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 2022（冊子）
Author(s)	
Citation	言語文化共同研究プロジェクト. 2023, 2022
Version Type	VoR
URL	https://hdl.handle.net/11094/91697
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

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

The University of Osaka

言語文化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
2022

菅原　裕輝

藤田　　郁

涌井　萌子

曹　　芳慧

田畑　智司

大阪大学大学院言語文化研究科

2023

言語文化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
2022

目　次

田畑　智司　プロジェクトの目的と活動 1

菅原　裕輝　南三陸の復興に関するハイパーリンク・エスノグラフィ . . 5

藤田　郁　A Study on Characteristic Sounds in Tennyson and Browning: Using Stylometric Approaches 19

涌井　萌子　匿名政治パンフレットの計量的分析
―「レ枢機卿のマザリナード」の帰属検証― 41

曹　芳慧　*Tess of the d'Urbervilles* の会話部による
キャラクタライゼーション 59

「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 彬尭講師が着任し，言語文化研究科と文学研究科の統合により人文学研究科が設立された 2022 年には人文学林より菅原 裕輝特任助教が加入し，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 6 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，岩根 久名誉教授），当研究科博士後期課程在学生 7 名（黄 晨雯，岡部 未希，徐 勤，福本 広光，藤田 郁，竹森 ありさ，涌井 萌子），博士前期課程在学生 4 名（小堀 彩夏，曹 芳慧，Vogatza Dimitra, 李 晨婕），研究生 1 名（肖 媛媛）に加え，OG の大阪医科薬科大学 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（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），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。

2022 年度は，前年度に引き続き，10 月までの研究会をオンラインで開催したが，11 月に待望の対面開催（オンライン併用のハイブリッド方式）に移行し，以後は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。

2022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　2022年4月15日開催
発表者・発表題目

全メンバー　2022年度の活動計画打合せ

第2回　2022年5月20日開催
発表者・発表題目

田畑 智司　「Using topic models to shed new light on body language in Dickens's fiction

第3回　2022年6月10日開催
発表者・発表題目

藤田 郁　「トピックモデルを用いたAlfred Tennysonの韻文研究」

第4回　2022年7月8日開催
発表者・発表題目

曹 芳慧　「Thomas Hardy, *Tess of the d'Urbervilles* の文体への語彙的アプローチ
徐 勤　「中国語作文における言語特徴の考察:
多次元分析による日本人中国語学習者と中国語母語話者の作文の比較」

第5回　2022年8月5日開催
発表者・発表題目

三野 貴志　「一般動詞を伴うthere構文
〜コーパス調査から見えてくること〜」
岩根 久　「Ronsardのソネの比較に向けて
Les Amours (1552) と *Les Amours* (1553)」

第6回　2022年9月2日開催
発表者・発表題目

八野 幸子　「教科等横断的視点を取り入れた英語教育支援ツール作成に向けた語彙研究」
山田 彬尭　「適用形の通時的構文交替：「させていただく」「させてもらう」「させてくださる」
「させてくれる」の選択に対する状態空間モデルを用いた時系列分析」
Hodoščeк Bor　「英語電子書籍テキスト処理
'Standard Ebooks' とspaCyを例に

第7回　2022年10月1日開催
発表者・発表題目

竹森ありさ　「色彩語whiteを含む強意直喩表現の 意味と媒体の背景について」
李 晨婕　「コーパスに基づくシャーロック・ホームズシリーズと
そのパスティーシュの文体的対比研究」

第 8 回　2022 年 10 月 2 日開催　英語コーパス学会第 47 回大会
発表者・発表題目

黒田 絢香	「トピックモデル可視化ツールの開発に向けて」
藤田 郁	“LDA Topic modelling of Tennyson’s Poetry”（学生優秀発表賞受賞）

第 9 回　2022 年 11 月 4 日開催
発表者・発表題目

菅原 裕輝	「研究紹介」
ハラルド　クマレ	Operationalizing topic models for writing conceptual history

第 10 回　2022 年 12 月 2 日開催
発表者・発表題目

小堀 彩夏	「英訳版 村上春樹作品における特徴の比較」
福本 広光	「米国一般教書演説に見られる分離不定詞の実態と効果について」

第 11 回　2023 年 1 月 6 日開催
発表者・発表題目

浅野 元子	「日本語論文の英語抄録における言語使用の調査： 日本語抄録の動詞的名詞に着目して」
王 鈺	「日本語における 2 種類の分離動詞の違い ー分離元と分離物の関係および構文的意味についてー」
Vogatza Dimitra	「イギリス文学における PITY の訳語の比較研究」

第 12 回　2023 年 2 月 10 日開催
発表者・発表題目

今尾 康裕	「CasualMallet の機能紹介」
竹内 綾乃	「登場人物のことばから探る人物造型論の可能性 ―源氏物語における六条御息所―」

第 13 回　2023 年 3 月 15 日開催
発表者・発表題目

涌井 萌子　「匿名政治パンフレットの計量的分析
—「レ枢機卿のマザリナード」の帰属検証—」
塚越 柚季　「サンスクリット文献『リグ・ヴェー ダ』の韻律分析用データベースと
『リ グ・ヴェーダ』の文書間類似度」
出川 英里　「KHCODER を利用した裁判判例の数量分析
ー 19 世紀エジプトの判例集を素材としてー」

2023 年　5 月
研究代表者　田畑 智司

南三陸の復興に関するハイパーリンク・エスノグラフィ

菅原裕輝

大阪大学大学院人文学研究科

〒 560-0043 豊中市待兼山町 1-8

E-mail: ysugawara.hmt@osaka-u.ac.jp

あらまし 本研究の目的は、東日本大震災の被災地である宮城県南三陸町の復興に関するオンライン上のグラスルーツな活動について、デジタルエスノグラフィ（とりわけ，ハイパーリンク・エスノグラフィ）の観点から質的かつ量的な側面を捉えたエスノグラフィを書き、南三陸の復興を巡るデジタルパブリックの様態を記述・分析することである。具体的には，オンライン上の活動の全体像の可視化と質的に仔細な記述を進め、南三陸の復興を巡るデジタルパブリックが様々なアクターの密な結びつきにより形成されていることを示す。

キーワード デジタルエスノグラフィ，デジタルメソッド，南三陸

Hyperlink Ethnography on the Reconstruction of Minamisanriku

Yuki Sugawara

Graduate School of Humanities, University of Osaka

1-8 Machikaneyama-cho, Toyonaka, Osaka, 560-0043 Japan

Abstract The primary objective of this research is to compose an ethnographic account that captures both qualitative and quantitative aspects of grassroots online activities related to the reconstruction of Minamisanriku, a town in Miyagi Prefecture that was devastated by the Great East Japan Earthquake. The study will be conducted from the perspective of digital ethnography, with a particular emphasis on hyperlink ethnography, in order to describe and analyze the nature of the digital public sphere surrounding the recovery of Minamisanriku. More specifically, the research aims to visualize the overall picture of online activities and provide an in-depth qualitative description, thereby illustrating that the digital public sphere associated with Minamisanriku's reconstruction is shaped by the intricate connections among various actors involved in the process.

Keywords Digital ethnography, Digital methods, Minamisanriku

1. 導入

本研究の目的は、東日本大震災の被災地である宮城県南三陸町の復興に関するオンライン上のグラスルーツな活動について、デジタルエスノグラフィ（とりわけ，ハイパーリンク・エスノグラフィ）の観点から質的かつ量的な側面を捉えたエスノグラフィを書き、南三陸の復興を巡るデジタルパブリック（digital public）(Boyd 2010) の様態を記述・分析することである。

南三陸町は町の中心機能の大半（5362 戸のうちの 62 %に及ぶ 3143 戸の建物や事業所）が津波によって流されてしまったために，町のデザインを新たに考え直す必要に迫られた町であり，町民会議や「かもめの虹色会議」などを通して町民の意見を町のデザインに反映させてきたという特色がある。山と海が近く，森・里・海の自然豊かな環境に惹かれた移住者も増えてきており，近年、南三陸において、移住者が中心となったオンライン上で様々な活動が行われてきている。オンライン上での活動は現地でのフィールドワークだけでは追いきれない部分であり、**オンライン上でのフィールドワーク**（online fieldwork）が必要となっている。

他方で、人類学分野においても、デジタル技術の社会への急速な浸透を背景とした**デジタル人類学**（digital anthropology）と呼ばれる新たな研究アプローチが現れている (Pink et al. 2015)。デジタル人類学においては、伝統的な人類学の手法を基にオンライン上でのフィールドワークを行う質的アプローチ (Hallett and Barber 2014) と、データサイエンスの手法を活用しオンライン上の構造を適切に踏まえた上でデータを大規模に収集し分析を行う量的アプローチ (Caliandro 2018, Rogers 2019) が存在する。本研究では両方のアプローチを組み合わせ (Marres 2017)、南三陸の復興に関するウェブページにおいてフィールドワークを行う。エスノグラフィの作成に際しては、5 つの観点から論争を可視化する**論争マッピング**（controversy mapping）(Venturini and Munk 2021) を採用する。具体的には、(i) どんなトピックについて議論しているか（What)、(ii) 誰が議論しているか（Who)、(iii) どのような関係性を形成しているか（How）を記述することで、研究対象についての closed な記述と distant な記述を両立させる。

ウェブページのエスノグラフィとして、本研究ではまず、(1) 南三陸に関する様々な情報を扱うローカルなメディアとなっているウェブサイト「南三陸なう」(`https://m-now.net`) を最初のフィールドサイトとし、そこでどのようなトピックが扱われているかを質的に分析する。具体的には、それぞれの記事に付いているタグをリスト化し、記事の内容を踏まえて、(a) 防災、(b) 福祉、(c) なりわい、(d) 移住、(e) まちづくり、(f) 復興、(g) 自然体験、(h) イベントという 8 つのカテゴリーがあることを発見した。そのうえで、(2) それぞれのカテゴリーの記事で対象としている人々を質的に分析し、(i) 先住者、(ii) 移住者、(iii) 観光者といった 3 種類の人々へ向けた情報共有を行っていることを同定したうえで、それぞれのアクターをリスト化した。(3) ウェブページのリストを基にして、それぞれのウェブページ内に含まれているハイパーリンクを抽出し、ハイパーリンク間の結びつきをネットワークグラフの形で示し，その結果を解釈する。

デジタル人類学の観点から南三陸の復興に関するウェブページ上のグラスルーツな活動を記述することを通して、オンライン上の活動の全体像の可視化と質的に仔細な記述を進め、南三陸の復興を巡るデジタルパブリックが様々なアクターの密な結びつきにより形成されていることを示す。

2.　背景：ハイパーリンク・エスノグラフィとは何か

2.1.　デジタルエスノグラフィ概説

ハイパーリンク・エスノグラフィは，デジタルエスノグラフィの一分野として位置づけられるため，ここではまず，デジタルエスノグラフィがどのような分野なのかを簡潔に説明する。

まず，デジタル・エスノグラフィは，二つの研究伝統の観点から整理することが可能である (Marres 2017:ch.3)。一つは**方法のデジタル化**（digitalization of methods）と呼ばれる研究伝統であり，そこではエスノグラフィの伝統的な方法や概念をオフライン空間だけでなくオンライン空間のなかでも実現させることが目指されている (Hallett and Barber 2014)。もう一つは**デジタル・メソッド**（digital methods）と呼ばれる研究伝統であり，そこではオンライン空間（ウェブページやソーシャルメディア）の特性を踏まえたうえでオンライン空間からデータを収集・分析することが目指されている (Rogers 2019)。前者はエスノグラフィの研究手法をオンライン空間に適用する方向性を持つが，後者はオンライン空間の特性を踏まえ伝統的なエスノグラフィの方法よりもデータサイエンスの方法がデータ収集やデータ分析で用いられるという方向性を持つ。両者の性格は異なるが，実際のデジタルエスノグラフィにおいては両者の視点や方法を組み合わせながら，**方法論的な学際性**（methodological interdisciplinarity）を体現させる形で調査が進められることが多いと思われる。

デジタルエスノグラフィにおいては，いくつか重要視されている鍵概念が存在するため，それらについても簡単に触れておきたい。第一に，**デジタルパブリック**（digital public）は「ネットワーク化されたテクノロジーを通して構築される社会的空間」(Boyd 2010:p.39) であるとともに，「人々・テクノロジー・実践が交差する結果として生じる想像的な集合」(Boyd 2010:p.39) として定義される概念である。自己をネットワーク化されるテクノロジーを媒介して成立するという認識に立ち，ヒト以外の様々なアクター（技術やモノなど）の存在も重要視されている。第二に，**自己表現**（self presentation）は，デジタル・パブリックの前でユーザーが自ら「自己表現をする」(Arvidsson and Caliandro 2016:p.7) ことにより社会的なアイデンティティを認識する行為である。ソーシャルメディアなどにおけるユーザーの活動の動機づけとして用いられることが多い概念である。第三に，**オンラインコミュニティ**（online community）は，ハッシュタグやリツイートを介して一時的に形成されるユーザー間の繋がり (Caliandro 2018) を指す概念である。オンラインコミュニティはしばしば，相互に結びついた社会的関係性として実現するものとしてのオフライン（対面）の「コミュニティ」と対比的に用いられることが多い概念である。第四に，**オンライン群衆**（online crowds）は，ハッシュタグやリツイートといったオンライン上の「場所」に情動的な繋がりの結果として生じる群衆 (Stage 2013) を指す概念である。オンライン群衆もまた，共通の活動を共有して物理的に集まる人々としてのオフラインの「群衆」と対比されることが多い概念である。第五に，**オンラインパブリック**（online public）は，ある特定の事象や政治的問題，ブランドといったものに対して向けられた注意の集まり (Arvidsson 2013) を示す概念である。オンラインパブリックもまた，直接的な相互作用によって結びついているものとしてのオフラインの「公衆」と対比されることが多い概念である。

デジタルエスノグラフィの方法に関しても分類がなされている。一つは，**文脈的フィールドワーク**（contextual fieldwork）であり，伝統的なエスノグラフィの方法をオンライン空間のフィールドワークにも適用することで，オンライン・サイトとオフライン・サイトの両方を探求し，マルチ・サイテッド・エスノグラフィを行うことを意図しているものである (Hallett and Barber 2014)。もう一つは，**メタフィールドワーク**（meta fieldwork）であり，「メタデータに基づいて組織・再配置された，脱文脈化したコミュニケーションの形跡の流れ」(Airoldi 2018:p.9) の中で行われるフィールドワークである。メタフィールドワークは，ハッシュタグやリツイートのネットワークなど，常に流動的な空間において調査を行うため，“unsited” なフィールドワークと呼ばれる場合もある。

2.2. デジタルエスノグラフィの一分野としてのハイパーリンク・エスノグラフィ

デジタルエスノグラフィには，ウェブページを結ぶハイパーリンク・ネットワークを追うフィールドワークである**ハイパーリンク・エスノグラフィ**（hyperlink ethnography）(Beaulieu 2005, Jacomy et al. 2016, Ooghe-Tabanou et al. 2018) と，ソーシャルメディア上でフィールドワークを行う**ソーシャルメディア・エスノグラフィ**（social media ethnography）(Postill and Pink 2012) という二つの分類がされる。オンラインに存在するデジタル・データを収集・分析する**デジタルメソッド**（digital methods）Rogers (2013, 2019) においても同様の分類がされており，デジタルメソッドにおいても，それらが扱うインフラストラクチャーの特性（infrastructural properties）に基づき**ハイパーリンク研究**（hyperlink studies）と**プラットフォーム研究** (platform studies）といった分類がされている。

デジタルメソッドにおいて，ハイパーリンク研究は，オンライン上のアクター（ウェブページ）の関係性をネットワーク上に記述・分析することが目指されている。そこで扱われるインフラストラクチャーの特性は，ウェブページという安定したオンライン空間であり，ハイパーリンクを介して密に結びついた存在としてコミュニティが捉えられている傾向がある。また，ウェブページを持つのは制度化された組織や，技術力のある個人に限られる傾向があるため，大きなアクター（組織・大きなグループ）を把握することに有効とされている。一方で，プロットフォーム研究は，twitter や instagram，facebook などのソーシャルメディア上で収集可能なデータを分析することが目指されている。そこで扱われるインフラストラクチャーの特性は，ソーシャル・メディアという一過性のあるオンライン空間であり，（ハイパーリンク研究と比べると）安定したコミュニティは存在せず，個人が分散して存在し，情動（affect）によって一時的に団結するといった傾向性がある (Papacharissi 2015)。また，ソーシャル・メディアは現在誰でもすぐに利用可能であるため，小さなアクター（個人・小さなグループ）を把握することに有効とされている。本研究で採り上げる方法は，ハイパーリンク研究に当てはまる方法であるが，分析したいインフラストラクチャーの特性を踏まえて方法を決める（あるいは，組み合わせる）のが望ましいと思われる。

ハイパーリンク・エスノグラフィを通して，(1) ウェブページがどのように関係しあっているかに関するネットワーク分析や，(2) ウェブページの内容に対する質的分析を基にし

て，ウェブページというアクターが社会的問題に対してどのような関与をしているかに関して綿密な記述を行うことが可能になる (Beaulieu 2005)。ウェブページのハイパーリンクを収集する際には，**スクレイピング**（scraping）と呼ばれる，ダウンロードしたウェブページから必要な情報を抜き出す手続きと，**クローリング**（crawling）と呼ばれる，ウェブページのハイパーリンクを次々に辿っていく手続きを行う。具体的には，(1）入力となるウェブサイト（root page）からハイパーリンク（URL）の情報を抜き出し（スクレイピング），(2）ハイパーリンク先のページの情報をダウンロードし（クローリング），そのページからハイパーリンクの情報を抜き出し（スクレイピング），(3）さらにそのハイパーリンク先のページをダウンロードし（クローリング），そのページからハイパーリンクの情報を抜き出す（スクレイピング）といった過程を指定した回数だけ実行することで，ウェブページとウェブページの間の一方向的な結びつきに関するデータを取得し，ForceAtlas2Jacomy et al. (2014) を用いてネットワーク（有向グラフ）として表現する。

ハイパーリンク・エスノグラフィを通して可能になることとして，以下の 4 つが挙げられる。

（1）**アクターの発見・同定**：スクレイピングとクローリングを用い，ウェブページ内に含まれるハイパーリンクを機械的に全て収集することで，人間では処理できない大量のウェブページを分析対象にすることができる。Guha et al. (2003) は，検索エンジンを用いた探索には，**ナビゲーション検索**（Navigational Search）と**研究検索**（Research Search）という二種類あることを指摘した。ナビゲーション検索は，すでに知っている情報がどこに位置するか（どう位置づけられているか）を把握することを目指す。研究検索は，情報を新たに集めようと試みる手段として検索エンジンを用いることを指す。まだ分かっていない情報がどこにあるかを探索することを目指す。(1）アクターの発見・同定は，この「研究検索」をクローリングとスクレイピングの手法を用いて行った形になる。

（2）**アクター間の関係性の記述**：引用・被引用・相互引用などの関係性の有無や程度を記述し，重要度の高いアクターを同定することが可能になる。(2）アクター間の関係性の記述は，検索エンジンを通して行われる「ナビゲーション検索」とは別の基準（ハイパーリンク）で関係性を記述し直したものだと捉えることが可能である。

（3）**コミュニティの抽出**：統計的な処理を基にして，ネットワーク全体にどのようなサブグループ（コミュニティ）が存在するかを記述することが可能になる。

（4）**ネットワーク全体の特徴の記述**：ネットワークがどれくらい密に結びいているかを記述することが可能になる。すべてのアクターが互いに結びついている状態が密度 1 の状態（密な状態），すべてのアクターが全く結びついていない状態が密度 0 の状態（疎な状態）である。

2.3. ハイパーリンク・エスノグラフィのツール：Hyphe

ハイパーリンク・エスノグラフィを可能にするツールとして，Hyphe が挙げられる。Hyphe はフランスの médialab Sciences Po によって開発されているデジタルメソッドの一つである。

Hyphe を用いることにより，ウェブページのハイパーリンクのスクレイピングやクローリングを容易に行うことができる (Jacomy et al. 2016:p.595)。Hyphe 以外にも，IssueCrawler(Rogers 2013) や SocSciBot(Thelwall 2009)，Voson(Ackland et al. 2006) といったような社会科学研究に用いられるツールも存在するが，スクレイピングやクローリングをより高い水準で行えるのは Hyphe だけだとされている (Jacomy et al. 2016:p.596)。

Hyphe は分析対象とするデータを**コーパス**（corpus）と呼ぶ (Jacomy et al. 2016:p.596)。収集したハイパーリンクなどのデータを「コーパス」に入れるかどうかについては研究者側で決めるため，研究者側でデータの内容や質を調整することが可能である。また，Hyphe は**ウェブ・エンティティ**（web entities）という単位でデータを扱う (Ooghe-Tabanou et al. 2018:pp.13-14)。ウェブページという静的な存在ではなく，ウェブエンティティという主体性を帯びた存在として捉え直すことは，研究対象であるウェブ上に存在するデータ（ウェブページ）を人間の社会活動と密接な繋がりを持ったモノ（ないしは，アクター）として扱うことを可能にすると考えられる。さらに，Wikipedia や YouTube のような巨大なウェブサイトはクローリングに大きな影響を与えるため，コーパスから外すことも可能であり (Jacomy et al. 2016:p.597)，逆に，最初に入力したリストに含まれないウェブサイトをコーパスに入れ，コーパスを段階的に拡大させていくことも可能である。整理すると，Hyphe は，データ収集から柔軟なコーパス構築，ネットワークグラフの視覚化などが可能であり，量的かつ質的にウェブページを分析することを可能にするツールである。

Hyphe を用いたハイパーリンク・エスノグラフィは，以下のような手続きで進められる。

（0）**Hyphe の環境構築**：Hyphe は Mac，Windows，Linux といった OS に環境構築を行うことが可能である。Hyphe の公式ウェブページでは，Mac や Windows の OS で Docker を利用した環境構築が推奨されている。環境構築の詳細については公式ウェブページを参照されたい（`https://github.com/medialab/hyphe`）。

（1）**ウェブページのリストの作成**：研究関心に基づき，ウェブページの名前と URL のリストを作成する（txt 形式や csv 形式のファイルの作成）。

（2）**ウェブページの質的分類**：ウェブページの内容に基づき，ウェブページの質的な分類を行う。

（3）**Hyphe でのクローリング（CRAWL 機能）**：Hyphe を起動し，(1) で作成したウェブページのリストを読み込み，クローリングを行う。

（4）**質的な分類の反映（TAGS 機能）**：（2）で行った質的な分類を TAGS 機能を用いて反映させ，Hyphe 上でグループ化を行う。

（5）**分類したグループ間の関係性の量的分析（TOOLS 機能）**：（4）で作成したグループの間の量的な関係性を TOOLS 機能を用いて確認することができる。

（6）**ウェブページ間の関係性の量的分析（NETWORK 機能）**：特定のウェブページとウェブページの間の量的な関係性を NETWORK 機能を用いて確認することができる。

（7）**ウェブページの内容や関係性の質的分析**：（5）や（6）で分かった量的な関係性を踏まえて，ウェブページの質的な内容や関係性を改めて分析する。

以上の7つの手続きをしばしば繰り返すことで（すなわち，(7）ウェブページの質的分析や（3）ウェブページのクローリングを行ったあとに，その結果を踏まえて，(1）ウェブページのリストの作成に戻るなど)，Hyphe内の「コーパス」を段階的に拡大させていき，ウェブページに関する質的かつ量的な分析を行っていく。なお，第3節で調査の内容を示す際は，以上の手続きと，論争マッピングで示されたいくつかの観点（(i）どんなことについて，(ii）誰が，(iii）どのように関わり合っているか）に対応させる形を採ることとする。

3. 南三陸の復興に関するハイパーリンク・エスノグラフィ

3.1. 南三陸の復興を巡る（i）Whatの記述

ウェブページのエスノグラフィとして、本研究ではまず、南三陸に関する様々な情報を扱うローカルなメディアとなっているウェブサイト「南三陸なう」(`https://m-now.net`）を最初のフィールドサイトとし、そこでどのようなトピックが扱われているかを質的に分析した。南三陸なうは，南三陸町の復興の過程を記録として残すためのアーカイブとして，南三陸へ移住してきた方々が南三陸町からの業務委託という形で運営されてきているウェブサイトである。2010年代中頃からブログでの情報発信を頻繁に行ってきており，2010年代後半からはYouTubeでの動画配信を中心に運営がされてきている。本研究では，南三陸なうの記事で「タグ」として使用されたトピックをリスト化し，その内容を質的に整理することを通して，南三陸の復興を巡るトピックの記述をするところから調査を開始した。具体的には、大きく分けて，(a）防災、(b）福祉、(c）なりわい、(d）移住、(e）まちづくり、(f）復興、(g）自然体験、(h）イベントという8つのカテゴリーがあることを発見した。なお，質的なコーディングを行う際は，MaxQDAという質的分析の際に用いるソフトウェアを使用した（図1を参照)。

3.2. 南三陸の復興を巡る（ii）Whoの記述

次に，南三陸の復興を巡るアクターの記述を行う。それぞれのカテゴリーの記事で対象としている人々を質的に分析し、(i）先住者（現地住民)、(ii）移住者、(iii）観光者といった3種類の人々へ向けた情報共有を行っていることを同定したうえで（図2を参照)、それぞれのアクターであるウェブページを調査し，リストを作成した（表1を参照)。なお，リスト自体はクローリング後に新たに発見されたウェブページを追加するという形で，クローリングを経て拡大していくという性質を持つ。

3.3. 南三陸の復興を巡る（iii）Howの記述

次に，南三陸の復興を巡るアクターがどのように繋がっているかを記述する。ここでは，ウェブページのリストを基にしてスクレイピングとクローリングを行い，ハイパーリンクの引用関係を表したネットワークグラフを出力した（図4を参照)。ノードの大きさは入次数（引用されている数）の大きさを意味している。紫色が地元住民へ向けたウェブページ，赤色が移住者へ向けたウェブページ，緑色が観光者へ向けたウェブページである。観光者へ向けたウェブページがグラフの下側に集まり，移住者へ向けたウェブページと地元住民へ向け

