



Title	テクストマイニングとデジタルヒューマニティーズ 2024 (冊子)
Author(s)	
Citation	言語文化共同研究プロジェクト. 2025, 2024
Version Type	VoR
URL	<a href="https://hdl.handle.net/11094/102217">https://hdl.handle.net/11094/102217</a>
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  
2024

黄 晨雯

藤田 郁

曹 芳慧

Dimitra Vogatza

田畠 智司

大阪大学大学院言語文化研究科  
2025



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  
2024

目 次

田畠 智司	プロジェクトの目的と活動	1
黄 晨斐	GraphRAGによる日本語文学テキストの知識グラフ分析 —江戸川乱歩『D坂の殺人事件』を事例として—	5
藤田 郁	Between Distant and Close: Sea and Inland in Tennyson by Topic Modelling	19
Dimitra Vogatza	Identifying Binary Oppositions in <i>the Lord of the Rings</i> : A primary quantitative approach to binary oppositions	43
曹 芳慧	HardyのWessex小説における伝達節と キャラクタライゼーション	69



# 「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 彰堯講師が着任・加入した。言語文化研究科と文学研究科の統合により設立された人文学研究科には「人文学林」という分野横断組織が置かれ、デジタルヒューマニティーズ振興の役割を担っている。その人文学林から、2022年に菅原 裕輝特任助教が、そして2023年に吉賀 夏子准教授が加入し、現在の陣容となっている。(職位はいずれも当時)。2016年度から、プロジェクトの名称を、当該リサーチコミュニティの名称としてより相応しい「テクストマイニングとデジタルヒューマニティーズ」にアップデートしたが、研究の系統は創始時より常に一貫している。

「テ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9名と名誉教授1名（今尾 康裕、菅原 裕輝、黄 晨斐、田畠 智司、Hodošček Bor、南澤 佑樹、三宅 真紀、山田 彰堯、吉賀 夏子、岩根 久名誉教授）、当研究科博士後期課程在学生6名（福本 広光、曹 芳慧、Camilleri Gabriele、陳 宇鎧、Vogatza Dimitra、王 翯影）、博士前期課程在学生3名（李 晨婕、肖 媛媛、于拙）に加え、OGの京都大学 徐 勤氏（2023年3月本研究科博士学位取得）、大阪医科大学 浅野 元子氏（2020年3月本研究科博士学位取得）・名古屋外国語大学 杉山 真央氏（2019年3月本研究科博士学位取得）、比治山大学 藤田 郁氏（2024年3月本研究科博士学位取得、本年4月に九州大学大学院システム情報科学研究院に異動）、帝塚山学院大学 八野 幸子氏（2018年3月本研究科博士課程修了）、国文学研究資料館の竹内 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実行している。

2024年度も、対面開催とオンライン併用のハイブリッド方式で開催した。2020–2022のパンデミック中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。

# 2024年度「テクストマイニングとデジタルヒューマニティーズ」研究会開催記録 およびメンバーによるDH関連学会での発表記録

第1回 2023年4月5日開催

発表者・発表題目

全メンバー 2024年度の活動計画打合せ

第2回 2024年5月10日開催

発表者・発表題目

于拙 「アイヌ語テキスト・コーパスを利用した  
アイヌ語の基本的な自然言語処理の試み」

第3回 2024年6月7日開催

発表者・発表題目

藤田 郁 “Was it Alfred, Charles, or Frederick?:  
Authorship questions surrounding Poems by Two Brothers (1827)”  
李晨婕 “A Stylometric Approach to the Sherlock Holmes Series and its Pastiches”

第4回 2024年7月5日開催

発表者・発表題目

カミッレーリ “Evaluating Word Alignment Strategies in a Japanese-Italian Translation Corpus”  
ガブリエレ  
竹森 ありさ 「英語色彩語を含む強意直喻表現における 含意と媒体の詳述について」

第5回 2024年8月30日開催

発表者・発表題目

岩根 久 「AI支援による16世紀フランス語テキスト研究」  
黄 晨斐 「日本語テキストに対するGraphRAGを用いたナレッジグラフ分析の試み」  
田畠 智司 “Dickens in vector space: Word vectors for literary semantics”

第6回 2024年9月13日開催

発表者・発表題目

山田 彰堯 「第一言語獲得・第二言語獲得と状態空間モデル」  
杉山 真央 「「ソビエト語」と「プーチン語彙」：  
戦勝記念日演説における「言説空間」の広がりと語彙による物語性の分析」  
八野 幸子 「理科・外国語科連携による教科等横断的教育での  
『種の起源』の活用可能性の検討2」

第7回 2024年10月5-6日開催 英語コーパス学会第50回大会於青山学院大学  
発表者・発表題目

立野 寛太 「Hurfordの制約に基づく“A or B”/“B or A”の形式差：  
COCAデータとR/Stanによるロジスティック回帰分析」

菅原 裕輝 「コーパス研究は仮説検証型の科学か？：形式概念分析を用いたメタ分析」

李 晨婕 “A Stylistic Approach to the Sherlock Holmes Series and its Pastiches”

肖 媛媛 「コーパスに基づく英語政治ニュース研究  
—英語母語圏と非母語圏の比較研究—」

曹 芳慧 「TEIを用いたHardy ウェセックス小説の会話部コーパス構築と可視化」

菅原 裕輝 (・ 神原一帆) 「コーパス研究の論理と倫理：学術的目的のコーパス利用に着目して」

杉山 真央 「大祖国戦争の記憶の形成と感情：プーチン戦勝記念日演説にみる人称代名詞」

第8回 2024年11月1日開催  
発表者・発表題目

曹 芳慧 「Thomas HardyのWessex小説の会話部によるキャラクタライゼーション」

肖 媛媛 「英語政治ニュースにおける主観的バイアスの検出と分析」

徐 勤 「多次元分析による中国語現代詩歌の言語特徴の考察」

第9回 2024年12月6日開催  
発表者・発表題目

王 簡影 「大規模言語モデルを利用した日本語母語識別/JNLI：  
従来の自然言語処理手法との比較」

南澤 佑樹 「スウェーデン語の前置詞framf'orの時間的意味について」

第10回 2024年12月21日開催 国際シンポジウム デジタルヒューマニティーズと研究基盤：  
欧州と日本の最新トレンド  
発表者・発表題目

Tomoji ポスターセッション  
Tabata, “Teaching Digital Humanities: As an Advanced Minor Programme”  
Natsuko  
Yoshiga,  
and Yuki Sug-  
awara

第11回 2025年1月17日開催  
発表者・発表題目

吉賀 夏子 「人文科学分野向け研究データ管理促進：  
デジタルヒューマニティーズ学習教材開発プロジェクト」

陳 宇錯 「専門家コミュニティと一般向けの科学技術日本語における  
レジスター・バリエーション—科学技術系の学術雑誌と「note」を対象とした多次元分析  
(MDA) —」

第 12 回 2025 年 2 月 7 日開催

発表者・発表題目

Hodošček Bor “Token and Dependency Grammar-based Pattern Matching for Japanese”

立野 寛太 「Hurford の制約に基づく “A or B”/“B or A” の形式差：

COCA データと R/Stan によるロジスティック回帰分析」

第 13 回 2025 年 3 月 7 日開催

発表者・発表題目

Dimitra Vogatza “Identifying Binary Oppositions in *the Lord of the Rings*: a primary approach”

浅野 元子 「A Vocabulary-Based Study Exploring Medical Students’ Interest in Research Abstracts

語彙的視点から見る医学生の論文抄録への関心」

第 14 回 2025 年 3 月 8 日開催 第 1 回 グローバル人文学 研究交流会

発表者・発表題目

曹 芳慧 「ハーディの Wessex 小説におけるキャラクタライゼーション—

文体意匠としての会話部—」

CHEN 「科学技術日本語は二つの場面においてどのように異なるのか？

YUKAI —科学技術系の学術論文とブログに対する Multi-Dimensional 分析—」

2025 年 5 月

研究代表者 田畠 智司

# GraphRAG による日本語文学テキストの知識グラフ分析 —江戸川乱歩『D 坂の殺人事件』を事例として—

黄 晨雯

大阪大学大学院人文学研究科  
〒 560-0043 豊中市待兼山町 1-8  
Email: huang.chenwen.hmt@osaka-u.ac.jp

**概要** 本稿では、Microsoft が 2024 年に提唱した検索拡張生成手法 GraphRAG を日本語文学に適用し、その応答生成および知識構造の可視化に基づく解釈支援の有効性を検討する。従来の RAG は外部ドキュメントを参照しながら生成を行う手法だが、非構造的情報に依存するため、複雑な文脈や人物関係、因果構造の把握には限界がある。これに対し、GraphRAG は知識グラフを統合することで、エンティティ間の関係を構造的に可視化し、文脈的に一貫した応答を実現する。

本稿では、江戸川乱歩の短編小説『D 坂の殺人事件』を対象とし、主題理解と犯人特定という二種類の問い合わせに対して GraphRAG を用いた応答分析を行った。さらに、Gephi および Neo4j によって生成された知識グラフを可視化し、登場人物間の関係性、構造的中心性、因果関係のネットワーク的把握を試みた。その結果、GraphRAG は日本語文学において意味的整合性のある応答を生成し、構造の視覚化を通じた物語解釈支援にも高い有用性を示した。

**キーワード** GraphRAG, ネットワーク可視化, 物語構造

## Knowledge Graph Analysis of Japanese Literary Texts Using GraphRAG: A Case Study of Edogawa Rampo's “The Murder on D-Slope”

Chenwen Huang

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

**Abstract** This study applies GraphRAG, a retrieval-augmented generation method proposed by Microsoft in 2024, to Japanese literary texts and examines its effectiveness in supporting textual interpretation through response generation and the visualization of knowledge structures. While conventional RAG relies on referencing unstructured external documents for generation, it faces limitations in capturing complex contexts such as interpersonal relationships and causal structures. In contrast, GraphRAG integrates knowledge graphs to structurally visualize the relationships between entities and generate contextually coherent responses.

In this research, Edogawa Rampo's short story The Case of the Murder on D. Hill is used

as a case study. Two types of queries—one targeting thematic understanding and the other aiming at identifying the perpetrator—are analyzed using GraphRAG-generated responses. Furthermore, the resulting knowledge graphs are visualized using Gephi and Neo4j to explore character relationships, structural centrality, and causal connections through network-based representations.

The results indicate that GraphRAG successfully generates semantically coherent responses in Japanese literary contexts and demonstrates high utility in supporting narrative interpretation through structural visualization.

**Keywords** GraphRAG, Network Visualization, Narrative Structure

## 1. 研究背景

近年、大規模言語モデルの発展は、自然言語処理の分野に大きな変革をもたらしている。中でも、検索拡張生成 (Retrieval-Augmented Generation, RAG) は、生成モデルに外部情報源を参照させることで、より正確で文脈に適した回答や文章の生成を可能にする手法として注目されている。RAG は、モデルが事前に獲得した知識に加え、外部の情報検索機能を組み合わせることで、ユーザーの問い合わせに対し動的かつ柔軟に対応する能力を備えている。

一方で、RAG のアプローチにはいくつかの課題も指摘されている。特に、日本語のように文法構造が複雑で省略や曖昧さの多い言語においては、RAG が生成する回答の文脈理解に限界が見られる。また、RAG は主に非構造的なドキュメント検索に依存しており、問い合わせが抽象的あるいは高文脈依存の問い合わせに対しては、関連情報の適切な取得が困難となる場合がある。

こうした背景の中で、2024 年に Microsoft が提唱した「GraphRAG」という新たなアプローチが注目を集めている。GraphRAG は、RAG に知識グラフを統合した新たな生成手法であり、複雑な文脈処理への対応力が期待されている。

本研究では、この GraphRAG を日本語文学テキストに適用する試みとして、江戸川乱歩の短編小説『D 坂の殺人事件』を対象に、その分析手法と成果を検証する。具体的には、Python を用いた GraphRAG ライブラリによりテキストからエンティティとリレーションを抽出し、その結果を Gephi や Neo4j で視覚化・分析することで、知識グラフの構築と解析を行う。

本稿の目的は、GraphRAG が日本語文学テキストに対してどの程度有効に機能するのかを実証することであり、同時に課題と今後の応用可能性についても考察を加えることである。生成 AI の日本語処理の今後を見据え、構造化情報を取り入れた生成モデルの新たな展開を探ることが、本研究の意義である。

## 2. RAG と GraphRAG の理論的基盤

### 2.1 RAG (Retrieval-Augmented Generation)

RAG は、大規模言語モデルと外部情報検索を組み合わせた生成手法である。RAG の主な特徴は、事前学習された知識に加え、ユーザーのクエリに応じて外部知識ベースから関連情報を

検索し、それを元により文脈に即した正確な文章を生成できる点にある。

この手法は、社内向けチャットボット、医療相談、法務支援など、正確な情報の参照が求められる場面で特に有用である。一方で、従来の RAG は主に非構造的な文書の検索に依存しており、情報の粒度や関連性が限定されることが多い。たとえば、「昨年の売上高はいくらですか？」のような具体的な質問には強いが、「昨年の経営的成功は何か？」といった抽象的な質問には対応が難しい。

このように、RAG には高い柔軟性と有用性がある一方で、抽象的、かつ関係性の強い問い合わせに対しては限界が存在する。

## 2.2 GraphRAG の概要と研究動向

このような RAG の限界を克服するために、Microsoft は 2024 年に GraphRAG (Graph-based Retrieval-Augmented Generation) と呼ばれる新たな手法を提案した。GraphRAG は、RAG に知識グラフ (Knowledge Graph) を統合することで、文書間の構造的な関係性を明示化し、それを検索および生成の過程に活用するものである。

GraphRAG のアプローチは、大きく「インデックスフェーズ」と「クエリフェーズ」の二つの段階に分けられる。まず、インデックスフェーズでは、テキストから人物、場所、出来事などのエンティティとその関係性を抽出し、知識グラフを構築する。この過程では、ノードとエッジの情報がグラフ構造として記録される。次に、クエリフェーズでは、ユーザーのクエリに基づいてグラフ構造が探索され、関連するノードやコミュニティ情報が取得される。これらの情報をもとに、言語モデルが回答文を生成する。

このように構造化データを活用することで、GraphRAG は非構造的な検索手法と比較して、文脈理解や関係性の把握において高い精度を実現する。実際、FAQ システムや顧客サポート、研究支援ツールなどの実務領域において、一部企業による GraphRAG の試験的導入が進められている。

Han ら (2024) は、GraphRAG の構成要素および検索生成の過程を整理し、従来の RAG との比較を通じて、その特性や課題 (例: 知識グラフ構築時のノイズやスパース性) を明らかにしている。さらに、問答やクエリ要約といったタスクにおける RAG と GraphRAG の性能差や、両者の統合的活用の可能性についても検討している。一方で、具体的なテキスト分析タスクへの応用には言及されていない。

このように、現時点では GraphRAG を用いた実証的な学術研究はまだ限られており、その応用可能性にはさらなる検討の余地がある。本研究はその初期的な応用例として、日本語のテキスト分析タスクにおける GraphRAG の有効性を検証することを目的とする。

### 3. 分析対象と方法論

#### 3.1 分析対象：江戸川乱歩『D 坂の殺人事件』

本研究の分析対象は、江戸川乱歩による短編推理小説『D 坂の殺人事件』である。本作は、名探偵・明智小五郎が初めて登場する作品として知られ、日本近代ミステリーの出発点の一つともされている。物語は、語り手である「私」が、友人の明智とともに事件現場となる古本屋に偶然立ち寄り、密室殺人の謎を追うという筋書きを有している。

『D 坂の殺人事件』を本研究の分析対象として選定した理由は、以下の三点に基づく。第一に、本作の日本語原文は構文的に複雑かつ多様であり、江戸川乱歩特有の文語的文体や繊細な心理描写を含んでいる。このような文体は、自然言語処理において一定の困難を伴うため、日本語における GraphRAG の性能評価に適した題材と考えられる。第二に、登場人物・空間・出来事の関係性が明確に記述されており、推理小説としての性質上、人物間の相互関係や時間的進行、行動と動機の因果関係が比較的明瞭に把握できる。そのため、知識グラフ構築に必要な情報構造を明示的に抽出しやすいという利点がある。第三に、本作品は著作権が切れており、青空文庫を通じて一般に入手可能である。このことは、研究の再現性や今後の応用展開を視野に入れた際の利便性という観点からも有用である。以上の理由から、本研究では『D 坂の殺人事件』を分析対象として採用する。

#### 3.2 GraphRAG の実装と応用

環境構築においては、Python 3.10 上に GraphRAG パッケージを導入し、プロジェクトディレクトリの初期化を行った後、OpenAI の API キーを設定し、使用モデルとして gpt-4o を選定した。特に、日本語による出力の自然性を担保するため、プロンプト設計において「日本語で答えてください」と明示的に指定するなど、言語面での調整を加えた。

