Random Forests を用いた男装作家 Alice Bradley Sheldon の計量文体分析

木村 美紀（明治大学大学院 文学研究科）

本研究では，正体不明・性別不明作家として著作活動を行っていた Alice Bradley Sheldon (1915-1987: 米国) の作品群を，この作家と同時代・同ジャンルで活躍していた女性作家作品群のコーパスを使用しながら，Random Forests という統計手法を用いて計量文体分析を試みた. Random Forests でテキスト分類を試みた後，この統計手法を利用して外れ値プロット, variable importance プロット等を描画した．ジニ係数の平均減少率に基づく分類に有効であった指標の提示を行い，言語指標に注目した後，先行研究との相違に関して検証した.

Employing Random Forests in Text Classification of Three Female Writers: Alice Bradley Sheldon, Octavia Butler, and Ursula K. Le Guin

Miki Kimura (Graduate School of Arts and Letters, Meiji University)

This study performs a quantitative text classification analysis on the works of Alice Bradley Sheldon (1915-1987), an American writer of feminist science fiction who primarily used the pen name James Tiptree, Jr., to disguise her gender. In this study, I performed a quantitative stylistic analysis of Sheldon's work by employing a supervised statistical method, Random Forests, comparing it with that of Octavia Butler and Ursula K. Le Guin, female science fiction writers whose career overlapped with Sheldon's.

1. はじめに

本研究では, Alice Bradley Sheldon (1915-1987: 米国) という作家の文体を，金・村上 (2007) において他の分類手法に比べて分類感度がよいと結論付けられている Random Forests (RF) という統計手法を用いて計量的に検証する．分類正確率の算出後, RF の近接性の出力に基づいて多次元尺度法 (Multi-Dimensional Scaling: MDS) プロットを図示する. MDS プロットの図示によって，誤分類された作品群の同定を行う．これらの提示によって，今まで文学研究の枠組みでしか研究されてこなかった男性偽装作家 Alice Bradley Sheldon の文体分析を進める.

2. 先行研究

2.1 文学研究における評価

Alice Bradley Sheldon は James Tiptree, Jr.と Raccoona Sheldon という男女 2 名義を使用しながら正体不明・性別不明の作家として約 20 年間著作活動を行っていた作家である．主に短編の SF 小説を執筆していた作家であり，文芸批評において Alice Bradley Sheldon の男性名義である James Tiptree, Jr. 名義作品群の文体は Ernest Hemingway 作品と比較されることが多く，この作家の文体はしばしば「男性的」と評価される.

Alice Bradley Sheldon の James Tiptree, Jr. 名義作品群に関する有名な文芸批評として Silverberg [1] が挙げられる. Silverberg [1] では, "It has been suggested that Tiptree is female, a theory that I find absurd, for there is to me something ineluctably masculine about Tiptree's writing. I don't think the novels of Jane Austen could have been written by a man nor the stories of Ernest Hemingway by a woman, and in the same way I believe the author of the James Tiptree stories is male." と, Ernest Hemingway に言及しながら「James Tiptree, Jr. 作品群の著者は男性である」と結論付けている．さらに, Silverberg [1] は "Hemingway was a deeper and trickier writer than he pretended to be; so too with Tiptree, who conceals behind an aw-shucks artlessness an astonishing skill for shaping scenes and misdirecting readers into unexpected abysses of experience. And there is, too, that prevailing masculinity about both of them" というように, Hemingway と比較しながら作品の類似性，男性性の共有ということに言及している.

また, 小谷 [2] では, 「ジェイムズ・ティプトリー・ジュニアなる作家は, (中略) その華麗な文体，ヘミングウェイを思わせるマッチョな作風で一躍 SF 界を魅了した. 時代はフェミ

ニズム SF 華やかなりし頃, そのなかでこの著者不明 = 正体不明の作品は, その作風から, 手堅い稀有の才能を持つ男性新人作家の書いたものと判断されていた」といったように, 特に James Tiptree, Jr. 名義の作品の作風・文体に関して主観的な評価を行っている.

