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　５ 年　 ６ 月　 ９ 日現在

機関番号：１２６１１
研究種目：若手研究
研究期間：2019～2022
課題番号：１９Ｋ２０６２７

研究課題名（和文）『源氏物語』の計量的な文体研究

研究課題名（英文）Statistical Analysis of Storyline Structure of "The Tale of Genji"

研究代表者
土山　玄（Tsuchiyama, Gen）

お茶の水女子大学・文理融合 AI・データサイエンスセンター・准教授

研究者番号：００７５５３９０

交付決定額（研究期間全体）：（直接経費）　　1,900,000円

研究成果の概要（和文）：『源氏物語』は平安時代に成立した日本を代表する古典文学作品である。本研究では『源氏物語』のテキストデータを対象とし、統計的な手法を用いて計量的な観点から研究を行った。まず、本研究では『源氏物語』における作者問題の解決のために、単語の出現率などについて分析を行った。その結果、『源氏物語』の最終13巻と他の巻との間に相違は認められず、『源氏物語』は一人の作者によって書かれた可能性が高いと結論づけることができる。次いで、単語の共起状況を分析した結果、一部の連語が『源氏物語』の終盤の巻に多く出現することが分かった。すなわち、ここに『源氏物語』54巻における文体的特徴の出現傾向の変化が認められる。

研究成果の学術的意義や社会的意義
近年の欧米では人文学領域の研究対象について、自然科学の手法、すなわち統計学や情報学の手法を用いるデジタル・ヒューマニティーズは独立した学問領域として急発展を遂げている。その一方で、日本の古典文学作品を対象としたデータサイエンスの手法を用いた研究は十分に展開しているとは言えない。本研究もデジタル・ヒューマニティーズの研究の1つとして位置づけられることから、テキストアナリティクスなどのデジタル・ヒューマニティーズに加えて、国文学や国語学などの人文学と言った幅広い学問領域の発展に寄与できる可能性を有することが本研究の意義である。

研究成果の概要（英文）："The Tale of Genji" is one of Japanese classical literary works established in the Heian period (794-1185). In this study, the text data of "The Tale of Genji" was the subject of research from a quantitative perspective using statistical methods. First, in order to solve the problem of authorship in "The Tale of Genji," we analyzed the relative frequency of words in the text. As a result, no differences were found between the last 13 volumes of "The Tale of Genji" and the other volumes, and it can be concluded that "The Tale of Genji" was most likely written by a single author. Next, we analyzed the co-occurrence of words, and found that some of the collocations appear more frequently in the last volume of "The Tale of Genji". In other words, a change in the tendency of stylistic features to appear in the 54 volumes of "The Tale of Genji" can be recognized here.

研究分野：計量文献学

キーワード：源氏物語　計量文献学　テキストアナリティクス　文化情報学　デジタル・ヒューマニティーズ

１．研究開始当初の背景

本研究では研究対象として『源氏物語』を採り上げた。『源氏物語』は日本を代表する古典文学作品であり、平安時代の代表的な女流作家である紫式部（930-1014 年頃）によって執筆されたとされる平安朝を舞台とした主人公である光源氏の恋物語を描いた長編物語である。また、『源氏物語』は鑑賞の対象としてだけではなく、古くから研究対象でもあった。20 世紀末に『源氏物語』のテキストデータは作成されたが[1][2]、このテキストデータを対象とした統計的な手法を用いた計量的な研究は、現状において十分に展開されているとは言えない。そこで、本研究では『源氏物語』のテキストデータを使用し、計量的な観点から『源氏物語』各巻の文体を分析することで『源氏物語』の作者問題及び成立過程について検討を加える。なお、本研究における文体とは、著者の文章表現における形式的、習慣的な特徴を意味する。