インデックス作成の工程では、対象となる日本語小説テキストからエンティティおよび関係性を抽出し、構造化された知識グラフを生成した。その後、この知識グラフを参照して、global モードおよび local モードの二通りによる質問応答を実施した。global モードでは作品全体に関わる構造的・意味的情報に基づいた包括的な応答が得られるのに対し、local モードでは特定の文脈に焦点を当てた詳細な情報の抽出が可能である。これにより、物語の主題に関する理解と、登場人物の行動・動機といった具体的要素の双方から分析を行うことができた。

#### 3.3 知識グラフの可視化と分析ツールの活用

知識グラフの構造的特徴を可視化し、複雑なエンティティ間の関係性を俯瞰的に把握するために、Gephi および Neo4j の二つのツールを併用した。Gephi は、GraphRAG によって出力されたグラフデータを視覚的に表示するために用いられ、ノード間の距離、ネットワーク密度、

クラスタリング構造などの可視化に寄与した。一方、Neo4j はグラフデータベースとして機能し、クエリベースの探索やノード属性（例：Person, Location, Organization）の分類、特定条件に基づくサブグラフの抽出に有用である。これにより、GraphRAG の出力結果に対する柔軟かつ構造的な分析が可能となった。

上記の分析手法を通じて、本研究では物語理解における GraphRAG の有効性、特に「主題把握」と「犯人特定」という二つの典型的な問い合わせに対して、どの程度意味的整合性のある応答が得られるかを評価した。

## 4. 分析結果と考察

第4章では、GraphRAG を用いて江戸川乱歩『D坂の殺人事件』を分析した結果に基づき、日本語テキストにおける意味構造の把握と知識抽出の有効性について考察を行う。

分析対象としては、第3章で述べた二種類の問い合わせ、つまり抽象的な主題分析と、具体的な犯人特定をもとに、それぞれ global モードおよび local モードによる応答結果を比較検討する。ここでは、生成された回答の精度、知識グラフの構造的特性、そして Gephi や Neo4j による可視化が分析に与える効果の三点を中心に論じる。

### 4.1 小説の主題に関するグローバルクエリ分析

最初に、「この小説のテーマは何ですか？」という問い合わせに対して、GraphRAG は以下のようないくつかの重要なテーマが存在します。それらのテーマは物語の進行やキャラクターの関係性を通じて深く掘り下げられています。

#### 1. 小説のテーマ

この小説にはいくつもの重要なテーマが存在します。それらのテーマは物語の進行やキャラクターの関係性を通じて深く掘り下げられています。

#### 2. 犯罪と捜査

犯罪と捜査のテーマは小説の中心的な要素です。明智や彼、そして警察などの主要なキャラクターが関与する様々な事件や捜査が描かれています。これらの要素は、犯罪を解決することの複雑さや人間行動の理解の難しさを強調している。

#### 3. 人間関係と相互作用

犯罪と捜査の文脈における人間関係と相互作用も重要なテーマです。明智と語り手、彼と私、そして様々な証人や容疑者との関係が物語に深みを与え、犯罪が個人的なつながりに与える影響を探求しています。

#### 4. 心理学的探求

心理学的探求もまた重要なテーマです。明智のようなキャラクターが『心理学と犯罪』といった作品を引用し、人間の観察や記憶の不確かさを説明します。このテーマは、コミュニティの知的深さや心理学理論が捜査に与える影響を強調しています。

#### 5. 社会問題と秘密

古本屋での事件や男と女が関与する殺人事件など、社会問題と秘密のテーマも見逃せません。これらの出来事は、隠された社会問題や虐待的な関係、秘密の欲望を明らかにし、複雑で多面的な物語を形成して

います。

## 6. 正義と法的手続き

正義と法的手手続きのテーマも重要です。警察、探偵、検察官が様々な捜査に関与することで、正義の追求や犯罪の法的な影響が強調されています。

## 7. 結論

この小説は、犯罪と捜査、人間関係、心理学的探求、社会問題と秘密、そして正義と法的手手続きといった多岐にわたるテーマを通じて、読者に深い洞察を提供します。それぞれのテーマが絡み合い、物語に豊かな層を加えています。

GraphRAG は、原文テキストに含まれる主題を自動的に分類し、それぞれに関連する文脈情報もとに解釈を付加する点で、高い有用性を示した。とりわけ、日本語における抽象的な表現や比喩的記述に対しても一定の精度で意味を捉え、それを明示的な言語として再構成している点は特筆に値する。また、テーマ間の重なりも適切に記述されており、知識グラフの構造によって各トピック間の関係性が視覚的かつ論理的に提示されることで、作品に内在する複雑な意味構造をより明確に理解できるようになっていた。

## 4.2 犯人に関するローカルクエリ分析

次に、「殺人を犯したのは誰ですか？」という問い合わせに対する GraphRAG の応答を検討する。

### 1. 殺人を犯したのは誰か？

殺人を犯したのは、旭屋の主人です。彼は物語の中で中心的な役割を果たしており、複数の証拠が彼を犯人として指し示しています。

### 2. 旭屋の主人の役割

旭屋の主人は、残酷なサディストとして描かれており、殺人罪を犯したと疑われています。彼は罪跡を隠すために便所を借りた男のことを話しており、その行動が彼の犯行を裏付けるものとされています [Data: Entities (93); Relationships (102, 127, 128, 130, 129)]。

### 3. 犯行の動機と手口

旭屋の主人は、サード卿や女のマゾッホとの関係を持ち、複雑な人間関係の中で犯行を行ったとされています。彼の行動は計画的であり、証拠を隠すために様々な手段を講じています。例えば、電燈のスイッチを操作して指紋を残さないようにしたり、便所を借りて証拠を隠そうとしたりしています [Data: Entities (93); Relationships (102, 128, 130)]。

### 4. 他の登場人物との関係

犯人である旭屋の主人は、被害者との間に悪意がなかったとされる関係を持っていました。被害者は犯人と同意の上で殺害された可能性があり、この点が物語の複雑さを増しています [Data: Entities (91); Relationships (100)]。また、明智は犯人の存在を否定し、自分の無罪を証明しようとしていますが、これも物語の緊張感を高める要素となっています [Data: Relationships (39)]。

### 5. 結論

以上の情報から、殺人を犯したのは旭屋の主人であることが明らかです。彼の行動や人間関係、そして証拠の隠蔽工作が彼を犯人として特定する要因となっています [Data: Entities (93); Relationships (102, 127, 128, 130, 129)]。

```
{`create_final_relationships.parquet.as.json` 1 ×
118 {"source": "\\"法廷\"", "target": "\\"会員\"", "weight": 1, "description": "\\"会員は、法廷に持ち出される問題に関する記録を書くように指示
119 \"source": "\\"薔薇屋の主人\"", "target": "\\"死んだ細君\"", "weight": 1, "description": "\\"薔薇屋の主人は死んだ細君の夫であり、彼女の身
120 \"source": "\\"薔薇屋の主人\"", "target": "\\"話者\"", "weight": 1, "description": "\\"話者は薔薇屋の主人に対して聯想診断を行い、彼の心理
121 \"source": "\\"薔薇屋の主人\"", "target": "\\"自首\"", "weight": 1, "description": "\\"薔薇屋の主人は自首したことで物語に影響を与えている
122 \"source": "\\"大岡越前守\"", "target": "\\"心理的方法\"", "weight": 1, "description": "\\"大岡越前守は心理学が発達していない時代において
123 \"source": "\\"ボオ\"", "target": "\\"デュパン\"", "weight": 1, "description": "\\"デュパンはボオの作品『ル・モルグ』に登場するキャラクタ
124 \"source": "\\"ボオ\"", "target": "\\"ドイル\"", "weight": 1, "description": "\\"ドイルはボオの作品に影響を受けており、ホームズにデュパン
125 \"source": "\\"ドイル\"", "target": "\\"ホームズ\"", "weight": 1, "description": "\\"ホームズはドイルの作品『レジメント・ベーシェント』に
126 \"source": "\\"ホームズ\"", "target": "\\"レジメント・ベーシェント\"", "weight": 1, "description": "\\"ホームズは『レジメント・ベーシェン
127 \"source": "\\"僕\"", "target": "\\"旭屋の主人\"", "weight": 1, "description": "\\"僕は旭屋の主人が殺人者であると考えている。\"", "text
128 \"source": "\\"旭屋の主人\"", "target": "\\"殺人罪\"", "weight": 1, "description": "\\"旭屋の主人は殺人罪を犯したと疑われている。\"", "text
129 \"source": "\\"旭屋の主人\"", "target": "\\"便所を借りた男\"", "weight": 1, "description": "\\"旭屋の主人は罪跡を隠すために便所を借りた男
130 \"source": "\\"旭屋の主人\"", "target": "\\"サード卿\"", "weight": 1, "description": "\\"旭屋の主人 is connected to サード卿 through
131 \"source": "\\"旭屋の主人\"", "target": "\\"女のマゾッホ\"", "weight": 1, "description": "\\"旭屋の主人 discovered 女のマゾッホ next to
132 \"source": "\\"旭屋の主人\"", "target": "\\"古本屋の細君\"", "weight": 1, "description": "\\"旭屋の主人 and 古本屋の細君 both have sa
133 \"source": "\\"古本屋の細君\"", "target": "\\"旭屋の細君\"", "weight": 2, "description": "\\"古本屋の細君と旭屋の細君は、共に生傷を持ち、同
```

図 1: GraphRAG 生成の関係データ（「旭屋の主人」に関する抜粋）

GraphRAG は、知識グラフ上に存在するエンティティ（旭屋の主人）とその関連性（犯行の動機、行動、他者との関係）を統合的に捉え、文中の因果構造を明快に説明する能力を示した。図 1 に示されるように、旭屋の主人は「殺人罪」（関係 127）や「便所を借りた男」（関係 128）、「女のマゾッホ」（関係 130）といった複数の関係を通じて描写されており、これらの情報は彼の犯行を示唆する重要な手がかりとなっている。

たとえば、関係 128 において旭屋の主人は「犯罪を隠すために便所を借りた男」として記述されており、これは物語内での行動（証拠隠滅）と結びつけられている。また、関係 130 では「女のマゾッホ」との発見的関係が記述され、加害と被害の単純な二元対立を超えた人間関係が浮かび上がる。さらに注目すべきなのは、これらの関係が自然言語で記述された説明文（description）と共に記録されている点である。たとえば、関係 127 には「旭屋の主人は殺人罪を犯したと疑われている」という記述が明示されており、モデルの解釈が単なる構造的リンクにとどまらず、意味的・語用論的含意を持つものであることがわかる。

このように、GraphRAG は従来のベクトル類似度依存の RAG とは異なり、知識の構造化と自然言語理解を組み合わせることで、物語内の因果関係や人間関係の重層的構造を的確に捉えることが可能である。

### 4.3 Gephi による知識グラフの可視化

GraphRAG によって自動生成された知識グラフは、物語内の登場人物、事件、場所などをノードとして構成し、それらの関係性をエッジで結ぶネットワーク構造を形成している。本節では、GraphRAG の出力ファイル `summarized_graph.graphml` を用いて、オープンソースの可視化ツール Gephi によりその構造を視覚化し、物語全体の意味的構造と登場人物の中心性を分析した。

データ - テーブル ×								
ノード	辺	設定	ノードの追加	辺の追加	検索/置換	スプレッドシートのインポート	テーブルのエクスポート	作業を継続
Source	Target	Type	Id	Label	Interval	Weight	フィルタ:	Source
・白梅軒	「明智小五郎」	無向	0		1.0	9d25989cccfbb6b0605864..	「明智小五郎は白梅軒で...	
・白梅軒	「古木屋」	無向	1		1.0	9d25989cccfbb6b0605864..	「古木屋は白梅軒の向か...	
・白梅軒	「小林軒事」	無向	2		1.0	0a61aca2a278f05bc5..	「小林軒事は白梅軒から...	
・白梅軒	「白梅軒」	無向	3		1.0	9d25989cccfbb6b0605864..	「白梅軒は白梅軒で...」	
・白梅軒	「スクリーミング」	無向	4		1.0	db91434e4a01011f10f4f489..	「白梅軒から-スクリーミング...	
・白梅軒	「君」	無向	5		1.0	20e8990a0f91d1f1f4f489..	「君は白梅軒の隣の...」	
・古木屋	「明智小五郎」	無向	6		1.0	9d25989cccfbb6b0605864..	「君は明智小五郎の隣の...」	
・古木屋	「古木屋の女房」	無向	7		1.0	5057023e0a1337f076f..	「古木屋の女房は古木屋...」	
・古木屋	「隅子」	無向	8		1.0	7db734b53849ebad-e4..	「古木屋の隅子は、万引...」	
・古木屋	「私」	無向	9		1.0	0f6532cf2d2cd3b4d..	「彼の古木屋で発生した...」	
・古木屋	「安」	無向	10		1.0	0f6532cf2d2cd3b4d..	「彼の古木屋で発生した...」	
・古木屋	「本沢沢」	無向	11		1.0	0f6532cf2d2cd3b4d..	「本沢沢は古木屋で発生した...」	
・古木屋	「上野の広小路」	無向	12		1.0	d73d62d62d07e1b1d..	「古木屋は、上野の広小...」	
・古木屋	「事件」	無向	13		2.0	229d95850962ab0a..	「古木屋の事件は通じて...」	
・古木屋	「加屋」	無向	14		2.0	25243ab34bdc7144deef..	「古木屋の加屋は隣接して...」	
・古木屋	「時計屋」	無向	15		1.0	25243ab34bdc7144deef..	「古木屋は時計屋の隣に...」	
・古木屋	「裏子屋」	無向	16		1.0	25243ab34bdc7144deef..	「古木屋は裏子屋の隣に...」	
・古木屋	「袋足屋」	無向	17		1.0	25243ab34bdc7144deef..	「古木屋は袋足屋の隣に...」	
・明智小五郎	「古木屋の女房」	無向	18		1.0	5087822e0a1337f076f..	「明智小五郎と古木屋の...」	
・明智小五郎	「宝物の山」	無向	19		1.0	ab4147228cc67f0281..	「明智小五郎の宝物の山...」	
・明智小五郎	「男」	無向	20		1.0	60c3250a091d942b126..	「明智小五郎は男の...」	
・明智小五郎	「腰り手」	無向	21		1.0	ec3852b9475a5de5ee..	「腰の手は明智小五郎の...」	
・明智小五郎	「障子」	無向	22		1.0	7db734b53849ebad-e4..	「九の高」の腰り手の時に...」	
・明智小五郎	「九月」	無向	23		1.0	b40af78f698abc9189..	「腰の障子を引き開け立...」	
・古木屋のお神さん	「加屋のお神さん」	無向	24		1.0	59761978f88a605864..	「腰の障子を引き開け立...」	
・谷崎潤一郎	「本屋」	無向	25		1.0	061360191d9d8050a29..	「谷崎潤一郎の作品『道...』」	
・谷崎潤一郎	「明智」	無向	26		1.0	7d79c947d7c51c01ec..	「明智は谷崎潤一郎の...」	
・谷崎潤一郎	「『達上』」	無向	27		1.0	7d79c947d7c51c01ec..	「『達上』は谷崎潤一郎...」	
・明智	「私」	無向	28		3.0	16330a5442d1b94124..	「明智と私(手札)は手...」	
・明智	「警察」	無向	29		2.0	a21b1e6711889940..	「明智は死を免れました後...」	
・明智	「事件」	無向	30		4.0	9c760ab590ea4ced9f2..	「明智は、事件の詳細を覗...」	
・明智	「智」	無向	31		1.0	165839478f78ab50c0b..	「智はinquired about...」	
・明智	「彼」	無向	32		1.0	185ed197672b0cb6b..	「彼(明智)の部屋に足を...」	
・明智	「神田伯龍」	無向	33		1.0	0661af124d74781688..	「明智は頭(さき)や声が...」	
・明智	「D坂の事件」	無向	34		1.0	99944ca67676a8501adec..	「明智はD坂の事件につ...」	
・明智	「君」	無向	35		1.0	187aab5e563a63cae43..	「明智は君の推理につ...」	

図2: エッジテーブル

Gephi では、知識グラフを構成するノードおよびエッジの詳細情報を表形式で確認することができる。エッジテーブル（図2）には、Source（始点ノード）、Target（終点ノード）、および両者間の関係の強さを示す Weight が記録されており、これらの情報が視覚的なネットワーク描画の基盤となっている。

