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

平成　３０　年　　６　月　　８　日現在

機関番号：　１４４０１
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ０２６００

研究課題名（和文）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

研究課題名（英文）Authorship attribution and stylometric analysis of Dickens's journals

研究代表者
田畑　智司（Tabata, Tomoji）

大阪大学・言語文化研究科（言語文化専攻）・准教授

研究者番号：１０２４９８７３

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，様々なマイニング手法を比較検討することにより，検出力の優劣や適性をメタ分析によって考察し，精度の高い計量文体学的著者推定方法論の確立を試みた。特に，共著テクストにおける文体変化に対する分析感度を最適化するために，テクストの分割サイズ，ステップサイズ，分析変数のタイプや項目数を変化させた条件下で，Support Vector Machine，Nearest Shrunken Centroids，Naive Bayesなどデータマイニングで用いられている分類手法をテクストチャンクの著者推定に応用した。本研究で開発した解析法により極めて高い精度で著者交代の箇所の特定が可能となった。

研究成果の概要（英文）：Exhaustive comparative analysis has been carried out on a wide range of data mining techniques to help develop a highly accurate style variation detector on texts of mixed authorship. The development of the analytical methods draws heavily on machine-learning approaches in an effort to identify subtle stylistic shifts or variations in texts. The stylometric authorship attribution methods studied in this research have achieved a high degree of accuracy, making it possible to pinpoint where one author takes over from the other in texts of mixed authorship.

研究分野：　コーパス文体論，デジタルヒューマニティーズ

キーワード：　Dickens　style markers　authorial takeover　data mining　machine learning　stylometry　author ship attribution　mixed authorship

1．研究開始当初の背景

本研究代表者がこれまでの研究から得た重要な知見の一つは，19 世紀中葉から 20 世紀へ向かう小説言語の諸特徴の先駆が，まずディケンズの英語に見出されるということである。また，ディケンズの作品群と 19 世紀の代表的作品群は，マイニング技術の応用により極めて高い精度で識別が可能な一方で,（18 世紀や 20 世紀の英語を含めて比較を行うと）ディケンズと他の 19 世紀作品に共通する，すなわち 19 世紀を特徴づける言語使用のパターンを抽出することができる。

他方，ディケンズが編集した週刊ジャーナル *Household Words* (1850‐9)と *All the Year Round* (1859‐95)は当時の英国において大量生産型の出版ビジネスの原型を作り上げ，読書人口の拡大に寄与するなど，大きな社会文化的なインパクトを与えたことは知られているが，19 世紀の英語書き言葉の発達にどのような影響を及ぼしているかについては，これまで十分な学術的関心が向けられてこなかった(Drew, 2003)。少なくとも量的観点からディケンズのジャーナルの言語を解析する取り組みはこれまでまだなされていなかった。

そこで本研究では，工学的テクストマイニングの技法とコーパス文体論の分析方法論を相補的に組み合わせ，精度の高い著者推定法やテクストの分類・識別方法論を確立することにより，ディケンズ・ジャーナルの文体に迫る取り組みを行った。

2．研究の目的

本研究は，最新の工学的データマイニング技術を高度に応用した計量文体分析モデルを確立することにより，ディケンズの言語を特徴付ける文体指標の抽出を行うとともに，精度の高い著者推定法につなげることを目指すものである。特に，非伝統的著者推定法やテクストの分類・識別方法論に関する国内外の知見を批判的に組み入れることにより，文体分析に最適化した著者推定アルゴリズムを開発し，ディケンズのジャーナルの計量文体分析に応用する。

本研究において，最もチャレンジングな課題は，ディケンズが他の作家と共著したテクストにどの程度ディケンズの貢献がなされているかを推定することである。著者推定法を精緻化して応用することで，従来定説と見なされてきた知見を検証・再検討する。また，共著テクストにおいて具体的にどの箇所で著者の交代が起こっているのかを高精度で推定するとともに，共著者の貢献の度合いを視覚的に明らかにすることによって，Stone, Thomas, Nayder (2002), Drew, Allingham (2011)らの先行研究を補完し，ディケンズ・ジャーナルの文体研究に新地平をもたらすことを目指す。

