科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１２　日現在

機関番号：　１４４０１
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５００２９８

研究課題名（和文）工学的テクストマイニング技術を応用した１９世紀英語の計量文体研究

研究課題名（英文）Stylometric investigation of 19th-century English through text-mining

研究代表者
田畑　智司（Tabata, Tomoji）

大阪大学・言語文化研究科（研究院）・准教授

研究者番号：１０２４９８７３

交付決定額（研究期間全体）：（直接経費）　　4,000,000円、（間接経費）　　1,200,000円

研究成果の概要（和文）：本研究では近年急速に進展している工学的なデータマイニングの技術を応用し，従来の英語史・英語文体論研究において焦点を当てられることが少なかった言語項目を分析した。19世紀の英語散文は従来18世紀英語の延長線上にあり，かつ20世紀の英語へ通底する特徴が見られることは確かであるが，マイニング技法を用いて19世紀英語の特徴を抽出すると，散文ジャンルの中でも特に小説の言語において18世紀から大きな進化が見られることが明らかとなった。中でも，身体部位の描写によって感情を表現したり，人物の特徴付けを行なう表現法は，この世紀に大きく多様性をもつようになったことを客観的なデータで示すことができた。

研究成果の概要（英文）：This grant-in-aid study has tried applying the state-of-art data-minig tools in stylistic investigation of nineteenth-century English prose. The linguistic features analysed in this research range from very common function words through mid- to lower-frequency items. While nineteenth-centurry English shares, in principle, similar linguistic features with eighteenth-century English as well as English in the early twentieth-century, it has proved to be possible to clearly differentiate nineteenth-century texts from texts written in the eighteenth as well as in the early twentieth centuries using machine learning techniques. Major distinguishing variables extracted though machine-learning mining algorithms include body-part languages, which function as descriptors of emotion or as device for characterisation. Of further interest is that Charles Dickens triggered the significant development of such stylistic devices in the nineteenth century.

研究分野：　情報学

科研費の分科・細目：　図書館情報学・人文社会情報学

キーワード：　19世紀英語　文体　マイニング　stylometry

１．研究開始当初の背景

欧州・北米を中心に近年急速な展開をしている digital humanities (人文情報学, 以下「DH」と表記)において, 統計学を応用したテクストの計量研究は中核的な位置を占める。計量的文体研究の歴史は DH の諸分野の中で最も長く，デスクトップコンピュータの出現より遥か以前，単語の語長分布がテクストの書き手を識別する手だてになると考えた Mendenhall (1887) までさかのぼることができる。以後，統計学と言語学の進歩に伴い，著者推定 (Ellegård, 1962; Mosteller & Wallace, 1964; Burrows, 1989), 文体模写(McKenna & Antonia, 1994, 1996, etc.)，刑事事件供述の真贋判定 (Svartvik, 1968; Coulthard, 2000)，聖書の成立過程に関する諸説の検証(Linmans, 1998; Mealand, 1999; Miyake *et al*., 2005) などの事例に統計学が応用され，テクストの識別，分類・類型化による問題解決や考察に貢献してきた。

筆者がこれまで科学研究費補助金の助成を受けて，精緻化を進めてきた多変量解析を応用した文体分析モデルによる近代英語の文体研究もこの潮流に属する。しかし，筆者の関心はテクストの識別，分類・類型化にとどまらず，変数とする語彙項目間の相互関係を言語文化学的に解釈し，その背後に通底する文体論的原理を説明することに，より大きなウェイトを置いてきた。このアプローチに基づく Dickens の文体研究や，近代英語のコロケーション(語の習慣的結合関係)の研究に関して，最近の 5 年ほどは毎年，発表枠をめぐる競争率が高い国際会議 Digital Humanities (2007–2012)や PALA (2008–2013) で成果発表を行い，高い評価を受けている。

他方，Web 2.0 の登場を起爆剤として爆発的な量的拡大をみせる電子テクストデータの普及は，テクストデータへの工学的関心を劇的に高め，データマイニング，テクストマイニング方法論の発展に繋がっている。中でも，自然言語処理技術を応用し，機械学習を用いたクラス分類，回帰分析，クラスタリングに基づく知識獲得，マイニング方法論の進展は著しく，大規模なテクスト集（コーパス）の文体分析に大きな威力を発揮し得るものである。工学的テクストマイニングと，DH の計量的文体研究は，テクストという対象を共有している。しかし，残念なことに，これまで双方の学際的交流は意外にも少なく，それぞれが蓄積してきた知見を共有し，相互に役立てる試みが充分になされていなかった。

