テキストの多様性をとらえる分類指標の体系化の試み

小磯 花絵　　田中 弥生　　小木曽 智信　　近藤 明日子
人間文化研究機構　国立国語研究所

{koiso,yayoi,togiso,kondo}@ninjal.ac.jp

1　はじめに

近年，大規模な言語コーパスが研究に利用できる環境が整ってきた。これによって，例えば，小説と新聞，インターネット上のテキストの文体がどことなく異なるという直観を，言語コーパスを用いた定量的分析を通して，言語特徴の差として具体的に捉えることが容易に行えるようになった。しかし一方で，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古くから行われており，様々な観点や指標が提案されてきた。例えば永野 (1968) は，日本語の文章文体研究で取り上げられてきた分類の視点を,「機能」「内容」「形式」という３つの分類基準とその具体的設定項目に整理し，体系化を試みている。また英語を中心とした研究では，Halliday が状況のコンテクストとして,「活動領域（field)」「役割関係（tenor)」「伝達様式（mode)」の３つを提案している（Halliday and Hassan(1985) など)。言語表現自体を直接分類するものではなく，言語選択に影響を与えるものとして状況のコンテクストを体系化したものであるが，書き言葉・話し言葉を体系的にとらえるうえで上で重要な指標と言える。

しかしこれらの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，少なくとも日本語の研究を見る限りあまり行われていない。そこで著者等は，従来指摘されてきた文章を評価・分類する指標を参考に８つの分類尺度を構成した上で，４００のテキストを対象に被験者に５段階で評定してもらうという実験を実施したが，少なくとも評定結果からは，多様なテキストを有効に分類することはできなかった。

そこで，理論的側面から類型化・体系化を試みるという方向を離れ，まず人が種々の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る。本報告ではこのうち，今までに実施した（１）テキストの印象を評定する表現の収集調査,（２）評定実験,（３）評定結果の分析に基づく分類指標試作版の構築について報告する。

2　評定語の収集

一般の人がテキストを描写･評価する際に用いる様々な表現を収集することを目的に，７名の被験者に４０のテキストを読んでもらい，それぞれのテキストから受ける印象を記述してもらうという調査を実施した。以下に調査の具体的手続きと収集した評定語の概略を記す。

2.1　方法

資料：　現在，国語研究所が中心となって構築している『現代日本語書き言葉均衡コーパス』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」情報が付されたコア（小椋ほか 2010）と呼ばれるデータセットから，新聞（5 サンプル），小説以外の書籍（10 サンプル），雑誌（10 サンプル），行政白書（4 サンプル），Yahoo!ブログ（11 サンプル），計 40 サンプルを選んだ。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は対象外とした。

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の文末まで）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，あるいは途中で切れたりしているものもある。

調査者：　７名の調査者（男性１名，女性６名）が調査に参加した。

手続き： 調査者には，二つのテキストの対が一つの頁に記された調査票２０頁（２０対４０サンプル）が渡された。サンプルの組合せや出現の順番はランダムに決定し被験者毎に異なる。

調査者は二つのテキストを読んだ上で，まずそれぞれのテキストから受ける印象やそれを描写する表現を思いつく限り自由に記述してもらった。次に，二つのテキストに共通する印象，対立する印象があれば，それを記してもらった。対立する印象については，例えば「硬い－柔らかい」のように，対語形式とするよう依頼した。テキストを比較して共通する印象，対立する印象を考えることで，個々のテキストを読むだけでは思いつかないような評定語が出てくる可能性があると考え，このような手続きを踏んだ。

なお調査者には，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を記すよう具体例を挙げて注意を促した。例えば殺人に関するテキストを読んで「怖い」と評価するのではなく，殺人に関するリアルな文章を読んで「臨場感あふれる」と評価するように，といった指示である。