2.2 計量文体分析

この作家に対する計量文体分析の先行研究としては木村 [3] が挙げられる. ここでは, Burrows [4] に基づき, 高頻度語彙上位 50 語を指標として Alice Bradley Sheldon 全 72 作品 (865,802 語) と Ernest Hemingway69 作品 (271,475 語) の分類を試みた. Random Forests を用いて分析を行った結果, 分類正確率は 92.20%であり, 文芸批評で主張されているようなこの 2 著者間の文体の類似は確認することができなった.

また, 木村 [3] では, Random Forests の近接性に基づいて外れ値プロットを図示した. その結果, Alice Bradley Sheldon 作品群の中でもデビュー直後の作品が Ernest Hemingway に誤分類されているということが判明した. これは, 小谷 [5] に基づいて Alice Bradley Sheldon 作品群の通時的文体変化を検証した木村 [6] の結果の一部と合致する. さらに, 木村 [3] では, 部分従属プロット (partial dependency plot) を用いて, 目的変数である高頻度語彙が Alice Sheldon 作品群と Ernest Hemingway 作品群の分類にどのように作用しているのかということを提示した. これによって, Burrows [4] で行われている主成分分析を用いたテキスト分類と都の分類に寄与している変数の提示を, より高度な統計手法を用いて追試可能になった.

3. 分析と結果

3.1 データ

木村 [3] で行った分析は, ジャンルの違いや執筆年代の相違など交絡因子が数多くある文体比較である. これらの問題を解決するため, 本研究では, Alice Bradley Sheldon と同性, 同ジャンル, 同時代に活躍した作家のコーパスを構築し交絡因子を排除した計量文体分析を行う. Alice Sheldon の作品群と女性作家との比較, 文体分析を行うことによって, 文芸批評で主張されているような Alice Sheldon の文体の男性性・独自性の存在の有無を検証していく. 本研究では, 論文執筆者自らが紙媒体から構築した Alice Bradley Sheldon コーパス (72 作品, 延べ 865,802 語) に加えて, Ursula K. Le Guin コーパス (45 作品, 延べ 589,481 語), と Octavia, E. Butler コーパス (93 作品, 延べ 867,396 語) を構築した. 指標としては Burrows and Hassal [7] や Rybicki [8] に基づいて, 実行の容易さとその実績から高頻度語彙上位 50 語を採用して分析を行った. 高頻度語彙を指標として用いることによって, 複数のテキストに遍在してる言語指標でのテキスト分類を行い, 文体指標の一般化を試みた. 本研究では目的変数のレマ化は行っていない.

3.2 結果と解釈

表 1 に RF の出力を表 1 として示す. RF での分類正確率は 91.90%であり, サンプルサイズの違いから算出した分類正確率の基準 44.29%を大幅に上回っており, これら 3 著者間の分類に成功していると結論付けられる.

クラスごとの分類正確率を確認する. Octavia Butler 作品群は, 全ての作品が正しく分類されている. 次に Le Guin 作品群の分類結果を検討すると, 45 作品中 32 作品が正しく分類されていることがわかる. このクラスでの分類正確率は 71.11%だった. 誤分類された 13 作品群のうち, 10 作品は Alice Sheldon 作品群へと誤分類されており, 3 作品は Octavia Butler へと分類されている. ここから, Le Guin の文体は, Alice Sheldon の文体に類似している可能性が存在しているといえる. さらに, Alice Sheldon 作品群の分類正確率は, 94.44%だった. 誤分類されている 4 作品中 3 作品は Le Guin と判断されており, ここからも Alice Sheldon と Le Guin 作品群の文体の類似性があるといえる.