２．研究の目的

当研究では，筆者が研究対象とする 19 世紀英語を収録するコーパスの分析に，筆者が進めてきた多変量文体分析モデル，および DH の分野で定評のある他の計量分析手法に，工学的なテクストマイニング技法を組み合わせることにより，計量的文体研究の方法論を補強し，19 世紀英語の文体的特徴を説得力の高い客観的な形で提示することを目的とした。

19 世紀は英語散文の発達史上重要な時代区分である。産業革命に続く中産階級の勃興，教育の充実やジャーナリズムの発展に伴い，Dickens に代表される小説や新聞，雑誌の読書人口が著しく増大し，書き言葉としての英語の多様性が広がった時代であるからである。その一方で, 19 世紀の英語は「見かけ上，現代英語に近い」(Kytö, Rydén, et al., 2009)ため，その共時的特徴，通時変化はまだ十分に研究されていない ‘an unexplored territory’ と言われる(Kytö, Rudanko, and Smitterberg, 2000)。この時代をカバーする大規模かつバランスの取れたコーパスがまだ少ないなど，研究基盤が十分に整っていなかったこともその理由の一つである。

そこで，本研究の目的は，これまでの科研費助成研究による知見の蓄積，King’s College London や統計数理研究所との共同（利用）研究の成果を踏まえ，多変量文体分析モデルと工学的テクストマイニング技法を相補的に組み合わせ，大規模コーパス解析に最適な計量文体分析方法論を開発することである。

３．研究の方法

当研究目的を達成するため，次のような 4 段階の遂行計画によって当研究を進めた。

(1) 一次資料としての 19 世紀英語コーパス(+対照コーパス)の整備, テクスト処理実験試行
(2) 統計学的文体分析アルゴリズムの研究および R による分析器プロトタイプ開発
(3) さまざまな分析手法によるデータ解析・視覚化，解析結果の比較検討
(4) 最適化した分析法による 19 世紀英語コーパス，Dickens コーパス分析結果の言語文化学的考察，有効性の検証

工学的テクストマイニングでは, *the, and, I, of,*…等，機能語を中心とする高頻度語は，一見，知識獲得に貢献しないと判断され，‘stop words’として分析から除外されることが多い。ところが，筆者の研究では，多変量解析による視覚化によって，これらの語彙の生起パターンもテクストの重要な語彙的・統語的特徴や談話特徴と密接に関連しているということがわかってきた。このようにコンピュータと統計学的解析法を駆使して初めて視覚化し得る言語事実を信頼度の高い客観的な形で提示するために，当研究では可能な限り多くの言語項目を分析変数として，機械学習の手法を応用して言語項目とテクスト間の複雑な相互関係を視覚化する方法を採用した。

４．研究成果

Random Forests (Breiman, 2001)は，ensemble learing（集合学習）による回帰・分類ツールであり，金・村上 (2007)，小林・田中・冨浦 (2011) などでテクスト分類に用いられ，高い

分類精度を誇る手法であるが，分類に加えてデータ分類に貢献度の高い変数を出力することができる。本研究では，研究対象の 18 世紀，19 世紀コーパスに生起する語彙項目のうち，Random Forests にかける語彙変数の数を上位 5,000 項目から上位 100 項目まで変化させて実験を行った。

上位 1,000 の語彙項目は高頻度で生起する機能語から中頻度の内容語などの語彙層をカバーしており，計量的観点から文体の記述を行なう上でバランスの取れた変数であることが判明した。上位 1,000 項目の語彙を分析変数として Random Forests を実行した結果，18 世紀のテクストと 19 世紀のテクストは平均 98.72%の精度で分類でき，19 世紀，18 世紀それぞれに一貫して共通する特徴，また両者を一貫して識別する言語特徴が存在することを物語っている。下記の 3 次元散布図（図 1）は，Random Forests 実行時に生成される近接距離行列をもとに多次元尺度構成法でデータ間の関係性を視覚化した結果である。グラフの左側に 19 世紀のテクスト群，右側に 19 世紀のテクスト群がそれぞれクラスターを形成していることが判る。

