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　６ 年　 ６ 月　 ３ 日現在

機関番号：１４４０１
研究種目：基盤研究(B)（一般）
研究期間：2018～2021
課題番号：１８Ｈ００６７５

研究課題名（和文）機械学習によるコーパス文体論分析モデルの提示とそれに基づく国際連携基盤の創成

研究課題名（英文）Machine-learning Approaches to Corpus Stylistics: Towards the Creation of International Collaborative Network

研究代表者
田畑　智司（Tabata, Tomoji）

大阪大学・大学院人文学研究科（言語文化学専攻）・教授

研究者番号：１０２４９８７３

交付決定額（研究期間全体）：（直接経費）　　12,100,000円

研究成果の概要（和文）：本研究では，トピックモデリングやワードエンベディングに代表される機械学習に基づく自然言語処理技術を後期近代英語期の小説作品コーパスの分析に適用し，新たなコーパス文体論研究モデルを提案した。それにより，従来のコーパス言語学の手法では統計の網から外れてしまうような，出現頻度の低い多数の単語によって構成されるトピックや語群にも新たな光を当てること，そして従来の手法ではアプローチすることが困難であった「意味の問題」を量的に考察する足掛かりを築いたと言える。この研究から得られた手法や知見をもとに，研究協力関係にある国内外の研究グループと大規模な共同研究ネットワーク構築へ向けた連携基盤形成を図った。

研究成果の学術的意義や社会的意義
この研究プロジェクトは，機械学習に基づく自然言語処理技術と言語学的文体分析を組み合わせた研究モデルを用いた点で特異な学術的意義をもつ。特に，「意味の領域」への量的アプローチの可能性を広げることで，後期近代英語散文の文体を従来とは異なる観点から分析を可能にし，人文学とデジタルの架橋を築いたと言えよう。社会的には，国内に加えて，国際的な研究連携ネットワークの構築により，知見や研究手法の共有，交換を可能にする基盤創成を行ったことは高く評価できると考えている。これにより，分野横断的な共同研究の萌芽形成や，人的交流に寄与し，学術的な洞察と技術的手法の融合から新たな研究領域が開かれる可能性がある。

研究成果の概要（英文）：In this study, we applied machine learning-based natural language processing techniques, notably topic modeling and word embedding, to analyze a corpus of Late Modern English novels, thereby proposing a new model for corpus stylistics research. This approach has illuminated topics and groups of words composed of low-frequency occurrences that traditional corpus linguistics methods might overlook, offering a quantitative foothold for exploring 'issues of meaning' that were previously difficult to approach with conventional methods. Based on the methodologies and insights derived from this research, we have initiated the formation of a collaborative foundation aimed at constructing a large-scale international research network with our academic partners both domestically and abroad.

研究分野：デジタルヒューマニティーズ，英語文体論，コーパス言語学

キーワード：コーパス文体論　国際連携　研究ネットワーク形成　自然言語処理　デジタルヒューマニティーズ　トピックモデリング　ワードエンベディング　言語学的文体論

１．研究開始当初の背景

本研究に先駆けて，筆者は，基盤研究(C)「マイニング技術を応用した著者推定法の開発とディケンズ・ジャーナルの計量文体研究」を実施した。この研究では random forests, Nearest Shrunken Centroid などの統計学的分類器を応用して，Dickens と他の作家，特に Wilkie Collins との共著作品において，文体変化，すなわち「著者の交代が起こったと疑われる箇所」を高い精度で特定する手法を開発し，テクスト中の動的文体変異の相を視覚的に提示する計量文体分析法を確立した。この方法では，特に，執筆時に著者の意識的制御が及びにくい機能語を中心とした高頻度語彙項目を分析変数とすることで, テクストの類同, 相異を高精度で識別することが可能となった。しかし，高頻度語の生起パターン情報だけで説明できる文体的特徴は極めて限定的であり，特にテクストの（潜在的）意味構造やトピック生起パターンを計量的に捉えることは極めて困難な課題として残った。他方，Blei (2012)による「潜在的ディリクレ配分法(Latent Dirichlet Allocation: LDA)」に基づく topic modelling や Mikolov et al. (2013)が提案した word2vec に代表される word embedding 法など，最新の機械学習アルゴリズムに基づく大規模テクストデータ解析法は，従来のコーパス研究が忌避敬遠してきた「意味」の問題にアプローチする可能性を切り開きつつある。

