

Title	「テキストマイニングとデジタルヒューマニティーズ」プロジェクトの目的と活動
Author(s)	田畑, 智司
Citation	言語文化共同研究プロジェクト. 2024, 2023, p. 1-4
Version Type	VoR
URL	https://doi.org/10.18910/97313
rights	
Note	

Osaka University Knowledge Archive : OUKA

<https://ir.library.osaka-u.ac.jp/>

Osaka University

「テキストマイニングとデジタルヒューマニティーズ」 プロジェクトの目的と活動

本共同研究は、自然言語処理、コーパス言語学・計量言語学、数理統計学、データマイニング、機械学習など、諸分野の知見を有機的に統合した方法論を開発し、テキストマイニングを応用して人文学、言語文化学の諸問題にアプローチする、すなわち「デジタルヒューマニティーズ (Digital Humanities)」の実践と理論的精緻化の可能性を探る営みである。このプロジェクトは、2001年度に岩根久教授、緒方典裕助教授、および筆者の3名でスタートした「電子化言語資料分析の方法論」を基礎とするが、2003年度から名称を一部改め、言語文化研究科の大学院生もメンバーに加わった。2006年度には三宅真紀助教の加入を得て、対象言語も英・仏・ギリシャ語に広がった。2011年には言語文化教育論講座に着任した今尾康裕講師が加入した。2014年度後期から、さらに Hodošček Bor 講師が加わった。そして、2019年度をもって退職された岩根久教授の後任として、2020年度に山田彬堯講師が着任・加入した。言語文化研究科と文学研究科の統合により設立された人文学研究科には「人文学林」という分野横断組織が置かれ、デジタルヒューマニティーズ振興の役割を担っている。その人文学林から、2022年に菅原裕輝特任助教が、そして2023年に吉賀夏子准教授が加入し、現在の陣容となっている。(職位はいずれも当時)。2016年度から、プロジェクトの名称を、当該リサーチコミュニティの名称としてより相応しい「テキストマイニングとデジタルヒューマニティーズ」にアップデートしたが、研究の系統は創始時より常に一貫している。

「テキストマイニングとデジタルヒューマニティーズ」プロジェクトは大きく分けて二つの層で構成されている。一つは研究基盤となるコーパス、テキストアーカイブの開発・構築、もう一つは構築したコーパス、テキストアーカイブからのデータ抽出法研究、並びに得られた高次元の言語データの計量分析である。前者には英・仏語の文学作品や、聖書(共観福音書)などの電子テキスト化、ロシア語政治演説コーパス、近代日本文学コーパスの編纂、マークアップ言語XMLによるTEI (Text Encoding Initiative: デジタル化したテキストの国際互換規格の枠組) に準拠したタグ付けなど、人文学資料のデジタル化やマークアップ法、データ符号化方法論の開発などが含まれる。一方、高次元人文学データ分析の事例として、語彙・語法、コロケーション、意味構造、語用論などのレベルにおける言語使用の実態研究、高度な数理モデルや機械学習を応用した言語分析やテキストマイニング、文学作品の言語特徴の特定や、使用域間の言語変異や文体識別問題の考察、著者推定法の精密化研究を挙げることができる。

本プロジェクト班は人文学研究科の専任教員7名と名誉教授1名(今尾康裕、菅原裕輝、田畑智司、Hodošček Bor、三宅真紀、山田彬堯、吉賀夏子、岩根久名誉教授)、当研究科博士後期課程在学学生6名(福本広光、藤田郁、竹森ありさ、涌井萌子、曹芳慧、Camilleri Gabriele)、博士前期課程在学学生5名(小堀彩夏、Vogatzá Dimitra、李晨婕、肖媛媛、于拙)に加え、OGの黄晨雯氏(2024年5月に本研究科言語文化学専攻助教着任)、京都大学徐勤氏(2023年3月博士学位取得)、大阪医科大学浅野元子氏(2020年3月博士学位取得)・名古屋外国語大学杉山真央氏(2019年3月博士学位取得)、本学非常勤講師の高橋新氏、南澤佑樹氏(2024年4月に本研究科外国学専攻助教着任)、帝塚山学院大学八野幸子氏(本研究科博士課程修了)、国立国語研究所の竹内綾乃氏を主たる参加メンバーとしている。研究を遂行するために、コアメンバー以外も自由に参加できる月例の研究會・討論會などを通して、研究情報の交換、論文や開発ツール、構想段階のプロジェクトや進行中のパイロットスタディのプレビューなどを行っている。

2023年度は、対面開催とオンライン併用のハイブリッド方式で開催した。オンラインでの研究会の開催を続けているうちに、学外からの研究会参加者が増加したこともあり、今後もハイブリッドでの開催を続ける予定である。