3．研究の方法

当研究計画は次の(1)から(4)へのフローで構成されている。

(1) 一次資料としてのディケンズ週刊ジャーナルコーパスの整備，テクスト処理実験の試行

(2) 統計学的文体分析アルゴリズムの研究およびR による分析スクリプトの開発

(3) さまざまな著者推定法，マイニング法によるデータ解析・視覚化，解析結果の比較検討

(4) 最適化した分析アルゴリズムによるコーパス分析結果の言語文化学的考察，有効性の検証

(1) まず,ディケンズの作品 28 点 484 万語，同時代作家ウィルキー・コリンズの作品 24 点 346 万語, 18 世紀の代表作品 23 点 416 万語，19 世紀の主要作品 31 点 512 万語を収録した大規模コーパス(総語数 1,758 万語)の編纂, 校訂を行い，これらを基にテクスト処理実験を試行する。

(2) 研究代表者がこれまでに活用してきた主成分分析，対応分析，Support Vector Machine, Random forests 等のツールに加えて，著者推定における識別力に優れていることが示された手法 Burrows's *Delta*, Craig's *Zeta*（Craig & Kinney eds., 2009; Hoover, 2010）のアルゴリズムを研究し，統計解析処理言語環境 R で分析スクリプトを開発する。

(3) 工学的テクストマイニングで評価の高い Naïve Bayes Classifier や，Nearest Shrunken Centroids 法（Jockers & Witten, 2010）等の分類器の他，知識抽出に有効な機械学習法のアルゴリズムを比較検討する。さまざまな手法の長所・短所を調査研究し，当課題の研究対象となるデータに適合した解析処理モデルを決める。

(4) ディケンズ･ジャーナルに掲載された作品のうち，特にコリンズとの共著作品に対して上記で取り上げた文体分析モデルを適用し，共著テクストの動的文体変化を捉える。

4．研究成果

以下では，語彙の生起パターンを統計解析することにより，ビクトリア朝の作家ディケンズとウィルキー・コリンズの作品を識別する指標語彙を抽出し，それをもとに二人の共著作品の特徴や他の作品との関係や位置づけを考察する。語彙頻度という internal evidence と，信頼度の高い書誌学的情報，伝記的記述などの external evidence を比較考察し，文体統計学的手法の妥当性を吟味するとともに，二人の共著作品への貢献度の推定を試みる。本研究では，まず，Random Forests を用いてディケンズとコリンズの文体識別に有効な語彙項目を

抽出する。そして，Random Forests によって抽出した作者判別マーカーを変数として，多変量解析を実行し，共著作品とディケンズおよびコリンズそれぞれの作品との関係を視覚化する。文体統計学的分析の結果からは二人の作家の不均衡な関係が見て取れる。

ディケンズは，主宰する雑誌 *Household Words* と *All the Year Round* の Christmas Numbers (クリスマス特集号)向けに，コリンズ他の‘staff writers’の手を借りて数多くの短編を著している。それらの作品の多くは，ディケンズが，自身の着想を基に，書き手にプロットやキャスト，描写の視点を細かく「指揮」(“conducted by Charles Dickens”)して執筆させた草稿に加筆修正を施して仕上げたものである。*Mugby Junction* と *No Thoroughfare* (1867) を除き，そうした短編作品は匿名で発表されているが，そのうち，*Household Words* の Christmas Numbers に掲載された作品については，会計帳簿の記録などからほぼ著者の特定が可能である (Thomas, 1982: 140)。Thomas (1982: 140–152)は会計帳簿および Stone (1968) を参考に，Christmas Numbers への寄稿者一覧を提示している。

以下，Stone (1968), Thamas (1982), Nayder (2002)によって明らかにされている書誌学的情報をもとに，ディケンズとコリンズ二人の協力によって著されたことが判明している作品 4 点は表 1 の通りである。

表 1　ディケンズ，コリンズの共著テクスト

No.	Date	Title	Part	Authorship
1	1857, 1866, 1874	*The Frozen Deep*		Collins & Dickens ⇒ Collins
2	1857	*The Lazy Tour of Two Idle Apprentices*	Chapter I	**Dickens & Collins**
			Chapter II	**Dickens & Collins**
			Chapter III	**Dickens & Collins**
			Chapter IV	Dickens
			Chapter V	**Dickens & Collins**
3	1857	*The Perils of Certain English Prisoners*	Chapter I	Dickens
			Chapter II	Collins
			Chapter III	Dickens
4	1867	*No Thoroughfare*	Overture	Dickens
			Act I	**Dickens & Collins**
			Act II	Collins
			Act III	Dickens
			Act IV	**Dickens & Collins**