潜在的ディリクレ配分法という，確率分布アルゴリズムに基づく topic modelling は，テクストが複数のトピックからなることを前提として，コーパスの局所限定的に出現する共起語群（すなわちトピック），あるいは逆に，広範囲に遍在して共起する語群を抽出する手法である。Topic modelling は，大規模コーパスに潜在するトピックを構成する語群を数量化して推定するとともに，コーパス内の各テクストにおけるそれらの語群の密度 (topic density)を指標として，テクストのクラスタリングを行うことを可能にする。最近では，topic modelling は，テクストデータだけでなく，遺伝子データ，画像，ソーシャルネットワークなど，いわゆるビッグデータのパターン解析にも応用されその真価を発揮し始めている。

一方，Mikolov らが開発した word2vec の名で実装された word embedding 法は，ニューラルネットワークの原理に基づき，コーパスにおいてある単語が使用された言語的コンテクスト（co-text, word co-occurrence）を再構築するモデルである。Word embedding 法は，コーパス中の個々の単語を，生起パターンが酷似した単語の最近傍に配置して視覚化することにより，コーパス中の単語を意味的関連性によってクラスタリングする手法であると換言できる。

これら最新の大規模テクストデータ解析技術は，従来のコーパス研究で困難とされてきた「意味」の問題に取り組む手掛りを提供するものである。また，これらの機械学習モデルは，コーパス研究のもう一つの死角，「低頻度語の問題」に新たな光を当てるのに寄与する。Topic modelling や word embedding 法は，コーパス中の単語を特定のトピックラベルの下に集約したり，生起傾向が類似する語彙項目をクラスタリングすることで，単独では生起頻度が低いがゆえに，（対数尤度比検定やカイ二乗検定など）従来の統計的特徴語抽出法では捉えることが困難な語彙的特徴を，計量文体分析の射程に含めることを可能にするという利点を備えていることを指摘しておきたい。本研究の重要な着眼点である。

２．研究の目的

本研究の目的は，機械学習による大規模コーパス解析アルゴリズムを応用することにより，申

請者がこれまでに進めてきた多変量文体分析モデル，工学的マイニング技術を用いたテクスト分析モデルを発展させ，潜在的意味構造の検出力と記述の粒度を高めたコーパス文体論の分析モデルを提示することである。特に，進展著しい LDA topic modelling (Blei, 2012)や word embedding 法 (Mikolov et al., 2013)等の機械学習の手法を文体研究に適化させ，これまでのコーパス研究では忌避される傾向にあった「意味」の問題をコーパス文体論の射程に統合する。当研究から得られる知見を随時，国際会議等の機会を活用し，協力関係にある海外の先端的コーパス文体論・計量言語学研究者の研究成果と比較検討するために，コロキアム，ワークショップを積み重ねて連携を強固にする。加えて，毎年，先端的研究者を大阪大学に招聘して，コーパス文体論研究の国際フォーラムを形成し，分析方法論，知見・知識の interoperability（共有，相互利用，相互補完）を可能にする国際連携基盤を創成することである。

３．研究の方法

本研究計画では，topic modeling と word embedding 法を活用した大規模コーパス分析方法論を開発し，18 世紀・19 世紀の代表的作品テクストを収録した後期近代英語フィクションコーパスを横断的に分析する。それにより，特定の作家やテクストと結びついている局所的トピック（共起語群）や特徴語，局所近傍語群（関連語クラスター）を特定する一方，コーパス全体を通して観察可能な遍在的トピックや近傍語群，作品ジャンルと密接に関連したトピックや関連語クラスターを明らかにする。マクロ的観点からは，コーパスに内在する通時的言語変異の相を反映する潜在的トピックや近傍語群を特定し，それらの分布，密度等を数量化したデータをもとに，樹状図やネットワークグラフ，ヒートマップなどの視覚化ツールを駆使して後期近代英語フィクションの系統を可視化する。開発した分析手法，得られた知見を，関連する分析手法やリサーチクエスチョンに取り組む先端的研究者の研究成果と比較検討，共有するための共同研究，ワークショップ，フォーラムを継続的に実施することにより，知見と洞察を相互に，そして相補的に提供し，活用することが可能なコーパス文体論の国際連携ネットワークを構築した。