表 1. RF 出力

	Butler	**Le Guin**	**Sheldon**
Butler	**93**	0	0
Le Guin	3	**32**	10
Sheldon	1	3	**68**

次に, MDS プロットを用いて RF の結果を図 1 において視覚化していく. ここでは, 誤分類された作品群にのみ作品の ID を表示している. 丸印が Butler 作品群を示し, 三角印が Le Guin

作品群を, 十字印が Sheldon 作品群を示すように描画の際の引数指定を行っている. このプロットから, Sheldon 作品群は x 軸で負の方向, y 軸で正の方向に位置している. 同様に, Le Guin 作品群は, x 軸で負の方向, y 軸方向でも負の方向に位置しているということがわかる. 最後に, Butler 作品群は, x 軸で正の方向, y 軸では 0.0 付近に存在している.

また, 誤分類された作品群を MDS プロットで確認していく. まずは, Ursula K. Le Guin の作品群において Octavia Butler 作品群へと誤分類されてしまった作品群を検証していく. 誤分類された作品群は, ID6 (Ether OR), ID39 (The Silence of the Asonu), ID45 (Very Far Away from Anywhere Els) であるということが同定可能になった. 次に, Le Guin の作品群で Alice Sheldon に誤分類されてしまった作品群は, ID3 (Brothers and Sisters), ID9 (Half Past Four), ID15 (Nine Lives), ID24 (The Left Hand of Darkness), ID27 (The Ascent of the North Face), ID28 (The Author of the Acacia Seeds), ID33 (The Lost Children), ID37 (The Rule of Names), ID40 (The Water Is Wide), ID42 (The Wife's Story) であるということが判明した. 特に, Le Guin の代表作であると言える ID24 (The Left Hand of Darkness) が Alice Sheldon へと誤判別されていることが興味深い.

最後に, Alice Sheldon 作品群における誤分類された作品群の同定を行う. ID103 (A Source of Innocent Merriment) という作品が Butler 作品群へと分類されてしまったということが判明した. また, ID49 (The Boy Who Waterskied to Forever), ID100 (Press Until the Bleeding Stops), ID106 (Amberjack) という 3 作品が Le Guin 作品群へと誤分類されてしまっていると同定可能になった.

次に, 木村 [3] や下川・杉本・後藤 [9] に基づいて RF の近接性を利用し, クラスの中心からはずれている標本の同定を図 2 において行った. ID1~45 までが Le Guin 作品群, ID46~119 までが Alice Sheldon 作品群, ID120~212 までが Butler 作品群を示している.

まずは, Butler 作品群中においてクラスの中心からはずれている作品を同定していく. ここでは, ID205 (Speech Sounds), ID146 (Parable of the Sower 1), ID170 (Parable of the Sower 9) という 3 作品が Octavia Butler 作品群の中でクラスの中心からはずれて独自の地位を占めている作品群であるということが判明した. これらの作品は, 他クラスへと誤判別されないまでも Butler の典型的な作品群とはかけ離れた特徴を有しているということが判明した.

