「テクストマイニングとデジタルヒューマニティーズ2019」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 年度に岩根教授，緒方助教授，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。2010 年にはサイバーメディアセンターの森助教が加わり，翌 2011 年には言語文化教育論講座に新たに着任した今尾講師が加入した。さらに，2014 年度後期から新メンバーとして Bor Hodošček 講師が加わり，現在の陣容ができあがった。(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，コロケーション，意味構造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言語文化研究科の専任教員 5 名（岩根　久，三宅 真紀，今尾 康裕，Bor Hodošček, 田畑 智司), 当研究科博士後期課程在学生 4 名（土村 成美, 浅野 元子 , 黒田 絢香, 三野 貴志), 博士前期課程在学生 2 名（岡部 未希，田辺 まりか)，科目等履修生 1 名（福本 広光）に加え，本学非常勤講師の高橋 新氏 , 南澤 佑樹氏（本研究科修了)，摂南大学後藤 一章氏（本研究科修了)，帝塚山学院大学八野 幸子氏（一昨年 3 月当研究科より博士学位授与)，本学基礎工学研究科の上阪 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，さらには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プレビューなどを行った。2019 年度の研究会の開催記録を以下に記す。

2019 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 2019」研究会開催記録

第 1 回　2019 年 4 月 26 日開催「2019 年度の共同研究計画についての打合せ」

第 2 回　2019 年 5 月 24 日開催

発表者・発表題目

田畑 智司　“Topic modelling and body language”

第 3 回　2019 年 6 月 21 日開催

発表者・発表題目

三野 貴志　「There speak 構文の対人機能と使用文脈 ~Loquor, Ergo Sum~」

岡部 未希　「「Emily Dickinson の詩における二人称代名詞」

第 4 回　2019 年 7 月 26 日開催

発表者・発表題目

岩根　久　「語彙計量的手法を日常のテキスト分析に」

黄　晨雯　「中国のミステリー小説におけるトピック解析の試み」

第 5 回　2019 年 8 月 9 日開催

発表者・発表題目

ホドシチェク ボル “NLP with spaCy: Short overview of recent advances and preliminary notes on the ‘Standard Ebooks’ corpus ”

第 7 回　2019 年 9 月 6 日開催

八野 幸子　「共起ネットワークによる昆虫の世界の可視化の試み」

福本 広光　「分離不定詞における splitter 副詞の意味的分類と通時的変遷」

第 7 回　2019 年 9 月 22–23 日開催（統計数理研究所共同利用研究班との合同中間報告会として開催）

発表者

岡部 未希　「Emily Dickinson の詩における Thou と You ーコ ロケーション分析を用いてー」

浅野 元子　「日本からの英語医学論文における collocation framework “the...of” の検討」

南澤 佑樹　「アイスランド語均衡コーパスを用いた前置詞の意味記述に向けて」

田畑 智司　“Visualising micro-topics in fiction”

今尾 康裕　「CasualConc への Stanford CoreNLP 機能実装について」

黒田 絢香　「Sherlock Holmes シリーズの語彙分析」

八野 幸子　「教科横断的視点を取り入れた小学校 外国語活動・外国語のための語彙研究 ー子供向け文章の中の昆虫関連語彙ー」

第 8 回　2019 年 10 月 11 日開催

発表者・発表題目

南澤 佑樹　「アイスランド語に見られる Semantic Prosody」

第 9 回　2019 年 11 月 15 日開催

発表者・発表題目

黒田 絢香　「LDA によるトピックモデルの概要と応用」

第 10 回　2020 年 1 月 10 日開催

発表者・発表題目

浅野 元子　「引用の仕方の予備研究―学術論文への橋渡しを目指して」

第 12 回　2020 年 2 月 14 日開催

発表者・発表題目

今尾 康裕

上阪 彩香

第 13 回　2020 年 3 月 13 日開催

発表者・発表題目

三宅 真紀

高橋　新

2020 年 6 月

研究代表者　田畑 智司