４．研究成果

本研究では分類器 support vector machine や，潜在的ディリクレ配分法(Latent Dirichlet Allocation: LDA)に基づくトピックモデリング(Blei, 2012)や Mikolov et al. (2013)が提案した word embedding 法など，機械学習に基づく手法を適用することで，研究対象とする Dickens の習作期から円熟期にかけての文体変化や，読者層が大きく拡大した 18 世紀，19 世紀の英国 classic fiction の系譜や影響関係，通時的発展の相にアプローチした。例えば，統計学的分類器 support vector machine を用いた文体の識別研究では，執筆時に著者の意識的制御が及びにくい機能語を中心とした高頻度語彙項目を説明変数とすることで，テクストの類同，相異を高精度で同定することが可能となった。もっとも，高頻度語の生起パターン情報だけで説明できる文体的特徴は限定的であり，それを補完するために，Blei (2012)による「潜在的ディリクレ配分法(Latent Dirichlet Allocation: LDA)」に基づく topic modelling や Mikolov et al. (2013)が提案した word2vec を用いた。これにより，コーパスに内在するトピックやテクスト中で重要な意味を担う語と共起する近傍語群の生起パターンを解析し，Dickens コーパスと比較対象の参照コーパスである英国 18 世紀，19 世紀の作品テクストとのトピック構成や近傍語群の相異に焦点を当てることで，これまで光を当てることが容易ではなかった Dickens の文体の側面を照らし出すこ

とが可能にになった。本研究の成果は，国際文体論学会(Poetics and Linguistics Association)の年次国際会議（PALA2018, 2019, 2023）や英語コーパス学会（JAECS 30 周年記念大会），日本英文学会・中国四国支部大会シンポジウム「デジタル時代の英語英米文学研究と英語教育—デジタル・ヒューマニティーズの有用性と可能性を考える」において発表したほか，デジタルヒューマニティーズの国際シンポジウム Gale Digital Humanities Day (British Library, 2019)や， Taiwanese Association for Digital Humanities DA/DH2019, Bridging Digital Humanities (Western Sydney University, 2022)，韓国比較文学会(POSTEC, 2019),韓国東国大学での Digital Humanities シンポジウム（2019），韓国英語英文学会(ELLAK, Seoul, 2023)等における招待講演等で公開してきた。中でも，2021 年に英国 Nottingham 大学で開催された PALA2021 では招待講演を行い，本研究の成果の一部を国際文体論学会に所属する連携研究者たちと共有したことは特筆に値すると考えている。これらの研究発表とともに，ポーランド・ヤギエウォ大学 (Jan Rybicki)および Polish Academy of Science (Maciej Eder, Joanna Byszuk)，オランダ・アムステルダム大学 (Karina van Dalen-Oskam)，ドイツ・ビュルツブルク大学 (Fotis Jannidis, Steffen Pielström)，同ハイデルベルク大学 (Beatrix Busse)，同トリーア大学 (Christof Schöch)，イタリア・ボローニャ大学 (Monica Turci)，ベルギー・アントワープ大学 (Mike Kestemont, Pieter Fivez)，英国・バーミンガム大学 (Michaela Mahlberg)，同ノッティンガム大学 (Peter Stockwell)，同アストン大学 (Marcello Giovanelli, Stephen Pihlaja)，スウェーデン・ウプサラ大学 (Dan McIntyre)，韓国・東国大学 (Kim Youngmin)，台湾・国立政治大学 (Liu Chao-lin)，米国・ニューヨーク大学 (David Hoover)など，コーパス文体論の先端的研究者との国際連携ネットワークを築くことができた。この人的ネットワークを基盤として本研究の延長線上に位置する研究計画を展開していく予定である。