『源氏物語』の作者問題とは、作者が単独であるか複数であるかという問題である。『源氏物語』は一般的に三部構成であると考えられており[3]、第 42 巻「匂宮」以降の 13 巻は光源氏没後の物語である。これら 13 巻は第三部と称されており、他の 41 巻とは作者が異なるという他作者説が以前から提起されている。第三部のうち第 42 巻「匂宮」、第 43 巻「紅梅」、第 44 巻「竹河」の 3 巻は匂宮三帖、第 45 巻「橋姫」以降の 10 巻は宇治十帖と称され、匂宮三帖および宇治十帖のどちらにおいてもそれぞれ他作者説が提起されている。匂宮三帖については、「竹河」巻末の官位昇進に記述が宇治十帖と矛盾することから「匂宮」「紅梅」「竹河」の 3 巻は別人の作であるという見解がある[4]。一方、宇治十帖は古くは一条兼良（1402-1481）によって著された『花鳥余情』に、宇治十帖を除く巻が紫式部の作であり、宇治十帖は紫式部の娘である大弐三位の作であると論じられている。

次に、『源氏物語』には現行の巻の配列と成立順序は相違するという見解がある[5]。『源氏物語』の成立過程についての考察は第一部と称される第 1 巻から始まる 33 巻に集中しており、登場人物の出現状況の調査に基づく客観的なデータから、第一部には「紫上系」と称される 17 巻と「玉鬘系」と称される 16 巻の 2 系統が混在しているという説が論じられている。これは、初出が「紫上系」である登場人物は「玉鬘系」においても登場するが、初出が「玉鬘系」である人物は「紫上系」に登場しないという事実に基づくものである。第二部においても、「玉鬘系」の人物が登場するのは第 34 巻「若菜上」、第 35 巻「若菜下」、第 36 巻「柏木」の 3 巻のみであると報告されている。また、第三部においては第 45 巻「橋姫」、第 49 巻「宿木」、第 53 巻「手習」の 3 巻の冒頭は共通して「そのころ」という発語によって開始されることから、第三部は構造上 4 つのブロックに分類されるという可能性が指摘されている[6]。

２．研究の目的

計量的な手法が有効であると考えられる『源氏物語』の研究課題は上述の作者問題および成立過程についての問題である。そこで、本研究では、データサイエンスの手法を用い『源氏物語』のテキストデータを分析し、文体的特徴の出現傾向について計量的な観点から分析を加えることである。これに加えて、これらの分析を通じて、『源氏物語』の作者問題及び成立過程の問題について計量的な議論に耐えられる透明性の高い分析結果および資料を提示することが本研究の目的である。

３．研究の方法

本研究では、(1) 『源氏物語』とその補作の文体を比較および検討、(2)『源氏物語』の文体的特徴の出現傾向に基づいて各巻の関係性を明らかにする、という段階を踏んで分析を行った。『源氏物語』の補作として本研究では、『山路の露』及び『雲隠六帖』のテキストデータを使用した。これら 2 作品は『源氏物語』と世界観を共有しているが作者が別人であると考えられている物語である。次に、文体という観点から『源氏物語』各巻の類似性を指摘する。これによって、『源氏物語』の作者が単独であるのかあるいは複数であるのか、また『源氏物語』54 巻の成立順序についての議論に耐えうる分析結果を示す。

分析においては、単語及び品詞情報などに関連した文体的特徴を特徴量として用いた。また、n-gram モデルなどの単語の共起情報も用いて分析を行った。n-gram とは文中において n 個の隣接する要素の組のことである。本研究では品詞の n-gram や単語の n-gram を特徴量として分析を行う。これに加えて、本研究の新規性の 1 つとして単語の文字数の n-gram を特徴量として分析に採り上げる。単語の文字数は古くから計量分析に用いられている。しかし、単語の文字数の n-gram はおよそ分析に用いられていない。本研究では語の長さを求める際に、全ての単語を仮名に変換することから、語の長さの n-gram は文章のリズムに関連する特徴量であると言える。計量的な文体研究において、文章のリズムの計量的な分析は十分に展開されておらず、本研究において提案する単語の文字数の n-gram を特徴量とし計量的に分析することで、単語の仮名文字数を分析に用いることで文章のリズムを近似的に把握できると考えられる。このような特徴量に対して、多変量解析の代表的な手法である主成分分析や機械学習の手法であるランダムフォレストなどを用いた。