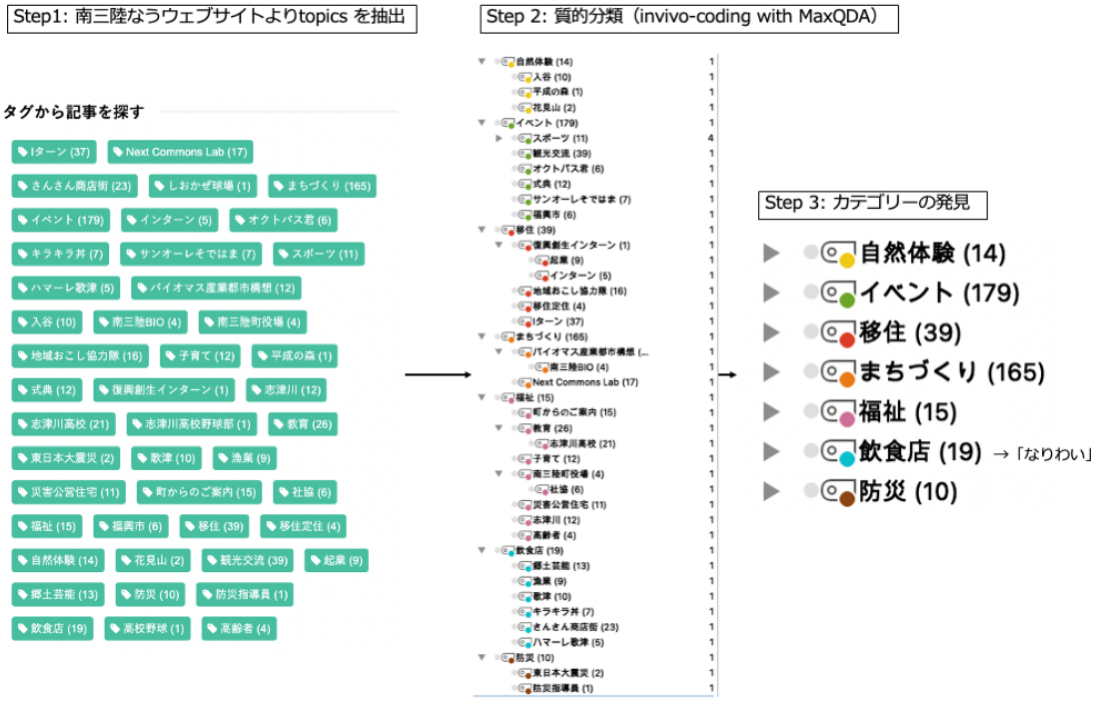


Fig. 1: 南三陸の復興を巡るトピックの整理

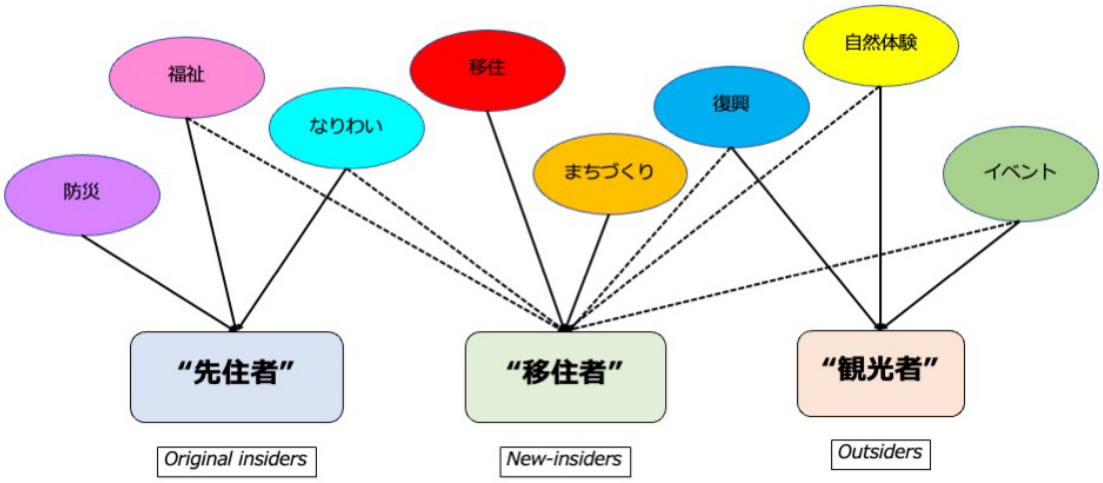


Fig. 2: 南三陸の復興を巡るアクターの種類

Table 1: 南三陸の復興を巡るアクター（ウェブページ）のリスト

ウェブページ名	分類	URL
南三陸観光協会	観光者	https://www.m-kankou.jp/
復興みなさん会	先住者	http://tohokuconso.org/common/minasan/
南三陸福興市	観光者	https://www.facebook.com/fukkouichi/
南三陸ふっこう青年会	先住者	https://www.facebook.com/msrfukkouseinen/
南三陸応縁団	観光者	https://www.minasan-ouen.com/
南三陸なう	移住者	https://m-now.net/
南三陸 Next Commons Lab	移住者	https://nextcommonslab.jp/network/minamisanriku/
ハマーレ歌津	観光者	https://hamare-utatsu.com/
サンマリン気仙沼ホテル観洋	観光者	http://www.mkanyo.jp/
宮城県南三陸町公式ウェブサイト	先住者	http://www.town.minamisanriku.miyagi.jp/
三陸新報	先住者	http://sanrikushimpo.co.jp/
ニュー泊崎荘	観光者	http://tomarizakisou.jp/pet.html
いりやど	観光者	https://ms-iriyado.jp/
南三陸商工会	先住者	http://www.m-shokokai.com/
多幸の社 入谷八幡神社	先住者	https://ocutopusjinja.webnode.jp/
南三陸モアイファミリー	先住者	https://www.moaifamily.com/
神割崎キャンプ場	先住者	https://www.m-kankou.jp/kamiwari-camp/
南三陸ワイナリー	観光者	https://www.msr-wine.com/
南三陸移住・定住支援センター	移住者	https://www.minamisanriku-iju.jp/
南三陸移住者バトン	移住者	https://www.instagram.com/minamisanriku.iju_baton/
南三陸 311 メモリアル	観光者	https://m311m.jp
南三陸さんさん商店街	観光者	https://www.sansan-minamisanriku.com
サステナビリティセンター	移住者	https://m-sustainable.org
いのちめぐるまち推進協議会	移住者	https://meguru.m-sustainable.org
南三陸戸倉っこかき	移住者	https://toguraquest.com
南三陸ネイチャーセンター友の会	先住者	https://m-inuwashi.jp
株式会社佐久	先住者	https://www.sakyu-minamisanriku.jp
株式会社 ESCCA	移住者	https://escca.jp
合同会社 MMR	移住者	https://m-mmr.amebaownd.com
宮城県漁業協同組合志津川支所	先住者	https://www.jf-miyagi.com
一般社団法人南三陸研修センター	観光者	http://ms-lc.org
一般社団法人さとうみファーム	観光者	http://satoumifarm.org
丸平木材株式会社	先住者	https://maruheitimber.com
株式会社ヤマウチ	観光者	https://www.yamauchi-f.com
株式会社はなぶさ	先住者	https://hanabusa-inc.jp
南三陸 YES 工房	先住者	https://yes-factory.jp
南三陸復興ダコの会	先住者	https://ms-octopus.jp/index.html
南三陸ミシン工房	先住者	http://www.mishinkoubou.org/
南三陸及善蒲鉾店	観光者	https://oizen.co.jp
FM 認証 宮城県 南三陸森林管理協議会	先住者	https://jp.fsc.org/jp-ja/Minamisanriku
南三陸森林管理協議会	先住者	https://mfsa.jp

たウェブページが媒介していることが確認できる。Hyphe 上で，特定のアクターを選択すると，そのアクターが他のアクターとどのような関係があるかを確認することができる 3。図 3 においては，赤線が被引用関係（「南三陸なう」が引用されていること），青線が引用関係（「南三陸なう」が引用していること），紫線が相互引用（「南三陸なう」と相互に引用しあっていること）を表している。

より具体的に，まずは，移住者向けのウェブページへ目を向けてみると，「南三陸なう」（1 の被引用，11 の引用，9 の相互引用）と「南三陸移住・定住支援センター」（4 の引用，4 の相互引用）がネットワークの比較的中央で様々なアクターとの間を媒介する役割を果たしていることが分かる。観光者向けのウェブページとしては「南三陸観光協会」が 15 の引用と 17 の相互引用をしており，ネットワークの中核にある。それに次ぐアクターとしては，「さんさん商店街」（3 の被引用，7 の引用，6 の相互引用），「ハマーレ歌津」（6 の被引用，4 の引用，2 の相互引用），「いりやど」（8 の被引用，1 の引用，2 の相互引用），「南三陸応縁団」（15 の引用，6 の相互引用）が大きいアクターとして存在する。地元住民向けのウェブページとしては「南三陸町」の公式ウェブページが一番大きい存在としてあるが（5 の被引用，6 の相互引用），他のアクターは比較的被引用および引用の数が少ない傾向性が確認できた。

図 5 は，モジュラリティ（modularity）と呼ばれる，ネットワークグラフ内のコミュニティ構造を評価する指標を用いて，グラフの色分けを行ったものである。ここで，コミュニティとは，ネットワークグラフのなかで似たような特性や機能を持つノードの集まりを意味し，ここでは 4 つ（4 色）に分けられている。質的な 3 分類（地元住民／移住者／観光者）とは重ならず，海／里／山といった地理的な分類とも対応しないが，4 つのサブグループがグラフの中央で混ざり合っている様子は確認することができる。質的な 3 分類とモジュラリティに基づく 4 分類の間に何らかの対応関係が見られないことは，質的に分類された 3 種類のアクターが混在する形で復興という実践を担っていると解釈することも可能であると思われる。

4. 結語

本研究では，デジタル人類学の観点から南三陸の復興に関するウェブページ上のグラスルーツな活動を記述することを通して、オンライン上の活動の全体像の可視化と質的に仔細な記述を進め、南三陸の復興を巡るデジタルパブリックが様々なアクターの密な結びつきにより形成されていることを示した。

ハイパーリンク・エスノグラフィを通して，現場におけるフィールドワークのミクロな視点での観察ではなく，マクロな視点に基づく観察が可能になる。ハイパーリンク・エスノグラフィにおける記述は，現地におけるフィールドワークの発見を裏付けるための資料として用いることも可能である。

しかしながら，ハイパーリンク・エスノグラフィには，いくつか限界がある。例えば，南三陸の復興に関わるアクターがみなウェブサイトやウェブページを持つわけではない。昨今では，Twitter や Instagram などのソーシャルメディアにおける活動は行うが，ウェブページを持たないといった団体も増えてきている。本稿の射程からは外れるが，ソーシャルメディ

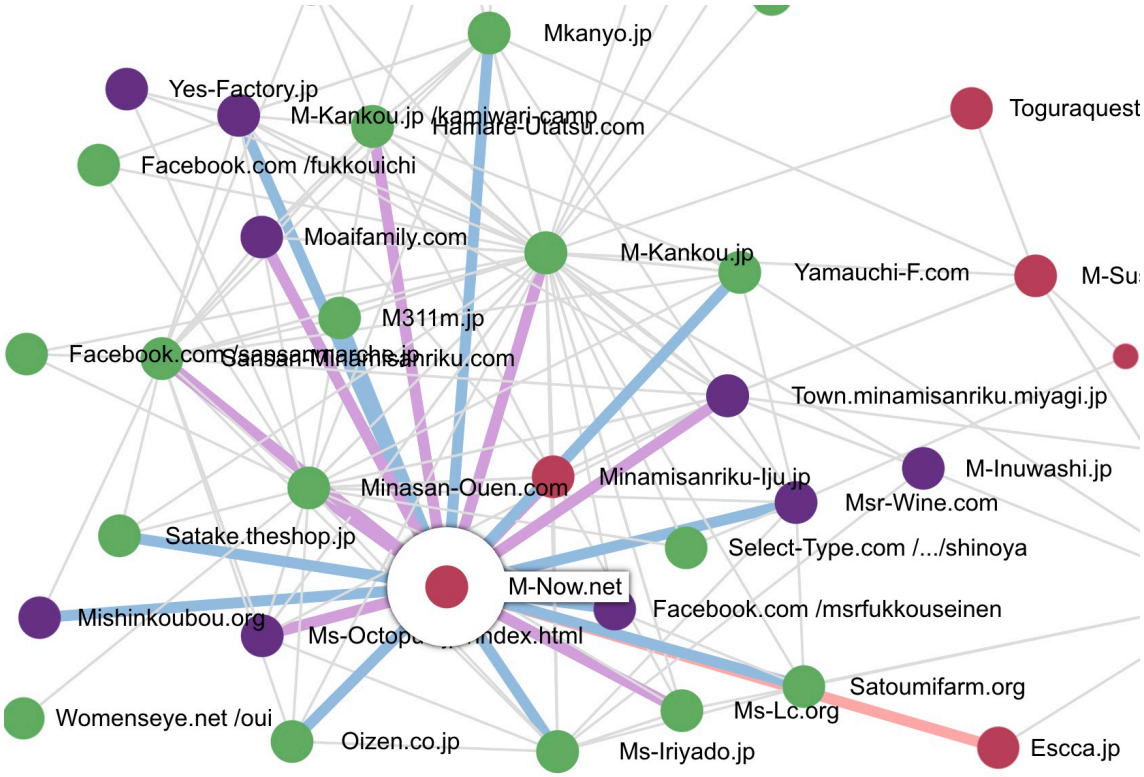


Fig. 3: 「南三陸なう」と関係するアクターの可視化（赤線が被引用／青線が引用／紫線が相互引用）

ア・エスノグラフィ（social media ethnography）も活発に行われてきており，ハイパーリンク・エスノグラフィとソーシャルメディア・エスノグラフィを効果的に組み合わせてオンライン空間の活動を記述する形が一番望ましいと思われる。今後は，ハイパーリンク・エスノグラフィやソーシャルメディア・エスノグラフィ，さらには，VR（virtual reality）空間におけるエスノグラフィなど，オンライン空間におけるエスノグラフィを有機的に組み合わせたデジタルエスノグラフィを行うことが益々重要になってくると思わえるため，そうした方法論的基礎を確立していくことを今後の課題としたい。

文　献

[1] Ackland, R., M. O'Neil, R. K. Standish, M. Buchhorn et al. (2006) VOSON: A Web services approach for facilitating research into online networks.

[2] Airoldi, M. (2018) Ethnography and the digital fields of social media. *International Journal of Social Research Methodology* 21(6): 661–673.

[3] Arvidsson, A. (2013) The potential of consumer publics. *ephemera* 13(2): 367.

[4] Arvidsson, A. and A. Caliandro. (2016) Brand public. *Journal of consumer research* 42(5): 727–748.

[5] Beaulieu, A. (2005) Sociable hyperlinks: An ethnographic approach to connectivity. *Virtual methods: Issues in social research on the Internet*: 183–198.

[6] Boyd, D. (2010) Social network sites as networked publics: Affordances, dynamics, and implications. In *A networked self*. 47–66. Routledge.

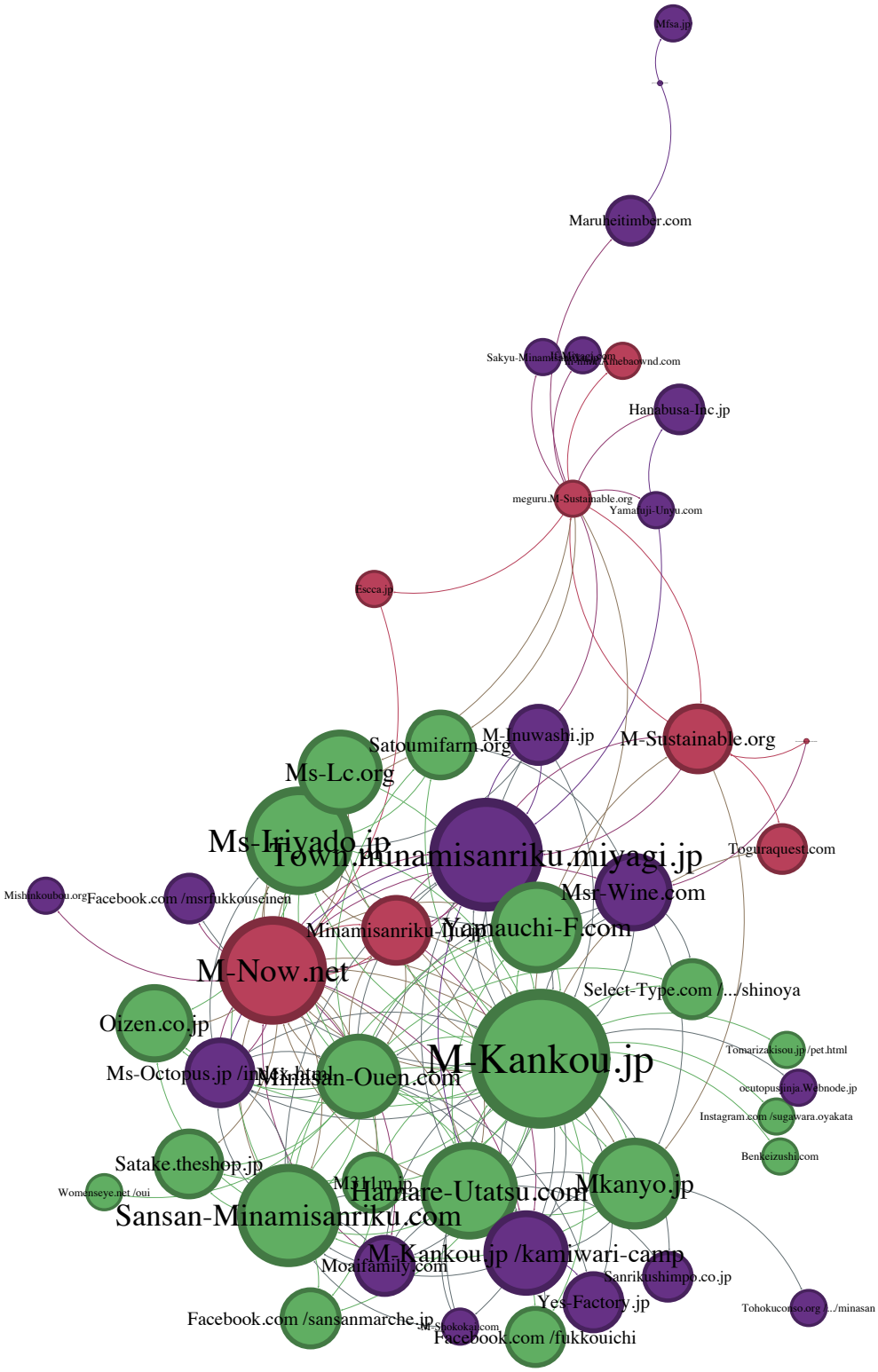


Fig. 4: 南三陸の復興を巡るアクターのネットワーク（質的分類を反映）

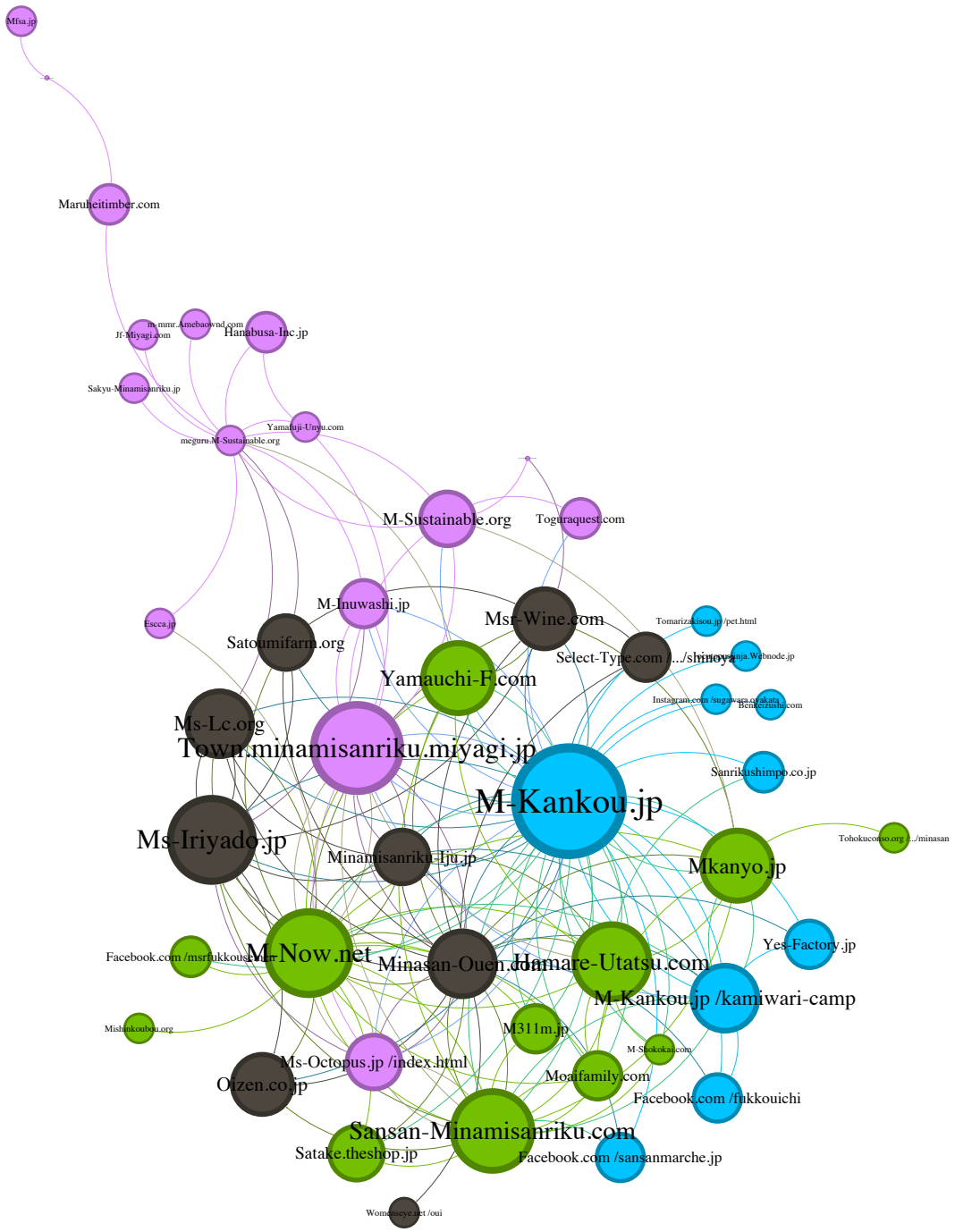


Fig. 5: 南三陸の復興を巡るアクターのネットワーク（モジュラリティによるサブグループの形成）

[7] Caliandro, A. (2018) Digital methods for ethnography: Analytical concepts for ethnographers exploring social media environments. *Journal of contemporary ethnography* 47(5): 551–578.

[8] Guha, R., R. McCool, and E. Miller. (2003) Semantic search. In *Proceedings of the 12th international conference on World Wide Web*. 700–709.

[9] Hallett, R. E. and K. Barber. (2014) Ethnographic research in a cyber era. *Journal of Contemporary Ethnography* 43(3): 306–330.

[10] Jacomy, M., T. Venturini, S. Heymann, and M. Bastian. (2014) ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one* 9(6): e98679.

[11] Jacomy, M., P. Girard, B. Ooghe-Tabanou, and T. Venturini. (2016) Hyphe, a curation-oriented approach to web crawling for the social sciences. In *Proceedings of the International AAAI Conference on Web and Social Media*. 595–598.

[12] Marres, N. (2017) *Digital sociology: The reinvention of social research*. John Wiley & Sons.

[13] Ooghe-Tabanou, B., M. Jacomy, P. Girard, and G. Plique. (2018) Hyperlink is not dead!. In *Proceedings of the 2nd international conference on web studies*. 12–18.

[14] Papacharissi, Z. (2015) *Affective publics: Sentiment, technology, and politics*. Oxford University Press.

[15] Pink, S., H. Horst, J. Postill, L. Hjorth, T. Lewis, and J. Tacchi. (2015) *Digital ethnography: Principles and practice*. Sage.

[16] Postill, J. and S. Pink. (2012) Social media ethnography: The digital researcher in a messy web. *Media International Australia* 145(1): 123–134.

[17] Rogers, R. (2013) *Digital methods*. MIT press.

[18] Rogers, R. (2019) *Doing digital methods*. Sage.

[19] Stage, C. (2013) The online crowd: a contradiction in terms? On the potentials of Gustave Le Bon's crowd psychology in an analysis of affective blogging. *Distinktion: Scandinavian Journal of Social Theory* 14(2): 211–226.

[20] Thelwall, M. (2009) Introduction to webometrics: Quantitative web research for the social sciences. *Synthesis lectures on information concepts, retrieval, and services* 1(1): 1–116.

[21] Venturini, T. and A. K. Munk. (2021) *Controversy mapping: A field guide*. John Wiley & Sons.

A Study on Characteristic Sounds in Tennyson and Browning: Using Stylometric Approaches

Iku Fujita

Graduate School of Language and Culture, University of Osaka

1-8 Machikaneyama-cho, Toyonaka, Osaka, 560-0043 Japan

Abstract This study investigates whether tabulated phoneme frequencies may reveal aspects of Alfred Tennyson's (1809–1892) and Robert Browning's (1812–1889) works. Tennyson and Browning are two of the most representative Victorian poets in the United Kingdom. Although studies on their topics, motifs, and styles, as well as comparative studies of Tennyson, Browning, and other poets, abound, only a few studies have divided the sounds in their works into phonemes and examined them. In addition, few studies have used quantitative methods to contrast the tones in Tennyson's and Browning's works. To perform a more in-depth analysis of Tennyson's and Browning's poetry collections, this study examines the top 50 phoneme frequency poems from each author while considering the limitations of earlier studies using cluster analysis and principal component analysis. This study finds that the tabulated international phonetic alphabet (IPA) data effectively identifies the differences between the two writers and captures the features present in their poetry using cluster and principal component analyses.

Keywords Alfred Tennyson, Poetry, Robert Browning, Stylometry, Sound

Tennyson と Browning の音声比較: 計量文体論手法を用いて

藤田　郁

大阪大学大学院言語文化研究科

〒 560-0043 豊中市待兼山町 1-8

E-mail: u256780k@ecs.osaka-u.ac.jp

あらまし 本稿は，19 世紀を代表する 2 詩人，アルフレッド・テニスン (1809–1892) とロバート・ブラウニング (1812–1889) の作品を国際音声字母 (IPA) に変換し，IPA 記号から得られる頻度表に計量文体論的アプローチ（クラスター分析及び主成分分析）を応用することにより，各詩人の作品の音声特徴を明らかにすることができるかを探索的に分析することを目的とする。テニスンとブラウニングは，イギリスを代表するヴィクトリア朝の詩人であり，彼らの詩に現れるテーマやモチーフ，各々の文体に関する研究や，テニスン，ブラウニングのそれと他の詩人のそれとを比較する質的研究は数多くあるものの，作品中の音を音素に分割し，分析した研究はごくわずかである。また，テニスンとブラウニング の作品に含まれる音の要素を定量的な手法で比較した研究も少ない。本研究では，テニスンとブラウニングの韻文についてより詳細な分析を行うため，先行研究に挙げられる制約を考慮しつ

つ，各作家の音素頻度上位 50 の詩を対象として分析した。クラスター分析及び主成分分析の結果により，2 人の作家の作品内で使用されている音の違いが効果的に識別され，各詩人の特徴を捉えることができることが明らかになった。

キーワード　アルフレッド・テニスン，韻文，音声，計量文体論，ロバート・ブラウニング

1. Introduction

Alfred Tennyson (1809–1892) and Robert Browning (1812–1889) are two of the most representative Victorian poets in the United Kingdom. While studies on their themes, motifs, and styles, as well as comparative studies of Tennyson, Browning, and other poets, abound, only a few have divided the sounds in their works into phonemes and analyzed them. Moreover, only a few studies have compared sounds in Tennyson's and Browning's works using quantitative approaches. Tawfiq (2020) indicated the number of phonological poetic devices (e.g., alliteration, onomatopoeia, consonance, assonance, and rhyme). While Tawfiq (2020) succeeded in highlighting the differences in the five poetic devices between Tennyson and Browning using decimal information, some issues remained because its target data were limited to five examples from ten randomly selected poems by Tennyson and Browning. Notably, detecting (phonological) poetic devices poses a significant challenge for analysts, as it fully depends on qualitative, close reading techniques. Therefore, the limitation of the data under investigation was inevitable. However, the question remains whether the data and results of Tawfiq could represent Tennyson's and Browning's entire works. Another question is whether the numerical differences between Tennyson and Browning were statistically significant. Tawfiq concluded that Browning's frequency of onomatopoeia was significantly higher than that of Tennyson. However, the difference in their frequencies of onomatopoeia was only two: Browning used onomatopoeia four times in his selected poems, while Tennyson used it just twice (Tawfiq, 2020: 32).

Plamondon (2005: 153) issued the problems with phonetic analysis, mentioning that the terminology, symbols, and rules, as well as the large amounts of phonetic data, made the phonetic analysis of literature uneasy for students. The latter issue can also overwhelm analysts and researchers because the difficulty in finding phonological devices increases with larger amounts of data. However, an automated model for detecting phonological poetic devices remains unestablished. Plamondon (2005) challenged the difficulties in analyzing phonetic elements in poetry using the program AnalysePoems, which is 'designed to assist students to enhance their interaction with poetry and assist scholars in accumulating data, phonetic and otherwise, about poems' (Plamondon, 2005: 154). Plamondon described some results of his explorative analyses of Tennyson's and Browning's poems, which were converted into phonetic alphabets, through AnalysePoems.

Based on previous research and criticisms on sounds in Tennysonian and Browning's poems, Plamondon's attempts were 'to determine what in Browning's poetry produces the 'clamorous chaos' and what in Tennyson's poetry produces the 'melodious tone' and 'to determine if the

phonetic makeup of (Tennyson's and Browning's) poetry differs in any substantial manner.' Plamondon hypothesized that the more 'musical' Tennyson's poetry is, the more pleasant-sounding vowels it should have; the more 'harsh' Browning's poems are, the more harsh-sounding consonants they should contain (Plamondon, 2005: 163–164). Notably, Plamondon obtained the results by counting the phonemes in several poems of Tennyson and Browning, tabulating them, and calculating the proportion of each phoneme in the poems. The outcome was that Tennyson uses more vowels than Browning while Browning employs a higher proportion of consonants in his writing. Furthermore, it suggested that Tennyson's poetry contained shorter vowel sounds, which were allegedly responsible for its melodic quality, whereas Browning's poetry had more plosive consonants, which are presumably responsible for its harshness. However, as Plamondon already argued, the proportions of phonemes in Tennyson's and Browning's works differed slightly. Plamondon mentioned that the restrictions on the number of works might have caused the insignificant divergences. Plamondon continued:

> It seems, therefore, that either the counting of individual phonemes and tabulating them by class is not sufficient to characterize the difference in the quality of their poetry, or the long held opinion about the relative 'musical' quality of their verse is flawed.
>
> (Plamondon, 2005: 166)

A mere tabulation of individual phoneme frequencies might be insufficient to explain the differences in poetry. However, other quantitative approaches exist for effectively finding characterizations in poems using tabulated information. Considering the limitations and issues of previous studies, this study aims to explain whether tabulated phoneme frequencies can demonstrate the characteristics of Tennyson's and Browning's poetry and, if so, which phonemes are features of each poet. To investigate the poems of Tennyson and Browning more widely, this study used the top 50 poems of phoneme frequencies from Browning and Tennyson. The 100 poems of Tennyson and Browning were further tabulated and analyzed via cluster analysis (CA) and principal component analysis (PCA). The results were visualized using `stylo` by `R` (Eder et al., 2016), `Python`, and `Gephi` (Bastian et al., 2009) to facilitate the reading of results.