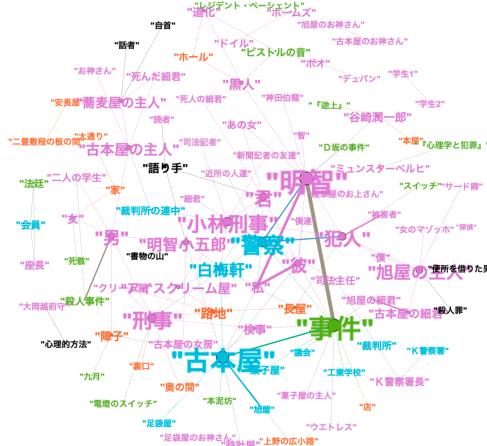


図3：全体ネットワーク図

図3は、Gephiによって可視化された知識グラフ全体の構造を示している。ノードのサイズは、そのノードに接続しているエッジの数に比例しており、サイズが大きいほど多くの関係がそのノードを中心に関連していることを意味する。また、エッジの太さ（weight）は関係の強度を示し、太い線ほど関係性が密接であることを表している。ノードの色は自動検出されたクラスタに対応しており、ネットワーク内のサブコミュニティ構造を視覚的に把握できる。図3で特に注目されるのは、「事件」「明智」「警察」「犯人」「旭屋の主人」といったノードがネットワークの中央付近に密集し、いずれも高い中心性と接続密度を有している点である。こ

これらのノードは、物語の中核的な構造を形成する重要な要素であることが視覚的にも確認できる。

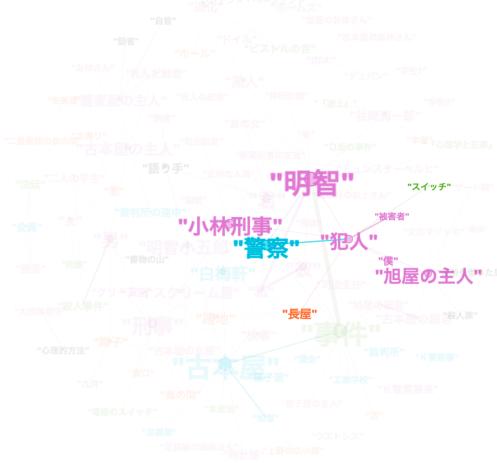


図4：犯人をめぐるネットワークグラフ

図4は、全体グラフから「犯人」に関するノードを中心として抽出・拡大したサブグラフである。この図からは、犯人が「明智」「警察」「小林刑事」「旭屋の主人」「被害者」「僕」などのノードと接続されていることが確認でき、特に旭屋の主人が主要な関係ノードとして現れており、物語中で犯人として疑われる立場にあることが構造的にも示唆されている。

このような知識グラフの可視化は、自然言語テキストの記述や構文解析だけでは捉えにくい人物間の関係性の強度や物語における構造的中心性を直感的に把握するうえで極めて有効であり、デジタル人文学における分析手法として大きな可能性を有している。

データ・テーブル×					
ノード	辺	設定	ノードの追加	辺の追加	検索/置換
		スプレッドシートのインポート			
Id	Label	Interval	source_id	description	type
「白梅軒」	「白梅軒」		0a67a2a52a278058bc549af7d58d8cc6...	白梅軒は、君が帰り道で立ち寄った場...	「ORGANIZATION」
「古本屋」	「古本屋」		0fa5352cf2f2ddcd3b3d031769e079911...	古本屋は、事件が発生した場所であり...	「ORGANIZATION」
「明智小五郎」	「明智小五郎」		50878206d0a133f7072f9298e7e92160...	明智小五郎は、主人公が白梅軒であり...	「PERSON」
「古本屋の女房」	「古本屋の女房」		50878206d0a133f7072f9298e7e92160...	「古本屋の女房は明智小五郎の副乳栄...」	「PERSON」
「障子」	「障子」		7db734bd538496ab8ed466872a4a6e4f...	障子は、専門家の間では無名と称され...	「GEO」
「九月」	「九月」		7db734bd538496ab8ed466872a4a6e4f...	「九月は、夏の終わりから秋の始まり...」	「EVENT」
「古本屋のお神さん」	「古本屋のお神さん」		597b7c6aaefc9919875265a87c7c159c2...	「古本屋のお神さんは純善な人が、身...	「PERSON」
「旭屋のお神さん」	「旭屋のお神さん」		597b7c6aaefc9919875265a87c7c159c2...	「旭屋のお神さんもよく備をしており、...」	「PERSON」
「旭屋の御神さん」	「旭屋の御神さん」		06136b019d9d8050a29712b500ea2a8792...	「古本屋一郎は日本の著名な作家であり...	「PERSON」
「谷崎潤一郎」	「谷崎潤一郎」		06136b019d9d8050a29712b500ea2a8792...	「谷崎潤一郎は、物語の中で登場人物が馴染んで...」	「GEO」
「本屋」	「本屋」		0661a124df74781a68b1254c20cc7bca...	「明智」は物語の登場人物であり、二階...	「PERSON」
「明智」	「明智」		0fa5352cf2f2ddcd3b3d031769e079911...	「明智」は、専門家の間では無名と称され...	「PERSON」
「私」	「私」		7d79c947d7c51cc04bd509f448ba...	「『途上』は谷崎潤一郎の小説で、作中...」	「EVENT」
「『途上』」	「『途上』」		0fa5352cf2f2ddcd3b3d031769e079911...	「探偵は犯事を解決する能力を持つ人物...」	「PERSON」
「探偵」	「探偵」		0fa5352cf2f2ddcd3b3d031769e079911...	「彼は、殺人事件の中心人物であり、犯...」	「PERSON」
「彼」	「彼」		0fa5352cf2f2ddcd3b3d031769e079911...	「0fa5352cf2f2ddcd3b3d031769e079911...」	「PERSON」
「本泥坊」	「本泥坊」		5963a919e0cd9299ac9e4b6e5d830d...	「本泥坊は古本屋で発生した事件である...」	「EVENT」
「警察」	「警察」		ca32c230a013731609a074cd24461ad...	「警察は指紋の検査を行なう組織です...」	「ORGANIZATION」
「細君」	「細君」		1bdb4bcfa5f332981d1b2e233cd3cd10...	「細君は細君の内に隠れて死んでいた女性で...」	「PERSON」
「二暮敷板の板の間」	「二暮敷板の板の間」		1bdb4bcfa5f332981d1b2e233cd3cd10...	「二暮敷板の板の間は、裏方に接して致...」	「GEO」
「安長屋」	「安長屋」		1bdb4bcfa5f332981d1b2e233cd3cd10...	「安長屋は、ごくありふれた間戸の建物...」	「GEO」
「死骸」	「死骸」		0b9a792b952c3ac3a93f93cb86fe996...	「死骸は、左側の壁寄りに店の間の方...」	「EVENT」
「女」	「女」		0b9a792b952c3ac3a93f93cb86fe996...	「女」は男性に襲撃された被害者であり...」	「PERSON」
「大通り」	「大通り」		0b9a792b952c3ac3a93f93cb86fe996...	「大通りは表にあり、往来が絶えない場...」	「GEO」
「家」	「家」		0b9a792b952c3ac3a93f93cb86fe996...	「家は障子一重で、一人の女の女が惨殺され...」	「GEO」
「上野の広小路」	「上野の広小路」		d73dbf62d07ee41b42ce95ea8fe1b77...	「上野の広小路は、夜店がよく出る場所...」	「GEO」
「足袋屋のお神さん」	「足袋屋のお神さん」		d73dbf62d07ee41b42ce95ea8fe1b77...	「足袋屋のお神さんは、勝家に任んでお...」	「PERSON」
「時計屋」	「時計屋」		2524343b4bc7f144de9e201744349...	「時計屋」は事件の際に足袋屋のお...」	「PERSON」
「古本屋の主人」	「古本屋の主人」		3fa5d5ab5e541242d2d5aa047b5459...	「古本屋の主人は、事件の中心にいる...」	「PERSON」
「お神さん」	「お神さん」		3fa5d5ab5e541242d2d5aa047b5459...	「お神さんは、隣家の足袋屋の人物で、時...」	「PERSON」
「K警察署長」	「K警察署長」		072a2e0073e931c94b1877c21dece186...	「K警察署長」は警察署の署長であり、...」	「PERSON」
「小林刑事」	「小林刑事」		072a2e0073e931c94b1877c21dece186...	「小林刑事は、指紋の検査を依頼された...」	「PERSON」
「裁判所の連中」	「裁判所の連中」		3fa5d5ab5e541242d2d5aa047b5459...	「裁判所の連中は警察からの急報で駆け...」	「ORGANIZATION」
「K警察署」	「K警察署」		3fa5d5ab5e541242d2d5aa047b5459...	「K警察署はK警察署が所属する警察...」	「ORGANIZATION」
「裁判所」	「裁判所」		072a2e0073e931c94b1877c21dece186...	「裁判所は法的な問題を扱う機関であり...」	「ORGANIZATION」
「司法記者」	「司法記者」		072a2e0073e931c94b1877c21dece186...	「司法記者は法的な事件を報道する記者...」	「PERSON」
「司法主任」	「司法主任」		072a2e0073e931c94b1877c21dece186...	「司法主任は法的な問題を担当する主任...」	「PERSON」

図5：ノードテーブル

一方、ノードテーブル（図5）では、各ノードに割り当てられた type（例：PERSON, GEO, EVENT）や description（自然言語による説明）などのメタ情報が確認できるが、これらの属性情報を活用した検索や条件付き可視化には、次節で述べる Neo4j による探索がより適している。

#### 4.3 Neo4j による知識グラフの可視化

Neo4j は、グラフデータを効率的に格納、可視化するためのグラフデータベース管理システムであり、エンティティとリレーションを高度に構造化された形式で保持することが可能である。本研究では、GraphRAG が output したファイルを Neo4j にインポートし、知識グラフを属性ベースで動的に探索する。

Gephi によるネットワークが主に関係構造の可視化とクラスタ分析に特化しているのに対し、Neo4j では各ノードやリレーションに付与された属性情報をもとに、クエリベースで特定のサブグラフや関係構造を抽出することができる。また、Cypher と呼ばれるクエリ言語を使用することで、複雑なパターンマッチングや階層的構造の探索も容易であり、文学作品内の因果関係や登場人物の役割分担といった分析に有用である。

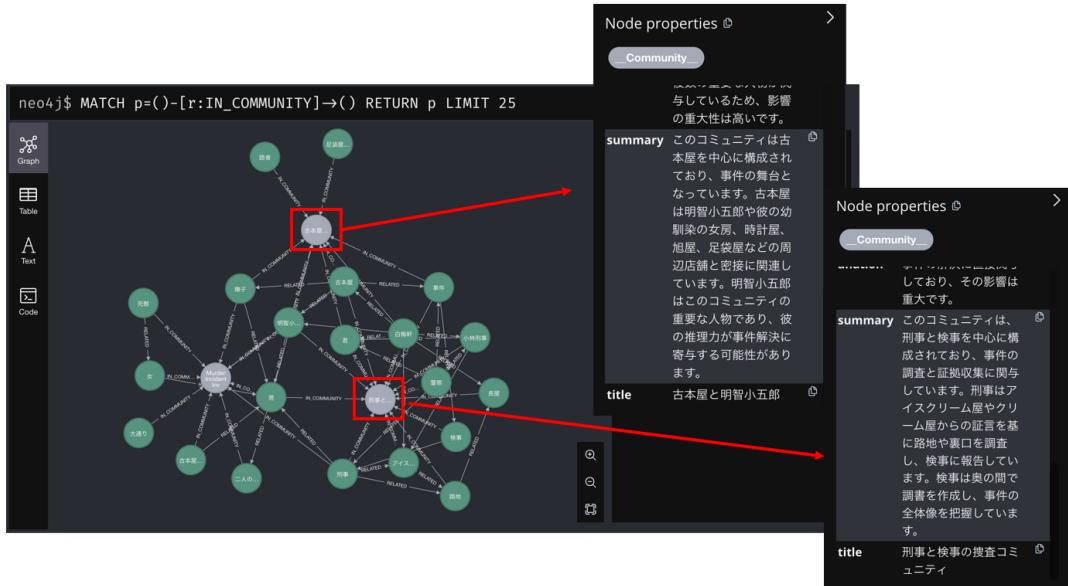


図 6：Neo4j におけるコミュニティ構造

図 6 は、Neo4j 上で IN\_COMMUNITY リレーションを用いて知識グラフのコミュニティ構造を抽出・可視化したものである。ここでは、登場人物や事件がどのような社会的・機能的なグループに属しているかを視覚的に把握することができる。

図中では、例えば「古本屋を中心に構成されたコミュニティ」と「刑事と検事の捜査に関連するコミュニティ」が確認され、それぞれのノードに付与された summary プロパティには、

そのコミュニティの構成要素や物語内での機能的役割が自然言語で記述されている。例として、古本屋を中心とするコミュニティには明智小五郎や旭屋、時計屋、足袋屋などの情報が含まれ、物語の舞台設定や犯行現場との関連性が明示されている。

一方、刑事を中心とするもう一つのコミュニティでは、アイスクリーム屋からの証言や捜査記録の整理が記されており、事件の調査と証拠構築における役割が強調されている。このように、Neo4j ではノードに付与された summary や title などの属性情報をもとに、各コミュニティの語り機能や構造的役割を定性的に把握することが可能であり、Gephi では実現しにくい意味論的補足を加える分析が可能となる。



図 7：人物ノードの関係構造

図 7 は、Neo4j において Person ラベルを持つノードを抽出し、犯人を中心とする人物ネットワークを表示したものである。図 7 では、「犯人」ノードが中心的な位置を占め、他の登場人物（警察、小林刑事、僕、彼など）と多数の関係性を有していることが視覚的に確認できる。

また、「旭屋の主人」に着目すると、そのノードには description 属性として「彼は殺人者であり、罪跡を隠すために便所を借りた男のことを話している」といった詳細な自然言語の記述が付与されている。これは、GraphRAG がテキスト中の人物描写や行動を解析し、意味的に解釈可能なプロファイルとして構造化していることを示している。

図 8 は、Neo4j において Geo ラベルを持つノードを抽出し、物語内に登場する地理的要素（場所）の構造を視覚的に表示したものである。図 8 では、「奥の間」や「裏口」「路地」などがノードとして確認され、それぞれの空間がどのような登場人物や事件と関係しているかが可視化されている。

たとえば「奥の間」は、図の中央付近に表示されており、「刑事と捜査のコミュニティ」や「犯人」などとの関係性を持つことが確認できる。右側の属性ビューに示されるように、このノードには description 属性として「奥の間は死体が発見された場所で、検事が陣取っていた部屋」といった事件における場所の役割が記述されており、物語内の犯罪構造を空間的に補完

する情報として機能している。このように、Geo ノードの探索を通じて、事件の発生場所や登場人物の行動軌跡を視覚的に把握することで、物語内の空間的ナラティブ構造をより立体的に読み解く手がかりが得られる。

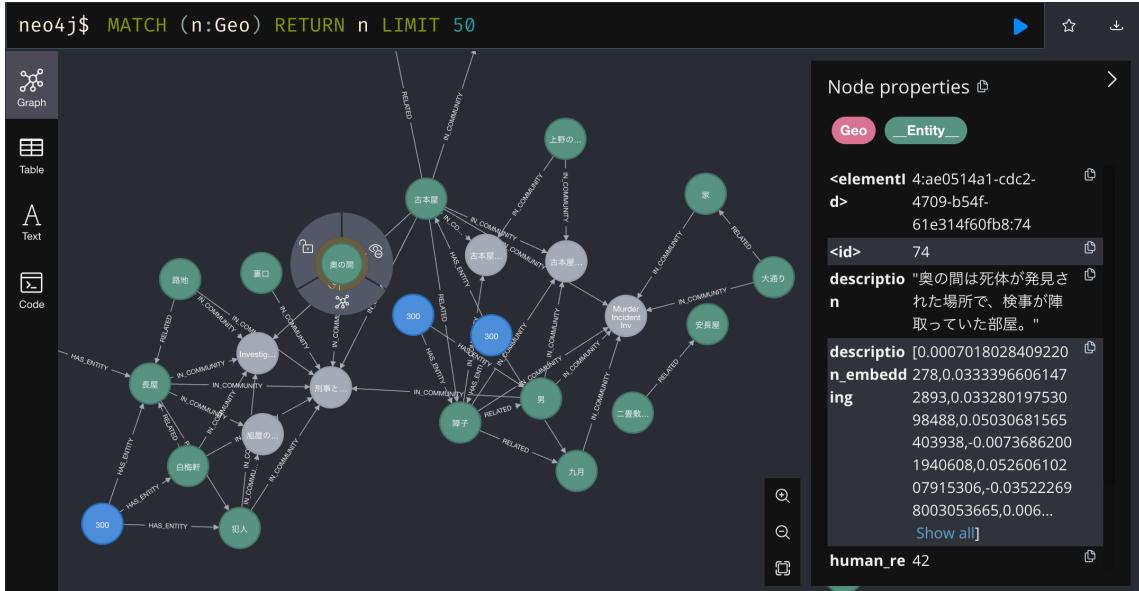


図8：地理ノードの関係構造

本節で示したように、Neo4j は GraphRAG によって構築された知識グラフを対話的かつ属性ベースで探索する上で、極めて有効な手段である。特に、ノードやリレーションに付与された type、description、summary といったメタ情報を活用することで、登場人物の役割、事件との関係性、物語内での空間的な意味づけなどを細部まで精緻に読み解くことが可能となった。