表 1 に挙げた作品のうち，*The Frozen Deep* と *No Thoroughfare* はもともと劇の脚本として書かれたものである。*The Frozen Deep* は，当時国民的関心を集めた Sir John Franklin 率いる 1845 年の南極探検隊遭難事件が契機となって生まれた作品である。消息が途絶えた探検隊の発見は叶わなかったものの，捜索の結果，先住民族から Franklin 一行が全滅したことを示す証拠品を入手し，食糧枯渇のため人肉食が行われた可能性を示唆した Rae 博士の報告に対して，探検隊は先住民に襲われ命を落としたと主張するディケンズが *Household Words* 誌上で Rae 博士と議論を戦わせた。*The Frozen Deep* はこれを発端として着想を得た劇である。この作品は，ディケンズがプロットおよびキャストを考案し，コリンズに行わせた下書きに大幅な改訂を施して 1857 年に上演された後，1866 年と 1874 年に(ディケンズとの間に距離のできた)コリンズが改訂した版があるほかに，短編小説に書き改められたものがあるなど，複数の異本が存在する (Nayder, 2002: 10)。

ト書きを除いて，ほぼ登場人物の dialogue だけで作品が構成される戯曲は，事物・情景・動作・心理などの描写を含む narrative と登場人物の dialogue を巧みに組み合わせた小説との言語的な差異が大きい。使用域間の言語的差異は往々にして作家間の文体差よりも大きくなることがあるため，分析対象に劇と小説を混在させてしまうと，テクスト間の差異が作家の文体差によるものか，使用域間の言語変異に基づくものかの判別が困難になることが予想される。そこで本研究では，*The Frozen Deep* については小説化されたものを用いた。

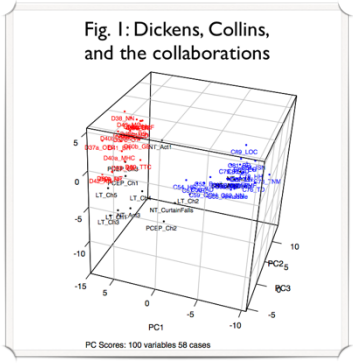
No Thoroughfare についても，*The Frozen Deep* 同様小説版を分析対象とした。この作品では Act II, IV をコリンズが，Act III をディケンズが，それぞれ単独で執筆しているが，Act I と The Curtain Falls は二人の共同執筆である (Thomas, 1982: 152)。

The Perils of Certain English Prisoners は，二人の分担章が明確である。Thomas (1982: 146) および Allingham (2011)によれば，Chapter I, III をディケンズが，Chapter II をコリンズが単独で執筆している。

The Lazy Tour of Two Idle Apprentices では，ディケンズが単独で執筆している Chapter IV を除き，他の章はディケンズとコリンズの共同執筆による (Nayder, 2002: 106)。

主成分分析によるテクストの空間布置

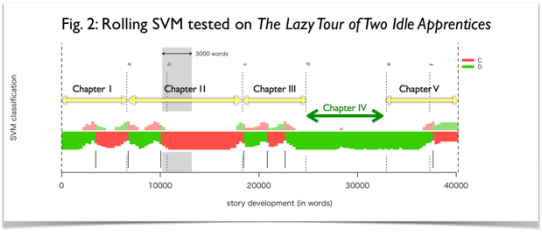
ディケンズの単著作品テクスト，コリンズの単著作品テクストと共に二者の共著作品テクストの関係性を主成分分析によって視覚化したものが Fig. 1 である。
ディケンズの単著作品，コリンズの単著作品は明確に分離されたクラスターを構成している一方，共著作品は両者の中間的な位置取りをしていることがわかる。



マイニングの応用による著者交代箇所推定

以下では，共著章を含んでいる *The Lazy Tour of Two Idle Apprentices* と *No Thoroughfare* を取り上げて動的文体変化の相を探る。

