OUKA
The University of Osaka
Institutional Knowledge Archive

Title	「テクストマイニングとデジタルヒューマニティーズ2017」 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. 2018, 2017, p. 1-4
Version Type	VoR
URL	https://doi.org/10.18910/69897
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

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

The University of Osaka

「テクストマイニングとデジタルヒューマニティーズ 2017」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, 囚畑 智司)，当研究科博士後期課程在学生 5 名（杉山 真央, 南澤 佑樹, 夏　天驕, 土村 成美, 浅野 元子)，博士前期課程在学生 3 名（黒田 絢香，藤原 いずみ，福本 広光）に加え，本学非常勤講師の高橋 新氏 , 摂南大学の後藤 一章氏（本研究科修了)，帝塚山学院大学の八野 幸子氏（本年 3 月当研究科より博士学位授与)，同志社大学（2017 年 10 月に本学データビリティフロンティア機構・特任助教に着任）の上阪 彩香氏を主たる構成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。2017 年度の研究会の開催記録を以下に記す。

2017 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 2017」研究会開催記録

第 1 回　2017 年 4 月 28 日開催「2017 年度の共同研究計画についての打合せ」

第 2 回　2017 年 5 月 26 日開催

発表者・発表題目

岩根　久　「テキスト理解のために計量的手法を―Mac 用フリーソフト CasualConc を用いたラブレーとモンテーニュの語彙の比較を例に」

第 3 回　2017 年 6 月 23 日開催

発表者・発表題目

土村 成美　“How Did Agatha Christie Describe Women?: Behaviour of *She* in Christie’s Novels”

第 4 回　2017 年 7 月 7 日開催

発表者・発表題目

黒田 絢香　“Machine-Learning Approaches to Literary Works: Novels of Sir Arthur Conan Doyle”

藤原いずみ「日本近現代文学作品の計量研究」

第 5 回　2017 年 8 月 18 日開催

発表者・発表題目

浅野 元子　“Linguistic Features of English in Medical Research Article Abstracts Written by Japanese Authors”

高橋　新　“Study on the Application of Stylometric Methods to Analysing English Translations of the Bible: Gospels of MARK and JOHN”

第 6 回　2017 年 9 月 4–5 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 於 大阪大学）

発表者

浅野 元子　「国際英文医学誌における日本人著者と中国人著者による言語使用の特徴」

高橋　新　「英語翻訳聖書間の計量的スタイル分析の考察―マルコ及びヨハネによる両福音書の分析を通して―」

田畑 智司　“A stochastic analysis of topic distribution across text sets”

黒田 絢香　「文学作品分析におけるトピックモデルの応用」

土村 成美　「トピックモデルを用いた Agatha Christie 作品の予備的分析」

八野 幸子

第 7 回　2017 年 9 月 8 日開催

南澤 佑樹　「怒りを表す類義語と概念メタファー」

第 8 回　2017 年 10 月 7 日開催

発表者・発表題目

夏　天驕「中国語擬音語の反復とアスペクト」

第 9 回　2017 年 11 月 10 日開催

発表者・発表題目

上阪 彩香　「テキストマイニングを用いた西鶴『万の文反古』版下の系列における相違点の検討」

杉山 真央　“Russian Presidents versus Russian press: Applying text-mining approaches to explore different points of view”

第 10 回　2017 年 12 月 8 日開催

発表者・発表題目

田畑 智司　（ポスター発表 1）“Mapping Classic Fiction in Networks: Key Words, Topics, and Distant Reading”

（ポスター発表 2）“‘Birds of a feather flock together ...’ : Literary Vocabulary in Vector Space”

第 11 回　2018 年 1 月 19 日開催

発表者・発表題目

Bor Hodošček “High/Low-Brow Corpus Subset from the Aozora Bunko Online Collection: Preliminary results from complexity, stylistics, and content analysis-based methods on the Japanese novels corpus”

第 12 回　2018 年 2 月 16 日開催

発表者・発表題目

今尾 康裕　「英日バイリンガルエッセイコーパスで遊んでみる」

第 13 回　2018 年 3 月 20 日開催

発表者・発表題目

三宅 真紀　「決定木モデルによる新約聖書ギリシャ語校訂本の分類ー言語研究と統計 2018 の発表に向けてー」

第 14 回　2018 年 3 月 29–30 日「言語研究と統計 2018」（於 統計数理研究所）として開催

発表者・発表題目

浅野 元子　「英語医学論文考察部における日本人著者と中国人著者の言語使用の特徴」

杉山 真央　「エリツィン大統領の年次教書と『独立新聞』の報道記事」
土村 成美　「イギリスにおけるミステリー黄金時代の女性作家作品の比較分析－ Christie, Sayers と Allingham －」
南澤 佑樹　「恐怖のメタファー・メトニミー」
上阪 彩香　「『万の文反古』における版下の系列と文章の特徴」
ホドシチェク ボル「青空文庫と DBpedia を合わせた近現代小説コーパスの書誌情報別分類実験 : 計量文献学における Julia 言語の可能性」
後藤 一章　「統語解析に基づく NS と NNS における名詞使用の差異」
三宅 真紀　「語彙多様性指標からみる新約聖書ギリシャ語校訂本の分類」
高橋　新　「英語翻訳聖書間の計量的スタイル及び語彙分析結果の考察―マルコ及びヨハネによる両福音書の分析を通して―」
黒田 絢香　「トピックモデルによる特徴語抽出の試み：アーサー・コナン・ドイルの作品から」
今尾 康裕　「日本の英語学習者によるエッセイでの文レベルの接続表現を探る―日本語エッセイ・英語母語話者によるエッセイと比較して―」
田畑 智司　“Dickens in Vector Space”

2018 年 5 月
研究代表者　田畑 智司