付録：

Table 1 (http://www.lang.osaka-u.ac.jp/~tabata/tables/50_topics_with_their_keywords.pdf):

本研究課題で構築したコーパス ORCHIDS (Osaka Reference Corpus for HIstrical/Diachronic Stylitics) に LDA トピックモデルを実行した結果得られた 50 点のトピックおよびキーワード群

Figure 1 (http://www.lang.osaka-u.ac.jp/~tabata/images/Fig_1_Network_of_50_topics.pdf):

ネットワークグラフによる ORCHIDS の主要トピックおよびトピックを構成するキーワード群の可視化

Figure 2 (http://www.lang.osaka-u.ac.jp/~tabata/images/Fig_2_association_between_topics_and_texts.pdf):

トピックとテクストとの関連を可視化したネットワークグラフ　コーパス内のトピックの通時的な変化だけでなく，著者のクラスターも示している

Figure 3 (http://www.lang.osaka-u.ac.jp/~tabata/images/Fig_3_heatmap.pdf):

階層クラスター分析を用いたヒートマップダイアグラム　コーパスを構成するテクスト群はトピックの濃度に基づき, 3 つの異なるクラスターに分類されている。ディケンズのクラスターは，他の 19 世紀や 18 世紀の小説と明確に区別されている。19 世紀および 18 世紀のテクストは，Thackeray の *Barry Lyndon* と *Vanity Fair* を除いて区別できる

Appendix (http://www.lang.osaka-u.ac.jp/~tabata/tables/Appendix.pdf):

18 世紀と 19 世紀のイギリスの主要な小説を収録した参照コーパス Osaka Reference Corpus for HIstorical/Diachronic Stylistics コーパスの詳細

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 田畑 智司	４．巻 2019
２．論文標題 英国Classic Fictionコーパスの潜在的トピック ： LDAによるテクストクラスタリング	５．発行年 2020年
３．雑誌名 言語文化共同研究プロジェクト『テクストマイニングとデジタルヒューマニティーズ2019』	６．最初と最後の頁 47～58
掲載論文のDOI（デジタルオブジェクト識別子） 10.18910/76991	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Tomoji Tabata	４．巻 2018
２．論文標題 Mapping Dickens’s Style in the Network of Words, Topics, and Texts	５．発行年 2018年
３．雑誌名 テクストマイニングと デジタルヒューマニティーズ 2017	６．最初と最後の頁 51--60
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 田畑 智司	４．巻 2018
２．論文標題 Digital Humanities: デジタルで拡張する言語文化学研究	５．発行年 2018年
３．雑誌名 テクストマイニングと デジタルヒューマニティーズ 2017	６．最初と最後の頁 61--90
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計18件（うち招待講演　11件／うち国際学会　11件）

１．発表者名 田畑 智司
２．発表標題 確率論的トピックモデリングによるBritish classic fictionの「遠読」（シンポジウム「デジタル時代の英語英米文学研究と英語教育」）
３．学会等名 日本英文学会中国四国支部第74回大会
４．発表年 2022年

１．発表者名
Tomoji Tabata

２．発表標題
Digital Humanities as/and computational Science

３．学会等名
Building Digital Humanities (Western Sydney University)（招待講演）（国際学会）

４．発表年
2022年

１．発表者名
Tomoji Tabata

２．発表標題
Different paths to the same peak: Digital humanities and Spitzerian stylistics

３．学会等名
The Poetics and Linguistics Association International Conference PALA 2021 Nottingham（招待講演）（国際学会）

４．発表年
2021年

１．発表者名
Tomoji Tabata

２．発表標題
Language Action Types and the Semantics of Texts: Using Rhetorical Annotation to Classify Texts into Meaningful Groups

３．学会等名
2020 Korea-Japan Symposium on Digital Humanities（招待講演）（国際学会）

４．発表年
2020年

１．発表者名
田畑 智司

２．発表標題
「ズームイン・ズームアウト―デジタルヒューマニティーズとテクストの「読み」―」

３．学会等名
Galeシンポジウム2020『第2回 デジタル人文学への誘い』（招待講演）

４．発表年
2020年

１．発表者名
Tomoji Tabata

２．発表標題
Digital Humanities as Non-Linear Reading: Style in classic British fiction