2. Methodology and Data

2.1. Methodology: Quantitative approaches in this study

This study employed CA and PCA to detect the characteristics of Tennyson's and Browning's poems. CA and PCA are statistical methods used to classify text variables (elements) into several groups. Stylometric studies often use these methods to reveal the stylistic features of their target texts or corpora and to detect authorship distributions (Burrows, 1987; Hoover, 2012; Tabata,

2004; Tabata, 2012; to name but a few). CA is a statistical method that uses the information in the data to form clusters by grouping data that are close to each other (Ishikawa et al., 2010: 163). The two main types of CA are hierarchical and nonhierarchical CAs. However, this paper focuses on the former, which can draw a dendrogram with results. In text mining and analyses, the groupable data can be the frequency of words in the texts in question. Relative frequencies are calculated based on raw frequencies from the texts; then, the computer locates each value, calculates the distance between each value, and groups close-distant data into the same groups.

There are multiple choices for calculating distance (e.g., Euclidean distance, Canberra distance, and Manhattan distance) as well as from where to calculate the distance (e.g., Ward and Complete methods). Further analyses were conducted using the Euclidean distance and Ward method, commonly employed in stylometric research. When Euclidean distance was employed, a mismatch of values occurred. Although the Euclidean distance value can be positive or negative, the relative frequencies of words would never have a negative value. The relative frequencies were z-scored before the analysis to avoid discrepancies in the values. The conditions of distance and method were applied to each analysis in this paper. PCA is a technique that compresses information into a small number of components when the number of variables is large. By compressing information into a small number of components, it is possible to efficiently extract values that are representative of the information in the data (Ishikawa et al., 2010: 193). Therefore, PCA is one of the multivariate analysis methods for data reduction, along with factor analysis and correspondence analysis. PCA facilitates visually grasping the overall picture of data by aggregating complex relationships among many variables and samples into a small number of comprehensive indicators (i.e., principal components) (Tabata, 1998: 195). In PCA, eigenvalues and eigenvectors can be obtained from a correlation matrix or covariance matrix (Tabata, 2004: 102). This study used a correlation matrix by referencing Tabata (2004) throughout all analyses.

2.2. Data: Original poems

This study focuses on two Victorian poets—Alfred Tennyson and Robert Browning. The original text data were compiled as corpora from the Delphi Poet Series and Ricks (1987). To build the corpora, all the texts were converted to plain text data using optical character read/recognition, which was then manually edited and proofread for problems, including unnecessary spaces, tabs, newlines, and garbled characters.

Table 1 shows the bibliographical data of Tennyson and Browning corpora. The Tennyson corpus contains 600 verse texts, with 423 taken from the Delphi Poet Series 'Alfred, Lord Tennyson' (2013) and the remaining 177 compiled from Ricks (1987). Browning's 204 poems are compiled from the Delphi Poets Series 'Robert Browning' (2012). In Table 1, Tennyson's number of works is almost three times that of Browning; however, Browning's total number of tokens is nearly double that of Tennyson. The data suggest that Browning wrote more lengthy/wordy poems than Tennyson did.

Table 1: Bibliographical data of Tennyson and Browning corpora

Authors	Num. of works	Num. of total tokens	Type tokens
Tennyson	600	355,235	21,661
Browning	204	590,315	36,159

The plain text data in Tennyson and Browning corpora are further converted to the data described in the international phonetic alphabets (IPA) following the method in Section 2.3.

2.3. Methodology: IPA conversion

To investigate the phonological elements in Tennyson and Browning, the original poetry data were converted to phonetic symbols—IPA. The online resource EasyPronunciation.com (https://easypronunciation.com/en/) was subscribed to and used for its conversion. The subscribed version of EasyPronunciation.com allowed us to translate 10 thousand alphabetical characters of English and other languages into phonetic transcription per run. For English translation, moreover, there are options for American, British, and Australian English accents. Hence, the English dialect of the target corpora is British English, and British English was chosen for the translation of this study (Figure 1). Some words, such as proper nouns, could not be converted because they are not in the dictionary of the tool, but these words are surrounded by voluntary symbols (e.g., #Camelot#). Unconverted words were excluded from further analysis.

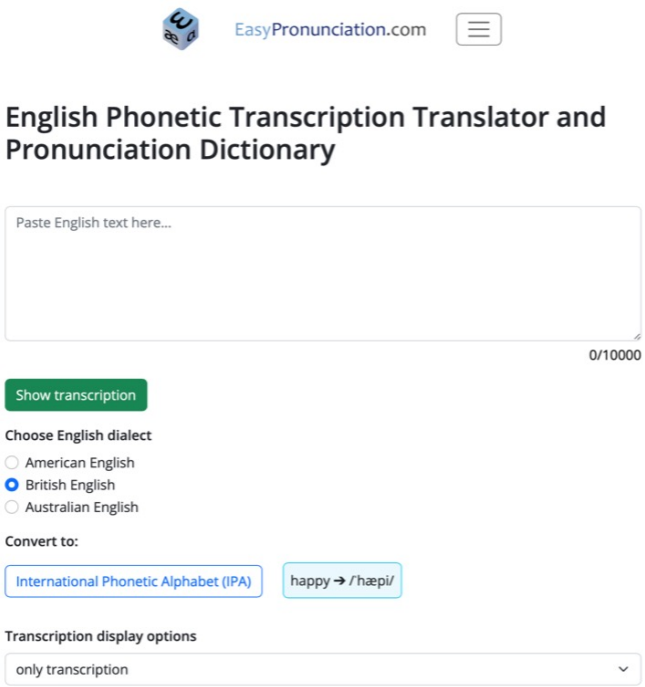


Figure 1: The top page of EasyPronunciation.com

2.4. Data: Converted to IPA

Table 2 lists the information on IPA-converted corpora. There were slight changes in the original text data. The difference in the data in Tables 1 and 2 is due to the deletions of unconverted words along with homophonic and heterophonic spelling.

Table 2: Bibliographical data of the IPA Tennyson and Browning corpora

Authors	Num. of works	Num. of total tokens	Type tokens
Tennyson	600	356,311	20,178
Browning	204	591,750	34,359

The IPA text data were further edited for more detailed analysis. Punctuations and other symbols for accents and quotations were deleted. Vowels and consonants were separated with spaces, but the contiguous vowels and consonants remained together. By doing so, we were able to investigate as small units of sound as possible while the group of vowels/consonants kept holding the sounds themselves, as explained with plural phonetic symbols (e.g., diphthongs /laɪk/ (like), long vowels /lɔːd/ (Lord), and consonant sounds expressed by a few alphabets /tʃɛːtʃ/ (church)). In this study, therefore, the word 'consonant(s)' does not necessarily mean that the legitimate phoneme is the smallest sound unit. However, it widely mentions the consecutive consonant group(s) found in a word. Here are the original poem lines (1) and the same lines of phoneme-divided IPA text (2):

(1) Strong Son of God, immortal Love,
Whom we, that have not seen thy face,

(2) str ɒ ŋ s ʌ n ə v g ɒ d ɪ m ɔː t ᵊ l l ʌ v
h uː m w iː ð ə t h ə v n ɒ t s iː n ð aɪ f e ɪ s

(Tennyson, *In Memoriam A.H.H.*, 1850: ll.1–2, underline added)

As can be seen in the IPA text of *In Memoriam A.H.H.* in (2), where underlines are added, the half (weak) vowel schwa /ə/ appears as superscripted in the first line, but two /ə/ in the second line are not superscripts. The conversion tool EasyPronunciation.com translates the schwa in nonstressed syllables into superscripts. Additionally, the syllabic (vocalic) consonant /l̩/ and nonsyllabic consonant /l/ are translated as separate sounds. Although there may be some disagreement about treating superscript /ᵊ/ and /ə/, syllabic /l̩/, and nonsyllabic /l/ as separate sounds, this paper treats them as different sounds each. Table 3 shows the bibliographical data of IPA text data, which has been divided into vowels (groups) and consonants (groups).

Table 3: Number and type of vowels and consonants of Tennyson and Browning corpora

Authors	Num. of vowels	Type tokens of vowels	Num. of consonants	Type tokens of consonants
Tennyson	487,139	20	597,193	293
Browning	828,936	19	1,025,789	266

Interestingly, the type tokens of Tennyson's consonants are 27 larger than Browning's, though Tennyson's total tokens and the number of consonants reach only about 60% of Browning's. The consonants found in Tennyson but not in Browning are listed in (3) with their raw frequencies in []:

(3) /ðh/ [1], /dθs/ [2], /fd/ [28], /fh/ [1], /ksd/ [34], /kslz/ [1], /ksp/ [1], /kstr/ [1], /kʃ/ [1], /lpd/ [4], /sh/ [1], /lʃ/ [2], /lθs/ [1], /mb/ [1], /mbr/ [1], /mfd/ [1], /mpd/ [5], /mpst/ [1], /mt/ [6], /mθs/ [1], /ndj/ [1], /ndlz/ [1], /nk/ [1], /nldz/ [2], /ntj/ [1], /ntld/ [1], /ntln/ [1], /ntʃd/ [22], /nw/ [1], /nzl/ [1], /ŋgd/ [1], /ŋgld/ [3], /ŋkd/ [17], /ŋr/ [1], /pb/ [1], /pd/ [28], /rl/ [28], /rlz/ [2], /skd/ [38], /spd/ [21], /sθ/ [1], /ʃd/ [207], /ʃm/ [1], /tb/ [1], /tʃd/ [133], /tt/ [1], /vl/ [1], /vm/ [1], /vnz/ [1], /θd/ [6]

On the contrary, the consonants found in Browning but not in Tennyson are shown in (4) with their raw frequencies in square brackets:

(4) /bld/ [1], /dnd/ [4], /dnz/ [1], /dʃ/ [1], /fθs/ [1], /ksθs/ [1], /ktr/ [1], /lb/ [2], /ldʒd/ [11], /lfθ/ [4], /lfθs/ [1], /ltʃ/ [8], /ltʃt/ [9], /mft/ [8], /mpts/ [34], /msn/ [6], /msnz/ [1], /ŋg/ [1], /ŋkst/ [1], /snt/ [11], /ss/ [1], /stlz/ [1], /vw/ [2]

The difference in vowel types causes the occurrence of the open/low-back unrounded vowel /ɑ/ in a Tennysonian poem, 'Stanzas' (1827, l. 14):

(5) In heav'n a more exalted throne,
ɪ n #heav'n# ə m ɔː ɪ g z ɑ l t ɪ d θr ə ʊ n

While the long open-back unrounded vowel /ɑː/ appears much more commonly in both Tennyson (raw frequency is 11,650) and Browning (raw frequency is 15,777), the short /ɑ/ appears only once in a work of Tennyson. Therefore, it seems reasonable to say that the use of extremely low frequency is more of a coincidence, or at least a feature of the work, than Tennyson's characteristics.

2.5. Trial PCA results

For further analysis, the IPA data were separated into the vowel and consonant folders of Tennyson and Browning. Before setting the number of target poems for the analyses, a trial PCA was conducted on the whole vowel and consonant data. Figures 2 and 3 show the results of PCA on vowels and consonants, respectively.

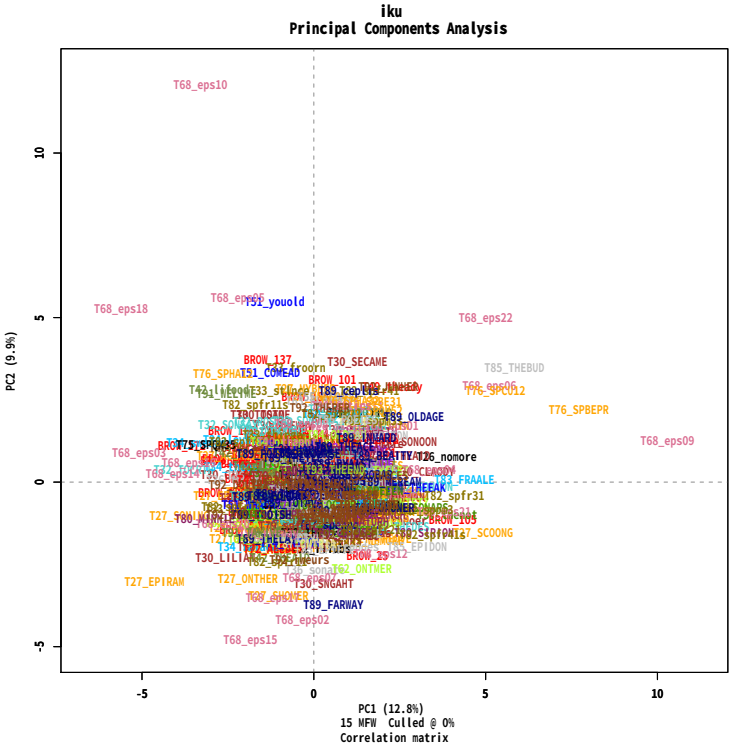


Figure 2: PCA results of the whole vowel data (804 works of Tennyson and Browning)

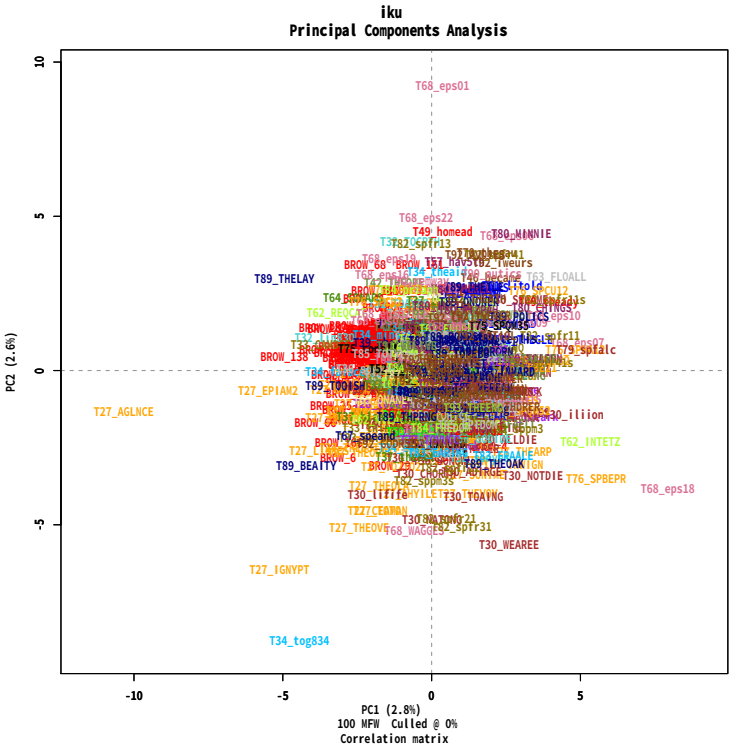


Figure 3: PCA results of the whole consonant data (804 works of Tennyson and Browning)

In both Figures 2 and 3, most of the variables (the title of works) congregate around the origins of the plots. The variables located apart from the origins are the short poems of Tennyson, whose total tokens are less than 100 words. Moreover, the contribution[1] ratios of each component were quite low (PC1 12.8% and PC2 9.9% in Figure 2; PC1 2.8% and PC2 2.6% in Figure 3); thus, it was estimated that calculating the entire 804 works at once was inapplicable. Considering the results that Tennysonian short poems scattered around while most of the other variables gathered around the origins, it is deemed that the number of poems for analysis should be reduced first, and second, the number of tokens of poems should not be too small and should not diverge too much. Therefore, the analyses in this study use 50 works from each poet, whose total tokens of vowels and consonants are the largest to 50th largest among the data.

3. Results and Discussion

This section illustrates the results of CA and PCA on the vowels (Section 3.1) and consonants (Section 3.2) of Tennyson's and Browning's top 50 poems.

3.1. Analysis of vowels in the top 50 works

This section presents the extracted results of CA and PCA for the vowels. Figure 4 shows the heatmap representation of the CA results. The shorter the length of each branch at the top and left

[1] Contribution (ratio) is a measure of how much of the variation in the data can be explained by this component alone. The higher the ratio is, the more a component describes the variation of the data.

sides of the plot, the shorter the distance between variables (i.e., the higher the similarity between variables). Conversely, when the length of each branch is longer, the distances between variables are farther; thus, the variables located farther are not alike. The heatmap is drawn by running CA on tabular data of vowel frequencies per work using CasualConc 3.0.4 (Imao, 2022). The poetry works are located horizontally, and the 20 vowels are located vertically. As the legend on the left top of the figure suggests, the color of a cell explains how significant the vowel in the poem where the cell interjects is. The legend has eight colors; the dark gray represents the most substantial, the first significant is brown, and the green explains the least significant. The dendrogram at the top of Figure 4 was broadly divided into two clusters with the two longest branches at the highest. A group on the left side of the figure is mostly clustered by Tennyson, while a group on the right side is mostly clustered by Browning (see Appendices 1 and 2 for the corresponding table of abbreviated file names and the original poem titles along with their publication years). However, some exceptions are located in the other poet's clusters, e.g., BROW_14 in the left, Tennyson's cluster: T30_SUPELF, T33_THEENT, T33_THETER, T42_THESIN, T42_WILOCK, T55_THEOOK, T62_ENODEN, T64_AYLELD, T68_LUCIUS, T80_SIRHAM, and T80_THEERS in the right, and Browning's cluster, which indicates that CA detected differences between the two authors' vowel usages. Figure 5 shows the results of PCA using the `stylo` package of `R`. The contribution ratio of principal component (PC) 1 is 22.0% and PC 2 is 12.6; therefore, it indicates that 34.6% of the data can be explained by the plot. The variables of Browning tend to scatter in the first and fourth quadrants; meanwhile, Tennysonian variables are inclined to be located in the second and third quadrants. Similar to the CA result, this PCA result shows the classification of the two authors' poems. Furthermore, it is possible to observe which vowels are the characteristics of which poets from the heatmap representation in Figure 4. The vertical black line in Figure 4 separates the plot from Tennyson's and Browning's clusters. The horizontal dotted line divides the vowels according to the dendrogram depicted on the left, which classifies the vowels in question. The heatmap indicates that the eight vowels, /aʊ/, /aɪ/, /ʊ/, /ə/, /ɑː/, /ɛː/, /iː/, and /ɔː/, are more significant in Tennyson than Browning. However, the other 11 vowels, /æ/, /ɒ/, /ɑ/, /ə/, /i/, /ɪ/, /u/, /ɔɪ/, /ʌ/, /uː/, and /e/, rather notably appear in Browning's poems. The result suggests that Tennyson tends to use diphthongs and long vowels, which are pronounced with a retracted and mid-retracted shape of the mouth, whereas Browning prefers short vowels, especially mid-retracted and rounded ones.

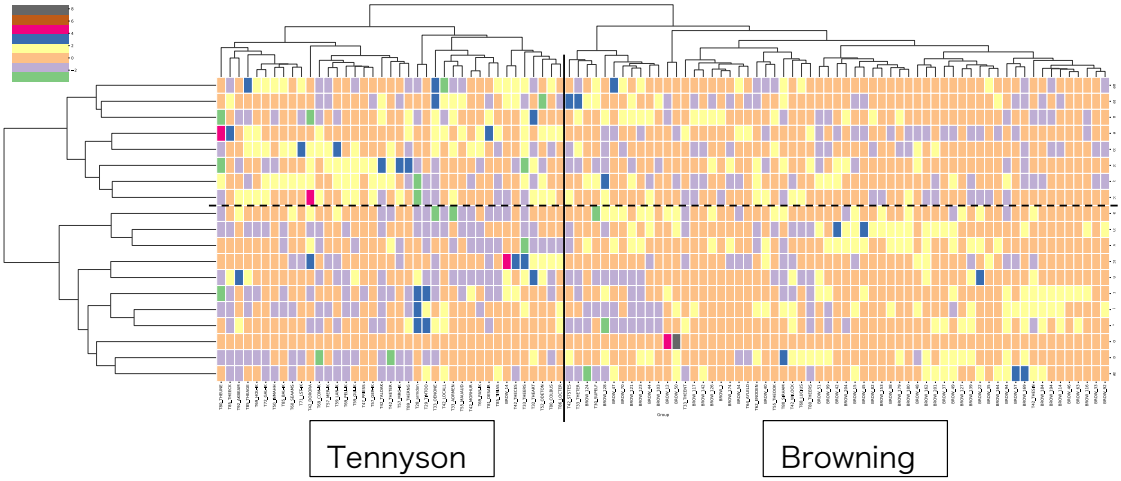


Figure 4: Heatmap representation of the results of CA on vowels

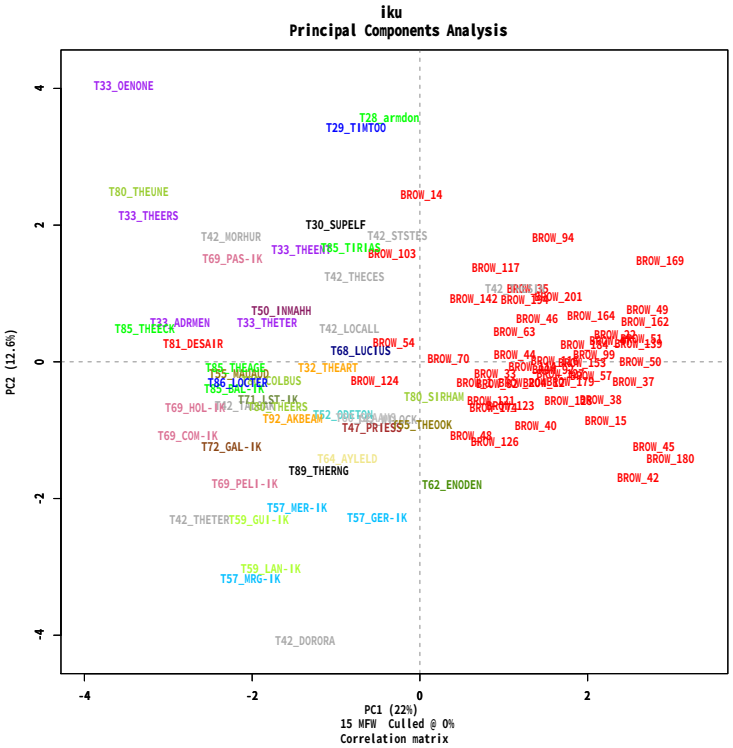


Figure 5: Results of PCA on vowels

Plamondon (2005: 163–164, italics added) assumed that 'musical' Tennysonian poem 'should have a greater proportion of pleasant-sounding *vowels*,' whereas 'harsh' Browning's poem 'should have a greater proportion of harsh-sounding *consonants*.' However, considering the results of CA, the musicalness and harshness of Tennyson and Browning may not be due only to the difference in vowels and consonants; rather, the musicalness and harshness of both vowels and consonants

have been caught by previous critics' and researchers' ears. That is, the long monophthongs and diphthongs introduce the musicalness of Tennyson, while the short vowels carry the harshness of Browning. Regarding musicalness and harshness, Section 3.2 discusses the consonant usages of the two poets. The `stylo` is also available to conduct CA. When CA is run on `stylo`, it draws a figure to facilitate reading the results and outputs the numerical results. Although the CA dendrogram is not used in this paper, the numerical result enables us to see how these variables (poems) relate and how similar they are by depicting a network graph (Figure 6).

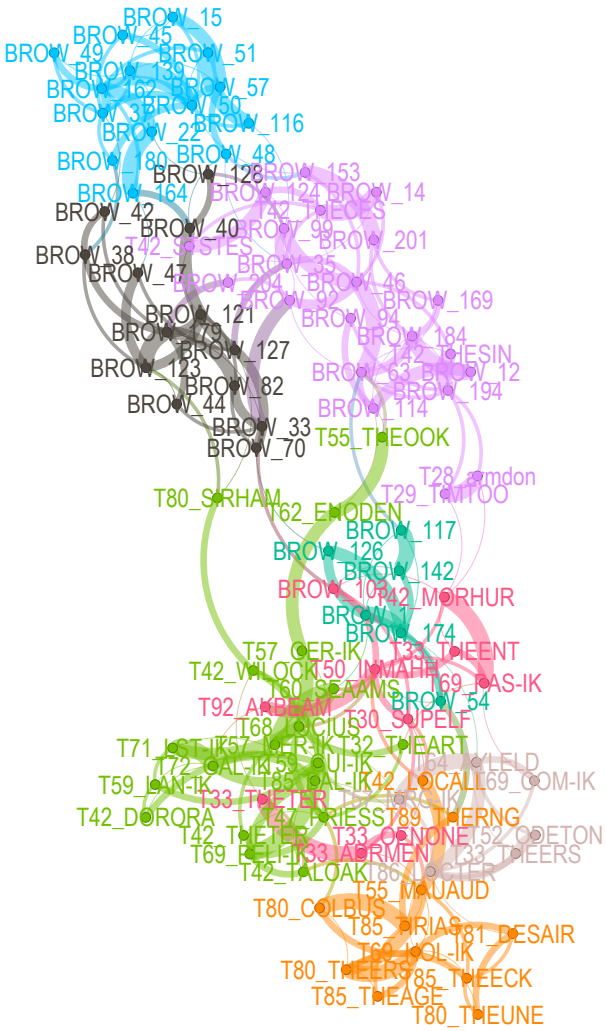


Figure 6: Network representation of vowels from the CA results

Figure 6 is drawn by `Gephi` using a numerical output of CA. The color differences depend on modularity, which indicates how unified the data constituents are. Thus, the elements colored the same are the close-knit poem groups based on the CA output. The degree of similarity is indicated by the thickness of the edges connected to each constituent. The thicker the lines, the more similar the elements are. Figure 6 illustrates that Tennyson's and Browning's poems are catego-

rized distinctly. Some works of Tennyson and Browning, as shown in the CA and PCA results, are confounded with the other poet's works; however, most of Browning's works are colored light blue, purple, brownish gray, and emerald, while Tennyson's are light green, reddish pink, beige, and orange. Tennysonian T28_armdon and T29_TIMTOO, colored purple, are connected with thick edges, one after another, but other lines from T28_armdon and T29_TIMTOO to Browning elements are narrow, suggesting that even though T28_armdon and T29_TIMTOO are categorized within the same group as Browning's, the two poems share more similarity than the other Browning poems. Tennyson's T42_MORHUR ('Morte d'Arthur, 1842) and T69_PAS-IK ('The Passing of the Arthur,' 1869) are edged thickly in reddish pink as both are the stories of the death of King Arthur, and thus, much of their words are distributed in common. Apart from T42_MORHUR and T69_PAS-IK, eight of the thirteen poems from the *Idylls of the King* series are classified in light green. The rest of the *Idylls of the King* series poems, T57_MRG-IK and T69_COM-IK, are in the same beige color group; T69_HOL-IK belongs to the category of orange alone. Although further detailed investigation is needed, the results suggest that the tabular data of vowels can distinguish not only author divergence but also the characteristics of each poem. A more in-depth investigation is encouraged so that the features found in a specific work or in the poetry of a specific poet can help reveal phonetic symbols or the effects that the authors rendered on specific sounds.

3.2. Analysis of consonants in the top 50 works

CA and PCA revealed intriguing results for consonants in Tennyson and Browning. Figures 7 and 8 show the CA and PCA results, respectively. In the heatmap plot (Figure 7) from CA, the top of Figure 7's dendrogram is roughly divided into two groups with the black line in the figure, with the two longest branches being the tallest. Most of Tennyson's works are gathered in a group on the right side of the picture, whereas Browning's works are clustered on the left. On the CA of the consonants, none of the poet's works merged into the other one's cluster; that is, only Tennyson's poems were classified into the groups at the right, and only Browning's works were grouped into the cluster at the left. In the PCA result in Figure 8, the contribution ratios of PCs 1 and 2 are 18.8% and 7.0%, respectively, indicating that the plot can account for 25.8% of the data. While Browning's variables are likelier to be found in the second and third quadrants, Tennyson's variables tend to be dispersed in the first and fourth quadrants. The consonants in question are roughly separated into three groups with the red, blue, and orange rectangles numbered one, two, and three, respectively, in Figure 7. According to the heatmap representation in Figure 7, the consonants in the red rectangle are the significant elements of Browning, and the consonants in the blue rectangle are the remarkable ones of Tennyson. The orange rectangle elements commonly appear in both poets' poems or only occur spottily in a few poems.