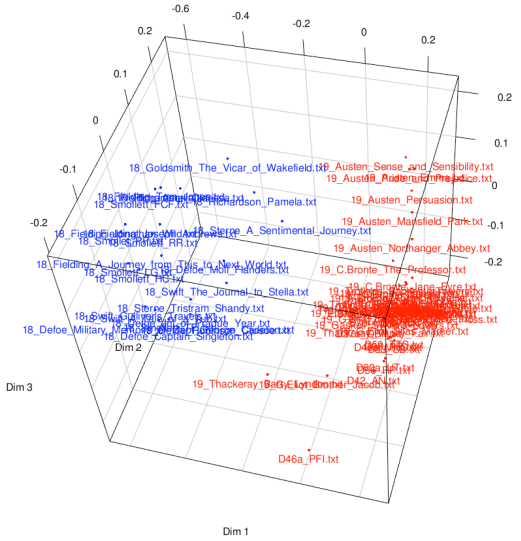

図 1　Random Forests による分析結果：3 次元空間に布置された 18, 19 世紀のテクスト群

表 1 に挙げる語彙項目は 18 世紀と 19 世紀のテクスト群における頻度が有意に異なる項目のうち，Random Forests で ‘important variables’として上位に列挙され，かつ，ウィルコクスンの *U* 検定の結果 *U* の効果量が 0.5 以上のものである。18 世紀のテクストを特徴付ける項目には *proper, resolved, order, justice, reason, liberty, virtue, service, advice* など騎士道精神に由来する価値観，道徳，規範を表す語が含まれる一方，*desire, danger* などそれに対立する概念を象徴する語が抽出されている。

他方，19 世紀のテクストを特徴付ける項目としては，*looking, looked, speak, spoke, spoken, speaking, said, talking, exclaimed, walked, walk, walking, turned* に代表される動作を表す語が多く含まれている。また，*feeling, feelings, look, face, smile, arm, hair, expression, eyes, mind, lips* など感情，表情，ならび身体の部位に関する語がランクインしている。KWIC コンコーダンスを用いて，これらの語の具体的な用法を調査すると，特に 19 世紀フィクションのテクストでは，身体部位を表す語は往々にして感情やその変化を描写する表現装置の一部を構成していることが判る。こうした表現法は 18 世紀には皆無であったわけではないが，Dickens の登場を機に 19 世紀のフィクションの言語において大きく活用の幅が広がった表現形式であると言えるだろう。

表 1　18 世紀，19 世紀テクストの識別語

Strongest 18th-Century markers:
ordered, this, part, desired, account, proper, resolved, hath, order, justice, reason, several, occasion, liberty, art, to-day, farther, desire, our, therefore, king, whole, least, virtue, condition, sent, nor, greatest, pleased, which, thus, neither, gave, thousand, danger, use, service, advice, carry, according, body, other
Strongest 19th-Century markers:
anxious, feeling, quite, nearly, look, looking, looked, waiting, quiet, feelings, slowly, feel, oh, tried, felt, round, pleasant, didn't, chance, couldn't, bright, thinking, it's, yes, wouldn't, spoken, tone, speaking, course, minute, voice, ah, position, suddenly, change, property, sitting, something, window, on, silent, talking, wrong, everybody, across, face, I've, don't, walked, moment, alone, you're, idea, over, smile, walk, arm, breakfast, back, room, hair, turned, expression, fact, question, likely, show, sat, home, once, like, again, early, old, everything, ma'am, anybody, right, hall, he's, I'm, tea, beautiful, getting, heavy, standing, through, dark, table, seat, still, strong, always, eyes, reply, walking, corner, else., speak, wonder, away, mind, beyond, there, sit, anything, door, opened, ask, been, proud, lips, silence, behind, never, red, chair, reached, outside, Mr, shook, clear, evening, its, to-morrow, try, when, liked, bit, rose, that's, spoke, exactly, down, there's, exclaimed, said, does, low