３．学会等名
DADH 2019: The Tenth International Conference of Digital Archives and Digital Humanities（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
Tomoji Tabata

２．発表標題
“Zooming in and zooming out”: Digital humanities and the (macro-/micro-) reading of texts

３．学会等名
Digital Humanities Lecture at National Chengchi University, Taipei, Taiwan（招待講演）

４．発表年
2019年

１．発表者名
Tomoji Tabata

２．発表標題
Dickens, Collins and their Collaborations: Pinpointing style change in collaborative texts

３．学会等名
Trans Media World Literature Institute International Colloquium: Transhumanism, Trans Media, World Literature, and Digital Humanities, Dongguk University, Seoul, South Korea（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
Tomoji Tabata

２．発表標題
Experimental Stylometry

３．学会等名
Stylometry workshop Amsterdam at Advanced Study in the Humanities and Social Sciences (NIAS, Amsterdam, the Netherlands)（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
Tomoji Tabata

２．発表標題
Tracing Thematic Transition in Dickens’s Literature and Journalism

３．学会等名
The Poetics and Linguistics Association International Conference PALA 2019 Liverpool（国際学会）

４．発表年
2019年

１．発表者名
Tomoji Tabata

２．発表標題
Reading texts non-linearly: Classic British fiction and Dickens

３．学会等名
Gale Digital Humanities Day at the British Library（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
Tomoji Tabata

２．発表標題
Dickens in Vector Space: Word Embeddings and Semantic Profiling of Style

３．学会等名
Poetics And Linguistics Association (PALA) 2018（国際学会）

４．発表年
2018年

１．発表者名
Tomoji Tabata

２．発表標題
Collaborative Texts under a Stylometric Microscope: Investigating Texts of Mixed Authorship

３．学会等名
英語コーパス学会第44回大会

４．発表年
2018年

１．発表者名
Tomoji Tabata

２．発表標題
Lexical Diversity in Classic British Fiction

３．学会等名
Osaka-Wurzburg Collaborative Workshop: Cross-Linguistics Perspectives on Complexity in Literary Texts（国際学会）

４．発表年
2018年

１．発表者名
田畑 智司

２．発表標題
Stylometry and Classic British Fiction

３．学会等名
日本文体論学会第114回大会（招待講演）

４．発表年
2018年

１．発表者名
Tomoji Tabata

２．発表標題
Corpus approach to semantic style: Body language, n-grams, and topics

３．学会等名
Osaka Symposium on Corpus Stylistics（国際学会）

４．発表年
2019年

１．発表者名
田畑 智司

２．発表標題
デジタルが変える「読み」― テクスト、データ、ディスタントリーディング ―

３．学会等名
Galeシンポジウム2018「デジタル人文学への誘い」（招待講演）

４．発表年
2018年

１．発表者名
田畑 智司

２．発表標題
Word Vectors and Semantic Style in Classic Fiction

３．学会等名
「言語研究と統計2019」

４．発表年
2019年

〔図書〕　計3件

１．著者名 田畑 智司（編）	４．発行年 2020年
２．出版社 大阪大学大学院言語文化研究科	５．総ページ数 96
３．書名 テクストマイニングとデジタルヒューマニティーズ 2020	

１．著者名 田畑 智司（編）	４．発行年 2021年
２．出版社 大阪大学大学院言語文化研究科	５．総ページ数 68
３．書名 テクストマイニングとデジタルヒューマニティーズ 2021	

１．著者名 田畑 智司, 杉山 真央, 土村 成美	４．発行年 2018年
２．出版社 大阪大学大学院言語文化研究科	５．総ページ数 90
３．書名 テクストマインニングとデジタルヒューマニティーズ2017	

〔産業財産権〕

〔その他〕

―

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

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1件

国際研究集会	開催年
Digital Humanities Workshop Osaka 2020	2020年～2020年

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