様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年 5 月 29 日現在

機関番号：62618
研究種目：基盤研究（C）
研究期間：2009～2011
課題番号：21520493
研究課題名（和文）　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
研究課題名（英文）　The Construction of the Classification Criteria for the Variation in Written Texts Based on the Written Corpus

研究代表者
小磯　花絵（KOISO HANAE）
大学共同利用機関法人人間文化研究機構国立国語研究所・理論・構造研究系・准教授
研究者番号：30312200

研究成果の概要（和文）：多様な書き言葉を分類するための基本的な尺度を，コーパスに基づく探索的研究を通して構築することを目的に，四回の評定実験を実施した。分析の結果，「スタイル」「抑揚・リズム」「文構成の明晰性」という三つの分類指標が得られた。また実験データの言語分析の結果，指標毎に特徴的な言語傾向を有することが分かった。この結果は，三つの指標によって多様な文体の書き言葉が分類可能であることを示唆する。

研究成果の概要（英文）：In order to construct the classification criteria for the variation in written text based on the written corpus, four experiments on evaluation of text were conducted. As the results of the analysis, three classification criteria, style of writing, rhythm of texts, and clarity of text structure, were extracted. In addition, the linguistic analysis reveals that each criterion has its own linguistic pattern, implying that the extracted three classification criteria could properly classify a diversity of texts.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２１年度	1,500,000	450,000	1,950,000
２２年度	1,000,000	300,000	1,300,000
２３年度	800,000	240,000	1,040,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：人文学
科研費の分科・細目：言語学・日本語学
キーワード：書き言葉コーパス・テキスト分類・分類尺度・文体・ジャンル

１．研究開始当初の背景

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たことにより，例えば，新聞と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，コーパスを用いた定量的分析を通して具体的に捉えることが容易になった。しかし，書き言葉の多様性は，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，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。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。書き言葉の頑健な分類尺度の構築は，コーパス日本語学にとって喫緊の課題と言える。

２．研究の目的

本研究の目的は，新聞や雑誌，インターネット上のテキストなど様々な書き言葉を多角的に分類するための基本的な尺度を，コーパスに基づく探索的研究を通して構築することである。具体的には，(1)従来提案されてきた観点や指標を参考に予備調査を経て分類尺度を選択，(2)様々な種類の書きことばを対象に尺度に基づき人手で評定値を付与，(3)評定値データを利用した多角的分析を通して個々の尺度の妥当性・有用性を評価した上で基本尺度を決定する，という手順を踏むことで，コーパス研究に有用な分類尺度を探索的，実証的に構築することを目指す。

３．研究の方法

書き言葉を多角的に分類するための基本的な尺度を，コーパスに基づく探索的研究を通して構築することを目標に，4 回の評定実験を実施した。

(1)評定実験 A：
『現代日本語書き言葉均衡コーパス』(以下，BCCWJ）より，テーマを社会に限定して 300 のサンプル（各 500 文字）を抽出した。その上で，評定者 3 名により，八つの評定語対（例：「改まった・くだけた」，「客観的・主観的」）にもとづき，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。評定語対については，先行研究を参考に文体や文章分類に関する指標・評定語を抽出・整理した上で，2 回の予備実験を通して選別した。

(2)評定実験 B：
評定実験 A の結果をふまえ，テキストと評定語対を増やした実験を実施した。具体的には，BCCWJ より，新聞，書籍，雑誌，行政白書，Yahoo!ブログから 40 のサンプル（各 300 文字）を抽出した上で，評定者 3 名により，20 の評定語対（例：「改まったーくだけた」，「整然としたー雑然とした」，「めりはりのあるー単調な」）にもとづき，SD 法による 5 段階評定実験を実施した。評定語対については，7 名による 40 のテキストを対象としたテキスト印象表現抽出調査に基づき作成した。

(3)評定実験 C：
評定実験 B で得られた三つの分類指標「スタイル」「抑揚・リズム」「文構成の明晰性」の妥当性を検証するため，被験者を 3 名から 30 名に増やした上で，評定実験 B と同じ手続き（同一テキスト 40 サンプル，異なる評定者 30 名，同一の評定語対 8 対に基づく SD 法による 5 段階評定）に従い実験を実施した。

(4)評定実験 D：
評定実験 C により三つの分類指標の妥当性が確認されたことを受け，サンプル数を 40 から 300 に増やした上で，評定者 3 名により六つの評定語対（スタイル指標から「改まったーくだけた」「硬いー柔らかい」，抑揚・リズム指標から「テンポのよいーテンポの悪い」「めりはりのあるー単調な」，文構成の明晰性指標から「簡潔なー冗長な」「整然としたー雑然とした」）に基づく SD 法による 5 段階評定実験を実施した。

４．研究成果

評定実験 A の結果を受け，評定語間の相関関係や評定者間の一致度などの観点から，調査の妥当性や問題点などについて検討した。また，品詞・語種・文特徴(文末文体など)と各評定語との関係についても予備的に調査した。

以上の調査の結果，先行研究を参考に選択した指標だけでは多様なテキストを有効に分類することはできないと判断し，評定実験 B の方針に切り替えた。評定実験 B は，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ものである。実験で得られた評定結果を対象に因子分析を行った結果，最終的に 12 の評定尺度が残り，「スタイル」「抑揚・リズム」「文構成の明晰性」という 3 つの因子が抽出された。

