp.1097–1100.

野元（1959）「話しことばの中での漢語使用」『ことばの研究』国立国語研究所論集 1 .

付記：本研究は，文部科学省科学研究費特定領域研究「日本語コーパス」及び基盤研究（C)「書き言葉コーパスに基づくテキスト分類尺度の探索的研究」の助成を受けている。

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved