「テク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，テク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・ギリシャ語に拡がった。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 彬尭講師が着任・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「テクストマ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。

「テクストマ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，テクストアーカイヴの開発・構築，もう一つは構築したコーパス，テクストアーカイヴ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）などの電子テクスト化，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：デジタル化したテク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クストマイニング，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 彬尭，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，藤田 郁，竹森 ありさ，涌井 萌子，曹 芳慧，Camilleri Gabriele），博士前期課程在学生 5 名（小堀 彩夏，Vogatza 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），大阪医科薬科大学 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。

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「テク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"
Hodoščcek,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」
Vogatza,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ー長文テキストを対象にー」

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・ 吉賀 夏子・ 菅原 裕輝・ ホドシチェク ボル	ポスター・デモンストレーションセッション 「大阪大学における共同研究プロジェクト 『テクストマイニングとデジタルヒューマニティーズ』の歩み」

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・吉賀 夏子・石田 友梨・ホドシチェク ボル　「デジタルヒューマニティーズの現在（いま）」―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―

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