その他，19 世紀のフィクションテクストでは会話部の占める割合が前世紀のテクストに比して格段に大きくなっていることは，19 世紀の特徴語に，会話部の導入や描写に使用される reporting verbs や，話し言葉を写し出す表現装置として使用されている短縮形などが数多く含まれることにも反映している。

修辞項目を変数に設定した分析の結果

上記は分析変数を単語とした場合の結果をまとめたものである。他方，より表現法，修辞的カテゴリーの分布を手がかりにテクストの特徴付を行なうことが必要である。そのため，本研究では米国カーネギーメロン大

学で開発された修辞プロファイリングを行なうソフトウェア，DocuScope を活用してコーパス中の全テクストに情報注釈を施した。修辞情報の注釈が埋め込まれたテクストから，98 項目に分類された修辞特徴の生起頻度を求め，Random Forests その他の分類器による解析実験を行った。

図 2 は図 1 と同様に Random Forests 実行時に生成される近接距離行列をもとに多次元尺度構成法でデータ間の関係性を視覚化した結果である。テクストの分類精度は 100%であった。単語レベルの分析以上に高い識別が可能であったことは特に注目に値する。

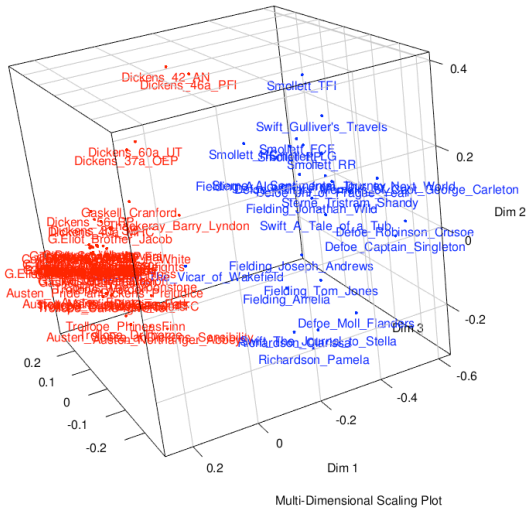

図 2　98 項目の修辞情報を基に Random Forests を実行した結果

図 3 は 18,19 世紀のテクストの識別に貢献度が高かった修辞項目を示している。

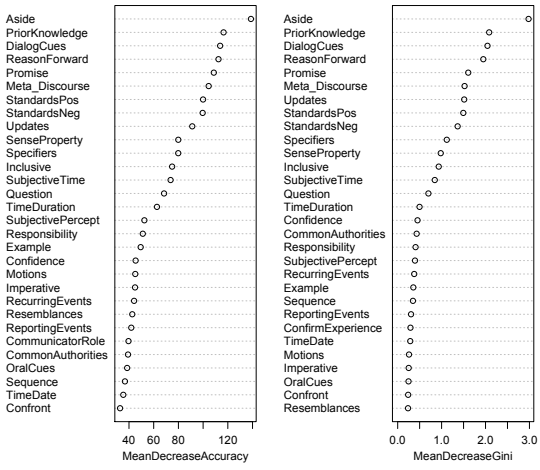

図 3　Important variables detected through Random Forests

本研究は，語彙項目ならびに修辞情報をもとに，マイニングの手法を応用してマクロ的，計量的な観点から 18,19 世紀のテクストの特徴を記述した。本研究で得られた成果は今後さらにミクロな視点で分析を深めるための重要な手がかりとして参照することができるであろう。

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　4　件）

田畑　智司「テクストマイニングからテクスト分析へ：Collins との共著作品における Dickens の文体」『電子化言語資料分析研究 2011-2012』大阪大学大学院言語文化研究科, 3–17.

Tomoji Tabata, ‘Approaching Dickens’s Style through Random Forests’, Digital Humanities 2012 Conference Proceedings, University of Hamburg, Germany, and The Alliance of Digital Humanities Organizations, 388–91.

田畑　智司「‘Key’ Words and Stylistic ‘Signatures’: Textometry with Random Forests」統計数理研究所共同研究リポート 264『マイニング技術を応用したテクスト分析研究』大学共同利用機関法人 情報・システム研究機構 統計数理研究所, 45–64.

