2024 年度中央大学共同プロジェクト　研究実績報告書

１．概要

研究代表者		所属機関	国際情報学部		2024 年度助成額
		氏名	橋本　健広		5,467,000 円
		NAME	Takehiro Hashimoto		
研究課題名	和文	文学を科学する：機械学習を用いて文脈にもとづくテクスト間の影響を調べる学際的研究		研究期間	2024～2025 年度
	英文	The science of literature: Interdisciplinary study of textual influence based on context-oriented machine learning			

２．研究組織

※所属機関・部局・職名は 2025 年 3 月 31 日時点のものです。

	研究代表者及び研究分担者		役割分担	備考
	氏名	所属機関/部局/職		
1	橋本　健広	国際情報学部・教授	研究全般	
2	木谷　厳	帝京大学・教育学部・教授	テクスト分析（同時代の新規の影響）	
3	小花　聖輝	国際情報学部・准教授	機械学習モデルの構築	
4	笹川　浩	商学部・教授	テクスト分析（後の時代の影響）	
5	矢島　壮平	国際情報学部・准教授	テクスト分析（多文化の影響）	
6	山西　博之	理工学部・教授	機械学習モデルの言語学的応用方法	
7	宮川　創	筑波大学・人文社会系・准教授	テクスト分析	
8	古屋　耕平	青山学院大学・経済学部・教授	テクスト分析	
9	佐藤　健	神戸大学・大学教育推進機構　国際コミュニケーションセンター・教授	応用言語学	
合計　　９名				

３．２０２４年度の研究活動報告　※行が不足する場合は、適宜、行を追加してご記入ください。

（和文）

本研究プロジェクトの目的は、1.特定の文学作品および研究成果を利用することで、イギリス文学研究に特化した機械学習のモデルおよび適用式を構築し、2. 機械学習を用いて意味や文脈にもとづく影響の析出を試みるテキスト分析を行うこと、また、3.実証研究の機会を増やすことで、文学の研究領域の拡大と、文学研究の再興を目指し、4.大学をデジタル・ヒューマニティーズ（以下DH）の人的ネットワーク拠点とし、イギリス文学、DH、自然言語処理の学会、研究所、研究者等と連携を試みることである。

具体的には、以下に示すように①機械学習のモデルおよび適用式の作成と、応用としてのテキスト分析である。

①　（機械学習のモデルおよび適用式の構築）

1）イギリス文学作品の韻文を中心とするテキスト群と意味が確定されたコンコーダンスを用いた、イギリス文学に特化した機械学習モデルの構築

2）文学テクストの分析に適した機械学習モデルの分析方法（適用式）の検討

②　（イギリス文学分野における機械学習を利用したテキスト分析）

1）文学テクスト間の影響分析において、これまで大まかな影響が指摘されただけのテクスト間から、具体的なテクストを選び示す

2）ある影響あるテクスト間において、これまで見落とされてきた別の箇所のテクストを選び示す

3）これまで影響があると考えられてこなかったテクスト間の影響を選び示す

１．研究計画の進行状況（研究内容）について

2024年度の活動は、2024年5月の検討会、2024年12月に行った報告会および国際シンポジウム、毎月開催する研究会（機械学習のモデルおよび適用式の構築、およびテキスト分析の実施）を通して行われた。4月にキックオフミーティングを行い、5月の検討会では生成AIと著作権に関する議論動向に関する講演（講師：中央大学教授小向太郎氏）を伺ったうえで、具体的な研究方法について話し合った。その後毎月テキスト分析の結果や評価データセット、新しいツールや大規模言語モデルのプラットフォームの紹介、影響のデータセットの作成などを検討した。12月の国際シンポジウム（Digital Approach to Literary Analysis 2024, 中央大学にて開催）では、本研究プロジェクトの紹介をするとともに、DHにおける批評とのかかわりやAIを使用したテキスト分析の講演を行った。

①　機械学習のモデルおよび適用式の構築　機械学習モデルの作成については、ファインチューニングのための元データとするためのコンコーダンスのデータ化、大規模言語モデルのファインチューニング、適用式の検討を行った。