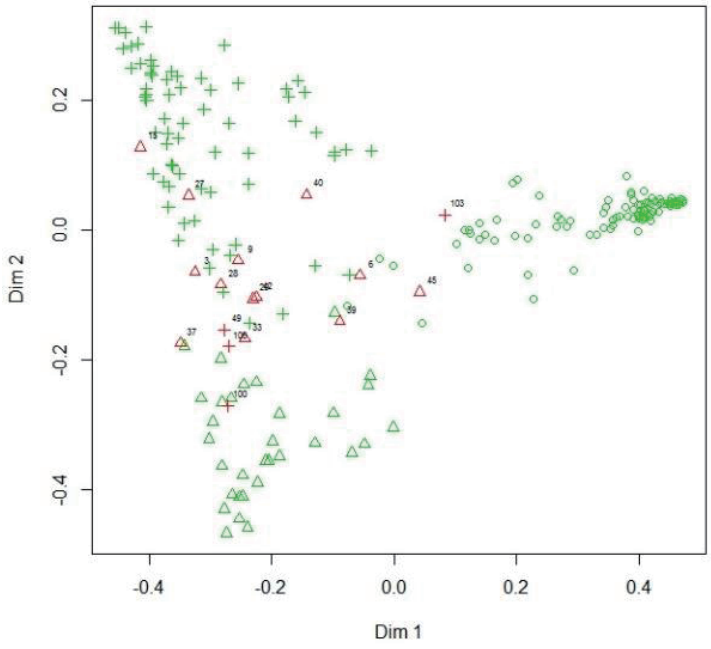

Fig. 1 MDS plot による誤分類作品群の提示

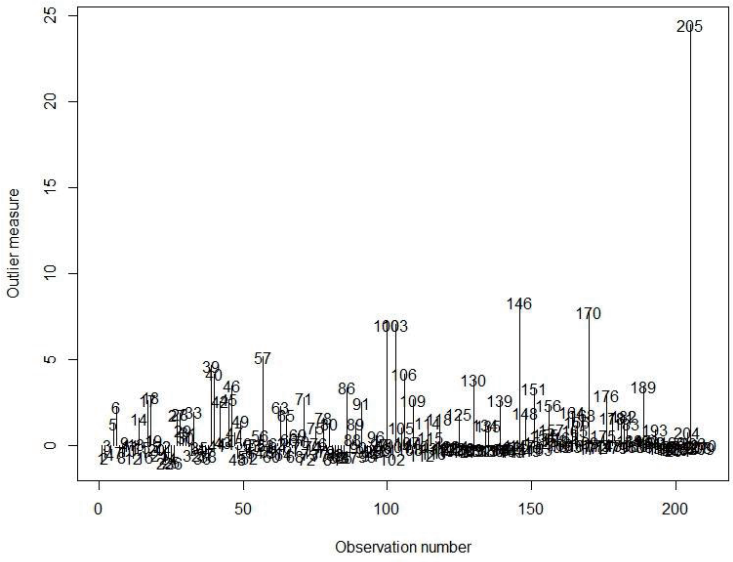

Fig. 2 外れ値プロット

次に, Alice Sheldon 作品群中においてクラスの中心からはずれている作品を同定していく. ここでは, ID100 (Press Until the Bleeding Stops), ID103 (A Source of Innocent Merriment), ID106 (Amberjack) という 3 作品が Alice Sheldon 作品群の中では特にクラスの中心からはずれている作品群であるということが判明した. これらの作品は, MDS プロットで算出した他クラスへと誤判別された作品群と一致している. さらに, ID57 (The Night-Blooming Saurian), ID46 (Second Going), ID86 (I'll be Waiting for You When the Swimming Pool is Empty) という作品群が, 他クラスへと誤判別されないまでも Butler の典型的な作品群とはかけ離れた特徴を有しているということが判明した.

最後に, Ursula K. Le Guin 作品群中においてクラスの中心からはずれている作品を同定していく. ここでは, ID100 (Press Until the Bleeding Stops), ID103 (A Source of Innocent Merriment), ID106 (Amberjack) という 3 作品が Alice Sheldon 作品群の中では特にクラスの中心からはずれている作品群であるということが判明した. これらの作品は, MDS プロットで算出した他クラスへと誤判別された作品群と一致している. さらに, ID57 (The Night-Blooming Saurian), ID46 (Second Going), ID86 (I'll be Waiting for You When the Swimming Pool is Empty) という作品群が, 他クラスへと誤判別されないまでも Butler の典型的な作品群とはかけ離れた特徴を有しているということが判明した.

