様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月　 ９ 日現在

機関番号：３４３１０
研究種目：基盤研究(C)（一般）
研究期間：2018～2020
課題番号：１８Ｋ００６２７

研究課題名（和文）データサイエンスに基づいた日本文体変化分析とその構造のモデリング

研究課題名（英文）A Study of Stylistic Change in Japanese Based on Data Science and Modeling of its Structure

研究代表者
金　明哲（KIN, Meitetsu）

同志社大学・文化情報学部・教授

研究者番号：６０２７５４６９

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究では，まず，100年以上にわたる膨大な小説の中から，毎年5～6人の代表的な作家の作品をサンプリングし，592人の作家による592編の小説（9557078文字）のコーパスを作成した。次に、コーパスに対して形態素解析と構文解析を行い、文体の特徴を分析した。分析は，教師なしの方法を用いて文体の特徴について概観したうえで，教師ありの学習方法で時間に伴って変化が顕著である変数を特定し，モデリングを行った．その結果，助詞や文末パターンには経年によって顕著に増減していることが明らかにされ，これらに対して言語学や文体学の視点で解釈を試みた．

研究成果の学術的意義や社会的意義
本研究では，日本語の現代文における文体および言語の経時的変化について機械学習やモデリングなどのデータサイエンスの手法で変化の要素を明らかにすると同時に，その現象の裏に潜んでいる要因を社会学，文体学，言語学の視点で究明を試みた．本研究の成果は，日本語文体および言語学の研究などに有益な学問的情報を提供するだけではなく，現代社会における人文社会科学の研究にデータサイエンスの方法を用いる有効性を示すに値する．

研究成果の概要（英文）：In this study, we first created a corpus of 592 novels (9557078 characters) by 592 authors with sampling the works of five to six representative authors each year from the vast collection of novels spanning over 100 years. Next, we performed morphological and syntactic analysis for the corpus to analyze the stylistic features. The analysis was conducted by using unsupervised methods to provide an overview of stylistic features, and then using supervised learning methods to identify and model variables that changed significantly over time. As a result, it was found that there was a marked increase or decrease in particles and sentence-final patterns over time. In addition, we've attempted to interpret them from the perspective of linguistics and stylistics.

研究分野：計量言語学，計量文体学，テキストマイニング

キーワード：文体の変化　モデリング　テキストマイニング　助詞　文末パターン

１．研究開始当初の背景

文体研究は，古典ギリシア時代から修辞学の一部として行われ，現在の形に発展してきた．日本では 1930 年代末から科学的アプローチによる文体論の研究が行われてきた．日本現代の文体に関して，樺島・寿岳(1965)は「現代小説」100 作品から各作品 80 文を無作為に抽出し，10 種類の文体項目についてデータを収集し，文章の性質の分析を行った．また，安本(1981)は 100 人の現代作家作品について，15 個の文体項目のデータを選出し，因子分析法を用いて 100 人の作家を 8 グループに分類した．村上・古瀬(1999–2001，科研番号 11878048)は，川端康成と三島由紀夫の作品のコーパスを作成し，文体について比較分析を行った．劉・金(2017)は，宇野浩二の戦前と戦後の文体変化について計量分析を行い，その変化の要素を突き止めた．

文体変化に関して，乾善(2012-2015，科研番号 24520514)は，平安和文と中世和漢混淆に関する初期の仮名文学作品の実態は，「和漢混淆文」という日本語の書記用文体（散文文体）の成立でもあったことを明らかにした．河野(2014–2017，科研番号 26370262)は，言文一致運動と近代における短歌表現について啄木短歌を中心に，動詞の終止形止めの歌について考察を行った．小西(2013–2016，科研番号 25770178)は，明治 20 年代の翻訳小説に限定してコーパスを構築し計量的に分析を行った．田中ら(2006–2007，科研番号 15K02585)は，「太陽コーパス」を活用して明治後期から大正期にかけて進んだ「言文一致」という出来事について，動詞を例に言文一致期に定着する語と衰退する語とを対比的に分析した．飛鳥ら(1994–1995，科研番号 06301047)は，19 世紀日本における言語文化の変容と異言語接触について総合的に研究を行った．