2023年度「テキストマイニングとデジタルヒューマニティーズ」研究会開催記録
およびメンバーによるDH関連学会での発表記録

第1回 2023年4月14日開催

発表者・発表題目

全メンバー 2023年度の活動計画打合せ

第2回 2023年5月12日開催

発表者・発表題目

黄晨雯 「ChatGPTとテキストマイニング」

第3回 2023年6月9日開催

発表者・発表題目

藤田 郁 “The Stylistics of Alfred Tennyson’s Poetry: A Stylometric Approach”
Hodošček, Bor “YoutuberLinguisticProfile: A toolbox for linguistic analysis of Youtube videos
Work-in-progress report on Tom Scott videos”

第4回 2023年7月7日開催

発表者・発表題目

杉山 真央 「戦勝記念日演説から見るプーチン大統領の人称代名詞使用」

第5回 2023年7月12–15日開催 Poetics and Linguistics Association International Conference (PALA)
2023, University of Bologna, Italy

発表者・発表題目

Tabata, Tomoji “Using topic models to explore body language in Dickens’s literature and journalism”
Fujita, Iku “‘How much I love this writer’s manly style!’:
Similarities and differences between Tennyson and other poets”

第6回 2023年8月4日開催

発表者・発表題目

曹 芳慧 “Characterization by dialogues in Hardy’s novels:
a preliminary quantitative study of three works”
塚越 柚季 「『リグ・ヴェーダ』の韻律情報を利用した文体比較」

第7回 2023年9月1日開催

発表者・発表題目

Camilleri, Gabriele 「因子分析を用いた文学作品翻訳コーパスにおける発話キャラクターの抽出」

第8回 2023年9月9-10日開催 英語コーパス学会第48回大会(学会創設30周年記念大会)
発表者・発表題目

- 後藤一章 「音声認識 AI を活用した音声・映像コーパス構築ツールの開発」
菅原 裕輝 「コーパス研究は仮説検証型の科学か? : 形式概念分析を用いたメタ分析」
Fujita, Iku “Extraneous to Allusions: Stylistic differences between Tennyson and 18th-century poets”
曹 芳慧 「Hardy 作品における会話部の感情分析」
杉山 真央 「戦勝記念日演説から見るプーチン大統領の“Мы (we)”と“Вы (you)”」
Tabata, Tomoji “Exploring body language in Dickens’s fiction through topic modelling”
鈴木 大介 「現代英語における worse の文副詞用法をめぐって」

第9回 2023年10月6日開催
発表者・発表題目

- 涌井 萌子 「『レ枢機卿のマザリナード』の帰属検証—分析方法と基準の検討—」
岩根 久 「Pierre de Ronsard (1524–1585) 作品集としての *Les Amours* の変容 (1552–1560)」

第10回 2023年11月3日開催
発表者・発表題目

- 福本 広光 「Clinton, G. W. Bush, Obama による分離不定詞再考：
コンコーダンスラインに見える特徴的用法に関する試論」
Vogatz, Dimitra “Investigating the translational equivalency of pity in EN & JP”
王 簫影 「漢字圏と非漢字圏日本語教科書リーダビリティ研究—長文テキストを対象に—」

第11回 2023年11月18日開催 国際シンポジウム デジタルヒューマニティーズと研究基盤：
欧州と日本の最新トレンド
発表者・発表題目

- 吉賀 夏子 パネルセッション
「人文学のためのデジタル研究基盤及び DH の人材育成に関する事例報告」
田畑 智司・ポスター・デモンストレーションセッション
吉賀 夏子・「大阪大学における共同研究プロジェクト」
菅原 裕輝・『テキストマイニングとデジタルヒューマニティーズ』の歩み」
ホドシチェク
ボル

第12回 2023年12月1日開催
発表者・発表題目

- 徐 勤
小堀 彩夏 「英訳版村上春樹作品における統計的手法を用いた文体分析」

第13回 2024年1月5日開催

発表者・発表題目

吉賀 夏子 「江戸時代の佐賀地域における情報基盤の構築とその可能性」
南澤 佑樹 「スウェーデン語の背中に関連する表現について」

第14回 2024年2月16日開催

発表者・発表題目

肖 媛媛 「コーパスに基づく英語政治ニュース研究 -英語母語圏と非母語圏の比較研究」
立野 寛太 「スポーツの社会言語学—結果の予測不可能性と実況の関わりについて—」

第15回 2024年3月6日開催 人文学林シンポジウム

発表者・発表題目

田畑 智司・ 「デジタルヒューマニティーズの現在（いま）」
吉賀 夏子・ 一方法論的学際性にもとづくデータ駆動型の人文知の追求へ向けて—
石田 友梨・
ホドシチェク
ボル

第16回 2024年3月8日開催

発表者・発表題目

菅原 裕輝 “Digital Sociology of Expectation”
今尾 康裕 「Mallet を統合した CasualConc の新機能」

2024年 5月
研究代表者 田畑 智司