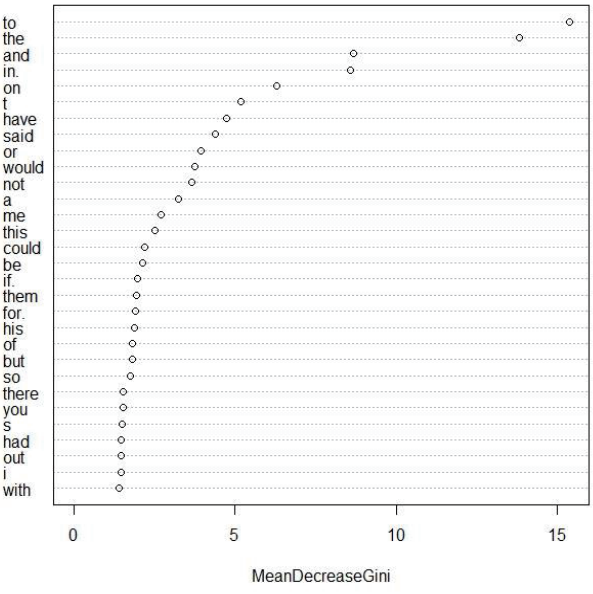

Fig. 3 Variable importance plot

誤分類された作品群の同定だけではなく, 分類に有効であった指標の提示を行い, 今回用いた高頻度語彙 50 語のうちどのような指標が「作家の文体」を示しているのか検証する. 本研究では, ジニ係数の平均減少率に基づいて, 図 3 と図 4 において分類に有効であった指標の提示を行った.

図 3 では, 今回指標として採用した高頻度語彙上位 50 語のうち, 分類に寄与している変数の上位 30 語を提示している. 分類に有効であった変数は具体的には, to, the, and in, on, t, have, said, or, would, not, a, me this could, be if, them, for, his, of, but, so, there, you, s, had, out, i, with であるということが判明した.

図 3 のみでは, 分類に寄与している変数がどの群に特徴的に出現しているのかは同定不可能である. 分類に寄与している変数の頻出する群を特定するために, 図 4 において箱ひげ図を図示した. 描画スペースの都合上, ここでは分類に寄与している変数の上位 10 語を用いて箱ひげ図を描画した.

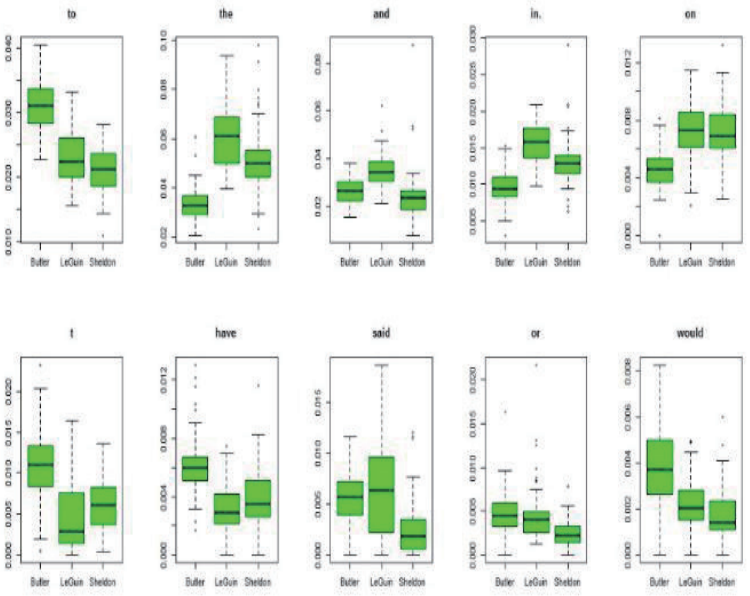

Fig. 4 箱ひげ図による分類に有効であった指標の使用率の提示

具体的には, 一番分類に寄与している程度の大きい変数である to は Butler 作品群に頻出し, Sheldon 作品群ではあまり使用されないということが判明した. また, 次に分類に寄与している変数である the は Le Guin 作品群に頻

出し, Butler 作品群ではあまり使用されないということが判明した. ここから, Le Guin 作品群における名詞類の使用率の高さが推定される. また, Butler 作品群における to という言語特徴の使用率の高さから, 動詞類の使用率の高さが類推される. また, Butler 作品群における否定辞の縮約形である t の使用率の高さや, 図 3 において否定辞の縮約形でない not が分類に寄与している変数であると判明した点から, 否定辞の縮約形とそうでない形での使い分けが存在し, このような言語特徴から作家の「文体」ということが探索的に定義できると考えられる.