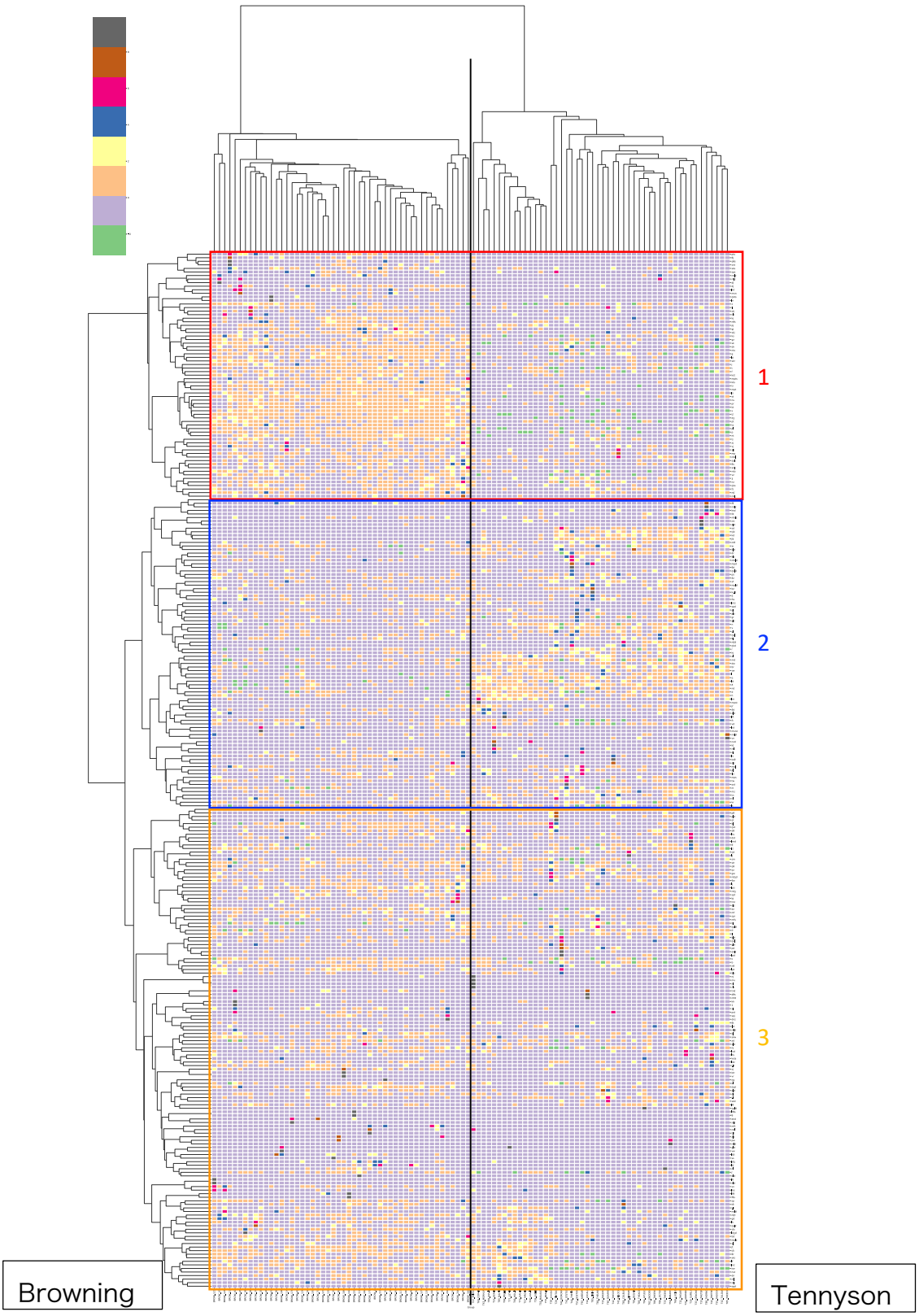


Figure 7: Heatmap representation of the CA result on consonants

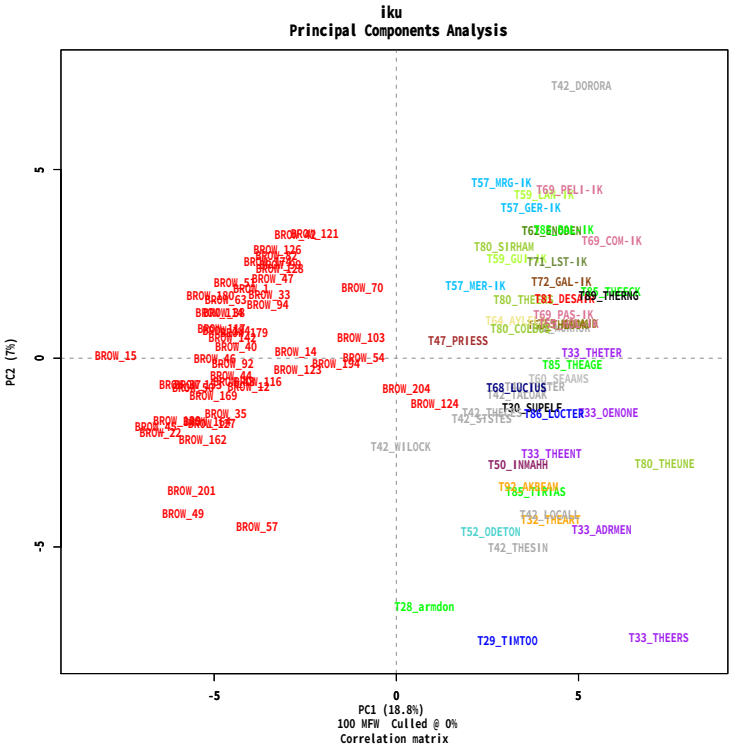


Figure 8: Results of PCA on consonants

Although Figure 7 indicates that CA detected differences in the consonants of the two poets, the dispersion of relative frequency values and the number of elements prevented the plot's readability. Moreover, the relative frequency can unnecessarily accentuate the value or mislead the differences hidden behind the calculated relative frequency value when the raw frequency of words is very low. For example, the word X occurs once in both Tennyson and Browning. When the total tokens of Tennyson's poem are 500, the relative frequency (per thousand words) is two. However, if the total tokens of Browning's poem were 1,000, the relative frequency of the word X would be one. Thus, even though the word X appears once in the two poets' works, the calculated relative frequency does not fully reflect the fact shown by the raw frequency. Notably, low-frequency words, such as those that appear only once or twice in the entire text, might not rise proportionately with the length of the target text, meaning that normalized values might not adequately reflect their usage. The crucial question for words that appear only once or a few times in an author's entire text data is identifying where, how, and what words they collide with. Hence, even though the frequency indicates the same value of 1, such words will be weighted differently depending on why they appear infrequently and inconsistently. To avoid incongruency, this study further applied CA to the top 24 raw frequency consonants, i.e., /t/ [108,233], /n/ [108,209], and /d/ [81,420] (values in brackets are the sum of the raw frequencies). As the raw frequency of 25th consonant, /dʒ/, is 8,890, and the digits get down from five to four, the elements whose raw frequency is lower than

10,000 were cut off. Figure 9 shows the heatmap representation of the top 24 consonants in the top 50 poems of Tennyson and Browning.

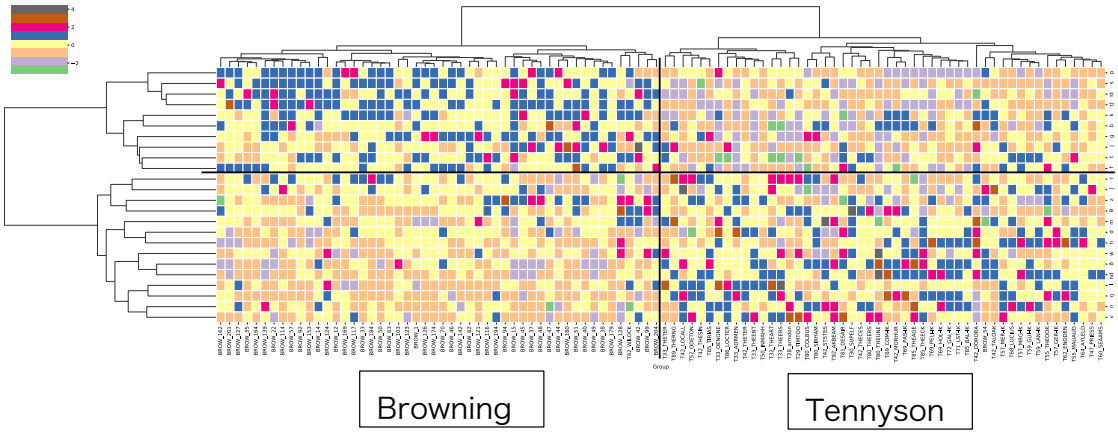


Figure 9: Heatmap representation of the CA result on the top 24 consonants

The top of Figure 9's dendrogram is separated into two groups by the black line in the figure. On the left side of the plot, Browning is primarily grouped beside Tennyson's poem, T42_WILOCK, whereas Tennyson is primarily grouped on the left, except for one of Browning's poems (BROW_54). The 24 consonants are generally divided into two groups by the dendrogram on the left side of the figure, where the horizontal black line is depicted. The consonants above the horizontal line appear more frequently in Browning's works because the color of cells at the left hand of the vertical line indicates higher values than Tennyson's works at the right hand of the vertical line. However, the consonants below the horizontal line appear more frequently in Tennyson because the color of the cells at the right hand of the vertical line indicates higher values than Browning at the left hand of the vertical line. The consonants that characterize each poet are listed in Table 4.

Table 4: Consonants classified by CA for each poet

Authors	Consonants
Browning	/t/, /s/, /f/, /k/, /b/, /p/, /st/, /g/, /j/, /tʃ/
Tennyson	/r/, /ʃ/, /z/, /θ/, /m/, /d/, /h/, /w/, /ð/, /nd/, /l/, /ŋ/, /n/, /v/

The tendency of the characteristic consonants of each poet listed in Figure 10 is more noticeable when the consonants are explained in tabular form in Figure 10. In Figure 10, the remarkable consonant symbols of Browning are in blue, while the prominent consonant symbols of Tennyson are in red. Figure 10 is depicted by the author of this paper by referencing Table 5 in Kubozono (1998: 43). The author of this work also added sonority to the right side of the figure by referencing Kubozono (1998: 51). Sonority is an indicator that measures how much the sound can be heard at a given distance; thus, for example, a weak sonority sound, the labial stop /p/, cannot be heard

while the strong sonority sound, the velar approximant /w/, can be heard at the same distance. This is not indicated in Figure 10, but the sounds with the strongest sonority are vowels.

Manner of articulation	Voiceless/ Voiced	fortis/lenis	Point of articulation: Labial	Dental	Alveolar	Post-alveolar	Palatal	Velar	Glottal	Sonority
Stop (Plosive)	Voiceless	fortis	p		t			k		Weak
	Voiced	lenis	b		d			g		
Affricate	Voiceless	fortis				tʃ				
	Voiced	lenis				dʒ				
Fricative	Voiceless	fortis	f	θ	s	ʃ			h	
	Voiced	lenis	v	ð	z	ʒ				
Nasal	Voiced	lenis	m		n			ŋ		
Lateral	Voiced	lenis			l					
Approximant	Voiced	lenis				r	j	w		Strong

Figure 10: Consonant map

The characteristics of Tennyson tend to spread below voiced fricatives, where the sonority of sounds is stronger than the stops, affricates, and voiceless fricatives in Figure 10. In contrast to Tennyson, Browning's features tend to be located on weak sonority consonants (i.e., stops, affricates, and voiceless fricatives). The harshness and musicalness in Browning and Tennyson, mentioned on vowels in Section 3.1, can be further explained with sonority here about consonants. Furthermore, Plamondon (2005) compared Browning and Tennyson and hypothesized that Browning should have more 'harsh' consonants, while Tennyson should have 'musical' vowels. Plamondon (2005: 165) further thought that Browning's significant 'harsh' consonants should have been plosives (stops). Plamondon also highlighted that the proportion of plosive consonants in Browning was only 0.6% higher than in Tennyson. Hence, the difference was 'probably too small to justify citing them as responsible for the relative musicality of the poems' (166). However, the CA results indicated that five out of six plosive consonants in English are significant in Browning. The shortage of proportion differences in the Plamondon could be caused by the small number of poems he used in his analyses. The new aspect suggested by CA in Figure 9 and Figure 10 supports what Plamondon has tried to reveal: Tennyson prefers to use 'less-harsh' consonants compared to Browning. Moreover, besides plosives, the postalveolar affricate consonant /tʃ/, labial fricative /f/, and alveolar fricative /s/ are notable consonants of Browning. The three consonants are voiceless, while the voiced sounds of the two—labial and alveolar fricative—among the three are remarkable in Tennyson. Although Plamondon has not acquired sufficient differences to support his hypotheses, the results of this study on both vowels and consonants support his hypotheses.

Before concluding this paper, possible compelling findings that should be intensified in future research are highlighted. `Gephi` was used to create Figure 11, utilizing the numerical output of the CA generated by stylo. Similar to Figure 6, the color differences rely on modularity, which represents how well the data's parts are united. Close-knit poetry groupings based on CA output are represented by components of the same color. The thickness of the edges connecting each item revealed the degree of similarity. More similarities between the components are thought to

exist when the lines are thicker. Figure 11 shows how the poetry of Tennyson and Browning is categorized in different ways. Future attention should focus on the nodes and edges, colored in orange and light blue in Figure 10. The large amount of work in orange and light blue corresponds to the variables in the first quadrant of Figure 8, including T59_LAN-IK, T57_MRG-IK, T57_GER-IK, T55_MAUAUD, and T47_PRIESS. On the one hand, the association of the works can be translated as the *Idylls of the King* series and some narrative poems by Tennyson chiefly converged in the classification when only the light blue network is considered. On the other hand, to dig further, the poems in the two orange and light blue groups will likely be characterized by their female characters. Although female characters are not necessarily the main, leading characters of the poems, they play an important role in the stories. In addition, some other works, such as T55_MAUAUD ('Maud,' 1855), T47_PRIESS (*The Princess*, 1847), and T42_DORORA ('Dora,' 1842), put female characters in key positions as protagonists. Thus, the CA and PCA results suggest that there is perhaps a feature of consonant usage differences in these poems with female characters. Although thorough research is necessary, this study findings suggest that sound features found in certain literary works or poets exhibit a degree of consistency with the poetic characters portrayed in the poems.

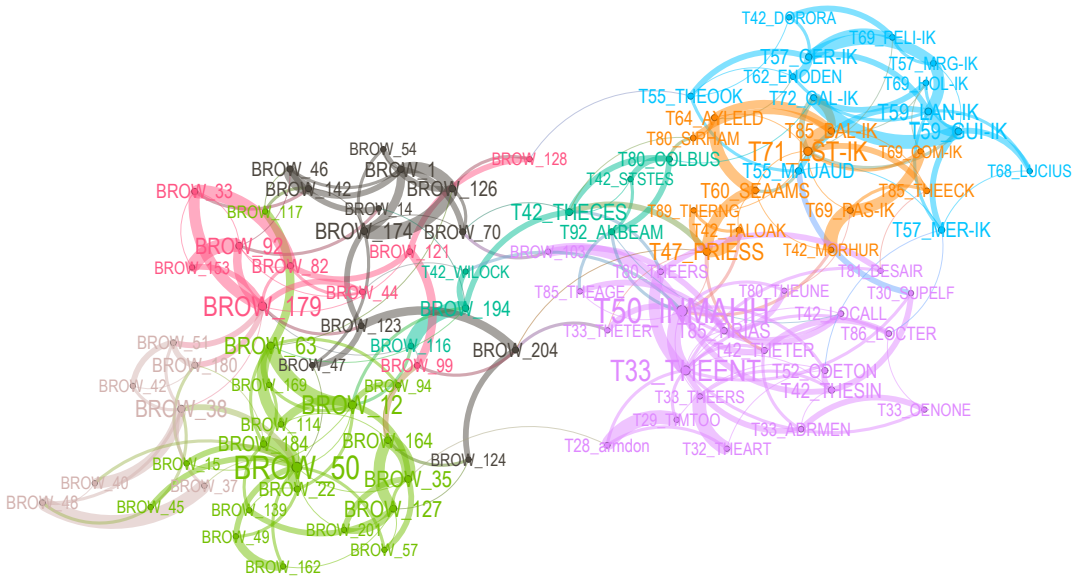


Figure 11: Network representation of consonants from the CA results

4. Conclusion

This study primarily aimed to clarify whether tabulated phoneme frequencies could show elements of Tennyson's and Browning's poetry and to highlight which phonemes are the characteristics of each poet in case the tabulated data detect differences between the two poets. Considering the limitations of prior studies, the top 50 phoneme frequency poems from Tennyson and Browning

were analyzed to thoroughly investigate their respective poetry collections. As determined by CA and PCA, the tabulated data of the IPA were efficient in finding the differences between the two authors and capturing the characteristics in their poetry. Although further in-depth research is necessary, this study findings corroborated the previous study's hypotheses while revealing a new prospect that the sex of characters in poems could coincide with the sound use of poets.

Acknowledgements

This work was supported by the Japan Science and Technology Agency Support for Pioneering Research Initiated by Next Generation [grant number JPMJSP2138].

Bibliography

[1] Bastian, M., M. Heymann and M. Jacomy. (2009) Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the Third International ICWSM Conference*, San Jose, California, USA, 3/1: 361–362.

[2] Burrows, J. (1987) *Computation into Criticism: A Study of Jane Austen's Novels*. Oxford: Clarendon Press.

[3] Delphi Poets Series (2013) *Alfred, Lord Tennyson*. East Sussex: Delphi.

[4] Delphi Poets Series (2012) *Robert Browning*. East Sussex: Delphi.

[5] Tawfiq, H. H. (2020) A Study of the Phonological Poetic Devices of Selected Poems of Robert Browning and Alfred Tennyson. *English Language and Literature Studies*. 10/4: 16–33.

[6] Eder, M., J. Rybick and M. Kestemount. (2016) Stylometry withR: A Package for Computational Text Analysis. *The R Journal*. 8/1: 107–121.

[7] Hoover, L. D. (2012) The Tutor's Story: A Case Study of Mixed Authorship. *English Studies*. 93/3: 324–339.

[8] Imao, Y. CasualConc (ver. 3.0.4). Accessed in December 30, 2022. https://sites.google.com/site/casualconc/

[9] Kubozono, H. (1998) *Onsei-gaku, On-in-ron. (Phonetics and Phonology)*. Tokyo: Kuroshio.

[10] Plamondon, R. M. (2005) Computer-Assisted Phonetic Analysis of English Poetry: A Preliminary Case Study of Browning and Tennyson. *TEXT Technology*. 14: 153–175.

[11] Ricks, C. (Ed.) (1987) *The Poems of Tennyson*. vol. I–III. (Second edition) London: Longman.

[12] Tabata, T. (1998) Corpus-ni Motoduku Buntairon Kenkyu (Corpus-based Stylistic Studies). In *English Corpus Linguistics*. Tokyo: Kenkyu-sha.

[13] Tabata, T. (2004) *–ly* Fukushi-no Seikihinndo Kaiseki-niyoru Bunntai Shikibetu: Koresupondensu Bunnseki-to Syuseibunn Bunnseki-niyoru Hikaku Kennkyuu (Style Identification of–*ly* Adverbs by the Occurrence Frequencies: Comparing the Correspondence Analysis and Principal Component Analysis. *Denshika Gengo Shiryou Bunseki Kenkyu 2003*. 97–114.

[14] Tabata, T. (2012) Textometry of Co-authorship: Dickens, Collins, and their collaborations. *Denshika Gengo Shiryou Bunseki Kenkyu 2011–2012*. 3–17.

Appendices

Appendix 1: Corresponding table of abbreviated file names and poem titles of Browning

Abbreviated file names	Poetry titles
BROW_1	'Tertium Quid'
BROW_103	'Saul'
BROW_114	'Juris Doctor Johannes-Baptista Bottinius Fisci et Rev Cam Apostol Advocatus'
BROW_116	'Easter-Day'
BROW_117	'The Book and the Ring'
BROW_12	'SORDELLO BOOK THE FIRST'
BROW_121	'Pompilia'
BROW_123	'Ivàn Ivànovitch'
BROW_124	'PAULINE'
BROW_126	'The Other Half-Rome'
BROW_127	'La Saisiaz'
BROW_128	'Caliban upon Setebos'
BROW_139	'WITH FRANCIS FURINI'
BROW_14	'Cleon'
BROW_142	'The Ring and the Book'
BROW_15	'WITH GEORGE BUBB DODINGTON'
BROW_153	'FIFINE AT THE FAIR'
BROW_169	'Pacchiarotto, and How He Worked in Distemper Pacchiarotto'
BROW_174	'Half-Rome'
BROW_179	'Guido'
BROW_180	'Mr Sludge, 'The Medium''
BROW_184	'RED COTTON NIGHT-CAP COUNTRY, OR, TURF AND TOWERS'
BROW_194	'Christmas-Eve'
BROW_201	'WITH CHARLES AVISON'
BROW_204	'James Lee's Wife I'
BROW_22	'FUST AND HIS FRIENDS'
BROW_33	'Count Guido Franceschini'
BROW_35	'Jochanan Hakkadosh'
BROW_37	'Pietro of Abano'
BROW_38	'THE INN ABLUM'
BROW_40	'Clive'
BROW_42	'In a Balcony FIRST PART'
BROW_44	'Fra Lippo Lippi'
BROW_45	'WITH DANIEL BARTOLI'
BROW_46	'The Pope'
BROW_47	'Ned Bratts'
BROW_50	'Aristophanes' Apology'
BROW_51	'Bishop Blougram's Apology'
BROW_54	'The Flight of the Duchess I'
BROW_57	'WITH GERARD DE LAIRESSE'
BROW_63	'Dominus Hyacinthus de Archangelis Pauperum Procurator'
BROW_70	'BALAUSTION'S ADVENTURE'
BROW_82	'Giuseppe Caponsacchi'
BROW_92	'PRINCE HOHENSTIEL-SCHWANGAU, SAVIOUR OF SOCIETY'
BROW_94	'An Epistle'
BROW_99	'A Death in the Desert'

Appendix 2: Corresponding table of abbreviated file names and poem titles of Tennyson

Abbreviated file names	Poetry titles
T28_armdon	'Armageddon'
T29_TIMTOO	'Timbuctoo'
T30_SUPELF	'Supposed Confessions Of A Second-rate Sensitive Mind Not In Unity With Itself'
T32_THEART	'The Palace Of Art'
T33_ADRMEN	'A Dream Of Fair Women'
T33_OENONE	'Oenone'
T33_THEENT	'The Lovers Tale A Fragment'
T33_THEERS	'The Lotos Eaters'
T33_THETER	'The Millers Daughter'
T42_MORHUR	'Morte Darthur'
T42_STSTES	'St Simon Stylites'
T42_TALOAK	'The Talking Oak'
T42_THECES	'The Two Voices'
T42_THESIN	'The Vision Of Sin'
T42_THETER	'The Gardeners Daughter'
T42_WILOCK	'Will Waterproofs Lyrical Monologue Made At The Cock'
T47_PRIESS	*The Princess*
T50_INMAHH	*In Memoriam A.H.H*
T52_ODETON	'Ode On The Death Of The Duke Of Wellington'
T55_MAUAUD	'Maud'
T55_THEOOK	'The Brook'
T57_GER-IK	'Geraint And Enid'
T57_MER-IK	'Merlin And Vivien'
T57_MRG-IK	'The Marriage Of Geraint'
T59_GUI-IK	'Guinevere'
T59_LAN-IK	'Lancelot And Elaine'
T60_SEAAMS	'Sea Dreams'
T62_ENODEN	*Enoch Arden*
T69_COM-IK	'The Coming Of Arthur'
T69_HOL-IK	'The Holy Grail'
T69_PAS-IK	'The Passing Of Arthur'
T69_PELI-IK	'Pelleas And Ettarre'
T71_LST-IK	'The Last Tournament'
T72_GAL-IK	'The Round Table Gareth And Lynette'
T80_COLBUS	'Columbus'
T80_SIRHAM	'Sir John Oldcastle Lord Cobham'
T80_THEERS	'The Sisters'
T80_THEUNE	'The Voyage Of Maeldune'
T81_DESAIR	'Despair'
T85_BAL-IK	'Balin And Balan'
T85_THEAGE	'The Ancient Sage'
T85_THEECK	'The Wreck'
T85_TIRIAS	'Tiresias'
T86_LOCTER	'Locksley Hall Sixty Years After'
T89_THERNG	'The Ring'
T92_AKBEAM	'Akbars Dream'

匿名政治パンフレットの計量的分析
―「レ枢機卿のマザリナード」の帰属検証―

涌井萌子

大阪大学人文学研究科
〒560-0043 豊中市待兼山町 1–5
Email: u738173g@ecs.osaka-u.ac.jp

概要　本稿は，レ枢機卿のマザリナードについて，デジタルヒューマニティーズを用いて，帰属の検証を行うことを目的としている。17 世紀フランスの匿名政治文書マザリナードは，虚偽虚飾が多く文体模写や偽称が認められるテキストであることから，帰属の確定には多くの困難がある。レ枢機卿の場合，その困難さは既刊の作品集・全集におけるマザリナードの収録数の違いに現れており，マザリナードの特徴だけでなく，編者の主観が介入する全集・作品集の性質も影響している。そのため，客観的かつ検証可能で再現性のある量的な観点から，デジタルヒューマニティーズの分析手法を用いて，帰属を検討する必要があると考えた。本研究では，レ枢機卿のマザリナードと考えられているテキストについて，クラスター分析および主成分分析を行い，帰属の再検証が必要なテキストの洗い出しと，それらのテキストに関して，質的な観点から帰属容認の可否を検証し，マザリナードを量的に分析する際の課題や問題点を提示する。

キーワード　マザリナード，クラスター分析，主成分分析，著者推定

Quantitative Analysis of anonymous political pamphlets:
The authorship verification of the mazarinades by Cardinal de Retz

Moeko WAKUI

Graduate School of Humanities, Osaka University
1-5 Machikaneyama-cho, Toyonaka, Osaka, 560-0043 Japan

Abstract　This paper aims to verify the attribution of “mazarinade” of Cardinal de Retz by using digital humanities. The Mazarinade, an anonymous political document of 17th century France, is a text with many falsehoods, stylistic imitations and pseudonyms, which makes it difficult to determine the attribution. In the case of Cardinal Re, this difficulty is reflected in the difference in the number of Mazarinades included in existing collections and complete works, which is influenced not only by the characteristics of the Mazarinade but also by the nature of the collections and complete works in which the subjectivity of the editor intervenes. Therefore, it was considered necessary to examine the attributions from an objective, verifiable and reproducible quantitative perspective, using the analytical methods of Digital Humanities. In this study, cluster and principal component analyses are carried out on texts considered to be the work of Retz. It identifies texts whose attribution needs to be re-examined, examines the acceptability of attribution from a qualitative perspective with regard to these texts, and presents the issues and problems in analysing the Mazarinade quantitatively.