2.2 収集した評価語

上記の手続きの結果，調査者あたりおよそ異なりで80〜180，７名全員で 441 の評価語が抽出された。その上で，次の手続きに従い評定語対を作成した。

まず,「練られた」「良く練られている」「文章が練られた」のような類似した表現の評価語をまとめ上げた上で,「練られた」「推敲された」「整理された」のように類似したカテゴリーのものをまとめ上げた。類似した表現によってカテゴリーを構成しないもの（例えば「こびるような」など）は，テキストの印象を表す表現として典型的ではないと判断し除いた。また，文章の内容や著者自身に関する印象と解釈される，あるいは解釈される危険性の高いもの（例えば「器の大きい」など）も省いた。その上で，各カテゴリーの中から対語形式の評定尺度を作成した場合に一番自然で曖昧性の少ないものを選択した。例えば「簡潔性」に関する評価語では,「簡潔な」「さっぱりした」「削ぎ落とされた感じ」「無駄のない」「まどろっこしい」「煩雑な」「ごちゃごちゃした」「冗長な」「簡潔－冗長な」「すっきりした－ごちゃごちゃとした」「まとまり感のある－まとまり感のない」などが分類されたが，この中から対語にした場合に表現として一番自然だと考えられる「簡潔な－冗長な」を選択した。

以上の手続きにより評定語の整理をした結果，次の２０の評定語対が構成された。

- 改まった－くだけた
- 型にはまった－個性的な
- よく練られた－練られていない
- 整然とした－雑然とした
- 簡潔な－冗長な
- 自然な－わざとらしい
- 直接的な－婉曲的な
- 客観的な－主観的な
- 臨場感のある－臨場感のない
- 具体的な－抽象的な
- 読み手に語りかける－語りかけの少ない
- 書きことば的－話しことば的
- 相手の理解を配慮した－相手の理解を無視した
- 重い－軽い
- 暗い－明るい
- 冷静な－興奮した
- 硬い－柔らかい
- めりはりのある－単調な
- テンポのよい－テンポの悪い
- 親しみやすい－とっつきにくい

3 評定実験

3.1 方法

安定した尺度を構成するため，得られた２０の評定語対をもとに，テキストに対する印象について SD 法による５段階評定実験を実施した。テキストサンプルとして，前節に示した評定語の収集調査と同じ４０サンプルを用いた。評定語の収集調査とは異なる３名の被験者（男性２名，女性１名）が実験に参加した。

被験者には，テキストを熟読した上で，２０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た。また評定尺度は適宜左右を反転させた。本番に先立ち練習問題として５つのサンプルを評定してもらった。被験者には，前節の評定語の収集調査と同様，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。

3.2 結果

得られた評定結果を対象に，被験者毎に各評定尺度毎の分散を求めたところ,「具体的な－抽象的な」と「直接的な－婉曲的な」については，いずれの被験者においても分散が小さい，つまり評定値が偏っていることが分かった。この偏りは，単純に評定対象としたサンプルに偏りがあったためとも考えられるが（例えば「婉曲的」なテキストがほとんどなかったために「直

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	**0.945**		-0.279
硬い－柔らかい	**0.919**	-0.147	-0.180
重い－軽い	**0.908**	-0.108	-0.307
型にはまった－個性的な	**0.895**		-0.129
書きことば的－話しことば的	**0.893**	-0.144	-0.217
冷静な－興奮した	**0.837**	0.323	-0.207
暗い－明るい	**0.707**	-0.181	-0.470
簡潔な－冗長な	0.120	**0.831**	
整然とした－雑然とした	0.124	**0.817**	0.362
自然な－わざとらしい	-0.325	**0.746**	0.397
めりはりのある－単調な	-0.241	0.285	**0.802**
テンポのよい－テンポの悪い	-0.197	0.208	**0.767**
寄与率	48.4%	18.6%	15.8%