文体は時代の変遷に伴って言語の使用傾向と共に変遷・変容・変化している．周知のように現代の日本語の文体は，明治時代に文学者による言文一致の改革運動を経て，文語体から口語体に変容してきた．また，国による国語政策も文体に大きい影響を与えている．なお，国民の文体に対する嗜好の時代性が文体の変化を後押ししている可能性もある．このような視点による近現代日本語の文学作品の社会的文体の変化に関するシステマティックな研究は管見の限りない．日本語の文体がどのように変化しているか，何が変化しているかに関する研究は，文体論に限らず，日本語の変遷にも密接に関係しているため，日本語・文体の現状を知り，文体の変化の行方を予測するのに非常に重要である．

２．研究の目的

本研究の目的は，小説文に焦点を当て，データサイエンスの手法を用いて近現代文学作品の文体変化およびその要素を分析する．また，その変化の構造をモデリングし，近未来の文体の変化の行方を予測することである．

３．研究の方法

本研究では，従来の内省による伝統的文体分析の結果を踏まえ，大量の文体要素およびその組み合わせをコーパスから抽出し，統計的方法および機械学習の手法などを用いて分析し，モデリングする．実現可能性を考慮し，小説に焦点を当てコーパスを作成して研究に用いる．

４．研究成果

本研究では，まず，100 年以上(1905～2015 年)にわたる膨大な近現代小説から，1 年ごとに約 5 名ずつの代表的な作家の作品をサンプリングし，コーパスを作成した．完成したコーパスは 592 人の 592 篇作品で，総文字数は 9557078 である．具体的には，青空文庫・電子文芸館から 175 人の 175 篇の作品，それ以外の 417 篇は紙媒体を OCR で電子化し，クリーニングを行った．次に，作成したコーパスについて，形態素解析や構文解析を行ったうえで分析を進めている．文体の変化を進化という視点で系統分析，モデリングの視点で正則化回帰モデル，ランダムフォレスト回帰と構造的トピックモデル，深層学習のアプローチからは BERT によるベクトル埋め込みなどを駆使して研究を行っている．その結果の一部として，各々助詞の中には経年によって顕著に増減しているものがあることが明らかにされた．その結果をまとめた論文は，専門学会の論文誌に採択された．また，文末の表現パターン，品詞の使用傾向などにも変化が見られ，その詳細な分析を進めている．

（1）助詞の使用変化について

①　助詞のモデリング

日本語は，叙述すべき事柄を請け負う語彙の一つひとつに助詞の形で添えて確めながら先へ先へと進めていく表現方式を取っている特徴を持つと言われる．このような構文的性格のため，日本語文章の表現特性及びその変化を助詞から窺い知ることができる．

本研究では，MeCabを用いて形態素解析を行った．形態素解析のための辞書は何種類

もあり，異なる特徴や利点を持つ．本研究では，単語長と曖昧性解消との観点から，デフォルトのIPA辞書を用いた．IPA辞書は学校文法をベースとしており，助詞が格助詞，係助詞，副助詞，接続助詞，終助詞，並列助詞，副詞化助詞，連体化助詞，特殊助詞という９つの種類に分類されている．今回はデータを集計する際，「ヲ_格助詞」「を_格助詞」のような，異表記は同じ項目としてカウントする．一方，「たり_並列助詞」「だり_並列助詞」のような，語彙の活用形の影響により清濁の違いが生じる項目については，異なる項目としてカウントした．その結果，コーパスから計131項目の助詞が抽出された．

助詞使用の歴史的変化を分析するに当り，まず系統樹を用いて，変化が発生しているかについて考察した．系統樹分析とは，言語や生物といった対象の集合の要素に対して特徴量のベクトルを抽出し，それを用いて，関連性を樹木の枝分岐の形式に示したものである．系統樹の生成手法は多く提案されており，距離行列法がその代表的な一つである．距離行列に基づいた系統樹の生成法は複数存在しているが，ここでは最も広く利用されている近隣結合法（Neighbor Joining，NJ 法）を利用する．各助詞の使用率は，標本の確率分布として見なすことができるため，距離行列の計算については，次に示す JSD(Jensen–Shannon Divergence）の平方根を用いた．作成された系統樹から，助詞の使用が時期の推移に伴って変化している傾向が窺えた．