Tomoji Tabata, Stylometry of Collaboration: Pinpointing style changes in the text of mixed authorship,『統計解析言語 R を活用したデジタルヒューマニティーズ研究』大学共同利用機関法人 情報・システム研究機構 統計数理研究所, 37–46

〔学会発表〕（計　15　件）

Tomoji Tabata, ‘Using random forests to identify Dickensian style’, Language Individuation: A symposium in honour of John Burrows, 4-8 July 2011, University of Newcastle, NSW, Australia.

Tomoji Tabata, ‘Investigating Dickensian Style through Random Forests’, Middle and Modern English Corpus Linguistics Conference 2011 (MMECL 2011), 平成 23 年 8 月 26－29 日, 大阪大学中之島センター.

Tomoji Tabata, Harold Short, Gerhard Brey, Maki Miyake, Yuichiro Kobayashi, José Miguel Monteiro Vieira, Matteo Romanello, ‘Statistical text-mining on English Woman’s Journal’, Osaka Symposium on Digital Humanities 2011, 12-14 September, Graduate School of Language and Culture, Osaka University.

田畑　智司「テクストマイニングからテクスト分析へ：Wilkie Collins との共著作品における Charles Dickens の文体を計る」『言語研究と統計 2012』, 平成 24 年 3 月 7 日, 統計数理研究所

Tomoji Tabata, 'Detecting Stylistic Differences in Collaborative Writings: Random Forests + Burrows' Delta on Dickens, Collins and their co-authored texts', Australasian Digital Humanities Conference 2012, 27–30 March 2012, Australian National University, ACT, Australia. (Long paper)

Tomoji Tabata, 'Approaching Dickens's Style through Random Forests', Digital Humanities 2012: International Conference of the Alliance of Digital Humanities Organizations, 16–22 July 2012, University of Hamburg, Germany. (Long paper)

Tomoji Tabata, 'Digital Enhancements to the Dickens Lexicon', The Bicentennial International Dickens Fellowship Conference, 9–14 August 2012, University of Portsmouth, UK. (Panel session)

Tomoji Tabata, 'Text-mining Linguistic Variations from a Diachronic Perspective: An experiment in textometry', JADH 2012 Conference "Inheriting Humanities", 15–17 September 2012, University of Tokyo, Japan. (Long paper)

田畑 智司「Key words and textometry: Are key words really "key" words?」『計量的言語研究の諸相』, 2012 年 9 月 19 日 北海道大学大学院メディア・コミュニケーション研究院

Tomoji Tabata, 'The State of Digital Humanities in Japan: Its history, development, and future perspective', International Conference of Digital Archives and Digital Humanities, National Taiwan University, Taipei, 29–30 November 2012.（招待講演）

田畑 智司 'Too many suspects, too much burstiness: A meta-analysis of key-word detection statistics for stylometry'「言語研究と統計 2013」2013 年 3 月 7–8 日 統計数理研究所

田畑 智司「テクストマイニングからテクスト分析へ」英語コーパス学会シンポジウム『私のコーパス利用』2013 年 4 月 27 日 大阪大学

Tomoji Tabata, 'Stylometry of Collaborations: Dickens, Collins and their collaborations', PALA 2013: International Conference of the Poetics and Linguistics Association, 31 July–4 August 2013, University of Heidelberg, Germany. (Long paper)

田畑 智司 'Burrows's Delta and the stylometry of collaborations'「統計数理研究所言語系共同利用研究班合同研究会」2013 年 9 月 9–10 日 西南学院大学

Tomoji Tabata, 'Opening up a New Perspective for Text Analysis in the Digital Age', Humanities Studies in the Digital Age and the Role of Buddhist Studies, 16–17 November 2013, University of Tokyo.（招待講演）

〔図書〕（計 3 件）
堀 正広 編『これからのコロケーション研究』（共著）「第 4 章 文体とコロケーション（田畑 智司）」ひつじ書房 （2012 年），107−152.

田畑 智司・岸江 信介（編）『言語研究のためのテキストマイニング』ひつじ書房（2014 年）

Bode, K. and Arthur, P. L. (eds.) *Advancing Digital Humanities*, Palgrave, 2014 (in press).（共著） (Chapter 3 Tomoji Tabata, Stylometry of Dickens's Language)

〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

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・准教授
研究者番号：10249873

(2)研究分担者
（ ）

研究者番号：

(3)連携研究者
（ ）

研究者番号：