接的」に寄ってしまった，など），評定尺度の設定自体に問題があった可能性も十分にあるため，分析対象外とした。

上記を除く１８の評定尺度を対象に，因子分析（最尤法，バリマックス回転）を行った。事前に主成分分析を行い，固有値などから３因子が一つの目安であると判断した（累積寄与率 80.2%）。因子分析の結果，因子負荷量がいずれも小さい値しか示さない評定尺度や複数の因子にまたがって高い因子負荷量を持つ評定尺度は削除した。これを繰り返した結果，最終的に１２の評定尺度が残った。各尺度の因子負荷量を表 1 に示す。

まず，因子３から見ていこう。因子３については「めりはりのある－単調な」と「テンポのよい－テンポの悪い」など，抑揚や速さ感に関する尺度が相対的に高い正の負荷を示しており，文章の抑揚・リズムに関する因子であると解釈することができる。

次に因子２を見ると，「簡潔な－冗長な」「整然とした－雑然とした」「自然な－わざとらしい」が相対的に高い正の負荷を示しており，文構成の明晰性に関わる因子であると解釈することができる。

因子１を見てみると，７つの尺度が正の負荷を示している。「改まった－くだけた」や「硬い－柔らかい」など発話のスタイルに関するものが含まれており，そのスタイルの違いにより，軽重，明暗，動静などの印象が派生したと考えると，これは文章のスタイルに関する因子と解釈することができる。

しかし残る「書きことば的－話しことば的」については，確かに「話し言葉的」といった場合に口語調のくだけた印象が喚起されることから，スタイルとの関連性があると考えられる一方で，「話しことば的」であっても改まり度の高い文章が容易に想像つくことからも分かるように，必ずしも同種の尺度ではない。

次に挙げる二つのテキストは，「改まった，かつ，話し言葉的」「くだけた，かつ，書き言葉的」と判断されたサンプルの一部を抜粋したものである。

【例１】では週別に販売数の推移を追い、どのように商品構成を変え、売れ筋の棚割と在庫を変え、販売数を伸ばしたかを**見てみよう**。（ID:PB46_00066）

【例２】お会計は現金で**おやぢ**とやり取りします。麺・だし・天ぷらといたって**フツー**ですが安心して食べられます。（ID:OY14_04336）

例１は，決してくだけた表現は用いられていないが，読み手に対する働き掛けの表現が含まれており，これが「話しことば的」という印象に影響を与えたと考えられる。実際，因子分析の過程で落とされたが「読み手に語りかける－語りかけの少ない」という評定尺度で語りかけの程度が高いと判断されたサンプルである。一方例２は，その種の働き掛けはないが（語りかけの程度が低いと判定），「おやぢ」や「フツー」といった表記の仕方がくだけた印象を与えていると考えられる。しかし，次の例のように，くだけた表現が多く使用されるものは，読み手への働き掛けの有無に関わらず話しことば的と判断されることも多い。

【例３】汗かいて帰ってきたら、シャワー浴びてすっきりして冷蔵庫から冷たいモノ出して、、、っていうのが超天国な気分。さすがに今日は、冷たいモノ食べたい気分なのでそうめんにしちゃいました。（ID:OY01_00848）

このように「書き言葉的－話し言葉的」という尺度は，改まりの程度や読み手に対する働き掛けの程度など，複数の観点が関与する多義的な意味合いをもつ尺度である可能性がある。今回のデータではスタイルに関わる因子と強い関係を示したが，今後もう少しサンプルのバリエーションを増やして慎重に検討する必要がある。

3.3 テキストの分類

最後に，実際に得られた三つの因子を用いて暫定的にテキストを分類し，どのような種類のテキストが分類されるかを概観する。ここでは単純に，各因子毎に，それに強く関係する評定尺度の平均を取るという方法で代表値を算出した。以下に具体例を示す。