ディケンズの単著作品，コリンズの単著作品をトレーニングデータとして，共著作品の冒頭部から 3,000 語のウィンドウを 300 語ずつ後ろにずらし，テクストの最後までを走査する。ウィンドウごとに，高頻度語上位 100 語の相対頻度をもとに，Support Vector Machine によって分類推定を行い，当該ウィンドウがディケンズによって書かれた確率，コリンズによって書かれた確率を求め，全ての結果を視覚化したのが Fig. 2 と Fig. 3 である。



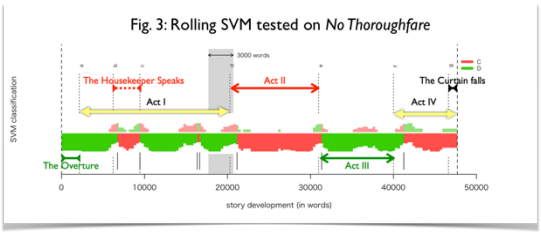


Fig. 2, Fig. 3 いずれも章の区切りはグラフに垂直に記した点線で示されている。書誌学的証拠よりすでに単著であることが明らかとなっている章は全てディケンズかコリンズを表す一色で表示されており，他方共著章では帯の色が途中で変わっている。そこが著者交代が起こった箇所であると推定される。交代が起こった箇所の前後では 3,000 語のウィンドウ内に両者のテクストが異なる割合で混合することになるため帯の幅が部分的に短くなり，その分が反対側に表示される。著者交代が起こることにより，テクストウィンドウの語彙生起パターンが変化し，その変化を Support Vector Machine がほぼ正確に捉えている。

共著章の著者の担当パターンはほぼ一貫していることが見て取れる。いずれの共著章でも，まずディケンズが筆を取り，3 分の 1〜半分程度進んだところで，コリンズに交代する様が語彙生起パターンの変化から観察できる。

本研究の結果は，計量文体分析モデルが共著作品の動的文体変化を検出するのに有効であることを示すものである。

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. Tomoji Tabata, Mapping Dickens's Style in the Network of Words, Topics, and Texts, *Practical Stylometry*: *Genres*, *Topics*, *and Key Words* (The Institute of Statistical Mathematics Cooperative Research Report 405), pp. 75–84, 2018.
2. 田畑 智司「修辞的特徴のマイニング: Dickens と 18–19 世紀英国小説の文体」『英語コーパス研究』（英語コーパス学会）pp. 101–122, 2017 年.
3. 田畑 智司「FLOB コーパスの意味構造 : 確率論的トピックモデルによる言語使用域の特徴付け」言語文化共同研究プロジェクト 2016『テクストマイニングとデジタルヒューマニティーズ』（大阪大学大学院言語文化研究科 言語文化共同研究プロジェクト 2016 成果報告書）pp. 5–22, 2017 年.

〔学会発表〕（計 20 件）

1. Tomoji Tabata, Mapping Classic Fiction in Networks: Key words, Topics, and Distant Reading, ポスター発表 第 2 回大阪大学豊中地区 研究交流会「文×理『知』の融合」2018 年 1 月 10 日大阪大学・南部陽一郎ホール.
2. Tomoji Tabata, "Birds of a feather flock together": Literary Vocabulary in Vector Space, ポスター発表 第 2 回大阪大学豊中地区 研究交流会「文×理『知』の融合」2018 年 1 月 10 日大阪大学・南部陽一郎ホール.
3. Tomoji Tabata, Through the Different Routes to the Same Landscape: What Do the Text-Clusterings Tell Us about Style? DARIAH-EU International Expert Workshop on Distant Reading in Literary Texts, 23–24 November 2017, The University of Würzburg, Germany. http://dariah-tda.github.io/meeting/activity/workshop/2017/11/23/Workshop-Distant-Reading.html（招待講演）
4. Tomoji Tabata, Applying Topic Models to Describe the Composition of the FLOB Corpus: Can the external criteria be associated with meaningful sets of internal evidence? JAECS 43rd Annual Conference（英語コーパス学会第 43 回大会）2017 年 9 月 30 日 関西学院大学.
5. Tomoji Tabata, Mapping Dickens's Novels in the Network of Words, Topics, and Texts: Topic-Modelling of a Corpus of Classic Fiction, JADH 2017 (The Japanese Association for Digital Humanities 7th International Conference), 11–12 September 2017, Doshisha University.
6. Tomoji Tabata, Topic Modelling Dickens's Fiction, PALA 2017 (Poetics And Linguistics Association Annual Conference), 19–22 July 2017, West Chester University of Pennsylvania, USA.
7. 田畑 智司, The Semantic Universe of Classic Fiction「言語研究と統計 2017」2017 年 3 月 27–8 日 大学等共同利用法人 情報・システム研究機構 統計数理研究所.
8. Tomoji Tabata, Sequencing a Literary Genome of Classic Fiction: When the Humanities Meets Digital ポスター発表 大阪大学豊中地区 研究交流会「文×理『知』の融合」2016 年 12 月 20 日大阪大学・大阪大学会館.