Keywords　mazarinade, cluster analysis, PCA, authorship attribution

1.　はじめに：マザリナードと著者推定

マザリナード mazarinade とは，フロンド La Fronde の時期に出回った大量の印刷物である。フランス史においては宗教戦争やフランス革命の時期にも印刷物が大量に出回ったが，短期間に集中して出版されたという点で，マザリナードは傑出している。

フロンドは，太陽王ルイ 14 世が成人する前の 1648 年から 1653 年に起こったフランス全土を巻き込んだ内戦である。乱は，三十年戦争が長引き，王権が軍事費をまかなうために増税を繰り返したことでフラストレーションが溜まっていたパリ市民が決起したことが直接的な原因として起こった。間接的な原因としては，ルイ 13 世とその宰相リシュリュー枢機卿以降，中央集権化や宮廷への権利集中が進んだのに対して，権力の監視者であり，宮廷の政策に意見することができたパリ高等法院や，血族であることによって恩恵を受けていた親王たちといった既得権益を持つ人々が，抵抗しようとしたことがある。これによってこの反乱は，増税に反対するパリ市民を中心とした市民レベルの小規模な暴動にとどまらず，パリ高等法院やコンデ親王，ラ・ロシュフコーなど王族や貴族を含む様々な政治家が参加し，宮廷，特に宰相マザラン枢機卿に反抗する政治闘争の性格を持つに至った。

マザリナードは「投石遊び fronder」を語源とするこのフロンドにおける「言葉の石」であるが，マザランに対してのみ投げられたのではなく，フロンドに関わる様々な党派に向かって投げられたものであった。対象や目的が多種多様であり，また攻撃を主眼とするために誹謗中傷や虚飾・潤色を多く含むこれらのテクスト群は，およそ 5600 種類がフランスを中心に現存する。

執筆者は，特にレ枢機卿 Jean François Paul de Gondi, Cardinal de Retz が執筆したと考えられているマザリナードに限定して研究を行っている。

1. 1. レ枢機卿 Jean François Paul de Gondi, Cardinal de Retz について

レ枢機卿 Jean François Paul de Gondi, Cardinal de Retz (1613-1679)とは，聖職者，回想録作家である。最も有名な著作『メモワール』で活写しているように，フロンド La Fronde(1648-1653)に参加した政治家の一人であり，レは武力による戦いにおいても，現在「マザリナード」と呼称されるパンフレを用いた言論による戦いにおいても，積極的にこの内戦に参加した。レ枢機卿が論争家 polémiste であるという側面は特にミリアム・ティンビディ[1]によって強調されており，その産物としての論争文書も多数現存している。

後ほど 2.2 で整理するが，レ枢機卿のマザリナードと考えられているテキストは，作品集・作品全集によってその数が様々である。これは，第一にレ枢機卿本人が「書いた」と言っていなくても，近しい人物が書いているテキストであれば，レ枢機卿に帰属する場合があること，第二にタイトルに「レ枢機卿が書いた」という文言が含まれていても，その文言が帰属の根拠として採

[1] Myriam Tsimbidy, *Le cardinal de Retz polémiste*, Université de Saint-Etienne, Renaissance et Age Classique, 2005.

用されるか否かは，当時のレ枢機卿を取り巻く政治情勢や，同時代人の証言に左右されることが影響している。また，この数の違いは，作者同定に大変困難があるマザリナードの性質だけでなく，ティンビディ（2019）[2]では「編者の主観が介入する」という全集・作品集の性質が生み出したものであると考えられている。

いずれにせよ，マザリナードが匿名文書であることが影響しており，テキストの帰属判明を難しくしている原因となっている。

1. 2. マザリナードとデジタルヒューマニティーズについて

テキストの帰属判明を難しくしている原因がマザリナードにあるにせよ，全集・作品集にあるにせよ，根本にはマザリナードが匿名文書であることが影響している。そのため，ある作家のマザリナードを扱う場合，必ずテキストの内実に迫る前に，コーパス選択と根拠を明示する必要が生じる。質ではなく量の面から，マザリナードの帰属を再検討した例に，シラノ・ド・ベルジュラックの例がある。

シラノ・ド・ベルジュラック Cyrano de Bergerac は，エドモン・ロスタンの韻文戯曲『シラノ・ド・ベルジュラック』の元となった実在の剣術家であり作家である。彼もまた，レ枢機卿と同様に，フロンドの時代を生きた人物だが，レ枢機卿の場合はフロンドの参加者であり，自分の政治家としての立場のためにマザリナードを書いたのに対し，シラノの場合は純粋に政治的態度を示す目的で書いたとは考えられていない。彼の執筆動機には，経済的な理由も含まれていると考えられ，当時流行のマザリナードを書くことで金策を講じたのだとされている。

シラノのマザリナードの場合, マザリナードに書かれた頭文字DBとBDというイニシャルが，帰属を支持する内外の唯一とされていたが，その証拠の脆弱性を指摘したマドレーヌ・アルコベールが量的に帰属の再検討を行なっている。

アルコベール(2012)[3]では，因果関係の表現について，シラノの他作品と比較し，他作品の傾向とマザリナードの使用傾向が合致しないことから，アルコベールは「「事実に基づいた結論が，糾弾しようとした批評家の「確信」と衝突」している」ことを提示した。このことによって，それまで，シラノ・ド・ベルジュラックは７編のマザリナードを書いたと考えられていたが，その量的な検討と質的な再検討の結果，１編も書いていなかったと結論づけられている。

---

[2] Myriam Tsimbidy, « Création et fabrications des *Œuvres complètes* du cardinal de Retz », dans *Éditer les œuvres complètes (XVIe et XVIIe siècles)*, sous la direction de Philippe Desan et Anne Régent-Susini, Classiques Garnier, p. 333-347.

[3] Madeleine Alcover, « Le *Cyrano de Bergerac* de Jacques Prévot », *Les Dossiers du Grihl* [En ligne], Les dossiers de Jean-Pierre Cavaillé, Libertinage, athéisme, irréligion. Essais et bibliographie, mis en ligne le 17 avril 2012. (http://journals.openedition.org/dossiersgrihl/5079)

2. 方法

2. 1. 目的

クラスター分析・主成分分析を行い，レ枢機卿の作品として一括されている作品群の中で，レ枢機卿自身が帰属を明示している作品とその他とでどれほど距離があるのかを探るとともに，これまで全集・作品集編者が示してきた見解と異なる結果について質的に再評価を行うことで，レ枢機卿のマザリナードの範囲を量的観点・質的観点の双方から明らかにすることを目指す。

2. 2. 使用したコーパス

本研究では，レ枢機卿のマザリナード作品群を人文情報学的アプローチから再評価するという目的のため，レ枢機卿のマザリナードとされている 21 点の作品群のほか，比較対象として他作家のマザリナードを含めた 23 点のコーパスを用いた（表 1 及び表 3）。

レ枢機卿のコーパスについては，既刊の全集・作品集に収録されたすべてのものを対象とし，テキストはシャンピオン社による最新版の全集（Cardinal de Retz, *Conjuration de Fiesque et Pamphlets*, tome VII, Paris, Honoré Champion, 2011.）を用いた。

表 1　レ枢機卿のコーパス（執筆年代順）

年代順	タイトル	表記
1	La vérité reconnue ou les intrigues de Saint-Germain	Doute1
2	La déroute des cabalistes au jardin de Renard	Doute2
3	Apologie des Frondeurs	Doute3
4	Avis important et nécessaire à M. de Beaufort et à Monsieur le Coadjuteur	Doute4
5	Défense de l'ancienne et légitime Fronde	Retz1
6	Le Solitaire aux deux désintéressés	Coed1
7	Avis désintéressé sur la conduite de Monseigneur le Coadjuteur	Retz2
8	Discours libre et véritable sur la conduite de Monsieur le Prince et Monseigneur le Coadjuteur	Coed2
9	Réponse du curé à la lettre du marguillier sur la conduite de Monseigneur le Coadjuteur	Coed3
10	Lettre de Monsieur le Prince de Conty écrite au Roi sur son voyage en Berry	Doute5
11	Les Contre-temps du sieur de Chavigny, premier ministre de Monsieur le Prince	Retz3
12	Manifeste de Monseigneur le Duc de Beaufort, par lequel il déclare se joindre à Son Altesse Royale, au Parlement et à la ville de Paris	Doute6
13	Manifeste de Monseigneur le duc de Beaufort, général des armées de son Altesse Royale	Retz4
14	Les intrigues de la Paix et les négociations faites à la cour par les amis de Monsieur le Prince depuis sa retraite ne Guyenne jusques à présent	Coed4
15	Le vrai et le faux de Monsieur le Prince et de Monsieur le cardinal de Retz	Retz5
16	Intérêts du Temps	Retz6
17	Le Vraisemblable sur la conduite de Monseigneur le cardinal de Retz	Retz7
18	Suite véritable des Intrigues de la Paix et les négociations de Monsieur le Prince à la cour jusques à présent	Coed5
19	Avis aux malheureux	Coed6
20	Remarques sommaires sur la maison des Gondy, par le sieur Hozier, gentilhomme ordinaire de la Maison du Roi, généalogiste de sa Majesté et juge général des armes	Doute7

et des blasons de France
21 Discours sur la conduite et l'emprisonnement de Monsieur le Cardinal de Retz Doute8

2. 2. 1. レ枢機卿が自分自身で執筆していることを言及しているコーパス

レ枢機卿が自分自身の信念によって執筆した７編のマザリナードは，現代のわれわれから最も想像しやすい「著者」としての関わり方である。政敵の文体を模写したテキストやこの関わり方によって数えられるマザリナードが７編あり，Retz1-7 として分析に加えたこの７編に関して，異論を唱える編者はいない。

表 2　レ枢機卿のマザリナードと書き手

	GEF[4]	Pléiade[5]	Bibliographie[6]	Champion[7]
レ自身が書いた	7	7	7	7
党派の人間が書いた	6	1	1	6
関与が考えられる	×	×	×	8
収録数	13	8	8	21

2. 2. 2. レ枢機卿が党派の友人や雇った文士に書かせたことを言及しているコーパス

作品全集にはこの７編にさらに６編加えられているが，ここで含まれているのは，パトルやデュ・ポルテル，ギー・ジョリといったレの党派に属する友人あるいは雇われ文士たちによる作品である。

17 世紀の文学場について『作家の誕生』を書いたアラン・ヴィアラ[8]によれば，正確に言えば「クリエンテリズモ clientélisme[9]」という語を用いるのが適切である。パトロンとその援助を受け

---

[4] Cardinal de Retz, *Œuvres*, dans *Les Grands Écrivains de la France*, direction de M. Ad. Renier, Nouvelle édition. rev. sur les autographes et sur les plus anciennes impressions et augmentée, tome V(Pamphlets), Paris, Hachette, 1880.

[5] Cardinal de Retz, *Mémoires ; la conjuration du comte Jean-Louis de Fiesque, Pamphlets*, textes présentés et annotés par Maurice Allem et Edith Thomas, Paris, Gallimard, « Bibliothèque de la Pléiade », 1956.

[6] Bertière, Simone. (2000). *Bibliographie des Écrivain Français ; Le Cardinal de Retz ; Bibliographie thématique des littératures francophones européennes*, 20, Memini, Paris-Roma.

[7] Cardinal de Retz, *Conjuration de Fiesque et Pamphlets*, textes établis, avec introduction, notes, bibliographie, index des noms de personnes, index des noms de lieux, reproduction de manuscrits, illustrations, par Jacques Delon, tome VII, Paris, Honoré Champion, 2011.

[8] Alain Viala, *Naissance de l'écrivain*, Paris, Éditions de Minuit, « Le sens commun », 1985, p. 51-57.

[9] クリスチアン・ジュオー『マザリナード』（嶋中博章，野呂康訳，水声社，2012 年. JOUHAUD

る著述家という意味でメセナと混同される「クリエンテリズモ」だが，メセナと比較してパトロンの介入度が高い，奉仕という社会的関係である。この「クリエンテリズモ」を結んだ書き手にもさまざま存在しており，モローが言及するところの「徒党に結びついた雇われ文士」と「信念も心的なきずなも持たない，雇われ文士」がいる。「クリエンテリズモ」はそもそも，中世における封建社会の性質を部分的に受け継いだものとされ，基本理念として忠義の精神があげられるが，その精神を持つのはもっぱら前者であり，後者のような文士は「クリエンテリズモ」を前にして悠々と二枚舌を用いることもあった。パトロンに雇われてパトロンの制限を受けた著述も存在した一方で，雇われていた方があっさりと凋落したパトロンを見捨て，「クリエンテリズモ」を渡り歩く例もあったことを，ヴィアラは紹介している。

パトルによる『大司教補殿の行動に関する教会管理人の手紙に対する主任司祭の返答』（Coed3）にはレ本人が加筆修正あるいは原案として関わった可能性も指摘されており[10]，同作品やギー・ジョリによる『ギュイエンヌ後退から現在に至るまで親王殿下の友人たちが宮廷で行った和平の陰謀と交渉』（Coed4）に関してはいくつかの断片がレ枢機卿のテクストと近似しているという指摘もある[11]。ただ，レがどれほど手を入れればそれがレの作品になるのかという判断は難しい。

2. 2. 3. 作品集編者によってレ枢機卿が関連した可能性が示唆されているコーパス

ドロンによって編まれた最新の全集でこれほどまでにレ枢機卿のマザリナードの数が増えたのは，彼が同時代の証言を丁寧に拾い上げたからである。ドロンによるコーパス作成の方針はこのように書かれている。

> コーパス作成にあたって，我々は当然，誤って彼に作者の資格が与えられ，後々本当の作者がわかったものを取り除き，同様に彼が自分の作家集団に着想を与えたが，遠くからしか参加しておらず，作品にするにあたっての気遣いを他に任せたままにしておいたものを取り除いた。反対に，彼に帰属させるのに正当性があると思われる幾つかの小冊子が見られた。それはその内容や形式の全体あるいは部分に対して，歴史的・文献的な観点からの確固たる根拠があった。(Cardinal de Retz, *Œuvres complètes*, *op.cit.*, tome VII, p. 350.)

ドロンは「遠くからde loin」の範囲に関して特に詳しくは言及していない。そのため，レ枢機

(Christian), *Mazarinades : la Fronde des mots*, Collection historique, Aubier, 2009. の訳本）ではこの語に「保護—被保護関係」という訳語を当てているのだが，その訳語であると「メセナ」にも通用してしまうので，ここではアラン・ヴィアラ『作家の誕生』（塩川徹也監訳，藤原書店，2005 年.）での訳語を採用した。

[10] Cardinal de Retz, *Mémoires ; la conjuration du comte Jean-Louis de Fiesque, Pamphlets*, *op.cit.*, « Bibliothèque de la Pléiade », p. 1181.

[11] シャンピオン全集における当該マザリナードについての解説参照（*Œuvres complètes*, *op.cit.*, tome VII, p. 522 et 582.)

卿が「書かせたfaire écrire」としている作品についてはもちろんのこと，マザランなど同時代の人物による証言を根拠に，比較的広範に採用しているのである。

ドロンによる可能性の検証という大仕事は素晴らしい成果であるが，前述の通り「彼が書いた」という偽称も政治的戦略として通用する時代なのである。レ枢機卿自身，かなりの量のパンフレに目を通すよう心がけていたことは『メモワール』に明らかだが[12]，出回っていた全てに目を通すことができたわけではなく，ましてや自分のものと偽って出されたもの全てに「本当の véritable」という形容詞をつけて反駁していたわけでもない。このような複雑な状況である以上，同時代の証言をどこまで事実として採用するかということには慎重さが求められる。

2. 2. 4. サラザンのコーパス

レ枢機卿のコーパスの比較対象として，同等の知的階級の書き手ジャン＝フランソワ・サラザンJean François Sarasinによる，単著かつ単語数の適度な散文のマザリナードを“autre_”としてコーパスに入れる[13]。サラザンは元々レ枢機卿の従者だった，同世代の人物である。フロンド勃発後，レ枢機卿に敵対するコンデ親王やコンチ親王の秘書になり，マザリナードを書いた人物である。

表 3　使用したサラザンのコーパス

年代順	タイトル	表記
1	Le Frondeur bien intentionné aux faux-Frondeurs	autre_Sa1
2	Lettre d’un marguillier de Paris à son curé sur la conduite de Mgr le Coadjuteur	autre_Sa2

2. 2. 5. 使用コーパス

以上により，最終的な使用コーパスは表 4の通りである。

表 4　使用コーパスと語数

表記	タイトル	語数
autre_Sa1	Le Frondeur bien intentionné aux faux-Frondeurs	1532
autre_Sa2	Lettre d’un marguillier de Paris à son curé sur la conduite de Mgr le Coadjuteur	4209
Coed1	Le Solitaire aux deux désintéressés	5010
Coed2	Discours libre et véritable sur la conduite de Monsieur le Prince et Monseigneur le Coadjuteur	6359
Coed3	Réponse du curé à la lettre du marguillier sur la conduite de Monseigneur le Coadjuteur	8321
Coed4	Les intrigues de la Paix et les négociations faites à la cour par les amis de Monsieur le Prince depuis sa retraite ne Guyenne jusques à présent	2887
Coed5	Suite véritable des Intrigues de la Paix et les négociations de Monsieur le Prince à la cour jusques à présent	2141

[12] Cardinal de Retz, *Mémoires*, *op.cit.*, « Bibliothèque de la Pléiade », p. 624.

[13] コーパスは以下の作品集を用いた。Jean-François Sarasin, *Oeuvres de J.-Fr. Sarasin*, rassemblées par Paul Festugière, Paris, É. Champion, 1926.

Coed6	Avis aux malheureux	1801
Doute1	La vérité reconnue ou les intrigues de Saint-Germain	1565
Doute2	La déroute des cabalistes au jardin de Renard	1097
Doute3	Apologie des Frondeurs	2669
Doute4	Avis important et nécessaire à M. de Beaufort et à Monsieur le Coadjuteur	6286
Doute5	Lettre de Monsieur le Prince de Conty écrite au Roi sur son voyage en Berry	1289
Doute6	Manifeste de Monseigneur le Duc de Beaufort, par lequel il déclare se joindre à Son Altesse Royale, au Parlement et à la ville de Paris	3212
Doute7	Remarques sommaires sur la maison des Gondy, par le sieur Hozier, gentilhomme ordinaire de la Maison du Roi, généalogiste de sa Majesté et juge général des armes et des blasons de France	5382
Doute8	Discours sur la conduite et l'emprisonnement de Monsieur le Cardinal de Retz	20286
Retz1	Défense de l'ancienne et légitime Fronde	1569
Retz2	Avis désintéressé sur la conduite de Monseigneur le Coadjuteur	2133
Retz3	Les Contre-temps du sieur de Chavigny, premier ministre de Monsieur le Prince	1998
Retz4	Manifeste de Monseigneur le duc de Beaufort, général des armées de son Altesse Royale	740
Retz5	Le vrai et le faux de Monsieur le Prince et de Monsieur le cardinal de Retz	8198
Retz6	Intérêts du Temps	2195
Retz7	Le Vraisemblable sur la conduite de Monseigneur le cardinal de Retz	3218

2. 3. 分析手順

CasualConc を用いて，全てのコーパスをクラスター分析および主成分分析にかけた。

クラスター分析は，異なるものが混ざりあっている集団の中から互いに似たものを集めてクラスターを作り、対象を分類する手法である。データの分類や分類の基準の発見、新しい仮説の検証などに幅広く使われ，レ枢機卿とテキストの関わりの多寡に関わらず一様に扱われているコーパス群を，分類するのに最適である。

主成分分析は，データの分散をより良く説明するという観点から，クラスター分析で得た大分類に対して，データの内部構造を明らかにするものであり，情報をより少ない次元に集約することでデータを視覚化できる。主成分分析（PCA）は，多変量データを分析するための統計的手法の一つであり，データの相関関係を調べ，それを元にデータの次元を削減することで，データの構造を簡潔に表現することを目的としている。そのため，主成分分析はデータの可視化や，多変量データの解析，データ圧縮などに利用される。

クラスター分析は，データのグループを見つけるために，似た傾向や属性を持つデータをクラスターとしてまとめためによく用いられる。一方，主成分分析は，元の多次元データの情報を縮約するために，最も重要な情報を持つ主成分を抽出することにより，次元を削減し複雑なデータの可視化や特徴の把握などに用いられる。

この二つの手法を用いることによって，クラスター分析でレ枢機卿の作品群に，単著のグループ、共著・共作のグループ、関連の少ないグループのまとまりを見つける。その後，主成分分析によって，クラスター分析によって発見されたグループの検証を行い，従来の先行研究による分類から大きく外れたテキストについて，個別に質的な検証を行う。

3.　分析と考察

本節では，前節で整理したコーパスについて，クラスター分析・主成分分析を行い，その結果を提示した上で考察を行う。

3. 1. Retz4 を含めた分析

まずは，対象となるすべてのコーパスをクラスター分析にかけた（図 1）。

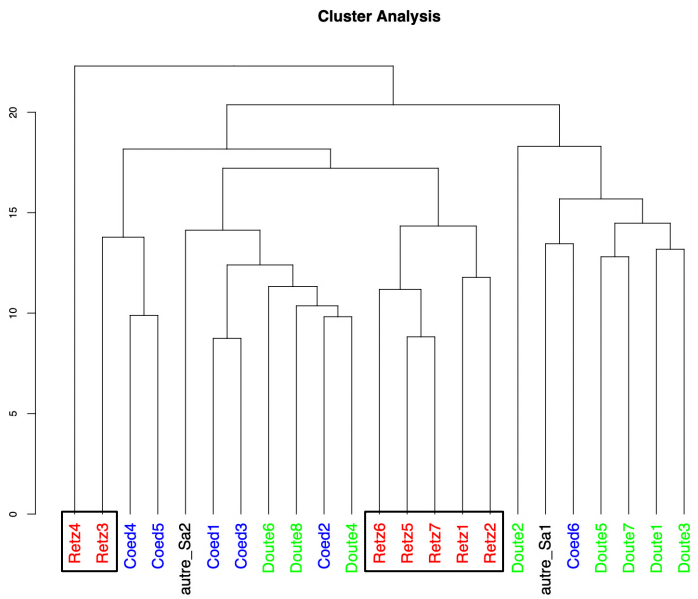


図 1　Retz4 を含めた場合のクラスター分析

結果を見ると，Ret3 及び Retz4 が他のレ枢機卿本人の作品群から離れた位置にあり，特に Retz4 に関しては他の作家（autre_Sa1,autre_Sa2）よりもさらに文体的距離が隔絶していることがわかる。この結果はクラスター分析のみに見られるものでは無い。同じコーパスを主成分分析にかけた（図 2）。

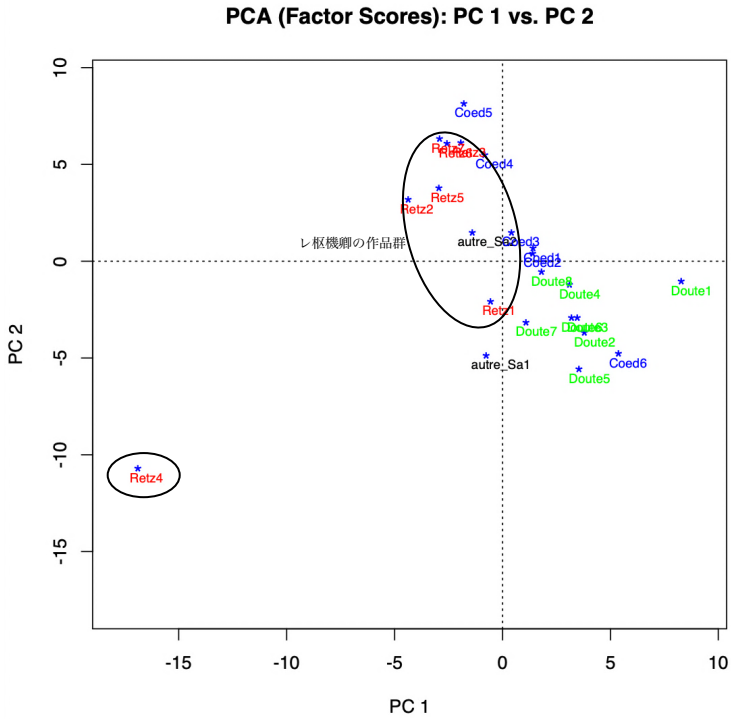


図 2　Retz4 を含めた場合の主成分分析

結果を見ると，Retz4 が他の作品群と比べて外れた位置にあり，この外れ値の存在によってレ枢機卿が書いた作品（Retz 群），共著や共作と考えられる作品（Coed 群），関連が示唆されるが疑わしい作品（Doute 群）の差異が明示されない他に，明らかに作者の異なる作品（autre_Sa2）がレ枢機卿の作品群の範囲に含まれてしまっている。

この結果は，マザリナードの分析においてクラスター分析と主成分分析が有用であることを同時に示す結果である[14]。というのも，レ枢機卿による第四のマザリナードである Retz4 は，レ枢機卿がボーフォール公の文体を模写して執筆したものであるからだ。他者の文体を模写したテクストであることに加えて，レ枢機卿が「彼の変な言葉遣い en son jargon」と形容した婉曲的で装飾に富んだ文体が，レ枢機卿やサラザンなどが用いた古典主義の先駆けと言われる簡潔かつ端正な文体といかに異なるかということがこれら二つの分析手法で示すことが可能だということがわかる。

しかしここでは，明らかに外れ値である Retz4 によって，結果が煩雑になってしまっているた

[14] 多変量解析においては Retz4 を Retz 群から離れ，autre 群とほぼ同じ距離である作品であるという結果を示していた。しかし Doute 群の一部の作品との比較が難しい結果を示したため，マザリナードの分析にあたってはクラスター分析と主成分分析を併用することが最も適していると判断した。

め，レ枢機卿の作品とされているテクスト群の帰属の量的再検証という本来の目的を考え，Retz4を外した 22 のテクストを再度，分析する（表 3）。

表 3　最適化したレ枢機卿のコーパス

年代順	タイトル	表記
1	Le Frondeur bien intentionné aux faux-Frondeurs	Autre_Sa1
2	Lettre d'un marguillier de Paris à son curé sur la conduite de Mgr le Coadjuteur	Autre_Sa2
3	Le Solitaire aux deux désintéressés	Coed1
4	Discours libre et véritable sur la conduite de Monsieur le Prince et Monseigneur le Coadjuteur	Coed2
5	Réponse du curé à la lettre du marguillier sur la conduite de Monseigneur le Coadjuteur	Coed3
6	Les intrigues de la Paix et les négociations faites à la cour par les amis de Monsieur le Prince depuis sa retraite ne Guyenne jusques à présent	Coed4
7	Suite véritable des Intrigues de la Paix et les négociations de Monsieur le Prince à la cour jusques à présent	Coed5
8	Avis aux malheureux	Coed6
9	La vérité reconnue ou les intrigues de Saint-Germain	Doute1
10	La déroute des cabalistes au jardin de Renard	Doute2
11	Apologie des Frondeurs	Doute3
12	Avis important et nécessaire à M. de Beaufort et à Monsieur le Coadjuteur	Doute4
13	Lettre de Monsieur le Prince de Conty écrite au Roi sur son voyage en Berry	Doute5
14	Manifeste de Monseigneur le Duc de Beaufort, par lequel il déclare se joindre à Son Altesse Royale, au Parlement et à la ville de Paris	Doute6
15	Remarques sommaires sur la maison des Gondy, par le sieur Hozier, gentilhomme ordinaire de la Maison du Roi, généalogiste de sa Majesté et juge général des armes et des blasons de France	Doute7
16	Discours sur la conduite et l'emprisonnement de Monsieur le Cardinal de Retz	Doute8
17	Défense de l'ancienne et légitime Fronde	Retz1
18	Avis désintéressé sur la conduite de Monseigneur le Coadjuteur	Retz2
19	Les Contre-temps du sieur de Chavigny, premier ministre de Monsieur le Prince	Retz3
20	Le vrai et le faux de Monsieur le Prince et de Monsieur le cardinal de Retz	Retz5
21	Intérêts du Temps	Retz6
22	Le Vraisemblable sur la conduite de Monseigneur le cardinal de Retz	Retz7

3. 2. Retz4 を除いた分析

まずは対象となる上記の再検討したコーパスをクラスター分析にかけた（図 3）。

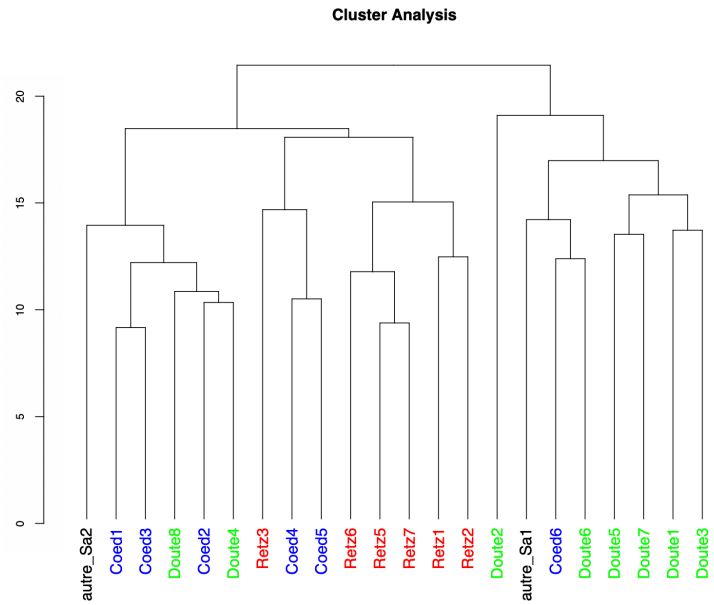


図 3　Retz4 を除いた場合のクラスター分析

結果を見ると明らかに大きく 3 つの群に別れていることが分かる。中央が Retz 群を中心としたグループである。このグループをさらに細かく，その枝分かれの様子を観察すると，クラスター分析による結果が論理構造や弁論の形態をよく反映していることが分かる。