さらに, Le Guin 作品群では said という変数が特徴的であるということが分かる. ここから, Le Guin の作品群において, 登場人物の発話の比率の高さが推測される. その一方で, Alice Sheldon の作品群においては, said という変数の過少使用が顕著である. ここから, Sheldon の作品群において, 登場人物の発話の比率の低さが推測される. また, Butler 作品群では would という変数が頻出しており, Sheldon 作品群においては, この言語特徴の過少使用が特徴的である. Butler 作品群では, would という助動詞を過剰に使用することによって Mitigating Strategies を確立していると考えられるのかもしれない.

Butler 作品群おける助動詞の過剰使用は, Kimura [10] で示した 6 著者の作品群における品詞使用率の提示においても確認されている. 図 5 では, Kimura [10] で検証した結果を示す.

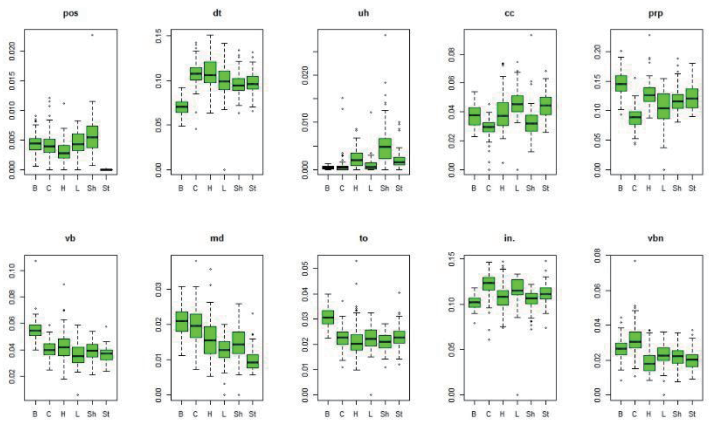

Fig. 5 箱ひげ図による分類に有効であった指標の使用率の提示 (Kimura [10])

Kimura [10] では, Alice Sheldon と同時代・同ジャンルで活躍していた女性作家 Octavia Butler, Ursula K. Le Guin 作品群に加えて, 同時代・同ジャンルで活躍していた男性作家 Arthur C. Clarke, Theodore Sturgeon 作品群と, 文学研究上比較されることの多い Ernest Hemingway を比較対象として, 指標としては品詞分布を採用し分析を行った. 分類に寄与している変数の上位 10 種を箱ひげ図として図 5 に示した. その結果, 高頻度語彙を指標として用いて分析を行った結果と品詞分布を指標として用いた検証において類似性が確認された.

具体的には, Butler 作品群は他の 5 著者作品群と比較して to という言語特徴の過剰使用が「文体」として挙げられる. また, Butler 作品群では, md (modal) という指標が過剰使用されている. これは, 高頻度語彙の分析で提示した, would という指標の過剰使用という現象を反映していると考えられる. また, これらの言語特徴に付随して, vb (verb, base form) という動詞の原形の使用率の高さが判明した. これは, 高頻度語彙という指標を使用した分析からの推測を根拠づける結果である.

また, Le Guin 作品群の言語特徴について検証していくと, 図 5 では cc (coordinating conjunction) という言語特徴の過剰使用が顕著である. 高頻度語彙での分析では, 図 4 に示すように, and や or という coordinating conjunction が特徴的である. ここから, Le Guin 作品群では, coordinating conjunction が多用されるような一文の長さの長い表現が多用されることが多いと推測できる. 今後の研究では, 語だけではなく, 文レベルでの検討が必要になると考えられる. このようにして, lexical な指標と syntactic な指標を同時に検証することによって, 各作家の「文体」を定義できる.