## 5. おわりに

本研究では、GraphRAG を江戸川乱歩の短編小説『D 坂の殺人事件』に適用し、従来の RAG では把握が難しかった構造的・意味的関係の抽出に成功し、物語内のテーマ分析や犯人特定といった問い合わせに対して、高い精度と多角的な応答を示した。また、Gephi や Neo4j といった可視化ツールとの併用により、抽出された知識グラフの構造を視覚的に把握することが可能となり、文学作品の解釈支援手法としての有用性も確認された。

また、GraphRAG は理論的にはより長大なテキストにも適用可能であり、たとえば江戸川乱歩の連作シリーズや現代小説を対象とした分析において、登場人物の変遷や物語構造の展開、主題の深化といった要素を構造的に追跡する応用が期待される。そのためには、段階的なインデックス生成や部分的再構築といった処理の最適化が今後の課題となる。また、複数作品間の知識グラフを統合し、物語世界を横断的に分析する読解支援手法への発展も視野に入れられるだろう。

## 文献

1. Chandak, P., Huang, K., & Zitnik, M. (2023). Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1), 67.
2. Chai, Z., Zhang, T., Wu, L., Han, K., Hu, X., Huang, X., & Yang, Y. (2023). GraphLLM: Boosting graph reasoning ability of large language model. *arXiv preprint arXiv:2310.05845*.
3. Chen, J., Lin, H., Han, X., & Sun, L. (2023). Benchmarking large language models in retrieval-augmented generation. *arXiv preprint arXiv:2309.01431*.
4. Cvitkovic, M. (2020). Supervised learning on relational databases with graph neural networks. *arXiv preprint arXiv:2002.02046*.
5. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
6. Gupta, S., Ranjan, R., & Singh, S. N. (2024). A comprehensive survey of retrieval-augmented generation (RAG): Evolution, current landscape and future directions. *arXiv preprint arXiv:2410.12837*.
7. Han, H., Shomer, H., Wang, Y., Lei, Y., Guo, K., Hua, Z., Long, B., Liu, H., & Tang, J. (2025). RAG vs. GraphRAG: A systematic evaluation and key insights. *arXiv preprint arXiv:2502.11371*.
8. Han, H., Wang, Y., Shomer, H., Guo, K., Ding, J., Lei, Y., Halappanavar, M., Rossi, R. A., Mukherjee, S., Tang, X., et al. (2024). Retrieval-augmented generation with graphs (GraphRAG). *arXiv preprint arXiv:2501.00309*.
9. Han, H., Xie, Y., Liu, H., Tang, X., Nag, S., Headden, W., Li, Y., Luo, C., Ji, S., He, Q., et al. (2025). Reasoning with graphs: Structuring implicit knowledge to enhance LLMs reasoning. *arXiv preprint arXiv:2501.07845*.
10. Hu, Y., Lei, Z., Zhang, Z., Pan, B., Ling, C., & Zhao, L. (2024). GRAG: Graph retrieval-augmented generation. *arXiv preprint arXiv:2405.16506*.
11. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., et al. (2021). Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv preprint arXiv:2005.11401*.
12. Ma, X., Gong, Y., He, P., Zhao, H., & Duan, N. (2023). Query rewriting for retrieval-augmented large language models. *arXiv preprint arXiv:2305.14283*.
13. Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., Zhang, Y., & Tang, S. (2024). Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*.
14. Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023). In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.
15. Xia, F., Sun, K., Yu, S., Aziz, A., Wan, L., Pan, S., & Liu, H. (2021). Graph learning: A survey. *arXiv preprint arXiv:2105.00696*.



# Between Distant and Close: Sea and Inland in Tennyson by Topic Modelling

Iku Fujita

Graduate School and Faculty of Information Science and Electrical Engineering,  
Kyushu University

744 Motooka, Nishi-ward, Fukuoka city, Fukuoka, 819-0395 Japan  
E-mail: tnnysn.annie.if@gmail.com

**Abstract** There have been many earlier studies on Alfred, Lord Tennyson, a representative poet of nineteenth-century English literature. Despite the fact that quantitative analyses have been used with prose texts, studies of Tennyson and other poets have rarely used such numerical approaches meanwhile they have employed the qualitative approach based on their close reading. This study reads Tennyson's poems from a distance by using a topic model, a quantitative technique. The analyses are based on the results of the latent Dirichlet allocation topic model (LDA), and examine how the poet depicts his themes by using imagery of the sea and inland environments, the latter including mountains and rivers. In the emerging results of LDA, the sea tends to convey the impressions of despair and eternity (after life) with the image of the broad sea itself, while inland landscapes of mountains connote a religious sense and rivers are used figuratively as life, but not hopelessness or perpetuity. The study's findings are not necessarily novel, but they do demonstrate how integrating quantitative and qualitative methodologies can be beneficial. Despite these components, neither the author nor the topic modelling stressed that a poem should only have one motif, theme or topic. The researcher 'reads' Tennyson's poems from a distance and up close while weighing the advantages and disadvantages of qualitative and quantitative approaches.

**Keywords** Alfred Tennyson, distant reading, LDA, poetry, topic model

## トピックモデリングを用いた テニスンにおける海と陸の考察

藤田 郁

九州大学大学院システム情報科学研究院

〒 819-0395 福岡県福岡市西区元岡 744 番地

**あらまし** 本研究は、定量的手法であるトピックモデルを用い、19世紀イギリス文学を代表する詩人一人、アルフレッド・テニスンの詩を分析、遠読するものである。テニスンやテニスンの作品に関する先行研究は数多く存在する。しかし、テニスンをはじめ、他の詩人や彼らの作品に関する先行研究では、定量的なアプローチが用いられることがあまり多くなく、精読に基づく定性的なアプローチが主として用いられてきた。本研究では、トピックモデルの中でも潜在ディリクレ配分トピックモデル (LDA) を用い、その結果に基づき、テニスンが作品内でどのように海と

内陸環境（山や川）のイメージを用い、テーマを描いているかを考察する。これまでの研究では、テニスンの作品における海と死の関連が指摘されていたが、LDAの結果、海は死だけでなく、広い海のイメージから、絶望や永遠性（死後の世界）の印象を伝える傾向があることが示唆された。また、内陸の風景である山は宗教的な意味を、川は人生を比喩的に表すために使われており、海に見られるような絶望や永遠といったイメージは含まれていないと解釈される。本研究によって得られた結果は、例えば海と死の関連など、必ずしも目新しいものではない。しかし、定量的手法と質的手法を融合することで、これまでの研究で指摘されてきたテニスンの作品におけるテーマや情景の描写が、特定あるいは单一の作品のみではなく、複数の作品に跨がっていることを明らかにしたほか、海や川といった要素が特定の概念だけではなく、様々な感情を含意、増幅する効果があ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、遠読、LDA、詩、トピックモデル

## 1. Introduction

This paper is a part of exploratory analysis and study of the verse texts of the nineteenth-century Victorian poet, Alfred, Lord Tennyson (1809–1892). The analysis of this study employs an approach of stylometry, also called computational stylistics, ‘is concerned with the quantitative study of writing style’ (Eder et al., 2016: 107). Numerous studies have examined Tennyson’s works by focusing on similarities and differences between Tennyson’s and other poets’ styles, syntax, or lexica. However, few studies have employed quantitative methods, whereas most studies take a qualitative approach to specific poems based on close reading.

Moretti (2013) indicates that the problem with qualitative literary studies is that they inevitably only rely on a few works or expressions of a writer. Furthermore, they only focus on a limited portion of a particular work by a writer, thus omitting the remaining large portion of the work. Moreover, it is not necessarily true that an exhaustive analysis is possible even if one does take the time to do so. In fact, more than 130 years after the publication of Tennyson’s last work, there has still been no macroscopic or comprehensive study of Tennyson’s work as a whole or a comparison of his works with those of other authors using qualitative methods. The term ‘distant reading’ was referred to by Moretti and is nowadays often used in contrast to ‘close reading’. Distant reading enables us ‘to focus on units that are much smaller or much larger than the text: devices, themes and tropes’ (Moretti, 2013: 48–49).

This paper employs the latent Dirichlet allocation topic model (Blei et al., 2003; hereafter LDA) as a stylometric approach. LDA is a method specifically designed for analysing large text data. Whereas previous studies that have employed LDA have mainly focused on prose texts, this study attempts to apply LDA to poetry data and obtain efficient results and insights. LDA enables distant reading by identifying underlying semantic relationships between words using probabilistic calculations. This study not only demonstrates the usefulness of the method but also presents new features produced by the method, thus reinforcing previous discussions. This study expands on the previous study (Fujita, 2023; see details in Section 2.2), which presented comprehensive results

using the same methodology as this study, but further focused on specific topics and broke down the findings. The slight alteration of the data set from Fujita (2023) has been made; however, the altered data set does not negate or contradict the findings of both this study and Fujita (2023).

It has been noted by several scholars that Tennyson had a particular affinity for bodies of water, such as the sea and the lake. However, comparatively little attention has been given to rivers in Tennyson's oeuvre by scholars. Instead, it is the case that rivers in Tennysonians attracted readers, and 'The Lady of Shalott', for example, has a notable presence and has been depicted in paintings and other visual arts. It is evident that scholars and readers' notions, as well as the following and Fujita's (2023) LDA results, have identified and classified the sea and river as discrete categories (topics). This study will therefore focus on the portrayal of the sea and the inland, including rivers, in Tennyson's works.

In addition to the stylometric approach, this study also employs qualitative methods in its discussion. Given the emerging results of the quantitative method, LDA, this paper conducts a close reading of Tennyson's poems, along with previous studies. By combining distant and close readings of Tennyson's poems, this study sheds light on new aspects of Tennyson's works. Computer analyses can provide clues to draw human attention to unintentionally ignored aspects of poetry, however, it is the responsibility of individuals to read, interpret, and engage in discussions about poetry. Neither computers nor humans are perfect. The author of this study therefore expects that quantitative and qualitative methods can be complementary in order to extend our knowledge and to identify new and unperceived aspects of literature. This study also aims to demonstrate that stylometric approaches and 'distant reading' can expand the scope of analysis and provide a bird's-eye view.

## 2. Literature Review

### 2.1. Previous studies on Tennyson

There are abundant previous studies of Tennyson's life and his poetry, style and prosody. *Tennyson: A Memoir* (H. Tennyson, 1897) and *Alfred Tennyson* (C. Tennyson, 1949) are undoubtedly the closest source biographies to Tennyson, having been written by Alfred's son, Hallam, and his grandson, Charles. Brooke (1894) is nominated an earlier comprehensive study on Tennyson's poetry that was published only two years after Tennyson's death. In addition, Dixon's (1896) survey ranges from *Poems by Two Brothers* (1827) to *Ballads and Other Poems* (1880) and included a detailed biography of Tennyson.

Ricks (1969, 1987a, b, c) never allowed other authors to overtake him in Tennyson annotations. Indeed, previous researchers such as Shaw (1973) and Hair (1991), and later studies including Thomas (2019) refer to Ricks' both works. Ricks identifies references to other authors in Tennyson's poems, which Shaw (1973) confirms later. Shaw also counts the number of such references and expands his own discussion of Tennyson's style. Hair (1991) identifies qualitative

features of Tennyson's poetry, such as its vocabulary, concepts or motifs, by comparing them with those of other poets. In a later critical study, Thomas (2019) highlights 'echoes' of Wordsworth's poems in Tennyson's works.

H. Tennyson (1897: 20) notes that from his childhood, Alfred Tennyson 'had a passion for the sea, and especially of the North Sea in wild weather', as well as emotional reactions to water: "Somehow," [Tennyson] would say, "water is the element I love best of all the four" (H. Tennyson, 1897: 49). Further, Okazawa (1969) explains Tennyson's love for, enthusiasm for, and exploration of waters such as seas, lakes, and rivers. Readers can find both seas and lakes in more than 20 poems, including 'Morte d'Arthur', 'Edwin Morris, or The Lake', and 'The Ring' and rivers in over 40 such as *The Princess*, 'The Lady of Shalott', and 'The Brook'.

For years, Tennysonian scholars have found the sea to be a more intriguing motif than rivers as well as a common motif in his poems. Keirstead (2019) broadly explains how Tennyson evokes the sea, citing works such as 'Mariana in the South', 'OEnone', 'The Lady of Shalott', 'Locksley Hall', 'Ulysses', *In Memoriam A.H.H.*, *Enoch Arden* and the *Idylls of the King* series. He describes 'Simply put, going to sea means almost certain death in Tennyson' (Keirstead, 2019: 75). However, he also suggests that when a beach or seashore is depicted in a poem, Tennyson is also expressing or connoting aspects such as comfort or unstableness that readers can receive from beach/seashore images (Keirstead, 2019: 84).

Keirstead is the most recent scholar to identify the sea motif in Tennyson and analyses a broad swath of his work with a qualitative approach. Besides, Fulweiler (1965) also examines what sea portrays, connotes, and relates to in Tennyson's poetry; he discusses, for instance, 'The Merman', 'The Mermaid', *In Memoriam A.H.H.*, 'Ulysses', 'Crossing the Bar', 'The Holy Grail' and 'Morte d'Arthur'. Although the sea where the dying Arthur heads for in 'Morte d'Arthur' and its later version, 'The Passing of Arthur', is a huge lake, Tennyson describes it with the words *sea*, *ocean*, *water* and *lake* in the two poems. Thus, Keirstead (2019) and Fulweiler (1965) discuss the lake in 'Morte d'Arthur' and 'The Passing of Arthur' as an inland sea. These previous studies are qualitative, which is a fairly standard approach in poetry research.

Whereas many of these researchers take a qualitative approach to studying Tennyson's poems, Plamondon (2005) is one of the few who employ a quantitative approach; he investigates differences in the frequency of some vowels and consonants between Tennyson's and Robert Browning's poems.

One of Tennyson's characteristics is his long-form poetry. He wrote over 400 lyrical poems, with word tokens ranging from 12 to 300, but critics can never pass over Tennyson's narrative and epic poems, which are often longer than the lyrical poems. For example, his longest poem is *The Princess*, whose word-token count is 26,526. Furthermore, his most representative poem, *In Memoriam A.H.H.*, is a lyrical poem that is also his second-longest work with 18,724-word tokens. Although Tennyson has more short lyrical works than long or narrative poems, these eye-catching

long poems are never neglected and are repeatedly observed by critics. Even though his works contain shared elements beyond individual poems, qualitative researchers can sometimes miss these elements when they focus on only representative but small subsets of Tennyson's works.

## 2.2. Stylistic approaches and topic model

The qualitative approach is intimately associated with the practice of close reading. This method is considered indispensable and consequential in literary research. Nevertheless, Moretti (2013: 48) states that 'the trouble with close reading [...] is that it necessarily depends on an extremely small canon'. He further explains that given that close reading focuses on a limited portion of the work, it leaves 'great unread' aspects outside of 'an extremely small canon'. Here, why he employs the phrase 'extremely small' is because the quantitative can handle a relatively larger amount of data at once while the qualitative finds it challenging to handle not only a small canon but non-canonical works as well.

'Distant reading' utilises quantitative approaches and can be effective in exploring beyond the canon, as Moretti suggests (2013: 48). However, neither I nor Moretti (2013) intend to criticise research based on close reading or the approach itself. As few researchers use statistical and stylistic approaches, this paper aims to identify the 'great unread' in Tennyson's poetry employing a quantitative method called LDA (Blei et al., 2003).

The basic idea of LDA is that 'documents are represented as random mixtures about potential topics, where each topic is characterised by a distribution over words' (Blei et al., 2003: 996). Briefly, LDA classifies words into groups called topics. The algorithm and emerging results reveal 'the semantic structure hidden in a corpus of text' (Tabata, 2018: 52). Additionally, LDA is a specialised method for analysing big text data (Iwata, 2015).

Several researchers have employed topic models to examine prose texts (e.g., Onodera et al., 2016; Kuroda, 2017; Tabata, 2017, 2018, 2020; Kiyama, 2018; Matsukawa et al., 2018; Huang, 2020a, 2020b). However, its applications for poetry are fewer than those for prose text studies (e.g., Rhody, 2012; Navarro-Colorado, 2018; Henrichs, 2019; Okabe, 2019). Although this can be seen as a consequence of difficulty in applying LDA to poetry, Fujita (2022) has shown that LDA is effective when examining Tennyson's poetical works. While Fujita (2022) focuses on the methodological conditions of applying LDA to Tennisonian works, Fujita (2023) claims the general results of exploratory analysis by LDA on Tennyson's poetry. Fujita's (2023) findings include the different depiction of sea and river as well as the association between the environmental materials and the emotion of characters, and the findings are closely related to this article, but it is more general suggestions that what LDA detected than that giving deep discussion on the themes. Thus, this study aims to gain more specific discussion on the specific themes that LDA found including more number of quotations and mentions on the poems themselves. With this earlier study as an empirical ground for justifying the use of LDA, I analyse 593 works of Tennyson, focusing particularly on the nouns he uses in this paper. The next section will show more detailed information about the data set of 593 poems, and about the methodology for the analysis.