４．研究成果

単語の共起を分析する場合は n-gram を用いることが多い。ただし、n-gram は隣接する単語を分析することには適しているが、隣接しない場合は容易ではない。そこで、本研究ではアソシエーション分析を行った。これによって、隣接しない場合の単語の共起も分析可能となる。アソシエーション分析はデータマイニングの手法であり、トランザクションというデータを用いる。トランザクションとは 1 度の商取引において扱われた商品を記録したデータのことである。本研究では 1 つのセンテンスを 1 つのトランザクションとみなし、単語を商品とみなして分析を行った。つまり、1 つのセンテンスにおいて共起する単語のパターンを抽出した。ただし、『源氏物語』の本文には句読点が付与されていないため、句点の情報については『日本古典文学大系 源氏物語』を利用した。このように、アソシエーション分析を行うことで、各巻において特徴的に頻出する単語の組み合わせを発見することが可能となると考えられる。その結果、一部の連語が『源氏物語』の終盤の巻に多く出現することが分かった。すなわち、ここに『源氏物語』54 巻における文体的特徴の出現傾向の変化を認めることができると考えられる。

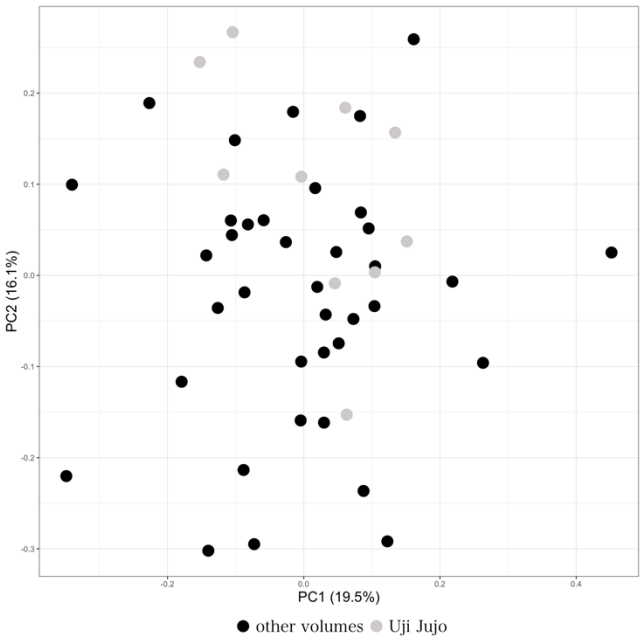

図 1　宇治十帖と他の諸巻についての主成分分析

また、『源氏物語』と補作である『山路の露』および『雲隠六帖』との比較分析を行った。分析においてはカイ二乗検定などの手法を用いた。補作の 2 作品は『源氏物語』と同じ舞台を描いていることから、単語の出現傾向など『源氏物語』と強い類似性が認められたものの、助詞などの文法的機能を担う語の出現率が『源氏物語』との間に顕著な相違が認められた。これらのうち、特に顕著な相違として、接続助詞の「ど」と「ども」の出現傾向があげられる。これら 2 語は同様な文法的機能を持つ助詞であるが、補作における「ど」の出現率は『源氏物語』における出現率より低く、反対に補作における「ども」の出現率は『源氏物語』における出現率よりも高いことが明らかになった。

最後に、本研究では『源氏物語』における作者問題の解決のために、単語 n-gram と単語の長さの n-gram を用いて分析を行った。その結果、図 1 に示すように『源氏物語』の最終 13 巻と他の巻との間に計量的な相違は認められなかった。『源氏物語』の作者問題は古くから議論されてきたが、本研究で行った分析により、『源氏物語』に複数の作者が存在した可能性があるという説を支持する根拠はないことが明らかになった。したがって、『源氏物語』は一人の作者によって書かれた可能性が高いと結論づけることができる。

＜引用文献＞

[1] 上田英代, 村上征勝, 今西祐一郎, 樺島忠夫, 上田裕一. (1994). 『源氏物語語彙用例総索引 自立語編』, 勉誠社.
[2] 上田英代, 村上征勝, 今西祐一郎, 樺島忠夫, 上田裕一, 藤田真理. (1996). 『源氏物語語彙用例総索引 付属語編』, 勉誠社.
[3] 池田亀鑑. (1951)「源氏物語の構成」(『新講源氏物語（上)』所収), 至文堂 .
[4] 石田穣二. (1961). 「匂宮・紅梅・竹河の三帖をめぐって」. 國文学 : 解釈と鑑賞, 26(12).
[5] 武田宗俊. (1954). 『源氏物語の研究』. 岩波書店.
[6] 加藤昌嘉, 中川照将. (2010). 「“『源氏物語』はどのように出来たのか？”を考えるために」(『紫上系と玉鬘系−成立論のゆくえ』所収), 勉誠出版.