1）　コンコーダンス　文学テキスト中での単語の意味が掲載されたシェイクスピア、シェリー、ミルトンの三種類のコンコーダンスをデジタル化する予定であったが、シェイクスピアはすでにデジタル化されていることが分かった。関係者に連絡を取り、データを入手済みである。現在シェリーのコンコーダンスの8割ほど(約8万語)のデジタル化を終えている。シェリーの残りとミルトンについては2025年度のデジタル化を予定している。

1.2）　モデルの作成　ファインチューニング用のデータを順次大規模言語モデル(BERT、OpenAI、Claudeなど)にファインチューニングさせている。当初Google Cloud Platformを使用する予定であったが、OpenAIやClaude等のプラットフォームが使えることがわかりそちらを使用している。主な発表論文等の欄のその他の項目に記載した[7][8][9]は本研究プロジェクトで作成中のモデルである。[7]はBERT、[8][9]はOpenAIを使用している。

単語の意味のデータセットだけでは満足のゆく性能を持つ英文学に特化したモデルの作成が難しいことがわかり、文学研究者（橋本、笹川、木谷、古屋、宮川）で分担してテキストの注釈を利用して影響するテキストペアのデータセットを作成することなった。時間のかかる作業で

あるが現在 440 件ほどのペアが集まっている。また作成したモデルを評価するための評価用データセットも必要とわかり、作成をした[10]。

2) 適用式の作成 英文学に特化したモデルと文学研究を接続できるような計算方法を考案中である。毎月開催する研究会において研究分担者とともに検討を行った。現在主流の NLP の手法とは異なるため展開が難しいが、文学批評の研究をし発表を試みるなど順次実施中である。[5]は本研究プロジェクトで考察した適用式に関する研究発表である。

② 機械学習を利用したテキスト分析 モデルを作成途中であるため、現在利用できる大規模言語モデルを利用して、毎月の研究会で分析を行っている。

1)2)3) 機械学習を使用したテキストの分析 Warton, Bowles, Coleridge, Wordsworth の川の詩のテキスト分析（笹川)、Hawthorn の”The Old Manse”と Melville の *The Piazza Tales*、Emerson の Nature, Essays の分析（古屋)、”Kubla Khan”と Paradise Lost 第４巻楽園の描写部分の影響（宮川）などが分析結果を発表した。[2]は本研究プロジェクトの成果を反映させた研究である。他のテキスト分析は今後発表につなげていく予定である。

計画通り遂行されていない場合の対応状況

2024 年度はモデルの作成に必要なデータセットの作成に当初想定していた以上の時間がかかり、また影響のデータセットや作成したモデルの評価用データセットの作成も必要とわかったため、本研究プロジェクトで作成したモデルを使用した本格的なテキスト分析は 2025 年度以降になる。影響のデータセットは、さまざまなテキストの注釈をもとに影響があると指摘されるテキストのペアを収集する作業であり、文学研究者で手分けして収集しているが、時間のかかる作業である。2025 年度はツールを作成して収集のスピードアップを図りたい。これまでに収集したデータの件数は現在 440 件ほどであるが、目標は 2000 件である。また英文学の影響関係を評価する評価用データセットが存在しないため、現在作成中である。[10]が作成途中の評価用データセットである。今後海外の研究者と協力してブラッシュアップを図りたい。

これらのデータセットをもとに、BERT を使用したり[7]、Claude、OpenAI[8][9]といったプラットフォーム上でファインチューニングを行って、今年度作成したモデルよりもより性能の高いモデルを作成していく予定である。

共同研究としての活動

今年度は毎月開催する研究会をベースに、各人が発表を行って、当該知識の理解を深め、研究活動をすすめていった。またデータセットの作成は文学の研究者で手分けして作成を行っている。後半は特に、宮川氏の協力を得て、ChatGPT、OpenAI、Claude、Dify、TRACER といった新規の技術やツールの用いたテキスト分析の試行錯誤を各人で繰り返した。

本学における教育活動への還元

2024 年 12 月 20 日中央大学において開催した報告会および国際シンポジウム（タイトル：Digital Approach to Literary Analysis 2024）において、研究者および学生に対してシンポジウムを公開し開催した。発表者は橋本、ヨーヘイ・イガラシ氏（コネチカット大学准教授)、ルシアン・リー氏（イリノイ大学ウルバマシャンペーン校大学院生）アルチョム・ススロブ氏（北海道大学大学院生）であった。橋本は本研究プロジェクトの内容を報告した。