### 3. Methods and materials

#### 3.1. Data sources and configuration

This study is an analysis of 593 poetic works written by Tennyson. Of the 593 poems, 423 were compiled from the *Delphi Poets Series* ‘Alfred, Lord Tennyson’ (2013), and the rest, 170, were from Ricks (1987a, b, c). All the texts were converted to plain text data by optical character read/recognition, and then it was manually proofread and edited for errors such as unnecessary spaces, tabs, newlines and garbled characters. Table 1 shows the statistical data of the Tennyson corpus, 593 digitalised poems.

**Table 1: Statistical data of Tennyson corpus**

The number of poems	593
Total tokens	351,063
The shortest poem in a number of words	12
The longest poem in a number of words	26,383
Mean tokens per poem	592.01
Standard deviation	1825.61

The texts of the works were given part-of-speech (PoS) tags using CasualConc 2.0.8 (Imao, 2022)<sup>1</sup>. The tagset is CLAWS5, the same tagset as used in the British National Corpus. To analyse components, which heavily reflect themes, topics and subjects of poetry, this analysis adopted only nouns that were suggested by the PoS tags as ‘\_NN1’ (singular nouns), ‘\_NN2’ (plural nouns) and ‘\_NN0’ (mass nouns). All other words were deleted before the LDA application after the works were separated into consecutive segments (See Section 3.2 for more details about segments).

#### 3.2. Modelling and data analysis methods

LDA calculates the probabilities of word frequencies and classifies words into groups, called topics, on the basis of the hypothesis that documents should be classified into several groups based on their tendency to co-occur (within the same document). Here, documents are defined as contiguous segments of textual data, and they are not necessarily equivalent to poetical or prose work units. Fig. 1 shows the sample results of running LDA (Blei et al., 2003: 1009). LDA analysed the documents in the lower part of the figure and classified words into topics. LDA outputs topics and other numerical results such as word-weight values for each word, alpha ( $\alpha$ ) values for topics and the density of topics in work, but did not produce the labels, ‘Arts’, ‘Budgets’, ‘Children’ and ‘Education’ in Fig. 1. The analysts of the study interpreted LDA results by combining their knowledge of the text data and assigning labels to each topic.

LDA employs a bag of words model, which ‘treats each document as a bag of words’ and ignores the order of words in documents (Jockers, 2014: 137). Regarding his own novel text

<sup>1</sup> Notwithstanding the perpetual updating both of CasualConc itself and its manual, it should be noted that this article is referring to the version of the program that was available at the time this study was conducted.

“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

**Fig. 1: Example text from the AP Corpus (cited by Blei et al., 2003: 1009).**

analysis, Jockers (2014) explains that LDA captures ‘some themes that run throughout and others that appear at specific points and then disappear in novels’, and ‘it is useful to divide novels (and other large documents) into chunks or segments and then run the model’. Further, LDA requires a significant number of documents to make its probability calculation sufficiently accurate.

In previous LDA studies on prose texts, works were split into segments of equal sizes, such as 1,000 (Kuroda, 2017; Tabata, 2020), 2,000, or more words (Huang, 2020b). On the other hand, poetry texts in studies employing LDA have not been sliced into segments; rather, scholars have considered poems themselves as segments. Navarro-Colorado (2018) and Henrichs (2019) apply LDA only to sonnets, whose tokens range from about 100 to 200. Meanwhile, Tennyson’s poems vary in length, as the standard deviations in Table 1 show. The large deviation of word tokens interferes with LDA results because a large population difference can affect the reliability of LDA calculations and results.

Further, the large disparities in segment size, instantly exert an influence on the results because LDA uses raw word frequency to identify the co-occurrence of words. Therefore, it is advisable to divide Tennyson’s poems into segments of possible equal size following the method of prose text studies considering the large deviation of tokens per poem. Jockers (2014: 137) additionally mentions the size of segments: ‘the bigger the bag, the more words that will tend to be found together in the same bag’. Nonetheless, careful segment sizing is necessary because in the reverse

of Jocker’s words, the smaller the bag, the fewer the words that will tend to be found together in the same bag.

Segment size should be set up for appropriate chunks that contain coherent topics in poems. The results will not fruitfully reflect the essences of the works when the poems are divided into too small segment size or concatenated individual poems. Therefore, segment size should be set to minimise differences between segments. Regarding these concerns, Fujita (2022) suggests the relevant segment size for LDA on Tennyson’s poems. This paper refers to Fujita and adopts a segment size of 592 words based on the mean number of tokens in the Tennyson corpus shown in Table 1.

After that poems were split into 592-word segment size, all words besides nouns (see Section 3.1) were deleted. LDA was applied to the segments using the Machine Learning for Language Toolkit (McCallum, 2002). The number of topics was set at 18 considering results of experimental trials, with the number of topics ranging from 10 to 200.

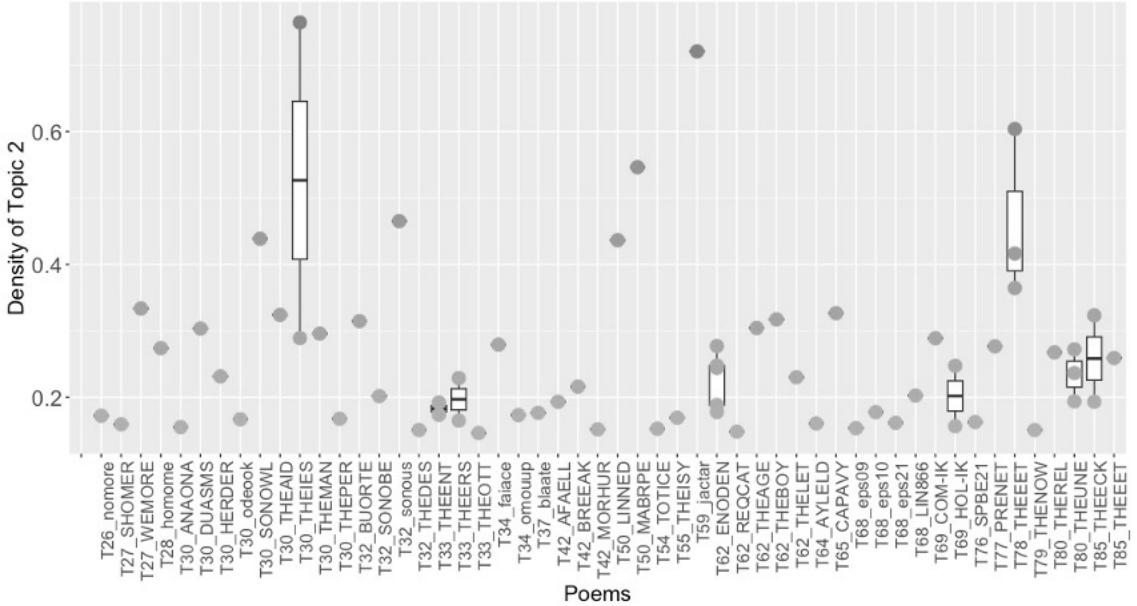
#### 4. Results

This section presents the results of LDA for the Tennyson corpus. Table 2 shows one set of output results, the  $\alpha$  value for the 18 topics, the 20 most salient keywords for each topic and the labels, which the author of this paper assigned manually. The keywords are ranked in descending order of their weights from the upper left to the bottom right of the table;  $\alpha$  value represents the universality of each topic. The higher the value, the more prominently the topic appears in multiple works; lower values indicate that a topic appears in fewer works, sometimes only one. Topic 1 in Table 2 had the highest  $\alpha$ , 0.8774, of the 18 topics, indicating its greater prominence than all the others; Topic 8 was the least prominent ( $\alpha = 0.0581$ ). Detailed analysis of the poems and the LDA output results suggested specific labelling for each topic. The labels show that both Topics 2 and 3 refer to natural environment. Specifically, Topic 2 pertains to seas and their surroundings, while Topic 3 pertains to natural ambience inland. The top 20 keywords for both topics include words related to bodies of water, seas, and rivers. Although previous studies have discussed seas in Tennyson’s work, rivers have been scarcely mentioned. The  $\alpha$  values of Topics 2 and 3 are similar, indicating that both topics are present in the corpus to a comparable extent.

**Table 2: Output results of LDA: 18 topics with their labels, alpha values and keywords**

Topics	Alpha Values	Labels	Keywords (1–20)
0	0.2177	Nobles	king queen knight knights hall man maid name field court face table day quest brother men realm hands sword prince
1	0.8774	Life (materials)	light sun day eyes heaven time voice earth shadow wind moon sea cloud sound stars hills land morn glory hill
2	0.1345	Environment (sea)	sea isle shore day boat sail seas ship wave side ocean home bells sand tide blast sails vessel storm beach
3	0.1266	Environment (in-land)	mountain river valley melody music city palm bridge pine mould tree prime water lake garden gleam fountain torrent song level
4	0.6847	Male	man life world things time men power years truth mind nature days voice friend heart peace earth change faith hands
5	0.2289	Natural scenery	year rose flower flowers summer garden birds roses air brook leaves feet tree spring leaf morning woodland day blossom wood
6	0.6871	Life (emotion)	love heart eyes life death tears hand soul face name lips beauty light cheek hope youth side dream blood smile
7	0.0592	Ladies	lady princess ringlet woman prince honour college goose gown head girls girl sister highness women letters echo web mirror books
8	0.0581	Discretions in life with seasons and nature	fool mark wine men leaf tongue cup swine ways hair innocence moment rhymes thing port lodge lies pity summers cups
9	0.0823	Attractiveness	eye air sense charms bloom face maze harmony lustre language depth horizon dye course ecstasy spell mysteries flash claim shrine
10	0.2844	Man and death	death man fire war god blood men hell gods way heaven hand earth battle head flesh foe flame king shame
11	0.0903	Woman and whereabouts	woman dream gold boon child men charm faith eye music belt books ruin town wrongs mood cave wife heaven street
12	0.0920	Knighthood	horse hall prince arms wood damsel knight shield lady knave pride walls horses squire bridge women armour charger lance town
13	0.3860	Family	mother child father heart man hand wife face day house boy men head bride woman word children will home daughter
14	0.1400	Battle and belief	land name men fame people war freedom hearts cause crowd sons friends kings throne queen shame peace health tyrant greatness
15	0.0505	Cheering finiteness	ring laughter t'amo ward finger bone harp bells month spring lisette cousin souls babble noise smile chest spirits landscape marge
16	0.0370	Religions and life outside the UK	children bread people chains church enemy moor heresy women colony chorus roof east banner temple goddess sepulchre island friend berries
17	0.1446	Power of nature	pride earth woe throne form storm eye course thunders ray lightning thunder bones riflemen wreath misery storms realms wing bow

Figures 2 and 3 show the densities, which are output results of LDA for Topics 2 and 3, respectively<sup>2</sup> <sup>3</sup>. The density values indicate the proportion of each topic within a segment. Therefore, it could be suggested that the lower the proportion of the topic, the more prominent the presence of other topics in these segments. For the next discussion, segments were filtered out the top 65 segments to reduce the count from 1,041 segments in 593 poems. While most of the poems listed in Fig. 2 for Topic 2 are lyrical works, several poems are narrative poems, such as T33\_THEOTT ('The Lady of Shalott', 1833), T62\_ENODEN (*Enoch Arden*, 1862), and T69\_COM-IK ('The Coming of Arthur', 1869). The number of narrative poems in Topic 3 is relatively lower. In addition, the proportion of Topic 3 in narrative poems, for instance T47\_PRIESS (*The Princess*, 1947) and T42\_AUDURT ('Audley Court', 1842), is also lower compared to other lyrical poems. The years of publication and writing of the poems in Topic 3 are scattered, but approximately half of the poems were published or written in the 1820s and 1830s.



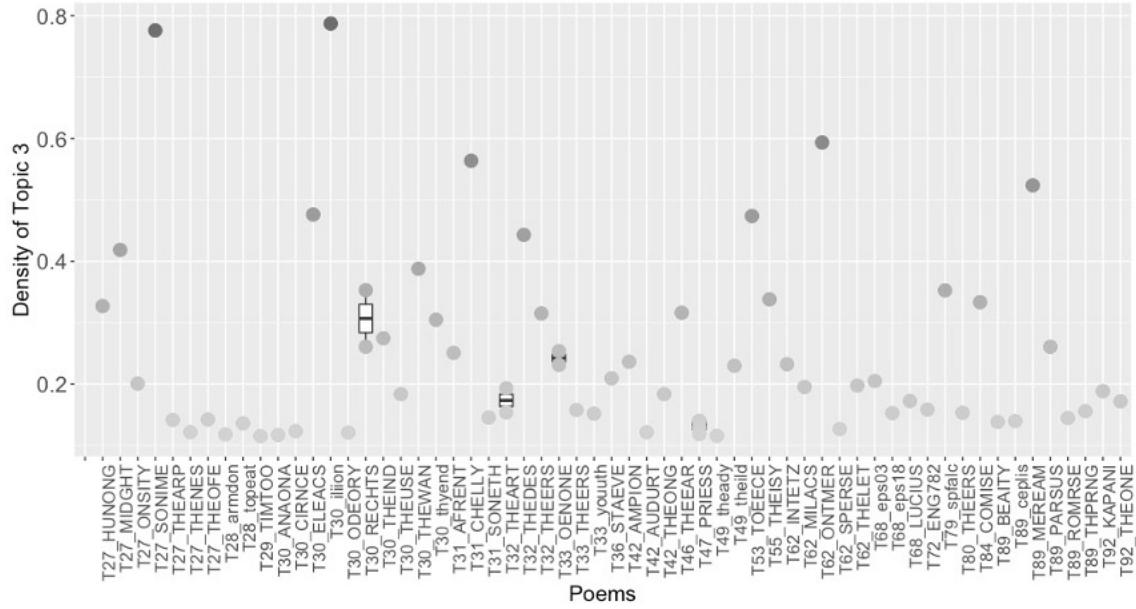
**Fig. 2: Density plot of Topic 2.**

Fig. 4 shows the mean density values of the top 65 segments for Topics 2 and 3. In cases where poems are duplicated in the top 65 segments, the mean density values were calculated to plot Fig. 4. It is worth noting that the density value for each poem is inversely proportional to the other. Specifically, as the density of Topic 2 increases, the density of Topic 3 decreases, or vice versa, in the poems for each topic. While most of the works tend to be allocated to only either Topic 2 or 3, there are some exceptions, such as T27\_THEARP, T32\_MINRIT, and T68\_eps22.

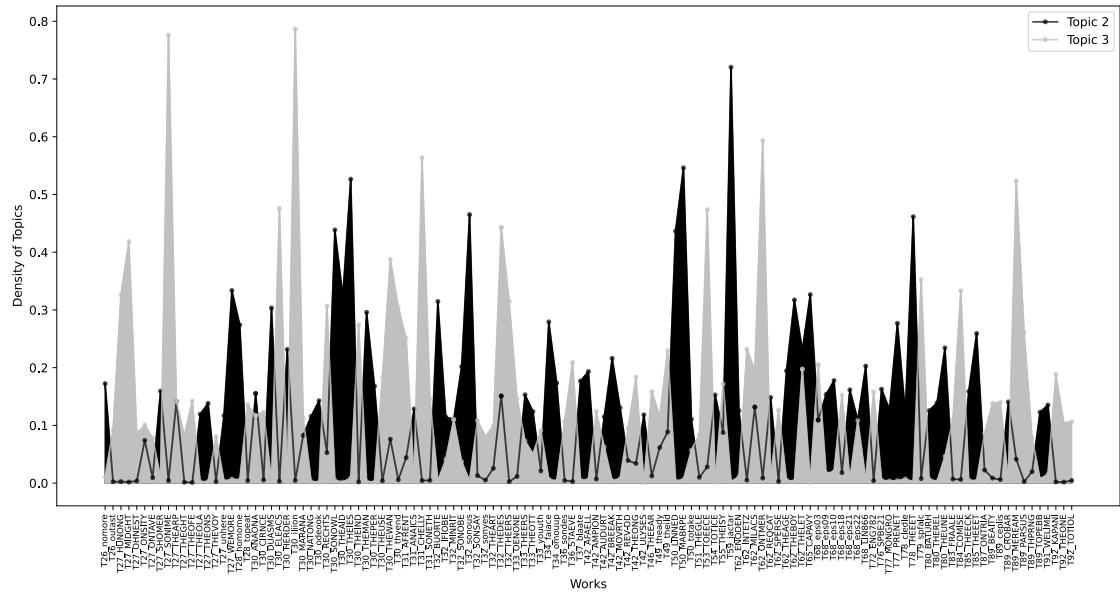
<sup>2</sup> The abbreviated titles shown in lowercase represent that the poems are not published, whereas uppercase titles indicate the opposite. A 'T' at the beginning of the title indicates the poem is written by Alfred Tennyson, and the following two-digit numbers indicate the last two digits of the year of publication (or, if unpublished, of its composition).