例４と例５は，いずれもスタイルと文構成の明瞭性が高いサンプルである。しかし例４は抑揚・リズム性が

低く，例５は高いという点において異なる。例４（行政白書），例５（新聞）ともに，過去に実施した事柄・過去の出来事を伝達しているテキストであるが，例４は全ての文が過去形で語られており変化がないのに対し，例５は，時制も過去形・現在形と変化があり，またアスペクトや伝聞形式が使用されるなど，バリエーションに富んでいることが分かる。次に挙げる例６・例７も共に抑揚・リズム性の高いサンプルであるが，やはり文末文体の変化や読み手に対する働き掛けなどが見られる。また例４では文の途中を省略したため分かりづらいが，全般的に長い文章が淡々と続くのに対し，後者は比較的短い文によって構成される。例５と同カテゴリーのテキストを概観すると，文長が相対的に短い，あるいは長い文と短い文が混在するものが多い傾向が見られる。これらの違いが抑揚・リズムの違いに結びついた可能性がある。

【例４ スタイル:高，文構成明晰性:高，抑揚リズム:低】
「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５ スタイル:高，文構成明晰性:高，抑揚リズム:高】
一方、公立高一年の少女（十六）＝殺人予備容疑で逮捕＝は、犯行に使うための文化包丁を学校帰りに百円ショップで**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、河内長野署捜査本部は、それぞれの殺意に関連があるとみて**調べている**。調べでは、少年は犯行の三日前の先月二十九日午後一時ごろ、自宅に近いホームセンターに一人で訪れ、刃渡り約二十センチの刺し身包丁を購入していた**という**。（ID:PN3d_00013）

例６と例７は，共にスタイルが低く抑揚・リズム性の高いサンプルであるが，前者は文構成の明瞭性が高いのに対し後者は低いという違いが見られる。例６は書籍のサンプル，例７はブログのサンプルである。被験者にはサンプルの出典に関する情報は一切与えられていない。同じスタイルが低めで抑揚・リズム性の高いテキストでも，執筆・出版過程でおそらく十分な推敲がなされたテキストと，相対的に十分な推敲がなされずその日の出来事を思い付く順に記したテキストでは，文構成の明晰性の観点でその差が出てくるということであろう。

【例６ スタイル:低，文構成明晰性:高，抑揚リズム:高】
おそらく、ご自分はそこまでひどくはないと思いながらも、みなさん多少は身に覚えがあることだからでしょう。それだけ世の母親というものは、わが子の欠点をリストアップすることに熱心で、わが子の悪い点に関しては権威でいらっしゃいます。だから、お子さんの悪いところをあげてくださいなどと言おうものなら、それこそいくらでも出てきて際限がありません。（ID:PB23_00051）

【例７ スタイル:低，文構成明晰性:低，抑揚リズム:高】
汗かいて帰ってきたら、シャワー浴びてすっきりして冷蔵庫から冷たいモノ出して、、、っていうのが超天国な気分。さすがに今日は、冷たいモノ食べたい気分なのでそうめんにしちゃいました。でも、それだけだとなんなので　夏野菜たっぷりサラダと激辛のフライドチキン。CMでレッドホットでしたっけ、やってるの見て「美味しそう?」と。。。（ID:OY01_00848）

4　おわりに

本研究では，一般に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。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。それに従い暫定的にテキストを分類してその違いを概観したところ，たしかにテキストの多様性（の一面）を捉えており，少なくとも全く妥当性・有効性のない指標ではないことがうかがわれた。

今後，今回得られた尺度によって分類されたテキストが共通してどのような特徴を有するかを詳細に分析すると同時に，これらの尺度では捉えきれないテキストの相違がないかを検討し，改良を重ねる必要がある。今後の課題としたい。

参考文献

永野賢 (1968)「文章の分類論」森岡健二ほか（編）『作文講座 4 文章の理論』, 明治書院.

Halliday and Hassan(1985) Language, context and text: A social semiotic perspective, Deakin University Press.

小椋秀樹ほか (2010)『『現代日本語書き言葉均衡コーパス』形態論情報規定集第３版』国立国語研究所内部報告書

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」の助成を受けたものである。