



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. 2021, 2020, p. 1-4
Version Type	VoR
URL	https://doi.org/10.18910/85005
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 彬亮講師の着任があり、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言語文化研究科の専任教員 5 名と名誉教授 1 名 (今尾 康裕, 田畑 智司, Hodošček Bor, 三宅 真紀, 山田 彬亮, 岩根 久名誉教授), 当研究科博士後期課程在学生 7 名 (黒田 絢香, 黄 晨雯, 岡部 未希, 徐 勤, 福本 広光, 藤田 郁, 三野 貴志), 博士前期課程在学生 3 名 (田辺 まりか, 竹森 ありさ, 王 鈺), 研究生 2 名 (曹 芳慧, 李 晨婕) に加え, OG の土村 成美氏 (2021 年 3 月当研究科より博士学位取得)・浅野 元子氏 (2020 年 3 月博士学位取得)・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 (本研究科博士課程修了), 摂南大学後藤 一章氏 (本研究科博士課程修了), 帝塚山学院大学八野 幸子氏 (本研究科博士課程修了), 数理・データ科学教育研究センターの上阪 彩香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, さらに統計数理研究所の言語系共同利用研究班との夏・春の合同セミナーの開催などを通して, 研究情報の交換, 論文や開発ツールのレビューなどを行っている。

2020 年度は, 新型コロナウイルスが世界中を席卷し, パンデミック下全ての研究会をオンラインで行わざるを得なくなった。前例のない研究会実施形式ではあったが, 次第に新たな形式に慣れていくとともに, 実は当プロジェクトの研究内容はオンライン発表に絶妙に適合しているのだということを実感した一年でもあった。もっとも, インフォーマルな雰囲気の中で情報交換を行う過程で思いがけない新たな着想, セレンディピティを生むオフライン, 対面での懇親の場を持つことができなかったのは残念であったけれども。2020 年度の本プロジェクト研究会開催記録を以下に記す。

2020 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

第 1 回 2020 年 5 月 29 日開催 (Zoom を介したオンライン)

発表者・発表題目

全メンバー 2020 年度の活動計画打合せ

第 2 回 2020 年 6 月 26 日開催 (Zoom を介したオンライン)

発表者・発表題目

Hodošček Bor 「言語処理とアノテーション」

第 3 回 2020 年 7 月 17 日開催 (Zoom を介したオンライン)

発表者・発表題目

田畑 智司 「R パッケージ Stylo によるテキスト分析」

第 4 回 2020 年 8 月 21 日開催 (Zoom を介したオンライン)

発表者・発表題目

八野 幸子 「理科との連携による教科等横断的視点を取り入れた英語教育のための語彙研究
—植物に関する文脈に出現する語彙を中心に—」

第 5 回 2020 年 9 月 11 日開催 (Zoom を介したオンライン)

発表者・発表題目

林 希和子 「留学生 (高等教育機関進学者) にむけた共通語彙の解明
—日本留学試験問題の語彙コーパスの作成と分析—」
藤田 郁 「Tennyson の -ly 副詞再考」

第 6 回 2020 年 10 月 9 日開催 (Zoom を介したオンライン)

発表者・発表題目

山田 彬堯 “Multinomial Mixed-Effects Models and Linguistic Variation:
Competitions among Japanese Subject-Honorific Constructions”
竹森ありさ “An Analysis of Intensifying Similes with Color Adjectives”

第 7 回 2020 年 11 月 4 日開催 (Zoom を介したオンライン)

発表者・発表題目

黄 晨雯 「Top2vec による小説の探索的研究」
岡部 未希 「ディキンソンの詩における鉱石の描写—ダイヤモンドに着目して—」

第8回 2020年11月29日開催 (Zoom を介したオンライン) Japan-Korea Collaborative Symposium
on Digital Humanities として開催
発表者・発表題目

Tomoji Tabata “Language Action Types and the Semantics of Texts:
Using rhetorical annotation to classify texts into meaningful groups”
Hodošček Bor “Visualization of Classical Japanese Poetic Vocabulary”
Iku Fujita “DH Approaches ‘to Know’ Tennyson”
Ayaka Kuroda “Exploring the Network of Topics, Words and Documents:
Machine Learning Analysis for Finding Key-Words”

第9回 2020年12月4日開催 (Zoom を介したオンライン)
発表者・発表題目

田辺 まりか 「日英機械翻訳における日本語省略表現訳出分析と事前学習モデルの改良
Wikipedia 日英京都関連文書対訳コーパスを用いて」
徐 勤 「日本人中国語学習者の叙述文における言語的特徴の分析—MF/MD 法を使って—」

第10回 2021年1月15日開催 (Zoom を介したオンライン)
発表者・発表題目

三野 貴志 「There comes の共時的・通時的・類型論的分析」

第11回 2021年2月5日開催 (Zoom を介したオンライン)
発表者・発表題目

三宅 真紀 「IIIF (International Image Interoperability Framework) 画像 API を用いた新約聖書
写本研究との連携」
浅野 元子 「研究論文抄録の英日対訳パラレルコーパスの言語使用と句読法の特徴について
—教育応用を視野に」
福本 広光 「19世紀アメリカ大統領スピーチにおける分離不定詞構造
—ABRAHAM LINCOLN の事例から—」

第12回 2021年3月5日開催 (Zoom を介したオンライン)
発表者・発表題目

王 鈺 「中国人日本語学習者と日本語母語話者における多義動詞「切る」の意味構造分析
—心理実験により意味分析の結果を検証する—」
高橋 新 “Study on the Application of Stylometric Methods to Analysing English Translations of the
Bible: 17+ x Gospels of Mark and John”
今尾 康裕 「次世代 CasualConc 開発版デモンストレーション」

第13回 2021年3月20日開催 (Zoom を介したオンライン) 統計数理研究所言語系共同研究班
合同セミナー「言語研究と統計 2021」として開催
発表者・発表題目

竹森ありさ	「英語色彩語を含む強意直喩表現の分析」
福本 広光	「アメリカ大統領演説における分離不定詞」
今尾 康裕	「ジャンルの違いはコロケーションにどのように反映されるのか」
山田 彬堯	「ラレル尊敬の特徴は何か？」
黄 晨雯	「Top2Vec による小説の探索的研究」
岡部 未希	「Dickinson の詩における鉱石」
黒田絢香	「トピックモデルによる文学作品分析」
浅野 元子	「国際英語としての英語学術論文のコーパス研究 —医学分野への日本からの貢献を例に—」
高橋 新	「英語翻訳聖書間の計量的スタイル及び語彙選択の分析 —マルコ及びヨハネによる 両福音書 17 + α 翻訳の分析」
王 鈺	「中国人日本語学習者と日本語母語話者における多義動詞「切る」の意味構造分析 —心理実験により意味分析の結果を検証する—」
八野 幸子	「理科との連携による教科等横断的英語教育のための語彙・表現研究」
徐 勤	「日本人中国語学習者の叙述文における言語的特徴の分析 —MF/MD 法を使って—」

2021 年 4 月
研究代表者 田畑 智司