<sup>3</sup> For the comparison tables of the abbreviated and full titles of the poems under Topics 2 and 3, see the Appendices.

Based on the density values, it can be inferred that Tennysonian poems tend to depict sea and inland environments separately rather than concurrently and to the same extent in one work. The following section will elaborate on the content and details of Topics 2 and 3 and investigate the association between topics and the content of poems.



**Fig. 3: Density plot of Topic 3.**



**Fig. 4: Comparison of the mean value of density for Topics 2 and 3.**

## 5. Discussion

Although the LDA analysis covers 18 topics that explain important and characteristic features of Tennyson's poems, this paper and subsequent discussion focus solely on his descriptions of the natural environment in his poetry. As previous studies have pointed out, Tennyson had predilection for describing and exploring bodies of water in nature. On the subject of bodies of water, Topic 2 specialises in the sea and Topic 3 has keywords related to water, *river* and *lake* (see Fig. 5 for keywords of Topic 2 and Fig. 6 for Topic 3). Furthermore, detailed descriptions of bodies of water are depicted in various natural settings. In Topic 2, geographic features, such as *isle*, *shore*, *sand*, and *beach*, closely related to the sea, exquisitely represent landscapes. On the other hand, in Topic 3, plants and sounds, for instance, *pine*, *tree*, and *melody*, are ranked in the top 20 keywords. With these detailed elements, Tennyson succeeded in describing and conveying highly detailed scenes using words to illustrate the elements as well as the characters' sentiments around the sea and river that appear in Topics 2 and 3.

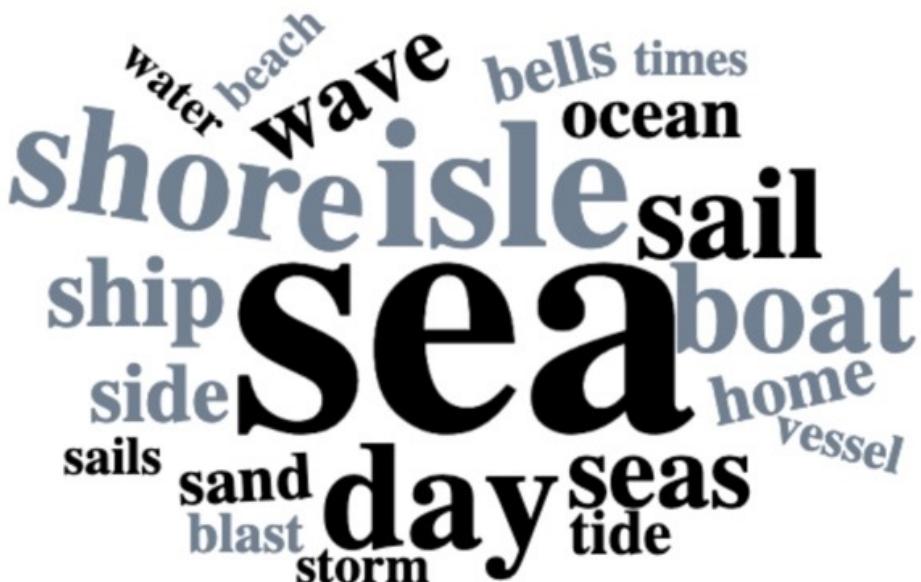


Fig. 5: Word cloud of the top 20 keywords (Topic 2).

### 5.1. Sea

As Keirstead (2019: 75) notes, 'Simply put, going to sea means almost certain death in Tennyson'. Death, dying people, and dying animals are mentioned directly, indirectly, and figuratively in 14 of Topic 2's 65 prominent segments. Specifically, the word *sea* is associated with the concept of death in eight poems. However, in the assigned poems of Topic 2, the sea is not necessarily synonymous with death. The investigation of this topic reveals that the sea is linked to several concepts present in the assigned poems, including death. In 'The Sailor Boy' (1862; T62\_THEBOY),

‘**danger** in the roaring **sea**’ (boldface added<sup>4</sup>) emphasises the dreadfulness of ‘a **devil**’ that ‘rises in my heart’, saying that this devil ‘in my heart’ is ‘[f]ar worse than any death to me’ (ll. 21–24). The final line of ‘Prefatory Sonnet’ (1877: l. 14; T77\_PRENET) uses a metaphor to convey a sense of depth and gravity: ‘In **seas** of Death and sunless gulfs of Doubt’. Though the word *sea* is not used in ‘Sonnet [But were I Loved as I Desire to be]’ (1832; T32 SONOBE), readers can easily connect the *surge* and *foam* to the sea in the following lines:

But were I loved, as I desire to be,  
What is there in the great sphere of the earth,  
And range of evil between death and birth,  
That I should fear if I were loved by thee?  
...

Apart upon a mountain, though the **surge**  
Of some new deluge from a thousand hills  
Flung **leagues** of roaring **foam** into the **gorge**  
Below us, as far on as eye could see.

“Sonnet [But were I Loved as I Desire to be]” (1832: ll. 1–4, 11–14; T32 SONOBE)  
(Boldface and underlines added)

The final line of the abovementioned poem, ‘as far on as eye could see’ (l. 14) implies a gap between *us* and the surface of the seawater. The theme of distance between objects is a recurring motif in poems of Topic 2.

Related to distance and watercraft, both *boat* and *ship* also appear. These seagoing vehicles are associated with regions and areas that are beyond the sea and with the action of travelling to those far places. Because Britain is an island country, there can be no doubt that Tennyson associated the sea with going abroad or establishing relationships with other regions. Ships take people abroad as well as to sail into battle over the seas.

In ‘Lines [Here Often, When a Child, I Lay Reclined]’ (1850; T50\_LINNED), a distant region (for Tennyson), Greece, is referred to in line 4 by the word *Grecian*, with the keyword of Topic 2, *ships*. Not only the description of Greece but also solitude, lamentation, and the distance of time are depicted in this poem.

In wartime in particular, the seas around the British Isles are the last line of defence for Great Britain. The most significant poem of Topic 2, the unpublished work ‘Jack Tar’ (1859 circa; T59\_jactar) describes ‘the noble pride of the Mistress of the Seas’ (l. 4) in a battle, and the phrase ‘the **pride** of the **Mistress** of the Seas’ repeatedly appear in lines 4, 7, 11, 14, 18, and 21. ‘Jack Tar’ is considered to be written while ‘anticipating war with France’ (Ricks, 1987b: 604). In addition, the term ‘the Mistress of the Seas’ is the sobriquet of Great Britain; thus, it forms a further compelling connection between war/battle and the sea.

Battle is the most significant scene in ‘The Revenge: A Ballad of the Fleet’ (1878; T78\_THEEET) and in ‘Buonoparte’ (1832; T32\_BUORTE), which is regarding the Battle of Trafalgar. In these

<sup>4</sup> Henceforth, boldface styling of terms in excerpts and quotations from poems will indicate the word to which the topic is assigned.

two poems, the sea is depicted as a site of battle. As with the images of shipwreck in *Enoch Arden* (1862; T62\_ENODEN) and battles, the sea is associated with an inescapable, unwilling, and lonely death<sup>5</sup>. In scenes where the sea appears, the sound of its waves is almost the only sound to be heard. The openness of the sea and the absence of any natural or artificial surroundings are also signified. This nothingness around the characters in the scenes near the sea further amplifies the feelings of loneliness, despair and fear, especially when the characters in poems lose their loved ones.

Some works, including ‘Mablethorpe’ (1850; T50\_MABRPE) and ‘To Professor Jebb’ (1889; T89\_TOPEBB)<sup>6</sup> describe foreign countries or regions without mentioning death. Not only are distant nations such as Egypt (‘To Professor Jebb’) conveyed in these poems but also their aloofness appears, specifically in the descriptions of archaic locations such as *Ilion*, recalling ancient Greece and Greek mythology.

Other poems under Topic 2 suggest mental, rather than physical, remoteness. ‘We Meet No More’ (1827; T27\_WEMORE, quotation below) is a poem that both mental and physical remoteness depicted by referring to the parting of *I* who is standing ‘on a distant shore’, and *Ellen* (ll. 7, 12), and what separates them is ‘angry seas’ (l. 6). Although death is not explicitly represented in ‘We Meet No More’, the *[f]arewell* in the final line, in conjunction with the term ‘no more’, conveys a sense of mental remoteness, loneliness, and an enduring sense of eternity. The phrase ‘no more’ is used repeatedly in this poem, yet it does not directly refer to eternity or nothingness. Instead, it suggests that a particular event or state has reached its conclusion at a specific juncture and is not capable of reoccurrence in the future.

According to Huddleston and Pullum (2002: 831), the function of ‘no more’ and ‘no longer’ is that of absolute negators, relating to the concepts of stillness and to ‘any more/any longer’. Whilst *still* is regarded as a positively-oriented polarity-sensitive item (PPI), preferring positive contexts, ‘any more/any longer’ is characterised as a negatively-oriented polarity-sensitive item (NPI), preferring negative contexts. Huddleston and Pullum posit the view that both ‘any more’ and ‘any longer’ are aspectual NPIs. Despite the absence of overt expressions of time continuity in the poem, it does exhibit a consistent time flow that can be inferred from specific lines such as ‘[o]ur every hope on earth is past’ (l. 3), ‘[a]nd lonely thoughts of days gone o'er’ (l. 9), and ‘[f]arewell for ever, Ellen!’ (l. 12). The poem’s repetitive depiction of events serves to underscore the notion of time’s progression and its constancy. The absolute negators ‘no more’ and ‘no longer’ in the correspondence relationship of the aspectual NPI imply the negation of the continuity of time. Consequently, events or actions that transpired during the period of continuous time will cease to persist once the consistency of time is negated by the use of these words. It

<sup>5</sup> The eponymous protagonist Enoch of *Enoch Arden* does not die in the shipwreck during his journey back home from the Far East but ends up dying in sorrow, deep love for his wife and family members, a profound sense of self-sacrifice and despair at the end of the story.

<sup>6</sup> This poem is not ranked in the top 65 segments of Topic 2, but it is 86th and its density of Topic 2 terms is 0.1228.

can thus be interpreted that the phrase ‘we meet no more’, which appears in lines 5 and 11, is as ‘we will never see each other again’ or ‘we will never meet again’. The interpretations suggest that the protagonists had had opportunities to see each other up until a certain moment. However, it is implied that the occurrence of a future meeting is rendered implausible. This interpretation suggests a connotation of everlastingness in the phrase ‘no more’.

WE meet no more — the die is cast,  
The **chain** is broke that tied us,  
Our every hope on earth is past,  
And there’s no helm to guide us:  
We meet no more — the roaring **blast**  
And angry **seas** divide us!  
And I stand on a distant **shore**,  
The **breakers** round me swelling;  
And lonely thoughts of days gone o’er  
Have made this breast their dwelling:  
We meet no more — We meet no more:  
**Farewell** for ever, Ellen!

‘We Meet No More’ (1827; T27\_WEMORE)

The poem entitled ‘The First Quarrel’ (1880; T80\_THEREL) also describes the separation of a man and a woman, and thier fates are sealed when the man dies in the sea, as alluded to in the final line of the poem: ‘An’ the **boat** went down that night—the **boat** went down that night’ (l. 92). The narrative of ‘The First Quarrel’ (1880; T80\_THEREL) is presented through the perspective of the *I*-narrator, who remains anonymous apart from being referred to as Harry’s wife. The narrative commences with the female narrator giving birth to a boy in challenging conditions. The narrator provides a detailed account of her circumstances to her physician, reflecting on her life in the context of her bereavement. Her husband, Harry, met his demise at sea. The narrator has been experiencing a state of unhappiness, precipitated by the demise of her spouse and her new born child. Nevertheless, the aspect that appears to be the most challenging for her is the fact that she lost her husband prior to having had the opportunity to reconcile with him following the very first and the last quarrel in their relationship. Their quarrel began when she suspected her husband of having an affair with another woman. Harry denied it, but the wife did not believe him. Before they become reconciled, Harry had to leave from the narrator’s presence for work. He had to travel to a location across the sea by boat. Following his departure, the wife realised that she was wrong, but it was too late, as the boat that Harry was on sank in the ocean, and she lost her husband and the opportunity to apologise to him at the same time. In ‘The First Quarrel’, therefore, the mental and physical distances, death, sea, loneliness and perpetuity, all of which are elements of Topic 2, are conveyed in a single poem. While ‘The First Quarrel’ serves to illustrate Keirstead’s (2019: 75) proposition, the poem also demonstrates that the sea in Tennyson is not exclusively associated with death; it is also linked to other elements, including mental and physical distances, loneliness, and eternity.

## 5.2. Mountain and River

Topic 2 associates the sea with distance, eternity, death, estrangement, foreign regions, and the past. In contrast, Topic 3 explores religious motifs and their impact on people's lives, as indicated by the keywords (Fig. 6). The top 20 keywords of Topic 3 associate with creeds as well as a focus on a moment or life before death but not everlasting or uncertain time and the afterlife. In terms of religious motifs, the most prominent keyword is *mountain* in Topic 3, which is more conspicuous than river. The religious mythologising of mountain-related terms in Tennyson, such as *pinnacles* in 'The Lotos-Eaters' (1833: 1. 16) and *mountain* in 'Youth' (1833 circa), are referred to by Paden (1971: 157) and Ricks (1987a: 637).



**Fig. 6: Word cloud of the top 20 keywords (Topic 3).**

In a poem, 'To E. L. on his Travels in Greece' (1853; T53\_TOEECE), in which Topic 3 densely appears, the *Gods* and a *mountain* are not directly linked. However, a sense of holiness or, more likely, a sense of solemnity from the surrounding items is expressed through Topic 3's keywords.

Despite the absence of *god(s)* in ‘Merlin and the Gleam’ (1889; T89\_MEREAM), the *mountain* are filled with *Wizard*, *Griffin*, *Giant*, *Fairies* and *dragons*, and other supernatural beings, giving the scene a surreal atmosphere. The definition of *Gleam*, a keyword in Topic 3 that frequently occurs in this work, is given by Tennyson himself as referring to ‘poetic imagination’ (Ricks, 1987c: 205). Neither *Gleam* nor ‘poetic imagination’ have any theological overtones or connections to the Almighty. However, *Gleam*, which appears on a mountain far from the real, human world, colours the otherworldly imaginary world of the work. In addition, the protagonist of this poem, Merlin ‘follow[s] The **Gleam**’ (l.10). A couple of stanzas later, ‘The Master whisper’d’ to Merlin, ‘Follow The **Gleam**’ (ll. 33–34). The presentiment of Merlin’s death, as well as ‘The Master’, who suddenly appears in the poem and murmurs to Merlin, further adjoin the sacredness for the

word *Gleam*. In addition to their presence on the mountain, Tennyson depicted *gods* seated on the clouds (floating) on the heights. These deities and their elevated positions also bear connections to religious concepts. Another religious term, *Heaven*, also connotes altitude, for instance in ‘Thy Soul is Like a Landskip, Friend’ (1830; T30\_thyend).

As the keywords in Fig. 6 indicate, *melody* and sound are widely spread in the poems of Topic 3. Within the most dominant works of Topic 3, ‘Ilion, Ilion’ (1830; T30\_ilion), ‘Elegiacs’ (1830; T30\_ELEACS), ‘The Dying Swan’ (1830; T30\_THEWAN) and others, are filled with melodies, sounds, and songs. In the 19 lines of ‘Ilion, Ilion,’ the word *melody* appears five times: ‘when wilt thou be **melody** born?’ (ll. 2, 9, 12, 19) and ‘ever onward to a **melody**’ (l. 5). Moreover, the first line of ‘Elegiacs’, ‘Lowflowing breezes are roaming’ leads to a passage in which living creatures make sounds: ‘Barketh the **shepherd-dog** cheerily; the grasshopper carolleth clearly; / Deeply the **turtle** coos; shrilly the **owlet** halloos’ (ll. 5–6). One theme of ‘The Dying Swan’ is undoubtedly death, but this poem is unlike the poems of Topic 2 with their scattering songs and sounds:

With an inner voice the **river** ran,  
Adown it floated a dying **swan**,  
And loudly did lament.  
...  
One willow over the **water** wept,  
And shook the **wave** as the wind did sigh;  
...  
The wild **swan**’s **death-hymn** took the soul  
Of that waste place with joy  
Hidden in sorrow: at first to the ear  
The **warble** was low, and full and clear;  
...  
With a music strange and manifold,  
Flow’d forth on a carol free and bold;  
As when a mighty people rejoice  
With **shawms**, and with **cymbals**, and harps of gold,  
And the tumult of their **acclaim** is roll’d  
...  
And the creeping mosses and clambering weeds,  
And the willow-branches hoar and dank,  
And the wavy **swell** of the southing reeds,  
And the wave-worn horns of the echoing **bank**,  
And the silvery **marish-flowers** that throng  
The desolate **creeks** and **pools** among,  
Were flooded over with eddying **song**.