アリストテレスは『弁論術』の中で弁論の形態を 3 つに大別しており，過去や未来の事象に対する議会での弁論を想定した審議的弁論，言動の正・不正を裁定する法廷的弁論，ある人物の言動の美醜や優劣に対する非難や賛美を想定した演説的弁論である。レ枢機卿のマザリナードとされる 7 編はこのうちの法廷的弁論と演説的弁論の特徴を持つが（表 4），その特徴をクラスター分析の結果が明示している点が興味深い。

表 5　レ枢機卿によるマザリナードと弁論の型

表記	弁論の型	内容	構造など
Retz1	演説	レの賛美	一人称二人称が多
Retz2	演説	レの賛美	
Retz3	演説	シャヴィニ卿の非難	感嘆文が多い
Retz4	文体模写		
Retz5	法廷	レとコンデの比較	第三者による二者比較
Retz6	法廷	レとコンデの比較	第三者による二者比較
Retz7	法廷	レとコンデの比較	第三者による二者比較

Retz4を外したことで，帰属の量的再検証をするにあたってよりわかりやすく結果が見られるようになり，今後検討すべきテキストの優先順位がわかりやすくつけられた。

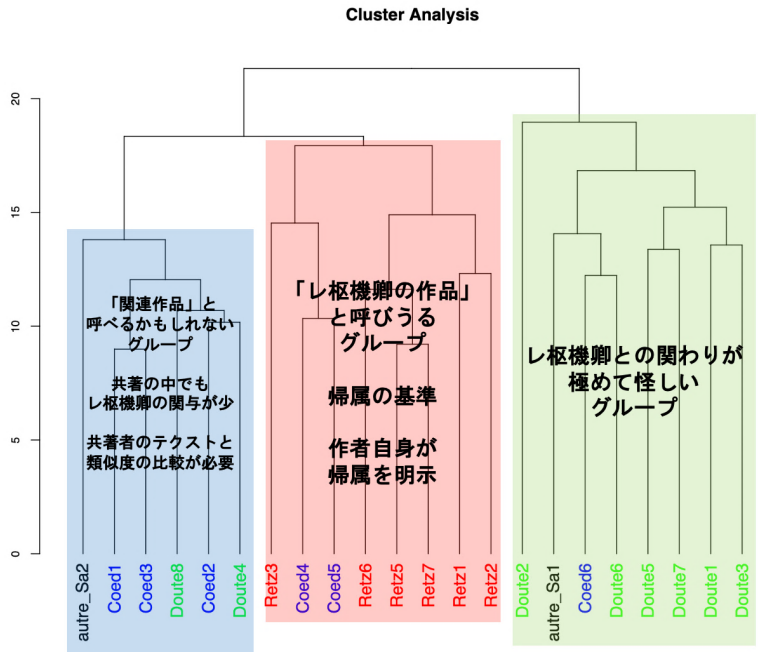


図 4　帰属のグループと優先順位

研究者・編者の主観の影響を受けない新たな分類であり，全ての作品集・作品全集・書誌に掲載されているもの，複数で認められているものについて，概ね従来に沿う結果になった。一方で，主成分分析の結果と合わせてみてみると，これまで先行研究者や全集編者の分類と異なるいくつかの例外の発見することができる。

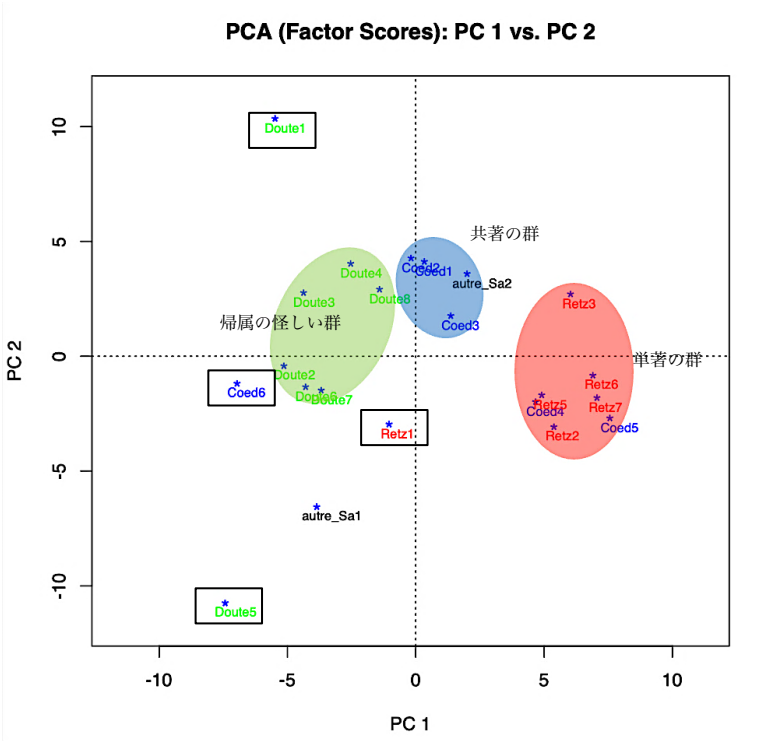


図 5　主成分分析によるマザリナードの分類

そこで，Retz1, Coed6, Doute1, Doute5 を質的に検討し，本来属すべき群から外れている原因を探る。

3. 3.外れテキストの質的な再検討と評価

クラスター分析・主成分分析によって見えてきた外れテキストを質的に再検討し，レ枢機卿の作品として扱うべきか，評価を行う。

3. 3. 1. Retz1: *Défense de l'ancienne et légitime Fronde*

まず Retz の群でありながら一つ離れている Retz1 のテキストだが，このマザリナードはレ枢機卿の他のマザリナードと比べてどのような特徴があるのか，CasualConc の WordCount における LogLikelihood 比較を用いて，その比較を行った。

あるいは目的語“vous”あるいは二人称複数形の所有格“votre”が多く用いられ，一人称複数系も多く用いられていることがわかる。逆に，三人称単数あるいは非人称主語として用いられる“il”や“on”は，他のマザリナードに比べてあまり用いられていない。これは、“Vous”による政敵コンデ親王陣営への語りかけや二人称複数形“-ons”による民衆への語りかけが多いことによって、三人称単数が少なくなっていることによる。

単語 実行 頻度のみ

retz1 含む

589 types 1,569 tokens in 1 files

0	Words	LL	頻度	相対頻度	割合
5	vous	79.22	31	19.758	1.98%
77	préfère	15.07	3	1.912	0.19%
43	votre	13.38	6	3.824	0.38%
18	nous	12.24	14	8.923	0.89%
64	propre	11.04	4	2.549	0.25%
64	tyrannie	11.04	4	2.549	0.25%
77	dessus	10.74	3	1.912	0.19%
77	légitime	10.74	3	1.912	0.19%
32	leur	10.33	10	6.373	0.64%
98	ancienne	10.04	2	1.275	0.13%
98	attaquez	10.04	2	1.275	0.13%
98	bienfaits	10.04	2	1.275	0.13%
98	brouillés	10.04	2	1.275	0.13%
98	cependant	10.04	2	1.275	0.13%
98	deviennent	10.04	2	1.275	0.13%
98	fauteurs	10.04	2	1.275	0.13%
98	fronde	10.04	2	1.275	0.13%
98	mazarins	10.04	2	1.275	0.13%
98	regarde	10.04	2	1.275	0.13%
98	temple	10.04	2	1.275	0.13%

単語 実行 頻度のみ

retz1 含む

589 types 1,569 tokens in 1 files

0	Words	LL	頻度	相対頻度	割合
26	il	-27.12	12	7.648	0.76%
183	retz	-9.00	1	0.637	0.06%
7	le	-8.57	30	19.120	1.91%
43	cardinal	-6.44	6	3.824	0.38%
77	se	-6.27	3	1.912	0.19%
98	sur	-4.93	2	1.275	0.13%
77	m	-4.79	3	1.912	0.19%
32	on	-3.88	10	6.373	0.64%
35	d	-3.80	9	5.736	0.57%
15	qu	-3.70	16	10.198	1.02%
98	toutes	-3.08	2	1.275	0.13%
77	une	-2.91	3	1.912	0.19%
183	était	-2.89	1	0.637	0.06%
29	a	-2.68	11	7.011	0.70%
183	intérêts	-2.28	1	0.637	0.06%
16	dans	-2.06	15	9.560	0.96%
40	un	-2.05	7	4.461	0.45%
40	monsieur	-1.98	7	4.461	0.45%
183	assez	-1.93	1	0.637	0.06%
183	beaucoup	-1.82	1	0.637	0.06%

図 6　Retz1 と Retz2,3,5,6,7 との使用単語比較

おそらく3.2で扱った弁論の型の違いの影響も含めた，使用語彙による違いと考えられ，レ枢機卿自身が『メモワール』でこのテキストの帰属を認めていることを上回る反論にはなり得ないため，従来通りレ枢機卿のテキストとして扱うのが良いと結論づける。

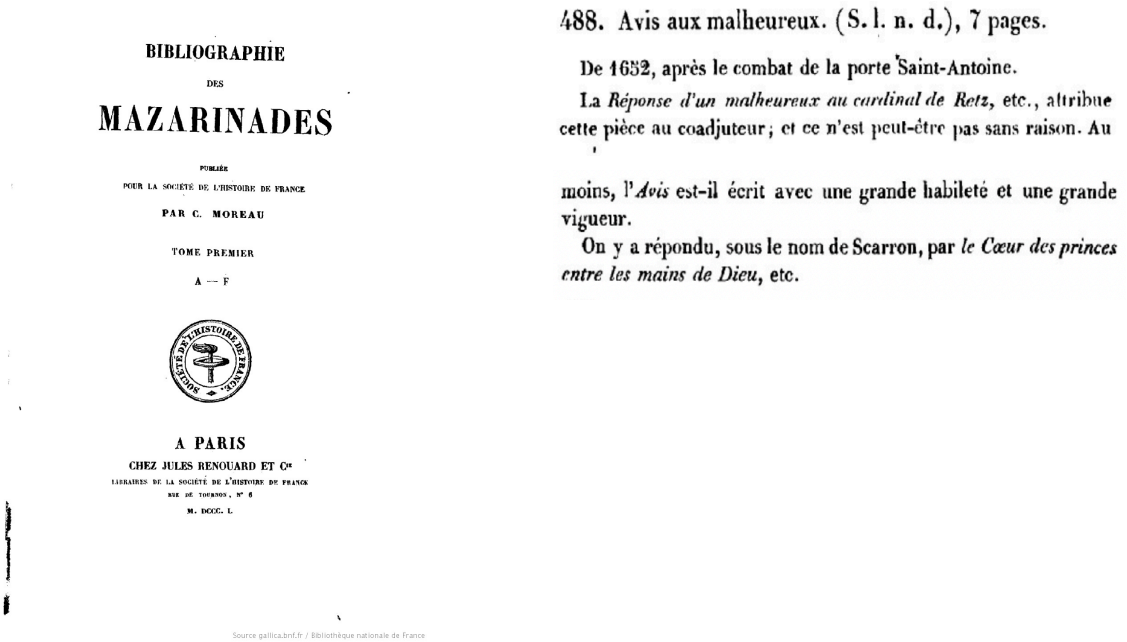
3. 3. 2. Coed6: *Avis aux malheureux*

レ枢機卿自身の言及あるいは先行研究者の指摘によって共著と考えられてきたが，極めて帰属が怪しいという結果が出た Coed6 について，論理構造や主張内容によってレ枢機卿の関連作品と呼びうるかどうか検討したい。

まず先行研究での扱いだが，Coed6 を関連作品に含めている全集編者のジャック・ドロンの根拠には，具体性がほとんどない[15]。ジャック・ドロンが根拠としているのは『マザリナード書誌[16]』という 19 世紀に編纂されたマザリナードを全てまとめたとされるカタログである。

[15] Cardinal de Retz, *Conjuration de Fiesque et Pamphlets*, *op.cit.*, tome VII, p. 643-645.

[16] Célestin Moreau, *Bibliographie des mazarinades*, tome I, Paris, la Société de l'histoire de France, 1850-1851.



BIBLIOGRAPHIE

DES

MAZARINADES

PUBLIÉE

POUR LA SOCIÉTÉ DE L'HISTOIRE DE FRANCE

PAR C. MOREAU

TOME PREMIER

A — F

A PARIS

CHEZ JULES RENOUARD ET Cie

LIBRAIRES DE LA SOCIÉTÉ DE L'HISTOIRE DE FRANCE

RUE DE TOURNON, N° 6

M. DCCC. L

488. Avis aux malheureux. (S. l. n. d.), 7 pages.

De 1652, après le combat de la porte Saint-Antoine.

La *Réponse d'un malheureux au cardinal de Retz*, etc., attribue cette pièce au coadjuteur; et ce n'est peut-être pas sans raison. Au moins, l'*Avis* est-il écrit avec une grande habileté et une grande vigueur.

On y a répondu, sous le nom de Scarron, par *le Cœur des princes entre les mains de Dieu*, etc.

Source gallica.bnf.fr / Bibliothèque nationale de France

図 7　『マザリナード書誌』表紙と Coed6 に関する記述

カタログ編纂者セレスタン・モローは、『レ枢機卿に対するある不幸な人からの返答』という他の作品がこの作品をレ枢機卿の作としており、「これはおそらく理由がないわけではないだろう」ということでレ枢機卿の作品に数えている。何か作品それ自体の中に具体的な根拠があったわけではないにも関わらず，この評価をもとに，全集やマザリナードの国際研究グループ RIM(Recherches internationales sur les Mazarinades)は，この Coed6 をレ枢機卿の作品として掲載している。

この作品の内容を見てみると，レ枢機卿の作品と考えられる根拠は，状況証拠でしかない。つまり、執筆当時の政治的立ち位置（国王派・反コンデ派）とは一致するものの，レ枢機卿の作品と断定できず，同じ政治的立場の他人の可能性が拭えない。

また，テキストの中で特に異質に思われるのは，パリ市民の扱いである。パリ市民とは，パリ司教座の枢機卿であるレにとって支持母体であり，帰属が明らかな Retz1-7 では「私たち nous」という形で総体として扱い、連帯を示している。これに対して Coed6 ではレ枢機卿が使わない「君 tu」という呼び方や命令形の多用が見られる。

さらに市民を分断する記述も見られ、第一・第二身分同様に政治変化に敏感な「大ブルジョワ gros bourgeois」「選ばれしもの élite」と今日明日の衣食住のみが興味の対象であり判断力に乏しく「哀れで不幸な民衆 pauvre malheureux peuple」と書かれている。レ枢機卿は，マザリナード以外の作品でも「民衆」支持の重要性を書いており，あえて「民衆」を分断する意義が見えない。

これらの理由から，Coed6 のテキストについて，量的にも質的にも，レ枢機卿の作品と呼ぶのは難しいのではないかと考えられる。

3. 3. 3. Doute1: *La vérité reconnue ou les intrigues de Saint-Germain* 及び Doute5: *Lettre de Monsieur le Prince de Conty écrite au Roi sur son voyage en Berry*

この 2 編はそもそもレ枢機卿本人による言及がなく，先行研究者や作品集編者の中でも，その帰属がほとんど認められていない Doute の群の中で，特に Retz 群から外れているテキストである。つまり，コーパスの中でも特にレ枢機卿が制作に関わった可能性が低いテキストと言える。

Doute 1 については，マザリナード研究者ユベール・キャリエが著書において，レ枢機卿が参加していたフロンド党派が掲げていた戦略との類似していることを根拠に，起草にレ枢機卿が非常に大きな役割を果たしたことを示唆している。全集編者のジャック・ドロンは，この記述を根拠に Doute1 を掲載しているが，証拠となりうる、他の同時代人による言及は確認されていない。また，宰相マザランと宮廷を激しく非難する内容は，レ枢機卿自身が『メモワール』で言及している「コンデ親王陣営が自分を攻撃するパンフレを書いたことに対抗するため」というマザリナードの制作動機に一致しないことから，更なる調査が必要だが，現時点で帰属はかなり疑わしいと言わざるを得ない。

Doute5 について，こちらも全く他人の作品である autre_Sa1 よりもさらに離れた位置にあるテキストである。この作品は Doute の中で唯一，政敵の名前を騙った作品であり，コンチ親王によるルイ 14 世への書簡の体裁をとり、外れ値として削除した Retz4 同様、文体模写の可能性がある作品である。レ枢機卿による文体模写なのか，政敵が書いたものなのかについては内容と背景を調べた上で慎重に検討する必要があるが、いずれにせよ，レ枢機卿の作品としてレトリックや道徳観を分析する，文学研究のコーパスと考えるのは厳しいことがわかる。

4. 終わりに

本研究では，クラスター分析・主成分分析を用いて，レ枢機卿のマザリナードコーパスについて，どの作品に帰属を認めるべきか，研究者・編者の主観の影響を受けない新たな分析結果を得た。全ての作品集・作品全集・書誌に掲載されているものや，複数で認められているものについては，概ね従来の帰属に沿う結果であったが，Retz1, Coed6, Doute1, Doute5 のテキストについてはこれまで先行研究者や全集編者の分類と異なる結果が出たため，それらを質的に再検証した。

今後は，主成分分析分析の結果にCoed1,2,3にautre_Sa2が含まれているので，autreの比較対象の作家を変えたときにどう結果が変わるのかを観察した上で，Coedを中心に「レ枢機卿の作品」に含めるかどうか検討する。また，コレスポンデンス分析を用いて，今回外れ値として除いたRetz4をRetz群に含める指標を模索し，他のテキストがどのような位置を取るか検証を進める必要があるだろう。将来的には，レ枢機卿の例から，デジタルヒューマニティーズによるマザリナードの帰属検証の方法を提示し，マザリナード全体の帰属問題に資する成果となるよう，進めていきたい。

文献

全集・作品集

Cardinal de Retz, *Œuvres*, dans *Les Grands Écrivains de la France*, direction de M. Ad. Renier, Nouvelle édition. rev. sur les autographes et sur les plus anciennes impressions et augmentée, tome III(Mémoires) et V(Pamphlets), Paris, Hachette, 1875 et 1880.

Cardinal de Retz, *Mémoires ; la conjuration du comte Jean-Louis de Fiesque, Pamphlets*, textes présentés et annotés par Maurice Allem et Edith Thomas, Paris, Gallimard, « Bibliothèque de la Pléiade », 1956.

Cardinal de Retz, *Mémoires ; précédés de La conjuration du comte de Fiesque*, texte établi avec introd., chronologie, notes, par Simone Bertière, Paris, Garnier, 1987.

Cardinal de Retz, *Mémoires*, Édition de Michel Pernot, Texte établi par Marie-Thérèse Hipp, Collection Folio classique (n° 3835), Paris, Gallimard, 2003.

Cardinal de Retz, *Conjuration de Fiesque et Pamphlets*, textes établis, avec introduction, notes, bibliographie, index des noms de personnes, index des noms de lieux, reproduction de manuscrits, illustrations, par Jacques Delon, tome VII, Paris, Honoré Champion, 2011.

Cardinal de Retz, *Mémoires*, textes établis avec introduction, notes, bibliographie, reproductions de manuscrits, illustrations, index des noms de personnes, index des noms de lieux, par Jacques Delon, tome VIII-IX, Paris, Honoré Champion, 2015.

研究書・その他

Alcover, M. (2012). « Le *Cyrano de Bergerac* de Jacques Prévot », dans *Les Dossiers du Grihl* [En ligne], Les dossiers de Jean-Pierre Cavaillé, Libertinage, athéisme, irréligion. Essais et bibliographie.

Bertière, Simone. (2000). *Bibliographie des Écrivain Français ; Le Cardinal de Retz ; Bibliographie thématique des littératures francophones européennes*, 20, Memini, Paris-Roma.

Sarasin, J-F. (1926). *Oeuvres de J.-Fr. Sarasin*, rassemblées par Paul Festugière, Paris, É. Champion.

Tsimbidy, M. (2005). *Le cardinal de Retz polémiste*, Université de Saint-Etienne, Renaissance et Age Classique.

Tsimbidy, M. (2019). « Création et fabrications des *Œuvres complètes* du cardinal de Retz », dans *Éditer les œuvres complètes (XVIe et XVIIe siècles)*, sous la direction de Philippe Desan et Anne Régent-Susini, Classiques Garnier, p. 333-347.

Viala, A. (1985). *Naissance de l'écrivain*, Paris, Éditions de Minuit, « Le sens commun ».

Moreau, C. (1850) *Bibliographie des mazarinades*, tome I, Paris, la Société de l'histoire de France.

アラン・ヴィアラ『作家の誕生』塩川徹也監訳，藤原書店，2005 年

クリスチアン・ジュオー『マザリナード』嶋中博章，野呂康訳，水声社，2012 年

アリストテレス『弁論術』戸塚七郎訳，岩波文庫，1992 年

参照サイト

Bibliographie des Mazarinades sur Bibliothèque Mazarine (https://mazarinades.bibliotheque-mazarine.fr/)

Bibliothèque municipale de Lyon (https://catalogue.bm-lyon.fr/)

Gallica(https://gallica.bnf.fr)

Recherches internationales sur les Mazarinades (http://mazarinades.org)

Tess of the d'Urbervilles の会話部による キャラクタライゼーション

曹 芳慧

大阪大学人文学研究科
〒560-0043 豊中市待兼山町 1–8
Email: u327503a@ecs.osaka-u.ac.jp

概要　Thomas Hardy は 19 世紀の有名なイギリスの小説家および詩人であり，著作が多く，作風に個性があり，非常に研究の価値のある作家である。その代表作である *Tess of the d'Urbervilles*（日本語訳：『テス』）についてすでに様々な解釈や論評がなされてきたが，本研究はこれまでの研究と異なる角度で小説の内容の分析を試みる。『テス』の会話部（直接話法）のみを対象として，各登場人物固有の言語的特徴や人物像，および話者間の人間関係に注目し，Hardy がいかに巧みに会話部を用いてキャラクタライゼーション（登場人物の特徴付け）を行なっているかを論じる。方法論として，TEI（Text Encoding Initiative）ガイドラインに基づき，小説テクストの構造化および会話部のアノテーションを行い，XML 解析によって登場人物間の対話ネットワークを考察し，主要な登場人物の会話に用いられている語の頻度パターン解析を対応分析で行う。

キーワード　Text Encoding Initiative (TEI)，XML 解析，会話部，キャラクタライゼーション，対応分析

Characterization by Idiolects in *Tess of the d'Urbervilles*

Cao Fanghui

Graduate School of Humanities, Osaka University
1-8 Machikaneyama-cho, Toyonaka, Osaka, 560-0043 Japan

Abstract　*Tess of the d'Urbervilles* is one of the representative works of Thomas Hardy, who was a famous British novelist of the 19th century. Most of the research on the novel *Tess of the d'Urbervilles* is qualitative. There are few studies conducted based on quantitative analysis on Hardy's literary works and quantitative research on the idiolects is also restricted due to technical limitations. This study focuses on the direct speech of the dialogue part and discusses how Hardy characterizes the characters of the work. In this paper, I will attempt to consider whether the findings shown by the quantitative research on the idiolects can generate new insights about Hardy's works, using methods that have not been implemented in the previous research. Firstly, I mark up and annotate the text of this novel based on the TEI (Text Encoding Initiative) guidelines to create a structured XML file. Secondly, I parse the file with

Python to extract the dialogue sections from the entire novel and represent the interaction network among the characters visually. Then, referring to the network, I analyze the frequency patterns of the words used in the idiolects of the main characters with CA (Correspondence Analysis) to clarify the linguistic features and personality traits of them, as well as the relationships within the speakers. In this way, this research overcomes the technical limitations in abstracting idiolects in novels and proposes a quantitative research approach for the dialogue part. By doing so, we are able to discuss how the author skillfully uses the idiolects to perform characterization in the novel.

Keywords Text Encoding Initiative(TEI), XML Analysis, Idiolects, Characterization, Correspondence Analysis

1. はじめに

Tess of the d'Urbervilles は，イギリスで創刊された絵入り週刊誌 *Graphic* に連載され，1891 年に James R.Osgood, Mc Ilvaine&Co.によって出版された小説であり，Thomas Hardy（トマス・ハーディ，1840 年 6 月 2 日～1928 年 1 月 11 日）の代表作の一つである。日本語訳ではよく『テス』の題名が使用されているため，以下は『テス』と称する。

1. 1. 『テス』の位置付け

本研究は，Hardy の代表作の一つ『テス』を研究対象に絞っている。名前からわかるように，『テス』の主人公は Tess である。Johnson T. (1968) で，「Tess は Hardy の最も巧妙に描き出した女性だ」[1]と述べられている。Laird J. T. (1975) で，「彼女（Tess）が Hardy の最も成功したヒロインの 1 人であるという判断に異議を唱える読者はほぼいないだろう」[2]と書かれている。また，1891 年に『テス』を，1896 年に『日陰者ジュード』を出版後，酷評されたこともあり，Hardy は執筆活動の中心を小説から詩作へと移し，以後は詩作に専念するようになった。ここからも窺えるように，『テス』は Hardy の数多くの著作の中で非常に重要な位置付けにある。

1. 2. 本研究の問題設定

これまでの文体や文学作品に関する研究は，作品全体をデータとして扱うものが多い。『テス』を会話部と地の文の二部分にはっきり分けて分析する質的研究もなかったようである。文学作品，特に発話順序が複雑な小説の会話部に関する量的分析もやや乏しい。

しかし，会話部が小説の成り立ちにおいては無視できないほど重要な役目がある。Page (1973) によれば，小説の対話（the dialogue in a novel）は多機能である。筋書きを立てたり、キャラクターを発展させたり、設定や環境を説明したり、道徳上の議論を提示したり、様々な話題についての議論を行なったりするのに対話が役立つ。これらの機能の任意の組み合わせも可能である。対話のさまざまな機能の中で、「キャラクターの提示と発展(the presentation and development of character)」が最も重要である。

そこで，本研究は，『テス』に関して，会話部のみに着目し，小説の会話部の語彙使用について

[1] 'Tess is Hardy's most masterly portrait of a woman'. Johnson T. (1968: 151).

[2] '…there are few readers of Hardy's fiction who would disagree with the judgement that she is one of Hardy's most successful heroines…'. Laird J. T. (1975: 118).

