



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. 2025, 2024, p. 1-4
Version Type	VoR
URL	https://doi.org/10.18910/102218
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

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

The University of Osaka

「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 9 名と名誉教授 1 名（今尾 康裕、菅原 裕輝、黄 晨雯、田畑 智司、Hodošček Bor、南澤 佑樹、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生 6 名（福本 広光、曹 芳慧、Camilleri Gabriele、陳 宇鎔、Vogatzá Dimitra、王 簫影）、博士前期課程在学生 3 名（李 晨婕、肖 媛媛、于 拙）に加え、OG の京都大学 徐 勤氏（2023 年 3 月本研究科博士学位取得）、大阪医科薬科大学 浅野 元子氏（2020 年 3 月本研究科博士学位取得）・名古屋外国語大学 杉山 真央氏（2019 年 3 月本研究科博士学位取得）、比治山大学 藤田 郁氏（2024 年 3 月本研究科博士学位取得、本年 4 月に九州大学大学院システム情報科学研究院に異動）、帝塚山学院大学 八野 幸子氏（2018 年 3 月本研究科博士課程修了）、国文学研究資料館の竹内 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、コアメンバー以外にも自由に参加できる月例の研究会・討論会などを通して、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。

2024 年度も、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。2020–2022 のパンデミック中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。

2024 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 5 日開催

発表者・発表題目

全メンバー 2024 年度の活動計画打合せ

第 2 回 2024 年 5 月 10 日開催

発表者・発表題目

于 拙 「アイヌ語テキスト・コーパスを利用した
アイヌ語の基本的な自然言語処理の試み」

第 3 回 2024 年 6 月 7 日開催

発表者・発表題目

藤田 郁 “Was it Alfred, Charles, or Frederick?:
Authorship questions surrounding Poems by Two Brothers (1827)”
李 晨婕 “A Stylometric Approach to the Sherlock Holmes Series and its Pastiches”

第 4 回 2024 年 7 月 5 日開催

発表者・発表題目

カミッレーリ “Evaluating Word Alignment Strategies in a Japanese-Italian Translation Corpus”
ガブリエレ
竹森 ありさ 「英語色彩語を含む強意直喩表現における 含意と媒体の詳述について」

第 5 回 2024 年 8 月 30 日開催

発表者・発表題目

岩根 久 「AI 支援による 16 世紀フランス語テキスト研究」
黄 晨雯 「日本語テキストに対する GraphRAG を用いたナレッジグラフ分析の試み」
田畑 智司 “Dickens in vector space: Word vectors for literary semantics”

第 6 回 2024 年 9 月 13 日開催

発表者・発表題目

山田 彬堯 「第一言語獲得・第二言語獲得と状態空間モデル」
杉山 真央 「「ソビエト語」と「プーチン語彙」：
戦勝記念日演説における「言説空間」の広がりと言彙による物語性の分析」
八野 幸子 「理科・外国語科連携による教科等横断的教育での
『種の起源』の活用可能性の検討 2」

第7回 2024年10月5-6日開催 英語コーパス学会第50回大会 於 青山学院大学
発表者・発表題目

- | | |
|------------------|--|
| 立野 寛太 | 「Hurford の制約に基づく “A or B”/“B or A” の形式差：
COCA データと R/Stan によるロジスティック回帰分析」 |
| 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」 |
| 李 晨婕 | “A Stylometric Approach to the Sherlock Holmes Series and its Pastiches” |
| 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—」 |
| 曹 芳慧 | 「TEI を用いた Hardy ウェセックス小説の会話部コーパス構築と可視化」 |
| 菅原 裕輝（・
神原一帆） | 「コーパス研究の論理と倫理: 学術的な目的のコーパス利用に着目して」 |
| 杉山 真央 | 「大祖国戦争の記憶の形成と感情: プーチン戦勝記念日演説にみる人称代名詞」 |

第8回 2024年11月1日開催
発表者・発表題目

- | | |
|------|---|
| 曹 芳慧 | 「Thomas Hardy の Wessex 小説の会話部による キャラクターライゼーション」 |
| 肖 媛媛 | 「英語政治ニュースにおける主観的バイアスの検出と分析」 |
| 徐 勤 | 「多次元分析による中国語現代詩歌の言語特徴の考察」 |

第9回 2024年12月6日開催
発表者・発表題目

- | | |
|-------|---|
| 王 簫影 | 「大規模言語モデルを利用した日本語母語識別/JNLI：
従来の自然言語処理手法との比較」 |
| 南澤 佑樹 | 「スウェーデン語の前置詞 framför の時間的意味について」 |

第10回 2024年12月21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

- | | |
|--|---|
| Tomoji | ポスターセッション |
| Tabata,
Natsuko
Yoshiga,
and Yuki Sug-
awara | “Teaching Digital Humanities: As an Advanced Minor Programme” |

第11回 2025年1月17日開催
発表者・発表題目

- | | |
|-------|--|
| 吉賀 夏子 | 「人文科学分野向け研究データ管理促進：
デジタルヒューマニティーズ学習教材開発プロジェクト」 |
| 陳 宇鎔 | 「専門家コミュニティと一般向けの科学技術日本語における
レジスターバリエーション—科学技術系の学術雑誌と「note」を対象とした多次元分析
(MDA) —」 |

第 12 回 2025 年 2 月 7 日開催
発表者・発表題目

Hodošček Bor “Token and Dependency Grammar-based Pattern Matching for Japanese”
立野 寛太 「Hurford の制約に基づく “A or B”/“B or A” の形式差：
COCA データと R/Stan によるロジスティック回帰分析」

第 13 回 2025 年 3 月 7 日開催
発表者・発表題目

Dimitra “Identifying Binary Oppositions in *the Lord of the Rings*: a primary approach”
Vogatza
浅野 元子 「A Vocabulary-Based Study Exploring Medical Students’ Interest in Research Abstracts
語彙的視点から見る医学生の論文抄録への関心」

第 14 回 2025 年 3 月 8 日開催 第 1 回 グローバル人文学 研究交流会
発表者・発表題目

曹 芳慧 「ハーディの Wessex 小説におけるキャラクターライゼーション—
文体意匠としての会話部—」
CHEN 「科学技術日本語は二つの場面においてどのように異なるのか？
YUKAI —科学技術系の学術論文とブログに対する Multi-Dimensional 分析—」

2025 年 5 月
研究代表者 田畑 智司