５．主な発表論文等

〔雑誌論文〕　計4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　4件）

１．著者名 土山玄	４．巻 33(2)
２．論文標題 言語統計学入門（２）：統計的尺度と言語データ	５．発行年 2021年
３．雑誌名 計量国語学	６．最初と最後の頁 100-109
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 オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 土山玄	４．巻 2020
２．論文標題 『源氏物語』第二部の文体的特徴についての計量的な検討	５．発行年 2020年
３．雑誌名 情報処理学会じんもんこん2020論文集	６．最初と最後の頁 129-134
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 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 土山玄	４．巻 2020(3)
２．論文標題 『源氏物語』及びその補作における特徴語句抽出の試み	５．発行年 2020年
３．雑誌名 研究報告人文科学とコンピュータ	６．最初と最後の頁 1-5
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 オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 土山玄	４．巻 33(7)
２．論文標題 夏目漱石の小説における文体の継時的変化について：文末表現についての計量的な検討	５．発行年 2022年
３．雑誌名 計量国語学	６．最初と最後の頁 481-492
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 オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計9件（うち招待講演　1件／うち国際学会　0件）

1．発表者名
土山玄

2．発表標題
『源氏物語』における語の共起情報の分析

3．学会等名
計量国語学会第66回大会

4．発表年
2022年

1．発表者名
土山玄

2．発表標題
文学作品のテキストデータを対象としたデータ分析

3．学会等名
シンポジウム「人文・社会科学研究におけるデータサイエンス」（招待講演）

4．発表年
2021年

1．発表者名
土山玄

2．発表標題
『源氏物語』第二部の文体的特徴についての計量的な検討

3．学会等名
情報処理学会人文科学とコンピュータシンポジウム「じんもんこん2020」

4．発表年
2020年

1．発表者名
土山玄

2．発表標題
文学作品のテキストデータを題材としたデータサイエンス演習

3．学会等名
数理・データサイエンス教育強化拠点コンソーシアム2020年度関東・首都圏ブロック第5回ワークショップ

4．発表年
2020年

１．発表者名
土山玄

２．発表標題
平安時代の文学作品における『源氏物語』の特徴語の抽出 ―『日本語歴史コーパス 平安時代編』を用いて―

３．学会等名
「通時コーパス」シンポジウム2020オンライン

４．発表年
2020年

１．発表者名
土山玄

２．発表標題
計量分析による『源氏物語』三部構成説の検討

３．学会等名
日本行動計量学会第47回大会

４．発表年
2019年

１．発表者名
土山玄

２．発表標題
助動詞の出現傾向に基づく『源氏物語』各巻の分類

３．学会等名
計量国語学会第63回大会

４．発表年
2019年

１．発表者名
土山玄

２．発表標題
『源氏物語』及びその補作における特徴語句抽出の試み

３．学会等名
第122回情報処理学会人文科学とコンピュータ研究会

４．発表年
2020年

１．発表者名
土山玄

２．発表標題
『源氏物語』における語の共起情報の分析

３．学会等名
計量国語学会第66回大会

４．発表年
2022年

〔図書〕　計4件

１．著者名 金 明哲、中村 靖子、上阪 彩香、土山 玄、孫 昊、劉 雪琴、李 広微、入江 さやか	４．発行年 2021年
２．出版社 岩波書店	５．総ページ数 248
３．書名 文学と言語コーパスのマイニング	

１．著者名 土山玄	４．発行年 2019年
２．出版社 勉誠出版	５．総ページ数 850
３．書名 文化情報学事典	

１．著者名 Gen Tsuchiyama	４．発行年 2022年
２．出版社 De Gruyter	５．総ページ数 229
３．書名 Quantitative Approaches to Universality and Individuality in Language	

１．著者名 土山玄	４．発行年 2022年
２．出版社 学術図書出版社	５．総ページ数 132
３．書名 文理融合データサイエンスの基礎	

〔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