視覚的な考察を踏まえて，時系列情報と助詞使用状況を表現する回帰モデルを構築し，モデリングに寄与する特徴的な助詞項目を特定し，分析を試みた．回帰分析は，現象の結果とそれに影響をおよぼすと考えられる複数の要因との相関関係のモデルである．

回帰モデルを構築するため，助詞データセットに，各年度の番号を表す1～105という数字列（y列と呼ぶ）を目的変数として付け加え，131個の助詞項目を説明変数とする．

集計した助詞データを使って，グリッドサーチにより α (0.9)を決め，elastic net回帰モデルの構築を行った結果，自由度調整済み決定係数は0.97であった．elastic net回帰モデル選ばれた8つの説明変数とその係数を表1に示す．係数の絶対値が大きいほど回帰モデルでの影響が大きいと判断される．

表1　elastic neモデルの説明変数の係数

1.の_格助詞	-2349.24
2.へ_格助詞	-2294.87
3.なんて_副助詞	1909.79
4.で_格助詞	1097.08
5.しか_係助詞	369.62
6.さえ_係助詞	-296.90
7.が_格助詞	196.26
8.よ_終助詞	87.81

作成されたelastic netモデルを確認するため，回帰木をアンサンブル学習するランダムフォレストによる回帰モデルを構築し，選択された変数の重要度の分析を行った．その結果，明らかに大きい上位７個の変数は，「へ_格助詞」「が_格助詞」，「で_格助詞」「の_格助詞」「けど_接続助詞」「なんで_副助詞」「しか_係助詞」であり，表1と共通しているのは6つである．本研究では，elastic netとランダムフォレストの結果の内，共通な項目について考察を行った．

選択された特徴的助詞，格助詞「へ」「で」などのような，意味（語義）が多様であり出現頻度が高いものもあれば，語義の種類が比較的に少なくそれほど頻繁には使われていないものもある．その変化は何を意味しているか，或いは文章の表現様式・文体にどのような影響をもたらしているかについて，助詞ごとの特性に応じて考察を進める必要がある．

語義も数も多い助詞に対し，周囲の語彙との関係を糸口にして，変化の詳細及びそれが言語様式･文体との関わりを浮き彫りにできる可能性がある．その方法の一つとして，共起ネットワークを通して対象助詞とその前後の語彙との関わりから考察することが挙げられる．一方，出現頻度が低いものに関して，一文一文を確認して考察することができる．

② 格助詞「へ」

まず，格助詞「へ」を含む語彙のbigram（助詞と一つ前/後の語彙による組合せ）を用

いた共起ネットワーク分析を行った．比較対照で変化を捉えるため，コーパス内の最初の10年間（1910～1919）及び最後の10年間（2005～2014）のデータから，格助詞「へ」及びそれと共起する語彙項目をbigramの形式で抽出して分析を行った．出現頻度が5以上の共起項目を用いて，「へ」を頂点としたネットワークを作成した．単語Xと単語Yの共起関係の強さを測る方法としてLLR（Log Likelihood Ratio，対数尤度比）を用いた．

1910～1919年及び2005～2014年という二つの区間の分析結果からみると，格助詞「へ」の出現頻度が減少するにつれ，格助詞「へ」に関わるbigram項目が大幅に減っている．すべてのbigram項目において「名詞＋へ」及び「へ＋動詞」といった共起が多数であるため，その減少は特に顕著に見える．動詞に着目すると，1910～1919年は，「落ちる」「置く」「出す」「入れる」「着く」「かける」など，＜着点＞を表す動詞との共起は多く見られるが，2005～2014年ではほとんど見られない．それに対し，「行く」「戻る」「入る」「出る」「向かう」などのような，＜方向＞を表す動詞は，両方に現れている．現在の言語感覚からみると，＜方向＞を表す動詞と共起する「へ」は，その多くが格助詞「に」で代用することができ，格助詞「に」のほうがより自然に聞こえる場合もある．

③ そのほかの助詞

格助詞「で」「の」「が」に対し，同じく最初の10年間（1910～1919）及び最後の10年間（2005～2014）におけるbigram項目をネットワークで分析した．