量的分析を行ってみる。会話部を対象とした量的研究が示した知見が Hardy の作品について新たな洞察を生み出しうるか，本研究では先行研究では（技術的な制約もあり）実施されていない手法も加えて考察を試みる。

2. 研究方法と目的

以下は，まず，『テス』本文の電子データをプレーンテクスト（.txt）として収集し，TEI（Text Encoding Initiative）に準拠したタグで，文学作品テクストの構造化，および文学作品の会話部と地の文を区別するマークアップを施し，XML（Extensible Markup Language）ファイルを作成する。次に，小説テクストの精読に基づく質的考察を行いながら，会話部に対してアノテーション（話し手と聞き手の情報付与）を行った上で，小説全体から会話部の直接話法を抽出する。そして，XML 解析により，ネットワークを描画し，主要な登場人物を特定する。それにより，主要な登場人物の個人語を詳細に分析する取り組みを行う。最後に，会話部の語彙について量的分析（対応分析）を行い，得られた結果に基づき，小説の各登場人物の発話の語彙的特徴および登場人物の間の関係性を明らかにしたい。

具体的なリサーチ・クエスチョン（RQ）は，（1）使用語彙を手掛かりとして登場人物の差が区別できるか，区別する場合，発話者をグループ分けする基準は何か，（2）会話で使われている語彙から，その登場人物の性格が見られるだろうか，見られる場合，どのような人物像が思い浮かぶか，（3）会話部の語彙から，登場人物の間の関係性は見られるだろうか，の 3 点である。

3. データ

3. 1. 小説本文の電子データの収集（TXT）

『テス』のテクストをデジタル化するに当たっては，CasualTextractor[3]というツールを使用し，電子書籍アーカイブ Project Gutenberg[4]から小説『テス』の本文をテキストファイル・プレーンテクスト（.txt）として収集した。小説テクストを抽出する際には，余分な記号（改行やタブ）を削除する，curly quotation marks（2 バイトの引用符）を ASCII quotation marks（1 バイトの引用符）に変換するなど，テクスト分析を行う前に，フォーマットを揃えておいた。抽出した小説テクストの総語数（Tokens）は 149,532 であり，異なり語数（Types）は 12,803 である。

3. 2. 小説のマークアップとアノテーション（TEI/XML）

3. 2. 1. XML と TEI

XML は，文章の見た目や構造を記述するためのマークアップ言語の一種である。XML はタグ名も属性も基本的に利用者が自由に設定できるため便利ではあるが，「各自が異なるルールでタグ付けをしてしまうと，共通のツールで利便性を高めたり，それぞれの成果を共有したりするこ

[3] https://sites.google.com/site/casualconcj/
[4] https://www.gutenberg.org/

とが難しくなってしまう」（石田他 2022: 128）。それを回避するために, TEI のガイドラインに沿ったテキストデータの構築が重要である。本研究は TEI/XML ファイルの形で小説テクストをデータ化することを試みる。

3. 2. 2.マークアップとアノテーション

『テス』は7つのフェーズと59の章に分かれている。『テス』の文書構造形式について, <text>, <body>, <div1>, <div2>, <head>タグを用いてマークアップを施した後, 会話部に対してアノテーション（話者の情報付与）を行う作業に入る。本研究は会話部の直接話法のみに着目するため, 小説の本文にある二重引用符を切り口とし, 二重引用符に囲まれた内容から分析対象となる会話データを選別する。

TEIガイドラインによると, 標準的なquotation marksには, 発話や思考（speech and thought）を表せる<q>と<said>, 引用（quotations）を表す<quote>, 書誌参照を伴う他の文書からの引用（citations）を表す（<quote>と<bibl>を伴う）<cit>, 言及された語句（words or phrases mentioned）を表す<mentioned>, 著者や語り手が責任を持ちたくない語句（words or phrases over which the author or narrator disclaims responsibility）を表す<soCalled>, 術語（terminology）を表す<term>, 注釈（glosses）を表す <gloss>がある[5]。本研究の会話部に関するアノテーションは, 引用内容が直接話法なのか間接話法なのかを示せる<said>タグを使用する。一方, 二重引用符には囲まれているが, 本研究の分析対象としない部分に対しては, より一般的な<q>タグで区別することにした。<said>タグと<q>タグの区別, および本研究で扱うタグに関する説明（図1）は以下の通りである。

<said>: (speech or thought). a more explicit element for the encoding of speech or thought, which allows the encoder to distinguish these from other quoted text

<q>: (quoted). the general element for quotation

```
1  <teiHeader>
2    <!-- ... -->
3    <encodingDesc>
4      <tagsDecl>
5        <namespace name="http://www.tei-c.org/ns/1.0">
6          <tagUsage gi="text">Contains a single unitary text.</tagUsage>
7          <tagUsage gi="body">Contains the whole body of a single unitary text, excluding
…  any front or back matter.</tagUsage>
8          <tagUsage gi="div1">Marks a phase. The <att>type</att> attribute gives the
…  division type. The <att>n</att> attribute gives the canonical phase number, where
…  appropriate.</tagUsage>
9          <tagUsage gi="div2">Marks a chapter. The <att>type</att> attribute gives the
…  division type. The <att>n</att> attribute gives the canonical chapter number, where
…  appropriate.</tagUsage>
10         <tagUsage gi="head">Provides the phase/chapter header.</tagUsage>
11         <tagUsage gi="said">Marks a speech within the text. The <att>who</att> and
…  <att>toWhom</att> attribute identifies the characters associated with that
…  speech.</tagUsage>
12         <tagUsage gi="q">Contains quoted sections of text and other speeches which
…  speaker and/or listener is/are unclear. The <att>who</att> attribute identifies the
…  characters associated with that speech. </tagUsage>
13       </namespace>
14     </tagsDecl>
15   </encodingDesc>
16   <!-- ... -->
17 </teiHeader>
```

図 1 『テス』TEI/XML ファイルに使用したタグに関する説明（<teiHeader>の一部）

[5] https://teibyexample.org/tutorials/TBED03v00.htm（2.5. Quotation を参照）
https://tei-c.org/release/doc/tei-p5-doc/ja/html/CO.html#COHQQ

『テス』の会話部に関するアノテーションの作業を終え，作られた XML 文書が well-formed（形式が整った，整形式）であるかどうか, valid（妥当的）であるかどうかを確認[6]してから次の作業に進む。作られたファイルの冒頭部分を図 2 で示す。

```
1   <?xml version="1.0" encoding="UTF-8"?>
2   <TEI xmlns="http://www.tei-c.org/ns/1.0">
3     <text>
4       <body>
5         <div1 type="phase" n="1">
6           <head>Phase the First: The Maiden</head>
7           <div2 type="chapter" n="1">
8             <head>Chapter I</head>
9             On an evening in the latter part of May a middle-aged man was walking homeward
…   from Shaston to the village of Marlott, in the adjoining Vale of Blakemore, or Blackmoor.
…   The pair of legs that carried him were rickety, and there was a bias in his gait which
…   inclined him somewhat to the left of a straight line. He occasionally gave a smart nod,
…   as if in confirmation of some opinion, though he was not thinking of anything in
…   particular. An empty egg-basket was slung upon his arm, the nap of his hat was ruffled, a
…   patch being quite worn away at its brim where his thumb came in taking it off. Presently
…   he was met by an elderly parson astride on a gray mare, who, as he rode, hummed a
…   wandering tune.
10            <said who="#John" toWhom="#Parson Tringham">"Good night t'ee,"</said>
11            said the man with the basket.
12            <said who="#Parson Tringham" toWhom="#John">"Good night, Sir John,"</said>
13            said the parson.
14  The pedestrian, after another pace or two, halted, and turned round.
15            <said who="#John" toWhom="#Parson Tringham">"Now, sir, begging your pardon; we
…   met last market-day on this road about this time, and I said 'Good night,' and you made
…   reply 'Good night, Sir John,' as now."</said>
16            <said who="#Parson Tringham" toWhom="#John">"I did,"</said>
17            said the parson.
18            <said who="#John" toWhom="#Parson Tringham">"And once before that-near a month
…   ago."</said>
19            <said who="#Parson Tringham" toWhom="#John">"I may have."</said>
20            <said who="#John" toWhom="#Parson Tringham">"Then what might your meaning be in
…   calling me 'Sir John' these different times, when I be plain Jack Durbeyfield, the
…   haggler?"</said>
21            The parson rode a step or two nearer.
```

図 2　TEI に準拠したタグでマークアップされた『テス』XML ファイル（一部）

紙幅のため，図 2 で示されているのは長い<teiHeader>の部分を省いたものである。実際に扱うファイルはヘッダ（<teiHeader>）を持っている。注意すべきなのは, TEI ではヘッダ（<teiHeader>）が必須である。全ての妥当な TEI ファイルは<TEI>エレメントの内側に，<titleStmt>と<publicationStmt><sourceDesc>を含む<fileDesc>を伴う<teiHeader>エレメントを追加しなければならない。石田他（2022: 49）によれば, <teiHeader>エレメントに沿って，データがいつ誰によってどのような資料に基づいて作られ，どのようなルールでデジタル化されているか，といったことを記述しておくことは，データの信頼性を担保する上できわめて重要である。

3. 3. 登場人物の interaction network 描画

登場人物の会話部に絞り込んで分析を行うためには, <said>要素を抽出し XML 解析を行う必要がある。XML 解析とこれから実行する登場人物の interaction network 描画に関しては, Karsdorp *et al.* (2021)を参照して Python で実行する。本研究は，『テス』のキャラクターインタラクションネットワークを次のように定義する。全ての<said>タグに対して, @who 属性にある人名 Name A を

[6] validation check: https://www.xmlvalidation.com/
syntax-check: https://www.w3schools.com/xml/xml_validator.asp

登場人物 i, @toWhom 属性にある人名 Name B を登場人物 j と見なし，登場人物 i と登場人物 j のペアが n 回現れると，A と B の間に interaction が n 回あると認識する。この定義に関する部分の Python コードは図 3 に示す。この定義に従って描いたネットワークの簡略化したものは図 4 である。

```
def character_network(tree):
    G = nx.Graph()
    # extract a list of speaker turns for the novel
    for said in tree.iterfind('.//said', NSMAP):
        # ... and extract pairs of adjacent speakers
        try:
            speaker_i = said.attrib['who'].replace('#', '')
            speaker_j = said.attrib['toWhom'].replace('#', '')
            # if the interaction between two speakers has already been attested, update their interaction count
            if G.has_edge(speaker_i, speaker_j):
                G[speaker_i][speaker_j]['weight'] += 1
            # else add an edge between speaker i and j to the graph
            else:
                G.add_edge(speaker_i, speaker_j, weight=1)
        except KeyError:
            continue
    return G
```

図 3　本研究で扱うキャラクターネットワークの定義（Python code block）

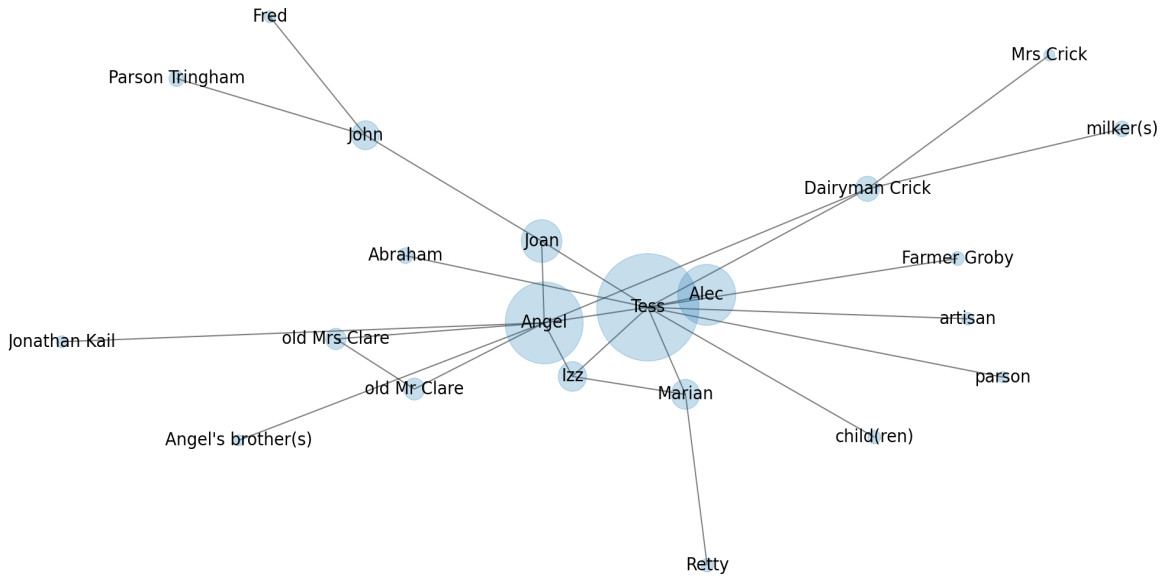


図 4　『テス』のキャラクターインタラクションネットワーク（簡略図）

図 4 で示した登場人物のネットワークはノード（node：節点・頂点，点）の集合とエッジ（edge：枝・辺，線）の集合から構成されている無向グラフである。nodes は小説の登場人物に対応しており，node labels は登場人物の名前を示している。登場人物 i と登場人物 j の間のインタラクション

（interaction：対話）の回数が多ければ node size が大きくなる。node と node を繋げている edge は登場人物 i と登場人物 j の間に interaction があることを表す。『テス』の全ての登場人物の対話ネットワークでは, 主要でない登場人物も数多く含まれており, 見づらいため, このように簡略化して見やすくした[7]。図 4 のネットワークでは, interaction の回数 が 10 回以上のキャラクターのみが含まれており, node 数は 22 で edge 数は 26 である。この図の中に, 幾つか三角形になっている部分もあるため, まずはそれらに注目して説明しておきたい。

Angel と Clare 夫妻は家族であり, この 3 人がネットワークで三角形になっているのは, Angel と old Mrs Clare, Angel と old Mr Clare, old Mrs Clare と old Mr Clare の間にそれぞれ対話があったことを意味している。しかし, 主人公の Tess を見ると, Tess と母 Joan, 父 John は一直線状に並んでいる。つまり, Tess と Joan, Joan と John の間に対話があったが, Tess と John は（同じ場面にいることはありながら）一度も一対一の会話をしたことがない。しかも, John は Joan に比べ, ネットワークの周辺辺りにある。このことから, 小説で Joan は John より重要な役割を果たしていることが推測される。同様に, Tess と Marian, Izz, Retty の四人は友人であるが, Retty だけが離れた位置にある。Marian と Izz は Retty に比べ, 物語の発展でより重要な役割を果たすことが言えよう。

他に, Tess-Angel-Dairyman Crick, Tess-Angel-Izz, Tess-Angel-Joan も三角形になっている。Tess と Angel は恋人で親密な関係にあり, 対話ネットワークでもこの二人は共通の知り合いが複数人いることが反映されている。Tess と Angel は共に酪農場で働いていたので, それぞれ酪農場主 Dairyman Crick と話したことがあるのは当然のことである。Angel と Izz の対話は『テス』の第 40 章にあり, Angel が Tess を見捨ててブラジルへ渡航する前に, Izz を誘って一緒にブラジルに行く部分にある。Angel と Joan の対話は『テス』の第 54 章で現れ, Angel がブラジルからイギリスへ戻り, Tess の居場所を探すために Joan のところに行った部分にある。

このように, 『テス』の会話部から登場人物間の人間関係をとらえることができ, インタラクションネットワークを描画することで, 登場人物の重要性が視覚化できたと考えられる。

3. 4. 主要な登場人物の会話部抽出

会話部の分析において対象となるキャラクターは, 簡略化したネットワーク（図 4）に含まれ（interaction の回数≧10）, 且つ発話語数が 500 語以上（Tokens≧500）の登場人物に絞り込んでいる。抽出された 10 人の登場人物の個人語に関する統計値は表 1 の通りである。10 人のうち, 男性（Angel, Alec, Dairyman Crick, John, Old Mr Clare）は 5 人で, 女性（Tess, Joan, Marian, Old Mrs Clare, Izz）は 5 人いた。男女比は正に 1 対 1 であった。分析対象となる会話テクストの総語数（Tokens）は 32,555 であり, 全小説の<said>タグに囲まれた会話テクストの総語数(36,186)の約 90%である。この 10 人の発話語数は『テス』の総語数（149,532）の約 5 分の 1 を占めている。なお, ここでいう「主要な登場人物」の概念は「発話語数が極めて少ない脇役」と対照的なものである。

[7] 簡略化せずにネットワークを描画すると, node 数は 48 で edge 数は 77 である。

表 1 主要な登場人物（10 人）の個人語に関する統計値[8]

	登場人物	性別	Tokens	Types	TTR	STDTTR	RTTR
1	Tess	女	9,125	1,280	0.14	0.43	13.40
2	Angel	男	7,547	1,566	0.21	0.48	18.03
3	Alec	男	6,634	1,352	0.20	0.48	16.60
4	Dairyman Crick	男	2,463	773	0.31	0.52	15.58
5	Joan	女	2,233	648	0.29	0.50	13.71
6	John	男	1,276	490	0.38	0.51	13.72
7	Marian	女	1,162	410	0.35	0.47	12.03
8	Old Mrs Clare	女	768	337	0.44	0.51	12.16
9	Old Mr Clare	男	675	334	0.49	0.53	12.86
10	Izz	女	672	280	0.42	0.43	10.80
	平均値	-	3,255.5	747	0.32	0.48	13.89

4. 結果と考察

このセクションでは，抽出された10人の主要な登場人物の個人語がいかに描き分けられているかに対して考察を行なっていく。まず，各キャラクターの発話の語彙に対して品詞タグを付与し，各登場人物の品詞の使用について探る（4.2）。次に，法助動詞に着目し，個人語の法助動詞の出現頻度で差が見られるかについて検討する（4.3）。最後に，『テス』で一番重要なTess, Angel, Alecの3人にさらに絞り込み，会話部の頻出語からこの3人の関係性が窺えるかを分析する（4.4）。分析に当たっては主に対応分析を使用するので以下はまず対応分析について説明する（4.1）。

4. 1. 対応分析とは

対応分析（Correspondence Analysis）は，データ表の行と列の対応（correspondence）関係に着目しており，行や列に含まれる多くのカテゴリの情報を少数の成分（次元［dimension］）に圧縮し，各次元を組み合わせて散布図を作成することで，分析結果を視覚化し，カテゴリ間の関係を直観的に解釈することを可能にする多変量解析手法の一つである。

石川他 （2010）によれば，対応分析はデータを分類するその他の統計手法[9]に比べ，以下のように多くのメリットを持っている。

[8] TTR, STDTTR と RTTR は語彙の多様性・豊富さを測定する指標である。TTR(Type/Token Radio)は延べ語数（Token 数）の中に占める異なり語数（Type 数）の割合である。TTR が基本的な考え方となるが，文章の長さに依存するため，長さの異なるテクストの比較には適切でない。STDTTR(Standardised Type/Token Ratio)は標準化 TTR で n 単語ごとに計算される。デフォルトでは, n=1,000 が，本研究では n=500 に設定した。RTTR（Guiraud Index ともいう）は Type 数を Token 数のルートで割ったものである。

[9] 対応分析以外に，データの分類手法として，判別分析，クラスター分析，主成分分析，因子分析等がある。

ア）変数とケースといったデータの性質の違いが問題にならない[10]。

イ）第 1 カテゴリ，第 2 カテゴリの同時分類が可能である。

ウ）分析の主目的が，データの合成や分解ではなく，カテゴリの整理と分類そのものにある。

エ）分析者が決定すべきオプションの幅が狭く，分析者の判断によって結果が左右されにくい。

（石川他 2010: 248）

また，多変量解析では，多くのカテゴリを少ない次元に圧縮するという分析の目的に照らして考えれば，およそ 5 個以上，通例 10 個以上のカテゴリを持つアイテムに対して分析を行うことが望ましいと言える（石川他 2010: 250）。今回は 10 人の登場人物の会話部を対象にして分析するので，対応分析を使用するのが適切であると考えられる。

4. 2. 品詞構成の比較

本節では, 10人の主要な登場人物の言語データを用いる。データ処理は，まず, CasualTagger[11]を起動し，抽出された10人の発話に対し, TreeTaggerのEnglish BNC tagset[12]で品詞タグ付けを行っておく。ここでは, TreeTaggerによる形態素解析の誤りについて特に修正を行わない。次に, CasualConc[13]を用いて，個々の登場人物の語彙リストを用意する。表1ですでに示したように，この10人の発話語数に尺度差がある。そのため，以下は個人語の品詞の出現頻度を500語あたり調整頻度に備えておく。得られた頻度表は表2に示す。計38種類の品詞タグがあったが，ここでは上位5位までの品詞のみを示す。

会話部全体で出現頻度の高い上位5位の品詞は, pnp（Personal pronoun：人称代名詞）, nn1（Singular common noun：名詞の単数形）, av0（General adverb：一般副詞）, vbb（The present tense forms of the verb BE：動詞の現在形）, vbi（The infinitive form of the verb BE：動詞の不定形）である。この点から会話部において人称代名詞が非常に重要な役割を果たしていることがわかる。

[10]他の多変量解析手法では，列方向（縦方向）に入るデータと行方向（横方向）に入るデータをそれぞれ異なる性質を持つものと考え，前者を変数，後者をケース（検体，個体，サンプル）として区別してきた（石川他 2010: 245）。しかし，対応分析では，行と列の入れ替え（転置）を行っても，分析結果は変わらない。

[11] https://sites.google.com/site/casualconcj/

[12] http://www.natcorp.ox.ac.uk/docs/c5spec.html

[13] https://sites.google.com/site/casualconcj/

表 2 主要な登場人物（10人）の個人語の品詞頻度表（相対頻度）[14]

	登場人物	Tokens	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
1	Tess	9,141	pnp (89.43)	nn1 (38.95)	vbi (35.66)	vbb (35.23)	av0 (34.73)
2	Angel	7,515	pnp (69.93)	nn1 (52.03)	av0 (35.40)	vbb (33.93)	vbi (30.87)
3	Alec	6,649	pnp (74.82)	nn1 (49.48)	av0 (36.55)	vbb (33.31)	prp (32.86)
4	Dairyman Crick	2,524	nn1 (67.55)	pnp (49.33)	av0 (34.67)	at0 (34.47)	prp (33.48)
5	Joan	2,298	pnp (59.62)	nn1 (54.61)	av0 (34.60)	vbi (30.46)	prp (28.29)
6	John	1,315	nn1 (65.78)	pnp (51.71)	prp (32.70)	av0 (32.32)	at0 (25.86)
7	Marian	1,197	pnp (84.38)	nn1 (41.77)	av0 (36.34)	vbi (31.75)	vbb (27.57)
8	Old Mrs Clare	769	pnp (66.97)	nn1 (64.37)	av0 (31.86)	prp (31.86)	aj0 (31.21)
9	Old Mr Clare	678	nn1 (64.90)	pnp (53.10)	prp (36.14)	aj0 (33.19)	at0 (29.50)
10	Izz	678	pnp (67.85)	nn1 (47.20)	av0 (44.99)	vbi (32.45)	prp (26.55)
	総語数	32,764	pnp (4,809)	nn1 (3,268)	av0 (2,300)	vbb (2,028)	vbi (2,021)

4. 2. 1. 品詞構成のクラスター分析

対応分析を行う前に，まずは，各登場人物の品詞頻度表に対してクラスター分析を用いて登場人物間の品詞使用の類似度を探る。クラスター分析（cluster analysis）とは，データが持つ情報を手掛かりにして，距離の近いデータ同士をまとめてクラスター（cluster：群，集落）を構成する統計手法である。言語コーパス研究では，語彙・品詞・ジャンルなどの分類の目的で幅広く使用される（石川他 2010: 163）。本研究では, 500語毎の相対頻度を基準に，全ての品詞種類の相対頻度

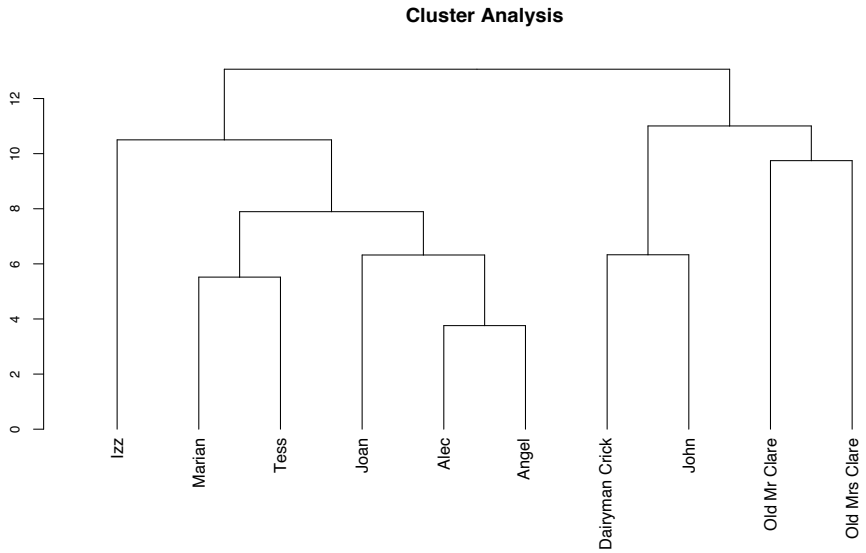


図 5 主要な登場人物（10 人）の個人語の品詞構成の樹形数

[14] こちらは品詞タグが付与されたデータなので，個人語の Token 数がタグのないデータに比較すると，数はやや増加か減少した。

をデータとして, Ward法, Euclidean距離を用いる（図5）。

クラスター分析の結果，この10人は大きく2つのグループに分けられた（図5）。左側の6人は正に図4のネットワークの中心に位置しているキャラクターで，右側の4人はより周辺的な場所に位置するキャラクターである。これは, Hardyが自分の小説で肝心な人物たちの発話とそれ以外の人物の発話に対してそれぞれ似たようなパターンで描いているかもしれない。また，樹形図の低い下位分類を見ると, MarianとTess, AlecとAngel, Dairyman CrickとJohn, Old Mr ClareとOld Mrs Clareの品詞使用が似ていることがわかる。MarianとTess, Old Mr ClareとOld Mrs Clareの品詞使用には類似性があるが，それは『テス』における設定を考えると，理解は容易かと思われる。一方，中央のAlecとAngelは共に物語の主要人物であるが，彼らは一度も会ったことがない。Dairyman CrickとJohnも知り合いではなく，同じ場面で登場したことがない。しかし，クラスター分析の結果を見ると，似たような品詞の使い方をしている。考えられるのは, AlecとAngel, Dairyman CrickとJohnが『テス』での重要性は同等レベルであることである。このように, Hardyが似たような設定にしたい人物の個人語では，使用する品詞も似通っている。

4. 2. 2. 品詞構成の対応分析

クラスター分析から読み取れたのは, Hardyが『テス』を作成する際に，関係が近い人物同士や似たような設定にしたい人物同士の会話部に用いる語彙に一定のパターンがあることである。しかし，クラスター分析で各登場人物間の類似度は見られたが，登場人物と品詞の間でどのような対応関係にあるのかは明らかにならなかった。そのため，次に，クラスター分析の結果に踏まえ，

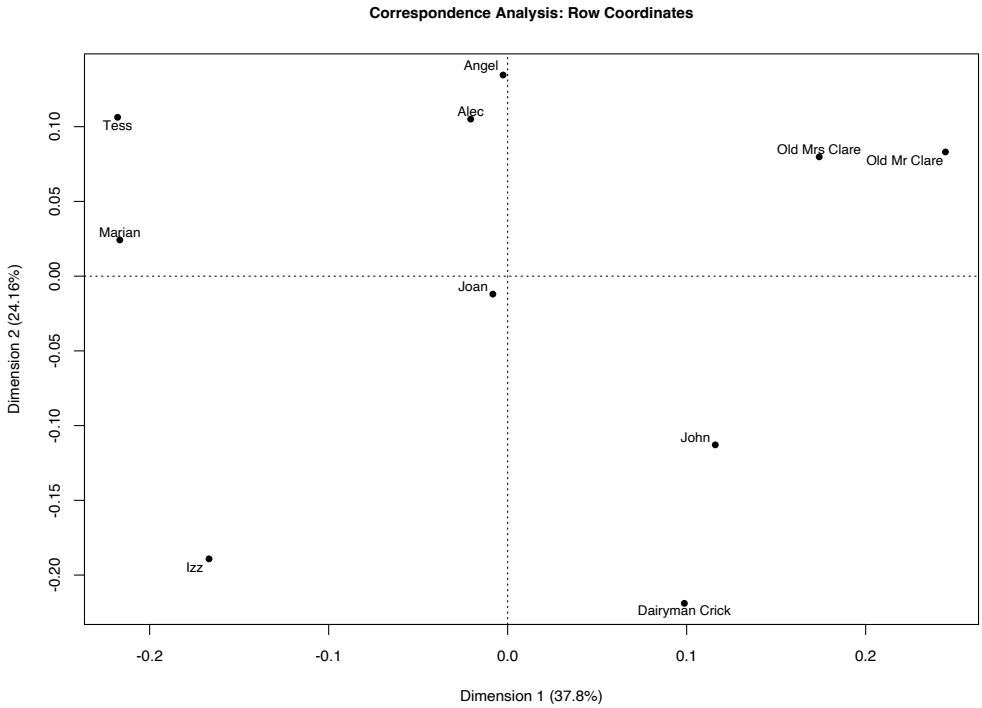


図 6　品詞の相対頻度による対応分析（第 1 アイテム）

品詞構成の対応分析を行う（図6，図7）。

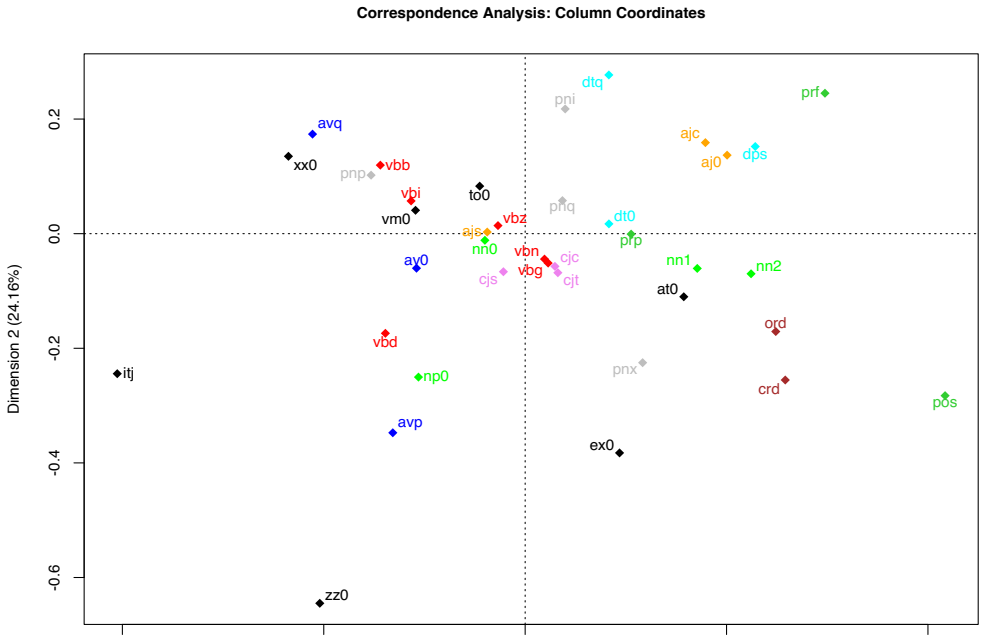


図 7　品詞の相対頻度による対応分析（第 2 アイテム）

第1アイテム（登場人物）がプロットされた散布図（図6）を見ると，クラスター分析の樹形図（図5）と同じ結果が得られているので，その説明を省略する。ここでは，第2アイテム（品詞）がプロットされた散布図（図7）に注目する。この図は第1アイテムの散布図（図6）と対応している。図6では，品詞の頻度パターンの近い人物同士が近く布置される。同様に，図7では，個人語における頻度パターンの近い品詞が近くに現れる。