２．研究費の執行状況

研究費はおおむね予定通りに執行した。データセット作成のための資料や機材、コンコーダンスのデジタル化、大規模言語モデルのプラットフォーム利用料、学会発表などに研究費を使用した。手数料は、当初 Google Cloud Platform の使用料として 100 万円を計上していたが、OpenAI、ChatGPT、Claude、Dify 等、他にも同内容でより簡便に利用できるサービスがあること

がわかり、そちらを使用した利用料となる。謝金・謝礼品は、コンコーダンスのデジタル化のために当初ﾊﾟｰﾄﾀｲﾑ職員給与での支出を考えていたが、委託報酬の枠で利用する方が簡便とわかりそちらで報酬を支払ったため、ﾊﾟｰﾄﾀｲﾑ職員給与が減り謝金・謝礼品が増えている。旅費・交通費については、海外での学会参加に費用が当初想定以上にかかったが、円安や物価高騰を反映した結果と考えられる。またデータセット作成のための資料をそろえるために機材等を購入したため、消耗品の利用が増加した。

３．研究成果の公表

[1][5]は橋本が行った評価用データセットおよび適用式についての論文および発表である。[2]は古屋が行った本研究プロジェクトでの分析を踏まえた論文である。また[3]についても同様に宮川による本研究プロジェクトでの成果が反映された論文となっている。また橋本は本研究プロジェクトで得られた知見をもとにした機械学習を使用したテキスト分析を[4]で発表予定である。[7][8][9]は影響分析のモデル、[10]は評価用データセットであり、公開されている。また本研究プロジェクトの紹介を[6]で行った。今後は、モデルを使用した文学テキストの影響分析の結果を、海外の学会で発表していく予定である。

４．研究分担者の活動

橋本は全体を統括し、コンコーダンスのデジタル化、新規に必要となったデータセットの作成、テキストの分析を行った。木谷、笹川、古屋は機械学習モデルを用いて文学のテキスト分析を行い、毎月の研究会で発表した。また影響のデータセットを手分けして作成中である。宮川、小花は大規模言語モデルや機械学習を使用したテキスト分析について適宜技術的な指導を行った。また宮川は大規模言語モデルや機械学習のツールを使用してテキスト分析を実施した。佐藤は毎月の研究会に参加し、英語教育の立場から研究に貢献した。

（英文）

This research project aims: 1. to build models and application formulas for machine learning specific to English literary studies by using specific literary works and academic resources; 2. to use machine learning to conduct text analysis that attempts to analyze context-based influences; 3. to increase opportunities for empirical research 4. to make the university a human network center for Digital Humanities (DH), and to attempt to collaborate with academic societies, research institutes, and researchers in the fields of English literature, DH, and natural language processing.

Specifically, as shown below, we will (1) create models and application formulas for machine learning and (2) analyze texts as an application of machine learning.

1- Construction of models and application formulas for machine learning

1) Construction of a machine learning model specific to English literature, using a collection of texts, mainly poetry, and concordances that show meanings of the texts.

2) Investigation of analysis methods (application formulas) for machine learning models suitable for analyzing literary texts

2- Text analysis using machine learning in the field of English literature

1) Analyzing the influence between literary texts, which was roughly mentioned in the previous research.

2) Finding other influenced texts within an influence text previously overlooked.

3) Finding influential texts that have not been previously considered as influences.

1. Progress of the research plan

Activities in FY2024 were conducted through the May 2024 meeting, the December 2024 international symposium, monthly meetings on building machine learning models and formulas, and performing text analysis. After listening to a lecture by Professor Taro Komukai of Chuo University on trends in discussions regarding AI and copyright, specific research methods for this project were discussed. During monthly discussions, we discussed text analysis results, evaluation datasets, the introduction of new tools and platforms for large language models, the creation of impact datasets, etc. We held an international symposium in December (Digital Approach to Literary Analysis 2024) at Chuo University. We introduced our research project at the seminar and lectured on the relationship with criticism in DH and text analysis using AI.

I. Construction of machine learning models and application formulas: To create machine learning models, concordance data were used as source data for fine-tuning, large language models were fine-tuned, and application formulas were examined.