二つの区間を比較した結果，格助詞「で」において，場所を表す名詞との共起が増えていることが確認された．「場」が明確に提示されることによって「コト的体験的把握」がよりしやすくなると尾野（2018）は述べている．つまり，より知覚的・感覚的に捉えやすくなり，そのことによって，「共感」の度合いが高まるのである．できごとの起きる場所や動きの空間範囲を明示することによって，読者に物語の内容をより感覚的に捉えさせ，共感を呼びやすくする効果が見込まれる．「で」格の増加には，空間変換による遠近感を現出させたり，読者に現実感覚を与えたりするような作家達の表現意識が働いていると推測される．

一方，格助詞「の」「が」について，ネットワーク分析により，二つとも主に「名詞」の後，「動詞」「形容詞」の前に現れており，この使用傾向は変わっていないが，格助詞「の」では減少し，「が」では増えていることが確認された．『広辞苑』によると，述語の表す内容をもたらした主体について，主文で「が」を使うのに対し従属文では「の」を使うとする考えもあったが，現代語では，従属文でも「が」で表すことが多い(新村, 2018)．今回の結果はこの点を裏付けている．とはいえ，格助詞「が」の増加量は格助詞「の」の減少量を大きく上回ることが見逃せない．

「が」文型の多くは，話し手自身の目に映り心に感じた事柄を直接に述べる現象文である．「が」格の現象文について，森田（1998）は「極めて自己中心的で場面依存型の表現形式」と指摘しているが，逆にいうと読み手を話し手の視点に立たせ，臨場感のある叙述に連れて行く効果がある．格助詞「が」の増加は，話者の心の内や目に映る情景をそのまま綴るという表現が増えていることを示唆している．

格助詞に比べ，同じく特徴的項目である副助詞「なんて」，係助詞「しか」，「さえ」及び終助詞「よ」は出現頻度が相対的に少なく，語彙内における語義差異もそれほど大きくない．これらの項目に対して，使用された文を確認した．本稿では，紙数の関係から全ての解説を記すことができないが，一二例の説明を簡略的に提示する．

係助詞「しか」は使用率が上がる傾向を示している．「しか」は，「花子しかいない」のように否定述語と拘束関係を持っている．この「しか」の拘束関係が出現したのは近世初期以降であることが指摘されている(山口, 1991)．そして，「しか」の類の助詞の成立は，否定表現の変革に大きく関わっており，その本質は日本語述語構造そのものの歴史的変革との指摘もある(宮地, 2007)．「しか」を始めとする否定呼応表現は，既に研究者に注目されていたが，質的，記述的に言及されるにとどまっていた．今回の分析結果は，先行研究の見解の定量的裏付けになると同時に，その意義を改めて認識し，再評価するための契機ともなりうる．

終助詞「よ」も時期の推移に伴って増加している．コーパスを確認したところ，「よ」のみではなく，終助詞の全体使用率は2000年前後に増えてくる傾向を示した．コーパスクリーニングの際，括弧やダッシュなどの記号で標記された会話文を削除したため，使われたコーパスにおける終助詞は，括弧などの記号が付けられない引用文，又は独話や心内発話文などに現れることが多い．「よ」を始めとする終助詞の増加は，内的モノローグの叙法やくだけた言葉遣い，喋り口調という特徴を持つ小説が多くなっていることを示唆している．その原因について深く検討する余地があるが，インターネットの普及に相まって発達してきたブログなど新メディアの表現による影響の可能性が考えられる．

（2） 深層学習 BERT を用いた格助詞「へ」の埋め込みとその言語モデル

さらに言語モデルの構築のため，深層学習による文脈の埋め込み方法による分析を試みた．文脈の埋め込みの最新の方法としては BERT がある．格助詞「へ」に関するデータを抽出して利用する．

コーパスから，格助詞「へ」が含まれる文を抽出し，BERT の学習モデルに適応させ，返された 768 次元の格助詞「へ」の埋め込みベクトルを非線形次元圧縮手法 t-SNE を使って 3 次元に圧縮し，視覚的に考察しながら，分類されたクラスターごとに言語のモデリングを試みている． t-SNE の三次元の動的グラフから，各々のクラスターを構成するポイントを確認すると，それぞれに顕著な特徴があることが分かった．これらに関しては，分析を進めている．

参考文献

樺島 忠夫・寿岳 章子(1965)文体の科学,綜芸社.