図7の動詞の分布に着目すると，上から下へと現在形（vbb），不定形（vbi），過去形（vbd）が分布している。動詞の-*s*形（vbz），過去分詞（vbn）や-*ing*形（vbg）は原点あたりにある。ここから，登場人物の間で動詞の時制の違いがあったことがわかる。第1軸（横軸）の上半分に分布している人物は現在形を，下半分に分布している人物は過去形を頻繁に使用している。図6と併せて見ると，第2象限にあるTess, Angel, Alec, Marianは現在形を多用し，第3象限にあるIzzは過去形を多用していることがわかる。コンコーダンスを調べた結果，Izzの個人語で過去形の使用は主に第40章でAngelとの長いやり取りに現れているが，全体を見ると，Izzの発話で過去にあった出来事の説明・伝達や状況の確認のために過去形が使われることが多数である。同じく対面でコミュニケーションをしているにも関わらず，過去形を会話で多用する人物は主に既に発生した出来事を他のキャラクター（と読者）に伝える役割を果たしており，現在形を多用し現在の物事を語る人物に比較すると重要度がやや低いことが考えられる。このように，『テス』で会話部の動詞の時制からもキャラクターの役割が反映されている。

また，他にも散布図から読み取れる情報ある。例えば，図7で副詞（avq, av0, avp）が第2軸（縦

軸）の左側に分布しており，図6の左側の6人（ネットワークの中央にある人物）が相対的に副詞をより頻繁に使用することがわかる。一方，形容詞や名詞は主に軸の右側に寄っている。最上級の形容詞（ajs），固有名詞（np0）と普通名詞（nn0）は原点近くにあるが，比較級（ajc）または一般的な形容詞（aj0），単数形の名詞（nn1）と複数形の名詞（nn2）は右側に分布している。つまり，図6の右側の4人（より周辺的な人物）の品詞使用は異なる傾向にある。さらに，形容詞以外，疑問限定詞（dtq），所有限定詞（dps），一般限定詞（dt0）と前置詞（prf）が第1象限に分布しているのに対し，所有格または属格のマーカー's または'（pos）は第4象限に分布している。Old Mr ClareとOld Mrs Clareの発話の文構造の複雑さ，JohnとDairyman Crickの文構造の簡単さがそれぞれ反映されている。

4. 3. 法助動詞の比較

品詞構成について分析した後，法助動詞に着目して，法助動詞の使用よる登場人物の分類ができるか試みる。語彙リストから法助動詞（*_vm0）を全て抽出し，10人の登場人物と全18種類の法助動詞の頻度表を作成する。ここでも，TreeTaggerによる形態素解析の誤りは特に修正しない。

4. 3. 1. 法助動詞の対応分析（粗頻度）

以下ではまず，頻度補正を行わない頻度表（粗頻度）[15]に対して対応分析を実行してみる（図8，図9）。

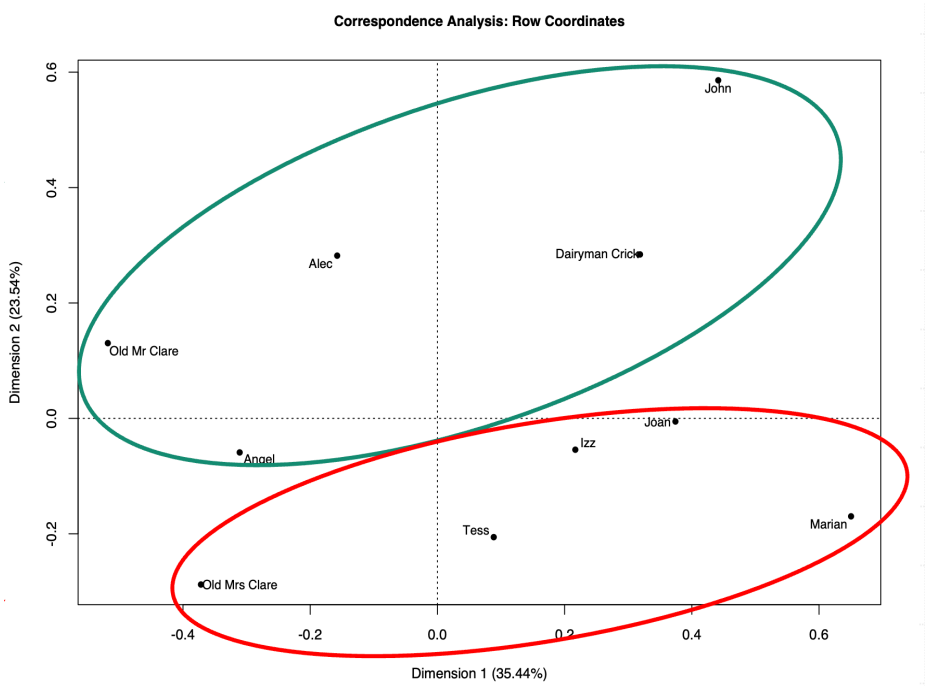


図 8　法助動詞の粗頻度による対応分析（第 1 アイテム）

[15] 一般に，ファイル間やコーパス間に総語数の差（尺度差）がある場合は○○語あたりの調整頻度によって尺度差を補正するが，対応分析では，補正前の生データを使うことも多い。

まず，第1アイテム（行：登場人物）がプロットされた散布図（図8）に注目すると，第1次元の正負によって，Angel一家（Old Mr Clare, Old Mrs Clare, Angel），Alecとそれ以外が区分されており，第1軸（横軸）はこれらを分ける軸になっていることがわかる。同様に，第2軸（縦軸）は，Angel（男性）以外，第2次元の正負によって男性群（Old Mr Clare, Alec, John, Dairyman Crick）と女性群（Old Mrs Clare, Tess, Izz, Joan, Marian）を区分する軸になっている。第1軸（横軸）に対してより詳しく分析してみると，図中の左から右に向かって登場人物の教養レベルが下がっていくという解釈が可能になる。

次に，第2アイテム（列：法助動詞）がプロットされた散布図（図9）を併せて考察する。この図は図8と対応しているため，教養レベルの高い人物に顕著な語が図中の左側にプロットされており，教養レベルの低い人物に特徴的な語は図中の右側にプロットされている。教養レベルの低い人物に特徴的な語は，*d*, *ca*, *ll*, *sha*, *wo*のような省略形の語である[16]。方言の影響があるともいえよう。それに対して，教養レベルの高い人物に特徴的な語として，*dare*, *may*, *will*, *shall*, *need*といった省略されていない形の法助動詞が現れている。

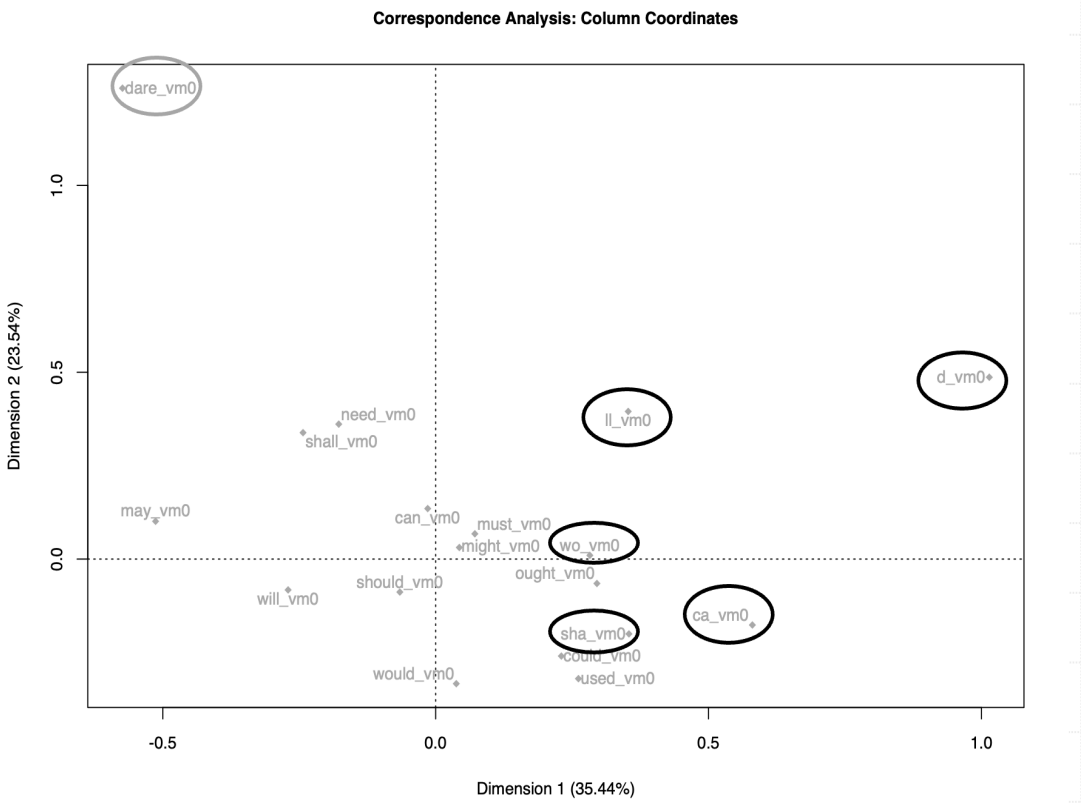


図 9　法助動詞の粗頻度による対応分析（第 2 アイテム）

[16] コンコーダンスを確認した結果，*d* は could, would, had に，*ca* は can の否定縮約形 can't の ca に，*ll* は will に，*sha* は shall の否定縮約形 shan't の sha に，*wo* は will の否定縮約形 won't の wo に，それぞれ対応している。なお，had は法助動詞ではないが，ここではタグの誤りに関して特に修正を行わない。

法助動詞の粗頻度についての考察では，『テス』の登場人物での男女別および教養レベル別から法助動詞の使用に差が見られたが，これは個人語で実際に使われる法助動詞の頻度による対応関係であり，異なるキャラクター間の発話語数の尺度差は考慮に入れなかった。そこで，相対頻度の対応分析を以下で実行してみる。

4. 3. 2. 法助動詞の対応分析（相対頻度）

粗頻度の図と比較すると，相対頻度の散布図（図10，図11）では，*need*, *sha*, *used*, *dare*の4つが（頻度が低いかつそれを用いる登場人物が少ないため）減少し，14種類の法助動詞のみが対応分析の対象になる。

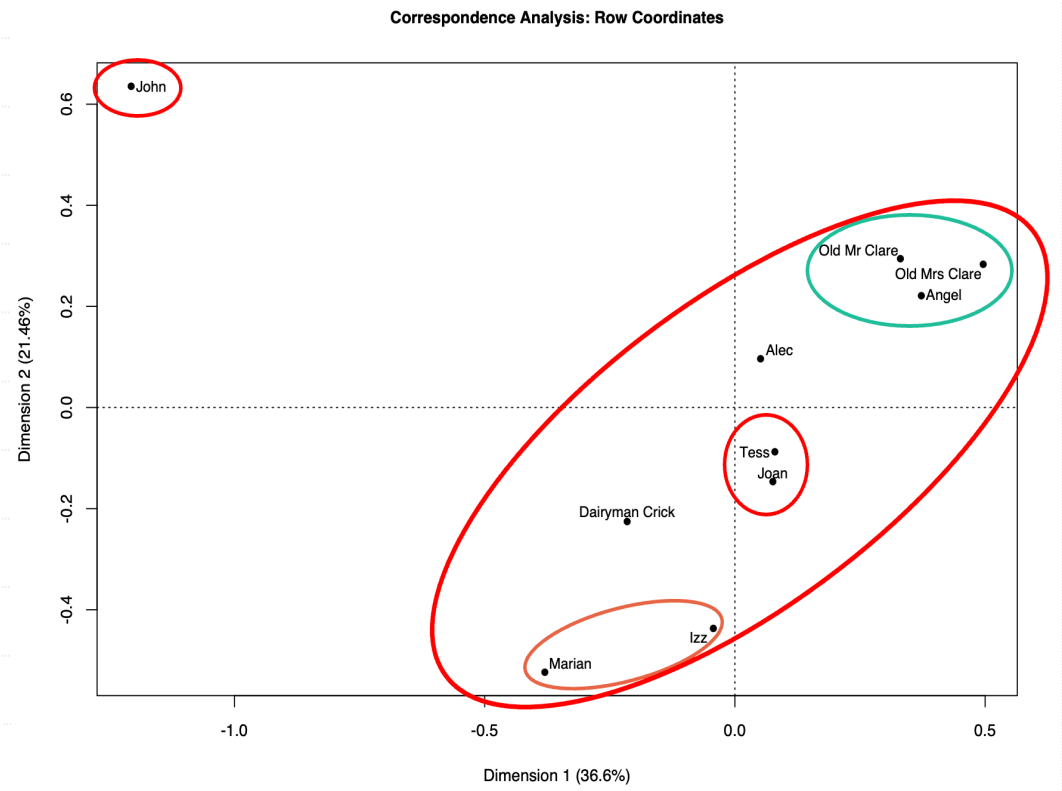


図 10 法助動詞の相対頻度による対応分析（第 1 アイテム）

図10で描かれた人物の分布から，人物同士の関係性を観察することができる。例えば，右上のClare夫妻（Old Mr Clare, Old Mrs Clare）とAngelの3人は家族なので，似たような場所に分布しており，使用する法助動詞も似ているという結論が得られる。原点あたりのTessとその母Joanも近いところに分布しており，底辺あたりのIzzとMarianは友人関係で，分布が似たところにあるので，法助動詞の使用も似ていることがわかった。しかし，この相対頻度に基づくデータでは，Tessの父Johnだけが遠く離れた場所（左上）にあることに注目すべきである。

図11と併せて見ると, *d*もそれ以外の法助動詞に比べると外れた場所（左上）にある。Johnは*d*や*wo*, *ll*, *ca*といった省略形の語を使用する傾向にあり, John以外の人物は省略されていない法助動詞を使う傾向にあることが読み取れる。これは, Johnの個人語における一つ顕著な特徴であると考えられる。

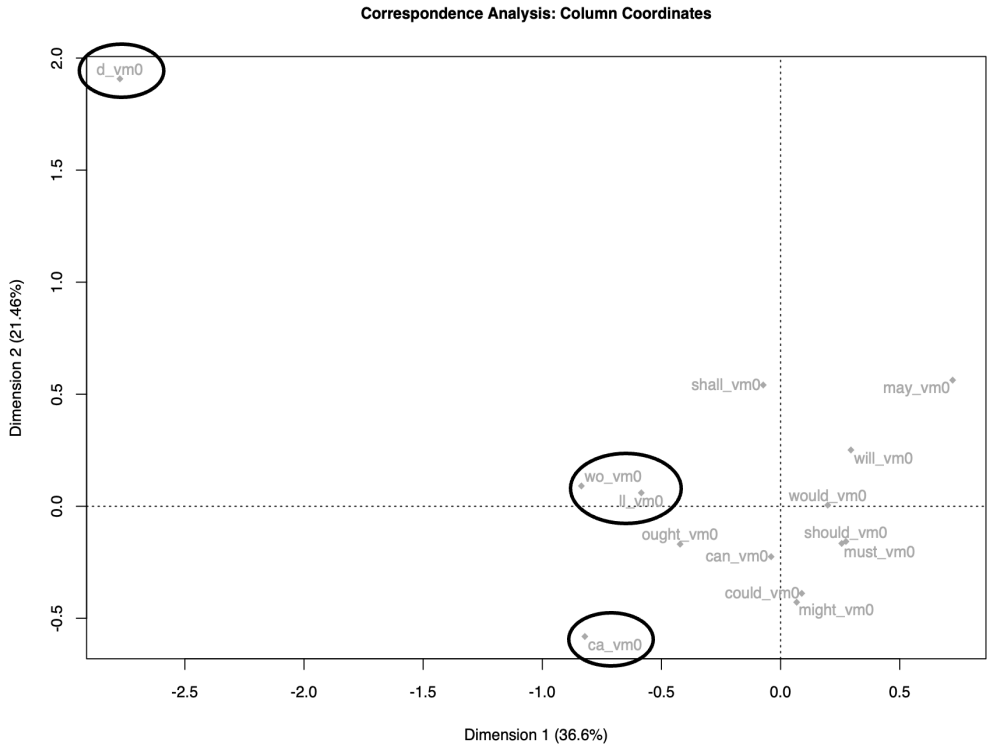


図 11 法助動詞の相対頻度による対応分析（第 2 アイテム）

以上をまとめると, 対応分析を用い, 個人語に出現した法助動詞（粗頻度）に注目して登場人物の会話部を分析した結果, 登場人物を男性と女性の2つのグループに分類することができた。また, 相対的に教養レベルの低い人物が法助動詞の省略形を使用する頻度が高いこともわかった。さらに, 個人語に出現した法助動詞（相対頻度）を考察した結果, 関係が近い人物同士の法助動詞の使用も似ていることが明らかになった。尤も, 個人語は常に身近な人に影響を及ぼされることから, 得られた結果はある意味では当然と言える可能性もあるが, 一方, 登場人物の中で, Johnの法助動詞の使用が特異な位置にあることが明らかになるなど, いくつかの興味深い知見も観察された。ただし, 言語データだけで言えることは限りがあり, これらの点についてさらに議論を進めるためには, 小説原文への考察を加えたより詳細な分析が必要となろう。

4. 4. Tess, Angel, Alec の発話における最頻出語の比較

次の作業では，既に述べたこと以外に, Tess, Angel, Alecの3人の個人語にある頻出語を中心として，登場人物間の関係性を調査・分析していきたい。ここではこの3人の頻出語（相対頻度）を降順で並び，上位55語を図12のワードクラウド（word cloud）で視覚化する。Word cloudでは，テクストに高い頻度で表れている語が大きいフォントで表示される。ワードクラウドから読み取れるのは, Tess, Angel, Alecの3人の特徴語である。

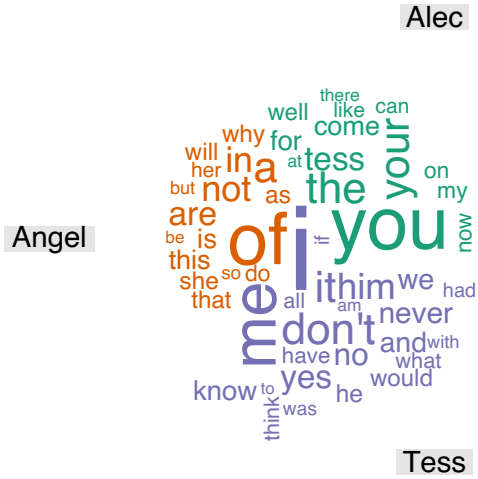


図 12 ワードクラウド

中央あたりに大きなフォントで表示されているのは*i*や*you*のような人称代名詞なので，まずは，この3人の頻出語にある人称代名詞の差を明らかにする。

図12でTessが*i*, *me*, *it*, *we*, *him*, *he*を他の2人より頻繁に使用しており, Angelは*she*と*her*を, Alecは人称代名詞*you*（と所有限定詞*your*, *my*）を多用している。コンコーダンスを考察すると, TessがAlecに対する発話での*him*と*he*はAngelを指すことが多く, TessがAngelに対する発話で*him*と*he*はAlecを指すことが多いのに対し，Angelの発話で*she*と*her*はほぼ全てTessを指す。会話で頻繁に三人称で互いのことを第三者に語ることが多いことから，この3人の感情が絡み合うことがわかり，物語の婉曲性・複雑性も反映されている。Alecの発話でも*tess*（Tess）が特徴語である。登場人物間の葛藤を発話部の語彙から読み取ることもできた。

また，TessとAlecの間で*i*と*you*の使用頻度で大きな区別があったが，その理由についても実際に使用されているコンテクストから分析してみた。結論から言うと, Tessは*i*を使って，自分の感情・考え・決断，自分の行動や状態に対する説明などを聞き手に伝えようとしている。*me*を使って，相手への依頼，他人の自分に対する言動やその言動が自分に与えた影響などを述べようとしている。さらに，Tessは*we*を多用し，他人と同じ立場に立ち,相手か第三者を自分の領域に受け入れる傾向にあり,より性格の優しい人物であることがわかる。それに対し, Alecが*you*, *your*, *my*のような相手と自分の間にきっぱりと境界線を引き，「私」と「あなた」を対立関係にするような言い方を好

む。より詳細にAlecの発話を見ると, Alecは主に*you*を使って, 自分の基準で聞き手に対して評価を下したり要望を伝えたりしている。性的に放埒で, Tessを凌辱した彼は, 会話部での用語のニュアンスもその攻撃的なイメージに相応しい。

次に注目したいのは, この3人の法助動詞の使用傾向である。Tessは*would*, Angelは*will*, Alecは*can*を相対的に高頻度で使用している。TessはAngelより法性の弱い法助動詞を使用しており, 性格がこの3人の中で最も従順的であることがわかる。最後は, ワードクラウドに現れたやや特殊な語について考えてみる。Tessの個人語で否定を表す語（*don't*, *never*, *no*）が上位に来ていることに注目に値する。Angelの個人語で*not*という否定辞が使用されているが, Alecの発話では否定の表現が相対的に稀である。Tessの個人語で否定の表現が数多くあることは, Tessの相手への抵抗感や彼女の性格で引っ込み思案のところを示している。

このように, 個人語の頻出語からでも登場人物の関係性や性格を観察することができた。

5. 終わりに

以上より, Hardyの代表作の一つである『テス』を対象として, TEIガイドラインに基づき, 文学作品テクストの構造化およびアノテーションを行った上で, 小説全体から会話部を抽出し, 登場人物間の対話ネットワークの考察および主要な登場人物の会話に用いられている語の頻度パターン解析を行った。このように, 本研究は技術的な制約を克服し, 会話部を対象とした量的研究の新しいアプローチを提案した。その結果, Hardyが『テス』の登場人物のキャラクタライゼーションを行う際に, 会話部に一定の工夫を入れたことがわかった。

RQ1（会話部の語彙による登場人物の分類の可否）については, 登場人物の語彙使用に違いがあると言える。キャラクターの性別や重要性による個人語の差が見られた。RQ2（登場人物の性格による個人語の差の有無）については, 会話で使われている語彙から, その登場人物の人物像が見えてくるという解答になる。法助動詞の使用に当たっては, 登場人物が会話部で*can*や*will*を好むか, *could*や*would*を好むかによってその人物の性格の強さや弱さを窺えることができた。また, 会話での頻出語から, その発話者は引っ込み思案の性格なのか, 自己中心のような人間であるのかもイメージがつくことが明らかになった。RQ3（人間関係による個人語の差の有無）については, 両者の強い結び付きがあると言える。クラスター分析と対応分析によって, 関係の近い人物同士や同じような設定の人物同士の個人語には似たようなパターンが潜在していることが確認できた。

本研究では, Hardyの代表作の『テス』を取り上げて会話部を抽出して分析したが, 地の文の語彙や小説全体の語彙的特徴についての検討は触れなかった。また, 試行的な研究でもあり, 品詞構成, 法助動詞と頻出語の3点のみに着目して考察を試みた。会話部分析の指標の汎用性が足りず, その他の分析指標の開発は途上である。今後は, より信頼性の高い計量手法の確立を目指している。さらに, Hardyの作品数を増え, 作品の出版年の変化につれて, Hardyの登場人物の特徴付けのパターンが終始一貫する特徴や傾向を示すものであるかどうかについての確認も行ってみると

良いであろう。

文献

Chapman R. (1990). *The Language of Hardy*. Basingstoke: Macmillan.

Ezawa K. (2015). Colours as Represented in *Tess of the D'Urbervilles* (1). 『東京成徳大学研究紀要—人文学部・応用心理学部』第 22 号, 129-137.

Ezawa K. (2016). Colours as Represented in *Tess of the D'Urbervilles* (2). 『東京成徳大学研究紀要—人文学部・応用心理学部』第 23 号, 141-152.

Johnson T. (1968). *Thomas Hardy*. London: Evans.

Hardy T. (1891). *Tess of the d'Urbervilles: A Pure Woman.* Toronto: Bantam Books.

Hardy T. *Tess of the d'Urbervilles: A Pure Woman*. https://www.gutenberg.org/cache/epub/110/pg110-images.html (参照 2022-05-26)

Howe I. (1967). *Thomas Hardy*. New York: Macmillan.

Imao Y. (2022). CasualConc (Version 3.0.4 beta) [Computer software]. Retrieved from https://sites.google.com/site/casualconcj/

Karsdorp F., Kestemont M., & Riddell A. (2021). *Humanities Data Analysis: Case Studies with Python*. Princeton University Press.

Laird J. T. (1975). *The shaping of Tess of the d'Urbervilles*. Oxford: Clarendon Press.

Leech G. (1987). *Meaning and the English Verb*. Harlow: Longman.

Leech G., Short M. (2007). *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. New York: Pearson Longman.

Page N. (1973). *Speech in the English novel*. Longman.

Tabata T. (2002). Investigating Stylistic Variation in Dickens through Correspondence Analysis of Word-Class Distribution. In T. Saito, J. Nakamura & S. Yamazaki (Eds.), *English Corpus Linguistics in Japan*. Amsterdam: Rodopi, 165-182.

Tabata T. (2004). Differentiation of Idiolects in Fictional Discourse: A Stylo-Statistical Approach to Dickens's Artistry. In Hiltunen, R. and Watanabe, S. (Eds.), *Approaches to Style and Discourse in English*. Osaka: Osaka University Press, 79-106.

石川慎一郎(2005). 「詩の言葉・小説の言葉—コーパス処理に基づくトマス・ハーディのテクスト構成語彙の分析—」富山太佳夫, 加藤文彦, 石川慎一郎（編）『テクストの地平 ： 森晴秀教授古稀記念論文集』英宝社, 153-168.

石川慎一郎, 前田忠彦, 山崎誠［編］(2010). 『言語研究のための統計入門』. くろしお出版, 東京.

石田友梨, 大向一輝, 小風綾乃, 永崎研宣, 宮川創, 渡邉要一郎（編）(2022). 『人文学のためのテキストデータ構築入門　TEI ガイドラインに準拠した取り組みにむけて』. 文学通信, 東京.

田畑智司(2006). 「米国歴代大統領就任演説の言語を計る—多変量アプローチによるテキスト類型化試論—」大阪大学大学院言語文化研究科（編）『電子化言語資料分析研究 2005-2006』. 言語文化共同研究プロジェクト, 57-67.

橋本史帆(2019). 『トマス・ハーディの小説世界. 登場人物たちに描き込まれた国際事情と「グレート・ブリテン島」的世界』. 音羽書房鶴見書店, 東京.

深谷修代(2013). 「コーパスを用いた『白雪姫』の会話分析」. 『江戸川大学紀要』第 23 号, 67-90.

米倉綽(2009).「Thomas Hardy の英語と文体——その統語法と語形成——」. 『英語史研究会会報研究ノート』第 14 号, 15-30.

執筆者紹介

田畑　智司（たばた　ともじ）	大阪大学大学院人文学研究科教授
菅原　裕輝（すがわら　ゆうき）	大阪大学大学院人文学研究科特任助教
曹　芳慧（そう　ほうけい）	大阪大学大学院言語文化研究科 博士前期課程
藤田　郁（ふじた　いく）	大阪大学大学院言語文化研究科 博士後期課程
涌井　萌子（わくい　もえこ）	大阪大学大学院人文学研究科 博士後期課程