最後に, Alice Sheldon の作品群に特徴的な言語指標に関して検証を行う. 図 5 では, Sheldon 作品郡中では uh (interjection) という言語特徴の過剰使用が顕著である. interjection の過剰使用という事実から, 作品中の narrative の形式に関して他の作家作品群と比較した際に独自の表現を用いていると推測できる. このような特徴から探索的に「文体」を定義できる可能性が存在している. また, Sheldon 作品群は他の 5 著者作品群と比較して to という言語特徴の過少使用が「文体」として挙げられる. このようにして, lexical な指標と syntactic な指標を同時に検証することによって, 作家の文体とい

う問題に関してより精密な検討ができるようになると考えられる.

4. 結論

本研究では, その文体の男性性が文芸批評の面から論じられてきた作家作品群の文体に対して, 高頻度語彙上位 50 語という指標を用いながら計量文体論の手法を用いて分析した. 先行研究における分析は交絡因子が数多く存在した. 本研究では, Alice Sheldon 作品群を彼女と親交のあった同時代, 同ジャンル, 同性の作家 Octavia Butler, Ursula K. Le Guin 作品群と比較することで交絡因子を排除した文体比較を試みた.

その結果, RF での分類正確率は 91.90%であり, サンプルサイズの違いから算出した分類正確率の基準 44.29%を大幅に上回っており, これら 3 著者間の分類に成功していると結論付けられる. また, RF の近接性に基づいて描画した MDS plot によって, これら 3 著者作品群の群間距離を視覚化した. これによって, 他クラスへと誤分類された作品群の同定が可能になった. また, ジニ係数の平均減少率に基づく分類に寄与している変数の提示を 2 種類の方法で行った. 特に箱ひげ図から, 各作家の文体を探索的に定義できるような言語指標の同定が可能になった.

今後は, Rybicki [8] や Jockers and Mimno [11] などで行われているような, 大規模コーパスを用いた著者の性別を基準としたテキスト分類を試みて, Alice Bradley Sheldon という長年男性であると思われていた男装作家の文体がどのように分類されるのかということを確認する.

参考文献

1) Silverberg, R.: Who Is Tiptree, What Is He?, *Warm Worlds and Otherwise*, pp. iv - x viii (1975).

2) 小谷真理 : 女性状無意識, 勁草書房 (1994).

3) 木村美紀: ランダムフォレストを用いた文芸作品の計量的分類と変数の特定の試みー Alice Bradley Sheldon と Ernest Hemingwayー, 英語コーパス研究, No. 24, pp. 41-54 (2017).

4) Burrows, J. F.: *Computation into Criticism: A study of Jane Austen's novels and an experiment in method*. Oxford: Clarendon Press (1987).

5) 小谷真理: 狭間の視線: メアリ・ヘイスティングス・ブラッドリー&ジェイムズ・ティプトリー・ジュニア母娘に見る passing の政治学, アメリカ研究, No. 33 (1999).

6) 木村美紀: Alice Bradley Sheldon 著者内変異と分類法の評価, 文学研究論集, No. 45, pp. 1-18 (2016).

7) Burrows, J. F., & Hassal, A. J.: Anna Boleyn and the authenticity of Fielding's feminine narratives, *Eighteenth Century Studies*, Vol. 21, pp. 427-453 (1988).

8) Rybicki, J.: Vive la différence: Tracing the (authorial) gender signal by multivariate analysis of word frequencies, *Digital Scholarship in the Humanities*. Vol. 31, No.4, pp. 746-761 (2015).

9) 下村敏雄・杉本知之・後藤昌司: 樹木構造近接法, 共立出版 (2013).

10) Kimura, M.: Quantitative Stylometry of Alice Bradley Sheldon: How Well Did Her Pseudonyms Hide Her True Identity?, In *Proceedings of the 8th International Conference of Digital Archives and Digital Humanities 2017*, pp.??-?? (2017).

11) Jockers, M. L., & Mimno, D.: Significant themes in 19th-century literature. *Poetics, 41*(6), pp. 750-769 (2013).