評定実験 C では，規模を増やした上で再度実験を行い，同種の三つの因子が得られた。これにより評定実験 B・C で得られた三つの分類指標の妥当性が確認された。

以上の 3 回に渡る予備調査を踏まえ，サンプル数を増やした本調査（評定実験 D）を実施し，各指標にいかなる言語特徴が関わるかを検討した。具体的には，文章の特徴を捉える上で関連が深いと考えられる言語特徴（品詞や語種など単語に関わる特徴量５種，文末の種類や文長など文に関わる特徴量８種，計 13 種）を先行研究･予備調査を踏まえて選択

し，各指標との関係を分析した。その結果，指標毎に特徴的な言語傾向を有することが分かった。

この結果は，三つの指標によってそれぞれ異なるタイプ･文体の書き言葉が分類されていることを示唆するものであり，書き言葉を多角的に分類する指標を探索的に構築するという本研究の目標に対しある一定の成果が得られたと言える。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 3 件）

① 佐野大樹・小磯花絵(2011)「現代日本語書き言葉における修辞ユニット分析の適用性の検証－「書き言葉らしさ・話し言葉らしさ」と脱文脈化言語・文脈化言語の関係－」『機能言語学研究』, 査読有, 6(2), pp. 59-81.
② 田中弥生(2011)「「質問－回答」における待遇表現の特徴 －書籍 QA、WebQA、Yahoo!知恵袋の比較から－」」『待遇コミュニケーション研究』, 査読有, 8, pp. 65-80.
③ 田中弥生(2010)「質問サイトにおける情報要求モデルと待遇コミュニケーション―「アットコスメ美容事典」の談話機能・談話構造の分析から」『待遇コミュニケーション』, 査読有, 7, pp. 33-48.

〔学会発表〕（計 30 件）

① 小磯花絵･田中弥生･小木曽智信･近藤明日子(2012)「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)」言語処理学会第 18 回年次大会(NLP2012), 2012 年 3 月 15 日， 広島市立大学
② 小木曽智信･相良かおる(2012)「医療分野で使われる複合語の語種構成」第 29 回社会言語科学会年次大会, 2012 年 3 月 11 日， 桜美林大学
③ 田中弥生(2011)「Q&A サイトにおける待遇意識の諸相―修辞ユニット分析を用いた脱文脈化の観点からの検討―」待遇コミュニケーション学会 2011 年度秋季大会, 2011 年 10 月 15 日, 早稲田大学
④ 田中弥生(2011)「修辞ユニット分析を用いた Q&A サイトの質問と回答における修辞機能の展開の検討」社会言語科学会第 28 回研究大会, 2011 年 9 月 17 日, 龍谷大学
⑤ 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, 2011 年 8 月 2 日， JA 共済ビルカンファレンスホール
⑥ 田中弥生・佐野大樹(2011)「Yahoo!知恵袋における質問と回答の分類－修辞ユニット分析を用いた脱文脈化-文脈化の程度による検討－」社会言語科学会第 27 回研究大会, 2011 年 3 月 19 日, 桜美林大学
⑦ 小磯花絵･田中弥生･小木曽智信･近藤明日子(2011)「テキストの多様性をとらえる分類指標の構築を目指して」文部科学省科学研究費特定領域研究「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 22 年度公開ワークショップ, 2011 年 3 月 16 日, 時事通信ホール
⑧ 小磯花絵･田中弥生･小木曽智信･近藤明日子(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(NLP2011), 2011 年 3 月 9 日, 豊橋技術科学大学
⑨ 田中弥生・佐野大樹(2011)「修辞ユニット分析からみた Q&A サイトの言語的特徴」 言語処理学会第 17 回年次大会(NLP2011), 2011 年 3 月 8 日, 豊橋技術科学大学
⑩ 田中弥生(2010)「Yahoo!ブログにおける待遇表現 ― 投稿に使用した機器による比較」待遇コミュニケーション学会 2010 年秋季大会, 2010 年 10 月 30 日, 早稲田大学
⑪ 小磯花絵・小木曽智信・小椋秀樹・宮内佐夜香(2010)「長単位情報に基づくジャンル間の文体に関する分析」特定領域研究「日本語コーパス」平成 21 年度公開ワークショップ, 2010 年 3 月 16 日, 東京工業大学
⑫ 小木曽智信(2010)「特定領域研究「日本語コーパス」平成 21 年度公開ワークショップ」社会言語科学会第 25 回研究大会, 2010 年 3 月 13 日, 慶応義塾大学
⑬ 田中弥生(2010)「Q&A サイトの「質問－回答」における結束性 ～ 省略の特徴分析 ～」電子情報通信学会 言語理解とコミュニケーション研究会 第 1 回集合知シンポジウム ～言語処理が紡ぎ出す未来～, 2010 年 1 月 25 日, 広島市まちづくり市民交流プラザ
⑭ 田中弥生(2009)「アットコスメの待遇意識表現」「言語と人間」研究会 11 月例会, 2009 年 11 月 9 日, 立教大学
⑮ 小木曽智信(2009)「『現代日本語書き言葉均衡コーパス』における可能表現のバリエーション」日本語学会 2009 年度秋季大会, 2009 年 11 月 1 日, 島根大学

６．研究組織

(1)研究代表者
小磯　花絵（KOISO HANAE）
国立国語研究所・理論構造研究系・准教授
研究者番号：30312200

(2)研究分担者
小木曽　智信（OGISO TOSHINOBU）
国立国語研究所・言語資源研究系・准教授
研究者番号：20337489

田中　弥生（TANAKA YAYOI）
神奈川大学・外国語学部・非常勤講師
研究者番号：90462811
（H22→H23 連携研究者）

(3)連携研究者
近藤　明日子（近藤　明日子）
国立国語研究所・コーパス開発センター・
プロジェクト奨励研究員
研究者番号：30425722

(3)連携研究者　井上　優
麗澤大学・外国語学部・教授
研究者番号：30213177