1) Concordance: We planned to digitize three concordances- Shakespeare, Shelley, and Milton- which list the meanings of words in literary texts. However, we found that Shakespeare had already been digitized. We have already contacted the relevant parties and obtained the data. We have now finished digitizing about 80% of Shelley's concordance (about 80,000 words). The remainder of Shelley's and Milton's will be digitized in FY2025.

1.2) Model creation: We fine-tune large language models (BERT, OpenAI, Claude, etc.) using concordances and other data. We initially planned to use the Google Cloud Platform, but we found that platforms such as OpenAI and Claude were available and used them instead of Google Cloud Platform. The models in [7], [8], and [9], which are listed in the "Others" section of the "Main publications, etc." column, are models that are being developed in this project. [7] used BERT, and [8][9] used OpenAI.

We found it challenging to create a model specialized for English literature with satisfactory performance using only a dataset of word meanings, so we decided to make a dataset of text pairs of influence mentioned in annotations of texts (Hashimoto, Sasagawa, Kitani, Furuya, and Miyagawa). This is a time-consuming task, and about 440 pairs have been collected. We also needed an evaluation dataset to evaluate the model we had created. The evaluation dataset we created is [10].

2) Creation of application formulas: We are devising a calculation method to connect a model specific to English literature with literary studies. We discussed this with the research members at monthly research meetings. Although it is challenging to develop the method because it is different from the current mainstream NLP method, we are researching literary criticism and attempting to present the results sequentially. [5] is a presentation of research on the application formula considered in this research project.

II Text analysis using machine learning: Since we are still creating a model for text analysis using machine learning, we are currently using a large language model and analyzing monthly research meetings. 1)2)3) Text analysis using machine learning Text analysis of river poems by Warton, Bowles, Coleridge, and Wordsworth (Sasagawa), Hawthorn's "The Old Manse" and Melville's *The Piazza Tales,* Emerson's Nature, Essays (Furuya), and the influence of "Kubla Khan" and the description of paradise in *Paradise Lost* Volume 4 (Miyagawa). [2] is a study that reflects the results of this research project. Other textual

analyses will be presented in the future.

Status of actions not carried out as planned

In FY2024, it took more time than initially expected to create the data set necessary to create the model. It was also necessary to create an impact data set and a data set for evaluating the created model. The full-scale text analysis using the model created in this project will be conducted in FY2025 or later. We hope to speed up the collection process by creating a tool in FY2025. The data collected so far is currently about 440, but the goal is to collect 2,000. In addition, we are creating an evaluation dataset to assess the influence of English literature. The dataset [10] is in the process of being created. In the future, we would like to brush up on the dataset in cooperation with overseas researchers.

Based on these datasets, we plan to create models with higher performance than those created this year by using BERT [7] and conducting fine-tuning on platforms such as Claude and OpenAI [8][9].

Activities as Collaborative Research

This year, based on the monthly research meetings, each member gave a presentation to deepen the understanding of the knowledge and to promote research activities. In addition, we are creating the influence dataset. In the latter half of the year, with the cooperation of Miyagawa, each researcher repeated the trial-and-error process of text analysis using new technologies and tools such as ChatGPT, OpenAI, Claude, Dify, and TRACER.

Contribution to educational activities at the University

The symposium was open to researchers and students at the debriefing and international symposium (title: Digital Approach to Literary Analysis 2024) held at Chuo University on December 20, 2024. The presenters were Hashimoto, Yohay Igarashi (Associate Professor, University of Connecticut), Lucian Lee (graduate student, University of Illinois at Ulverma-Champaign), and Artyom Suslov (graduate student, Hokkaido University). Hashimoto reported on the details of this research project.

Status of Research Funds

Research expenses were generally executed as planned. Research funds were used for materials and equipment for dataset creation, concordance digitization, platform usage fees for large language models, and conference presentations. Fees were initially set aside for using the Google Cloud Platform of 1,000,000 yen. Still, it was discovered that other services such as OpenAI, ChatGPT, Claude, and Dify were available for the same content and were more convenient to use, and the fees were used for those services. As for rewards and honoraria, we initially planned to use part-time staff salaries to digitalize the concordance but found that using them under the framework of outsourced remuneration was easier. Travel and transportation expenses were higher than initially expected for participation in overseas conferences, but this is thought to be a result of the weak yen and soaring prices. In addition, consumable supplies increased due to purchasing equipment and other materials to prepare data sets.