‘The Dying Swan’ (1830: ll. 5–7, 14–15, 21–24, 29–33, 36–42; T30\_THEWAN)

As can be seen, the anthropomorphic plants and animals talk, sing, and surround the people in the poems of Topic 3. Unlike the cases of Topic 2, living creatures are present in the poems and make various sounds. These works do not convey any impression of fear or despair of the type

felt in the works of Topic 2. Furthermore, both in ‘The Dying Swan’ and in the aforementioned ‘Merlin and the Gleam’, the final moments of the characters’ deaths are not described. These dying animate beings are surrounded by sounds and other living beings. These situations do not necessarily convey utter solitariness or despair to readers. This therefore suggests not only the reflection of the characters’ emotions but also Tennyson’s intention to sketch differently in relation to the natural surroundings, evoking these emotions not directly but in an indirect way.

Not only incidental sounds but also those of people, animals, and apostrophised nature, including plants and water, crowd the poems of Topic 3; this contrasts with Topic 2 poems whose scenes highlight the sea. Even the difference in altitude alluded to by mountains and valleys influences how people perceive their surroundings. In some cases, the objects under discussion are located distant in the physical realm of poems; however, in the case of Topic 3, these objects are rendered visible through their depiction in poems. This visibility, whether clear or vague, can be seen by the narrators, the characters, and the readers, even if they are distant. Such visibility does not convey the sense of despair felt in the poems of Topic 2. This suggests that the objects not only mirror the characters’ emotions but also that Tennyson’s intention was to sketch and evoke these emotions indirectly rather than directly.

In Topic 3’s poems, Tennyson employs rivers as a metaphor for life, for example in ‘Compromise’ (1884; T84\_COMISE):

Steersman, be not precipitate in thine act  
Of **steering**, for the **river** here, my friend,  
Parts in two channels, moving to one end—  
This goes straight forward to the **cataract**:  
That streams about the **bend**;  
But though the **cataract** seem the nearer way,  
Whate’er the crowd on either **bank** may say,  
Take thou the ‘**bend**,’ ’twill save thee many a day.

‘Compromise’ (1884; T84\_COMISE)

As can be seen in the quotation of ‘Compromise’ above, rivers wind through bends and cataracts until they finally reach a broad, soundless, open sea. This suggests that lives with ups and downs ultimately come to an end. Poems in Topic 3 do not extend rivers to the ocean, focusing instead on the present moment rather than the eternity or death of the sea, through bends. Religious connotations pertaining to death, such as the term *Heaven*, are not necessarily taken as an indication of a character’s death within a poem or of characters’ eternal life after death. Rather, the concept associates with a belief that one may hold while still alive. Hence, the poems in Topic 3 centre on the moments when the characters are alive. The belief and the focus on the living moment serve as the pivot in the poems of Topic 3.

### 5.3. Did LDA overlook a canonical poem?

Discussion of Tennyson's concept of the river cannot overlook several works, for instance, 'The Lady of Shalott'. The scene in which the Lady is on a boat floating into Camelot is illustrative. The river in this poem is one of the impressive elements in many paintings of the scene, such as those by J. W. Waterhouse and Henry Peach Robinson (Cheshire, 2009: 57, 90–91). However, this poem does not appear in the discussion of the river in this section or in the top 65 significant segments of Topic 3. Did LDA miss a prominent and representative theme in the Tennysonian canon, while it was handling and analysing large amounts of poetry, including the 'great unread' dimensions?

'The Lady of Shalott', belongs in fact to Topic 7 (the eighth most significant poem/segment) rather than to Topic 3. The reason why this poem was assigned to the topic is that it emphasises the extent of the Lady as well as the Lady's despair and her remembrance, even following her death. Here again, the river metaphorically suggests the Lady's life. In the poem, the river does not reach a vast, silent sea after all of its bends, windings and cataracts through Camelot; nevertheless, rivers generally reach the sea finally as vicissitude-filled lives inevitably come to an end. The river in 'The Lady of Shalott' unquestionably attracts readers' attention; however, what LDA detected here was not the Lady's life rendered as the river but her melancholy, forlornness and despair.

Tennyson, as a 'landscape-painter in words, a colourist' (Hair, 1991), lays out his multiple themes in detailed word portraits of nature. Although some LDA results contradict earlier scholarly readings and readers' intuitions, as shown in this section, LDA unfolded and spotlighted latent semantic connections between words spread over multiple poems that are hidden behind canonical or eye-catching themes or motifs. However, this does not amount to a cancellation by LDA of the outcomes of the close readings found in previous studies.

## 6. Conclusion

In this study, I observed poetic depictions of sea and inland landscapes, mountains and rivers based on the results of the LDA topic model. What were mentioned and discovered in this study were not necessarily new and groundbreaking findings; nevertheless, they revealed that quantitative and qualitative approaches combined can support each other and help to discover hitherto unknown dimensions of Tennyson. Of course, I was not insisting that poems should have only one topic. The topics in LDA are interrelated; in particular, in the works with two topics, they are often complexly intertwined rather than occurring independently within each work; this is not a problem but rather a testimonial to how complex Tennyson's depictions are in his works and how accurately the topic model captures this complexity.

In addition to the interpretations of the sea in Section 5.1, eternity can interweave with immortality, which is another prominent theme in Tennyson. Although LDA facilitated my reading of Tennyson's poems beyond his canon and revealed new dimensions, this paper does not provide a comprehensive analysis of all their elements. Future research should explore Tennyson's works

in greater depth, using both qualitative and quantitative approaches to investigate the motifs of immortality and others that emerged from the LDA results. By doing so, it will be possible to prevent leaving a gap of the ‘great unread’ in Tennyson’s works.

## Acknowledgements

This paper is partly based on the author’s presentation at PALA 2022 *Style and Sense(s)*, July 8, at Aix-Marseille Université, and the 118th conference of Japan Society of Stylistics, June 18, at Kyorin University. I would like to thank the chairs and audience members for useful discussion.

This paper is in part based on the author’s doctoral dissertation submitted to Osaka University in November 2023.

The research was in part supported by the Japan Science and Technology Agency Support for Pioneering Research Initiated by Next Generation [grant number JPMJSP2138], to which the author of the present article was granted funding from October 2021 to March 2023.

## Bibliography

- [1] Blei, M.D., Ng, Y.A. and Jordan, I.M. (2003) Latent dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022.
- [2] Brooke, A.S. (1894) *Tennyson: His Art and Relation to Modern Life*. London: Isbister.
- [3] Chesire, J. (Eds.) (2009) *Tennyson Transformed: Alfred Lord Tennyson and Visual Culture*. Burlington: Lund Humphries.
- [4] Dixon, M.W. (1896) *A Primer of Tennyson*. London: Methuen & Co.
- [5] Fujita, I. (2022) On segment size in poetry analysis: using the latent dirichlet allocation method. *The Japanese Journal of Digital Humanities*, 3(1): 3–15.
- [6] Fujita, I. (2023) ‘To spread the Word by which himself had thriven’: Analysis of Alfred Tennyson’s use of language based on the LDA topic model. *English Corpus Studies*, 30: 3–26.
- [7] Fulweiler, W.H. (1965) Tennyson and the “summons from the sea”. *Victorian Poetry*, 3(1): 25–44.
- [8] Hair, S.D. (1991) *Tennyson’s Language*. Toronto: University of Toronto Press.
- [9] Henrichs, A. (2019) Deforming Shakespeare’s sonnets: topic models as poems author(s). *Criticism*, 61(3): 387–412.
- [10] Huang, C. (2020a) Chugoku-no misuterii shousetsuniokeru topikku kaiseki-no kokoromi. (Experimental topic analysis on Chinese mystery prose texts). *Studies in Language and Culture*, 29 : 1–17.
- [11] Huang, C. (2020b) Chugoku-no misterii shousetsu-wo meguru keiryouteiki bunseki: Tei Shou Sei to Ki Ba Sei-no sakuhin-wo chuushinn-ni. (Quantitative analysis of Chinese mystery novels: focusing on the works of Cheng Xiao Qing and Gui Ma Xing). *Text mining and Digital Humanities*, 2020: 31–45.
- [12] Huddleston, R., & Pullum, G. K. (2002) *The Cambridge grammar of the English language*. Cambridge University Press.
- [13] Imao, Y. (2022) CasualConc, version 2.0.8. Available (newer version) at: <https://sites.google.com/site/casualconcj/casualconc/CasualConc>.
- [14] Iwata, T. (2015) *Topic Model*. Tokyo: Kodan-sha.
- [15] Jockers, L.M. (2014) *Text Analysis with R for Students of Literature*. Heidelberg, New York, Berlin, London: Springer Cham.
- [16] Keirstead, C.M. (2019) Sea dreams and realities: Tennyson, travel, and the shifting currents of littoral space in Victorian culture. *Victorian Poetry*, 57(1): 73–99.
- [17] Kiyama, N. (2018) How have political interests of U.S. Presidents changed?: a diachronic investigation of the State of the Union addresses through topic modeling. *English Corpus Studies*, 25: 79–99.
- [18] Kuroda, A. (2017) Quantitative analysis of literary works: novels of Sir Arthur Conan Doyle. *Text mining and Digital Humanities*, 2017: 23–41.

[19] Matsukawa, H., Oyama, M., Negishi, C., and Arai, Y. (2018) Analysis of the free descriptions obtained through course evaluation questionnaires using topic modeling. *Japan Journal of Educational Technology*, 41(3): 233–244.

[20] McCallum, A.K. (2002) Mallet: a machine learning for language toolkit. Available at: <http://mallet.cs.umass.edu>.

[21] Moretti, F. (2013) *Distant Reading*. London: Verso.

[22] Navarro-Colorado, B. (2018) On poetic topic modeling: extracting themes and motifs from a corpus of Spanish poetry. *Frontiers in Digital Humanities*, 5: 5–15.

[23] Okabe, M. (2019) *Thou and You* in Emily Dickinson's poems using topic modeling: reconsideration of interjections. *Proceedings of the Japanese Association for Digital Humanities Conference 2019*: 125–131.

[24] Okazawa, T. (1969) A study of Tennyson's predilection for waters. *Bulletin of Chofu Gakuen Women's Junior College*, 2: 1–12.

[25] Onodera, D., Huang, L. and Yoshioka, M. (2016) Classification of new article by using facet-biased topic model and distance metric learning. *The 30th Annual Conference of the Japanese Society for Artificial Intelligence 2016*: 1–4.

[26] *Oxford English Dictionary*. Available at: <https://www.oed.com> (Last accessed 22 October 2023).

[27] Paden, W.D. (1942) *Tennyson in Egypt: A Study of the Imagery in his Earlier Work*. University of Kansas.

[28] Plamondon, M.R. (2005) Computer-assisted phonetic analysis of English poetry: a preliminary case study of Browning and Tennyson. *The Journal of Computer Text Processing*, 14(2): 153–175.

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

[30] Rhody, M.L. (2012) Topic modeling and figurative language. *Journal of Digital Humanities*, 2(1): 19–35.

[31] Ricks, C. (1969) *The Poems of Tennyson*. London: Longman.

[32] Ricks, C. (1987a) *The Poems of Tennyson*. 2nd ed, vol. I. London: Longman.

[33] Ricks, C. (1987b) *The Poems of Tennyson*. 2nd ed, vol. II. London: Longman.

[34] Ricks, C. (1987c) *The Poems of Tennyson*. 2nd ed, vol. III. London: Longman.

[35] Shaw, W.D. (1973) *Tennyson's Style*. Ithaca and London: Cornell University Press.

[36] Tabata, T. (2017) Mapping Dickens's novels in a network of words, topics, and texts: topic modelling a corpus of classic fiction. *Japanese Association for Digital Humanities Conference 2017*, Doshisha University.

[37] Tabata, T. (2018) Mapping Dickens's novels in a network of words, topics, and texts. *Text Mining and Digital Humanities*, 2017: 51–60.

[38] Tabata, T. (2020) Latent topics in British classic fiction: using LDA to classify texts into meaningful groups. *Text Mining and Digital Humanities*, 2019: 47–58.

[39] Tennyson, C. (1949) *Alfred Tennyson*. London: Macmillan.

[40] Tennyson, H. (1897) *Alfred Lord Tennyson: A Memoir by His Son*. vol. I–III. London: Macmillan.

[41] Thomas, J. (2019) *Tennyson Echoing Wordsworth*. Edinburgh: Edinburgh University Press.

## Appendices

### Appendix I: Comparison table of abbreviated and full-length titles of poems: Topic 2.

Abbreviated Titles	Year of Publication	Poem Titles
T26_nomore	1826	‘No More’
T27_SHOMER	1827	‘Short Eulogium on Homer’
T27_WEMORE	1827	‘We Meet No More’
T28_homome	1828	‘Home’
T30_ANAONA	1830	‘Anacaona’
T30_DUASMS	1830	‘Dualisms’
T30_HERDER	1830	‘Hero to Leander’
T30_odeook	1830	‘Ode: O Bosky Brook’
T30 SONOWL	1830	‘Song: The Owl’
T30_THEAID	1830	‘The Mermaid’
T30_THEIES	1830	‘The Sea Fairies’
T30_THEMAN	1830	‘The Merman’
T30_THEPER	1830	‘The Grasshopper’
T32_BUORTE	1832	‘Buonoparte’
T32 SONOBE	1832	‘Sonnet [But were I Loved As I Desire to be]’
T32_sonous	1832	‘Sonnet [Conrad! Why Call Thy Life Monotonous?]’
T32_THEDES	1832	‘The Hesperides’
T33_THEENT	1833	‘The Lovers Tale: A Fragment’
T33_THEERS	1833	‘The Lotos Eaters’
T33_THEOTT	1833	‘The Lady of Shalott’
T34_faiace	1834	‘Fair is that Cottage its Place’
T34_omouup	1834	‘O Mother Britain Lift thou Up’
T37_blaate	1837	‘Black Bull of Aldgate’
T42_AFAELL	1842	‘A Farewell’
T42_BREEAK	1842	‘Break, Break, Break’
T42_MORHUR	1842	‘Morte D’Arthur’
T50_LINNED	1850	‘Lines [Here Often, When a Child, I Lay Reclined]’
T50_MABRPE	1850	‘Mablethorpe’
T54_TOTICE	1854	‘To the Rev. F.D. Maurice’
T55_THEISY	1855	‘The Daisy’
T59_jactar	1859	‘Jack Tar’
T62_ENODEN	1862	Enoch Arden
T62_REQCAT	1862	‘Requiescat’
T62_THEAGE	1862	‘The Voyage’
T62_THEBOY	1862	‘The Sailor Boy’
T62_THELET	1862	‘The Islet’
T64_AYLELD	1864	‘Aylmer’s Field’
T65_CAPAVY	1865	‘The Captain (A Legend of the Navy)’
T68_eps09	1868	‘Epigrams’ (1868, ix)
T68_eps10	1868	‘Epigrams’ (1868, x)
T68_eps21	1868	‘Epigrams’ (1868, xxi)
T68_LIN866	1868	‘Lines (1865–1866)’
T69_COM-IK	1869	‘The Coming of Arthur’
T69_HOL-IK	1869	‘The Holy Grail’
T76_SPBE21	1876	‘Songs from the Plays [‘Becket’ Act II Scene II]’
T77_PRENET	1877	‘Prefatory Sonnet’
T78_THEEET	1878	‘The Revenge: A Ballad of the Fleet’
T79_THENOW	1879	‘The Defence of Lucknow’
T80_THEREL	1880	‘The First Quarrel’
T80_THEUNE	1880	‘The Voyage of Maeldune’
T85_THEECK	1885	‘The Wreck’
T85_THEEET	1885	‘The Fleet’