9. 田畑 智司「デジタルが拡張するテクスト分析と文体論 : Stylometry の現在」『英学シンポジウム―文体論で極める文学とコミュニケーション―』2016 年 11 月 19 日兵庫県立大学・姫路キャンパス（招待講演）

10. Tomoji Tabata, 'Rhetorical annotation and its application in text analytics', Osaka Workshop on Corpora and Language Analytics: The State-of-Art and Future Perspectives, 6 August 2016, Graduate School of Language and Culture, Osaka University.

11. Tomoji Tabata, 'Devising Corpus Queries to Explore Textual Features', The 57th Summer Seminar of the English Research Association of Hiroshima, 7–8 August 2016, Hiroshima University Kasumi Campus

12. Tomoji Tabata, 'Rolling stylometry and Dickens's collaboration with Collins', PALA 2016 (Poetics And Linguistics Association Annual Conference), 27–30 July 2016, University of Cagliari, Italy.

13. Tomoji Tabata, 'Experimental Stylistics: A Meta-analysis to Evaluate Rolling Stylometry', The Digital Literary Stylistics Workshop, Digital Humanities 2016, 12 July 2016, Jagiellonian University, Krakow, Poland.

14. Tomoji Tabata, 'Pitfalls in key word analysis: corpus queries and approaches to style, International Workshop on Cognitive Grammar and Usage-Based Linguistics, 18–19 June 2016, Graduate School of Language and Culture, Osaka University.

15. 田畑 智司「修辞項目のアノテーションを活用したテクスト分析」英語コーパス学会第 42 回大会シンポジウム『コーパスアノテーション（タグ付け）の功績と課題』2016 年 10 月 1–2 日 成城大学

16. 田畑 智司「Body-part expressions を通して見る fiction の言語」「統計数理研究所言語系共同研究班 2016 年度夏季合同研究発表会」2016 年 8 月 29–30 日 神戸大学

17. 田畑 智司 「機能カテゴリーに基づく計量文体研究」「統計数理研究所言語系共同利用研究班合同研究会」2015 年 9 月 27‐8 日 大阪大学

18. 田畑 智司 'Digital Humanities, Distant Reading, and Dickens's Style' 「デジタル・ヒューマニティーズと英文学の未来」2015 年 9 月 23 日 東京女子大学（招待講演）

19. Tomoji Tabata, 'Mining Rhetorical Features of Charles Dickens: A study in rhetorical profiling of style', JADH2015: The 5th International Conference of the Japanese Association for Digital Humanities, 1–3 September 2015, Kyoto University. (Long paper)

20. 田畑 智司 'Mapping Dickens's Fiction: Distant reading & text analysis'（シンポジウム「Dickens の言語と文体」のパネリストとして）「ディケンズフェローシップ日本支部 2015 年春季大会」2015 年 6 月 13 日 関西外国語大学

〔図書〕（計 2 件）

1. 田畑 智司「共著作品における Dickens の文体」堀 正広 編『コーパスと英語文体』（ひつじ書房）pp. 53–71, 2016.

2. Tomoji Tabata, 'Stylometry of Dickens's Language: An Experiment with Random Forests', in P. L Arthur and K. Bode (eds.) *Advancing Digital Humanities*: *Research, Methods, Theories*, Palgrave Macmillan: 28–53, 2015.

6. 研究組織

(1)研究代表者
田畑 智司（Tabata, Tomoji）
大阪大学・大学院言語文化研究科・言語文化専攻・言語情報科学講座・准教授
研究者番号：10249873