Publication of Research Results

[1][5] are papers and presentations on the evaluation datasets and application formulas conducted by Hashimoto. [2] is a paper based on the analysis conducted by Furuya in this research project. [3] is a paper reflecting the results of Miyagawa's research project. Hashimoto will present a text analysis using machine learning based on the findings of this research project in [4]. [7][8][9] are models for impact analysis, and [10] is a dataset for evaluation, which is publicly available. We presented our research project at [6]. We plan to present the influence analysis results of literary texts using the model at conferences abroad.

4. Activities of Research Assignees

Hashimoto oversaw the entire project, digitizing the concordance, creating the newly required dataset, and analyzing the texts. Kitani, Sasagawa, and Furuya conducted a text analysis of literature using machine learning models and presented their results at monthly meetings. They are also working on a dataset of influences manually. Miyagawa and Kobana provided technical guidance on text analysis using large language models and machine learning as needed. Miyagawa also conducted text analysis using large language modeling and machine learning tools. Sato participated in monthly research meetings and contributed to the research from the standpoint of English education.

４．主な発表論文等（予定を含む）

２０２４年度に行った共同研究プロジェクトの研究課題としての成果内容についてご記載ください。

※行が不足する場合は、適宜、行を追加してご記入ください。

【学術論文】（著者名、論文題目、誌名、査読の有無（査読がある場合は必ず査読有りと明記してください）、巻号、頁、発行年月）＜発表予定を含む。但し、投稿中、投稿準備中のものは除く＞

[1] Hashimoto, Takehiro. “English Poetry Dataset for Evaluating Large Language Models Used for Analyzing the Influence of Poetry.” *Japanese Journal of Global Informatics*, vol. 5, March 2025, pp. 163-169.

[2] 古屋耕平「Herman Melville と十九世紀のコレラ流行 」.『青山經濟論集』, 第 76 巻, 第 1 号, 2024 年 6 月, 143-69 頁.

[3] Mika, Hämälainen, Öhman, Emily, Miyagawa, So, Alnajjar, Khalid, Bizzoni, Yuri, Rueter, Jack, Partanen, Nico. “The Growing Importance of Humanities for NLP in the Era of LLMs ” *Lightning Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, 査読有り, 2–6, November 2024.

【学会発表】（発表者名、発表題目、学会名、開催地、開催年月）

[4] Hashimoto, Takehiro. “How does war affect Romantic literature? Topic modeling Romantic documents.” Paper presented at DH2025, Universidade NOVA de Lisboa (Jisbon), July 2025.

[5] 橋本健広. 「文学批評から大規模言語モデルへ―単語埋め込みの組み換えによる文学テクスト解釈の試み」言語処理学会第 31 回年次大会, 出島メッセ長崎, 2025 年 3 月 11 日.

[6] Hashimoto, Takehiro. “A lightning talk proposal for literary text analysis using LLM.” Lightning Paper presented at DH2024 Workshop SIG-DLS, George Mason University (Virginia, the US), August 5th, 2024.

【図　　書】（著者名、出版社名、書名、刊行年）

【その他】(知的財産権、ニュースリリース等)

[7] Takehiro Hashimoto. hast2/all-mpnet-base-v2-influencev1. Hugging Face. 2024 年 12 月 7 日. [影響のデータセットをファインチューニングしたモデル].

[8] Takehiro Hashimoto. ft:gpt-4o-mini-2024-07-18:personal::BCSmwIqZ. OpenAI API. 2025 年 3 月 18 日. [コンコーダンス中の 77954 件の単語の意味をファインチューニングしたモデル]

[9] Takehiro Hashimoto. ft:gpt-4o-mini-2024-07-18:personal::AcLPVcba. OpenAI API. 2024 年 12 月 9 日. [404 件の影響するテキストペアをファインチューニングしたモデル].

[10] Takehiro Hashimoto.hast2/KublaPL. Hugging Face. 2025 年 1 月 1 日. [文学の影響関係の評価用データセット]

[11] 宮川創. 科学技術への顕著な貢献 2024（ナイスステップな研究者）, 最新テクノロジーを駆使したエジプト学およびアジア・アフリカの消滅危機にある言語の研究, 文部科学省科学技術・学術政策研究所（NISTEP）