安本 美典・本多 正久(1981)因子分析法,培風館.

劉 雪琴, 金 明哲 (2017). 宇野浩二の病気前後の文体変化に関する計量的分析. 計量国語学, 31(2), 1-16. 査読有り

尾野 治彦：『「視点」の違いから見る日英語の表現と文化の比較』，開拓社， 2018.

新村 出（編）：『広辞苑』第7版，岩波書店，2275p.，2018.

森田 良行：『日本語の視点――ことばを創る日本人の発想』第 4 刷，創拓社，54p.，1998.

山口 堯二：「副詞『しか』の源流――その他を否定する表現法の広がり」，日本語語源探求委員会『語源探求 3』，明治書院，pp. 34-48，1991.

宮地 朝子：『日本語助詞シカに関わる構文構造史的研究 ： 文法史構築の一試論』, ひつじ書房, 2007.

期間内の論文

李 広微・金 明哲(2021)． モデリングから見る小説における助詞の経時変化, 情報知識学会．31(3)．査読付き，採択決定済み， ページ数未定

学会発表

(1) 李 広微, 金 明哲 (2020)．構造的トピックモデルによる近現代小説の文体変化の考察．計量国語学会第 64 回大会予稿集，25-30，9 月 19 日オンライン大会.

(2) 李　広微, 金 明哲(2020)．トピックモデルに基づいた現代小説の接続表現の分析．第48回日本行動計量学会抄録集，152-153，9月1日-4日．早稲田大学戸山キャンパス(オンライン).

(3) 李 広微, 金 明哲 (2019)．過去百年間における小説の文体変容についての定量的分析.第47回日本行動計量学会，大阪大学豊中キャンパス，9月3-6日，大阪.

(4) 李 広微, 金明哲 (2018)．現代日本語小説の文体的特徴の変化について-大正・昭和の作品を中心として-．第46回日本行動計量学会抄録集，384-385，9月3-6日，慶應義塾大学，東京.

(5) 李 広微, 金明哲 (2018)．戦前・戦後の日本小説の分類とその特徴分析．日本分類学会シンポジウム，11月24-25日，沖縄県青年会館，沖縄.

(6) G. Li, M. Jin (2020). Diachronic Changes of Sentence-final Expression in Modern Japanese Novels. International conference on Language and Literature. at 2020 International conference on Language and Literature, International Islamic University, Jan.15-17, Malaysia. *refereed

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

〔学会発表〕　計6件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 李 広微，金 明哲
２．発表標題 トピックモデルに基づいた現代小説の接続表現の分析
３．学会等名 第48回日本行動計量学会
４．発表年 2020年

１．発表者名 李 広微，金 明哲
２．発表標題 構造的トピックモデルによる近現代小説の文体変化の考察
３．学会等名 計量国語学会第64回大会
４．発表年 2020年

１．発表者名 Guangwei LI, Mingzhe. JIN
２．発表標題 Diachronic Changes of Sentence-final Expression in Modern Japanese Novels
３．学会等名 International conference on Language and Literature（国際学会）
４．発表年 2020年

１．発表者名 李 広微，金 明哲
２．発表標題 過去百年間における小説の文体変容についての定量的分析
３．学会等名 第47回日本行動計量学会
４．発表年 2019年

１．発表者名
李 広微，金 明哲

２．発表標題
現代日本語小説の文体的特徴の変化について-大正・昭和の作品を中心として-

３．学会等名
第46回日本行動計量学会

４．発表年
2018年

１．発表者名
李 広微，金 明哲

２．発表標題
戦前・戦後の日本小説の分類とその特徴分析

３．学会等名
日本分類学会

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

テキストマイニング2018
https://www1.doshisha.ac.jp/~mjin/lab/TM2018.html

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	山崎　誠 (Yamazaki Makoto) (30182489)	大学共同利用機関法人人間文化研究期機構国立国語研究所・言語変化研究領域・教授 (62618)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関