## Appendix II: Comparison table of abbreviated and full-length titles of poems: Topic 3.

Abbreviated Titles	Year of Publication	Poem Titles
T27_HUNONG	1827	‘Huntsmans Song’
T27_MIDGHT	1827	‘Midnight’
T27_ONSITY	1827	‘On Sublimity’
T27 SONIME	1827	‘Song it is the Solemn Even Time’
T27_THEARP	1827	‘The Exiles Harp’
T27_THENES	1827	‘The Vale of Bones’
T27_THEOFE	1827	‘The Dell of E—’
T28_armdon	1828	‘Armageddon’
T28_topeat	1828	‘To Poesy [O God, Make This Age Great]’
T29_TIMTOO	1829	‘Timbuctoo’
T30_ANAONA	1830	‘Anacaona’
T30_CIRNCE	1830	‘Circumstance’
T30_ELEACS	1830	‘Elegiacs’
T30_illion	1830	‘Ilion, Ilion’
T30_ODEORY	1830	‘Ode to Memory’
T30_RECHTS	1830	‘Recollections of the Arabian Nights’
T30_THEIND	1830	‘The Poets Mind’
T30_THEUSE	1830	‘The Deserted House’
T30_THEWAN	1830	‘The Dying Swan’
T30_thyend	1830	‘Thy Soul is Like a Landskip, Friend’
T31_AFRENT	1831	‘A Fragment’
T31_CHELLY	1831	‘Check Every Outflash Every Ruder Sally’
T31 SONETH	1831	‘Sonnet [Me My Own Fate to Lasting Sorrow Doometh]’
T32_THEART	1832	‘The Palace Oo Art’
T32_THEDES	1832	‘The Hesperides’
T32_THEERS	1832	‘The Sisters’
T33_OENONE	1833	‘Œ none’
T33_THEERS	1833	‘The Lotos Eaters’
T33_youth	1833	‘Youth’
T36_STAEVE	1836	‘St Agnes’ Eve’
T42_AMPION	1842	‘Amphion’
T42_AUDURT	1842	‘Audley Court’
T42_THEONG	1842	‘The Poet’s Song’
T46_THEEAR	1846	‘The Golden Year’
T47_PRIESS	1847	‘The Princess’
T49_theady	1849	‘The Little Lady’
T49_theild	1849	‘The Losing of the Child’
T53_TOEECE	1853	‘To E.L. on His Travels in Greece’
T55_THEISY	1855	‘The Daisy’
T62_INTETZ	1862	‘In The Valley of Cauteretz’
T62_MILACS	1862	‘Milton (Alcaics)’
T62_ONTMER	1862	‘On Translations of Homer’
T62_SPERSE	1862	‘Specimen of a Translation of the Iliad in Blank Verse’
T62_THELET	1862	‘The Islet’
T68_eps03	1868	‘Epigrams’ (1868, iii)
T68_eps18	1868	‘Epigrams’ (1868, xviii)
T68_LUCIUS	1868	‘Lucretius’
T72_ENG782	1872	‘England and America in 1782’
T79_spfalc	1879	‘Songs From The Plays [‘The Falcon’]’
T80_THEERS	1880	‘The Sisters’
T84_COMISE	1884	‘Compromise’
T89_BEAITY	1889	‘Beautiful City’
T89_ceplis	1889	‘Cephalis’
T89_MEREAM	1889	‘Merlin and the Gleam’
T89_PARUS	1889	‘Parnassus’
T89_ROMRSE	1889	‘Romneys Remorse’
T89_THPRNG	1889	‘The Progress of Spring’
T92_KAPANI	1892	‘Kapiolani’
T92_THEONE	1892	‘The Death of Œnone’



# *The Lord of the Rings* における二項対立の識別 二項対立への主要な定量的アプローチ

Vogatza Dimitra

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: dvogatza@yahoo.com

**あらまし** 本論文は、*the Lord of the Rings* における二項対立を特定するための主要な定量的アプローチを提示する。Tolkien の作品のみに基づいて構築されたカスタム Word2Vec モデルを使用し、BERTopic を統合してトピックモデリングを行い、意味論的な関係が Tolkien の物語構造とテーマの対立をどのように反映するかを探求する。分析の結果、癒しと苦悩、自然と戦争といった明確なテーマの二項対立が明らかになり、これらは文脈レベルおよび語彙レベルの両面から検討される。トピック間の高コサイン類似度および低コサイン類似度のスコアは、意味論上における距離の指標として使用され、整合したカテゴリ内および概念的に異なるカテゴリ間の対立を明らかにする。結論として、エンベディングモデルが文学分析にもたらす意味について議論し、Tolkien の伝説体系全体にわたって二項対立の広範な影響を理解するための更なる探求を提案する。

**キーワード** Tolkien、二項対立、エンベディングモデル、コサイン類似度、BERTopic

## Identifying Binary Oppositions in *the Lord of the Rings*:

A primary quantitative approach to binary oppositions

Dimitra Vogatza

Graduate School of Language and Culture, University of Osaka

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

**Abstract** This paper presents a primary quantitative approach to identifying binary oppositions in *the Lord of the Rings*. Using a custom-built Word2Vec model trained exclusively on Tolkien's works, and integrating it with BERTopic for topic-modelling, the study explores how semantic relationships reflect Tolkien's narrative structure and thematic oppositions. These oppositions were revealed through analysing the results. The analysis showed binaries such as healing and trauma, and nature and war, which are examined both at a contextual and lexical level. High and low cosine similarity scores between topics are used as indicators of semantic distance, identifying oppositions both within the same thematic categories and across conceptually distinct ones. The study concludes with a discussion on the implications of an embeddings model for literary analysis, proposing further exploration across Tolkien's legendarium to understand the broader impact of binary oppositions in his work.

**Keywords** Tolkien, binary oppositions, embeddings model, cosine distance, BERTopic

## 1. Introduction

### 1.1. Binary Oppositions and *The Lord of the Rings*

J.R.R. Tolkien's (1892-1973) *the Lord of the Rings*' primary and overall concept that might come to the mind of any reader of the work is the triumph of good over evil—a very apparent binary opposition found in many an epic story. Yet, one of the structural observations that are made on Tolkien's work is in fact a form of a story-long established recurring opposition: havens in perilous situations (Holmes, 2014: 147). He creates these moments of safety in increasingly more dangerous situations as the narrative progresses. This is not all however, there is also deliberate linguistic and stylistic contrast between the Hobbits—plain and modern—and other more archaic peoples (Turner, 2007; Turner, 2014: 392). Additionally, on a more lexical level perhaps, Tolkien plays with words for colours, nature, emotions and moods to express opposed states. He even coined a word himself to express opposed states: *eucatastrophe*. He states (*Letters 131*):

In the scene where all the hosts of the West unite to do honour and praise to the two humble Hobbits, Frodo and Sam, we reach the 'eucatastrophe' of the whole romance: that is the sudden joyous 'turn' and fulfilment of hope, the opposite of tragedy, that should be the hallmark of a 'fairy-story' of higher or lower tone, the resolution and justification of all that has gone before.

Along the varying voices that discuss the style in *the Lord of the Rings*, one has given a quite intriguing view on binary oppositions. It is Le Guin's (2001) appraisal of the rhythmic quality of the work that inspired me to explore if and how binary oppositions in *the Lord of the Rings* function as rhythm. She claims that they appear in a pattern of stress and relief, much like a contextual trochee on a multitude of levels (from chapter to book to volume). Other scholarly works on binaries in *the Lord of the Rings* is from Pretorius (2002) focusing on a few specific binaries through a literary analysis approach, Doyle's (2020) cultural and moral approach on Tolkien's good and evil, and Khan's (2024) approach focusing on linguistic and narrative structures on binary oppositions. Verlyn Flieger, has also explored oppositions in *the Splintered Light* (1983), stating:

No careful reader of Tolkien's fiction can fail to be aware of the polarities that give it form and fiction.

This paper focuses on a primary quantitative approach to identifying binary oppositions through the creation of a Tolkien embeddings model and its incorporation in topic-modelling techniques. The model consists solely of Tolkien's writings that are stylistically and contextually relevant to *the Lord of the Rings*. The works and manuscripts that are included in the embeddings model are listed in Tables 1 and 2.

Table 1: Published works by Tolkien included in the model

No.	Date	Title	Classification	Word-tokens
1	1935	<i>The Hobbit</i>	Legendarium	95,968
2	1945	<i>Leaf by Niggle</i>	Short Story	7,467
3	1949	<i>Farmer Giles of Ham</i>	Short Story	16,476
4	1954-1955	<i>The Lord of the Rings</i>	Legendarium	496,764
5	1962	<i>The Adventures of Tom Bombadil</i>	Legendarium	8,738
6	1967	<i>Smith of Wootton Major</i>	Short Story	10,237
<b>Sum of word-tokens: 635,650</b>				

## 1.2. Tolkien’s Textual Background

The contextually relevant writings along with *the Lord of the Rings* are part of Tolkien’s legendarium. They start as early as 1914 and continue on well into the 1960s, yet throughout his life only three relative works were published (see Table 1). From the numerous unpublished works that are related to the legendarium some were complete yet abruptly ended (i.e. *the Children of Húrin*), but most were either incomplete, revised a multitude of times or existing in a few different versions of varying discrepancy (i.e. works published in *The Fall of Gondolin* see Table 2). Along his work on his legendarium, he also wrote a few different short stories and poems or verses, a small number of them is also included in the model based on their relevance.

Inspired by the Finnish and Celtic language when he was still a teen, he started creating his own languages, and they needed a world to exist in—the legendarium (*Letters* 219-220). It could actually be considered a ‘side project’ and his languages a ‘secret vice’ as he called it (*Letters* , *Fimi* 20), as he was a professor of English Language and Literature at the University of Oxford—not that his academic interests and his writings do not have common roots. His translations of Old and Middle English epics or romances like *Beowulf* and *Sir Gawain and the Green Knight*, influential and analytical lectures on said works (*Beowulf: The Monsters and the Critics*), and his background as a lexicographer for the Oxford English Dictionary mirror his own romances and epics and language creation. Yet it is because of this that he did not have the luxury of time to dedicate to his creations, which leaves us with numerous unfinished manuscripts, hastily and infamously illegible writings on loose pieces of paper, or poems written on the margins of books.

After Tolkien’s death the vast majority of his writings has been edited and published by his son Christopher Tolkien (henceforth Christopher). This process entailed many challenges, as it would often be the case where Tolkien would cross out words from manuscripts and make emendations, or create a whole new version of that writing. Sometimes the style was different, some other times the story would change, or some others the character names would change—more drastic as Tevildo to Sauron, or simpler like Melko to Melkor. The problem that arises here is that due to this multiplicity of versions and emendations, and manuscript completeness the process of identifying the appropriate texts for the model became significantly more time consuming than expected.

Table 2: Posthumously published works by Tolkien included in the model

No.	Date Approx. (Pub. Date)	Title Published In	Category	Word-tokens
1	1912-1914 (2015)	<i>The Story of Kullervo</i> <i>The Story of Kullervo</i>	Finnish Tale	8,593
2	1916-1917 (2018)	<i>The Tale of the Fall of Gondolin</i> <i>The Fall of Gondolin</i>	Legendarium	23,459
3	1917-1950s (1977)	<i>The Silmarillion</i> <i>The Silmarillion</i>	Legendarium	130,757
4	1917 (2017)	<i>The Tale of Tinúviel</i> <i>Beren and Lúthien</i>	Legendarium	15,230
5	1920-1943 (1976)	<i>Letters from Father Christmas</i> <i>Letters from Father Christmas</i>	Letters	15,966
6	1920s-1937 (2007)	<i>The Children of Húrin</i> <i>The Children of Húrin</i>	Legendarium	60,124
7	1925-1931 (2017)	<i>Lay of Leithian</i> <i>Beren and Lúthien</i>	Legendarium	18,163
8	1926 (2018)	<i>The Fall of Gondolin (Sketch)</i> <i>The Fall of Gondolin</i>	Legendarium	1,976
9	1930 (2017)	<i>Beren and Lúthien (Quenta Noldorinwa)</i> <i>Beren and Lúthien</i>	Legendarium	6,164
10	1930 (2018)	<i>The Fall of Gondolin (Quenta Noldorinwa)</i> <i>The Fall of Gondolin</i>	Legendarium	4,226
11	1930 (2017)	<i>Lost Tale of Nauglafring</i> <i>Beren and Lúthien</i>	Legendarium	2,505
12	1937 (1998)	<i>Roverandom</i> <i>Roverandom</i>	Short-Story	25,091
13	late 1950s (1993)	<i>Athrabeth Finrod ah Andreth</i> <i>History of Middle-Earth: Morgoth's Ring</i>	Legendarium	8,708
<b>Sum of word-tokens: 320,962</b>				

*Note: Dates reflect the estimated time of original composition. Publication dates refer to editions edited by Christopher Tolkien or later editors. The works are represented with the original composition followed by the book it was published in beneath it.*

Some years after the initial publication of *The Silmarillion*, a work intended by Tolkien himself to be published along with *The Lord of the Rings*, Christopher went on and published a series of 12 volumes called *The History of Middle Earth* including as much of his father's unedited writing as he could with ample commentary and discussion. Primarily *The History of Middle-Earth* follows closely the contents of *The Silmarillion*, the latter having received considerable editing to be able to be published as a complete narrative. However, due to the sheer size of the volumes and the overall time required for separating editor from author, these writings are not included in this paper's model construction, apart from *Athrabeth Finrod ah Andreth* finished some time around 1959. *The Silmarillion* in this case is included, albeit with slight caution.

## 2. Quantitative Methods for Identifying Oppositions

As discussed in 1.2. the model was constructed with the texts listed in Tables 1 and 2. The model was trained on sentence level granularity. This is due to the fact that there is no one way

the oppositions appear, there are cases where they appear in between a couple of pages (5.08.011-5.08.020)<sup>1</sup> or just within a paragraph (4.08.004)<sup>1</sup> or on chapter titles (i.e. chapters 1 and 2)<sup>2</sup>. So I chose finer granularity for this paper (Ding et al., 2021). Furthermore, I maintained the hyphenated words, i.e. when I removed the punctuation after the sentence breaks. This is to ensure Tolkien's characteristic hyphenation (Alden, 2021; Tauber 2025) as part of his stylistic choice. Similarly, accents to words belonging to Tolkien's languages are also maintained, i.e. Lúthien. Finally the sentences were broken on periods, exclamation marks and question marks. Only point of note with breaking the texts into sentences is the poems and songs found throughout his texts. I chose to maintain sentence level breaking and not break them per lyric, so that each line of the corpus would be a linguistically sound sentence.

## 2.1. Training the Tolkien Model

To train the word embeddings for the model on around 950,000 tokens of my Tolkien corpus, I used Word2Vec's skip-gram architecture. Even though it is primarily directed towards considerably larger datasets than mine, it suited mine because of its ability to capture nuances. Skip-gram has been shown to perform better with infrequent words (Mikolov et al, 2013; Golberg et al, 2014), which is particularly relevant in a literary corpus compared to other architectures like CBOW (Continous Bag of Words). This choice was also influenced by prior comparative work in NLP textbooks (Albrecht et al., 2020), as well as my own preliminary tests, where skip-gram produced more coherent clusters for the analytical goals of this study.

The skip-gram model was trained using a vector size of 100 and a window of 5. Words occurring fewer than five times were excluded from training to reduce noise. The model was trained over 5 epochs, providing sufficient convergence given the corpus' size and thematic consistency. Additionally, following the approach introduced by Mikolov et al. (2013), I employed negative sampling (negative=5) to improve efficiency.

I would like here to note that at an early stage I tested GloVe's applicability in identifying oppositions in *the Lord of the Rings*. GloVe—gigaword 50d variant—is trained on Wikipedia and newswire text, which can be considered as non-literary English usage (Pennington et al., 2014; Parker et al., 2011). It has a vocabulary size of approximately 400,000 words. This preliminary evaluation was conducted by measuring high-similarity word pairs using cosine similarity. Yet the results were quite unsatisfactory. It was with this incentive that I created the Tolkien model. While the Tolkien model proved to be more useful, I needed to confirm whether it was Tolkien's language that made it *better* or if this could be contributed to it being a literary model. With this in mind I constructed two comparative Word2Vec models: one trained on British literary texts

---

<sup>1</sup> Citation format taken from the Digital Tolkien Project. *The Lord of the Rings* is broken down into 6 books throughout 3 volumes, 2 books per volume. The main text format is <book>.<chapter>.<paragraph>.

<sup>2</sup> Chapter 1: *A Long-Expected Party*  
Chapter 2: *The Shadow from the Past*

from Tolkien’s period (1900s-1950s), and another on texts by authors known to have inspired Tolkien. While both literary models performed better than GloVe, the model based on Tolkien’s inspirations yielded the highest opposition-to-vocabulary ratio (see Fig.1). Simply put, aligning style and theme in training embeddings for literary analysis is highly important. Although the Tolkien-trained model did not produce the absolute highest ratio, it significantly outperformed GloVe, suggesting that domain-specific embeddings offer more precise semantic representation when applied to texts of similar register and lexicon. Refer to the Appendix for a high-similarity lexical comparison between GloVe and Tolkien models, based on a few keywords relative to the *the Lord of the Rings*.

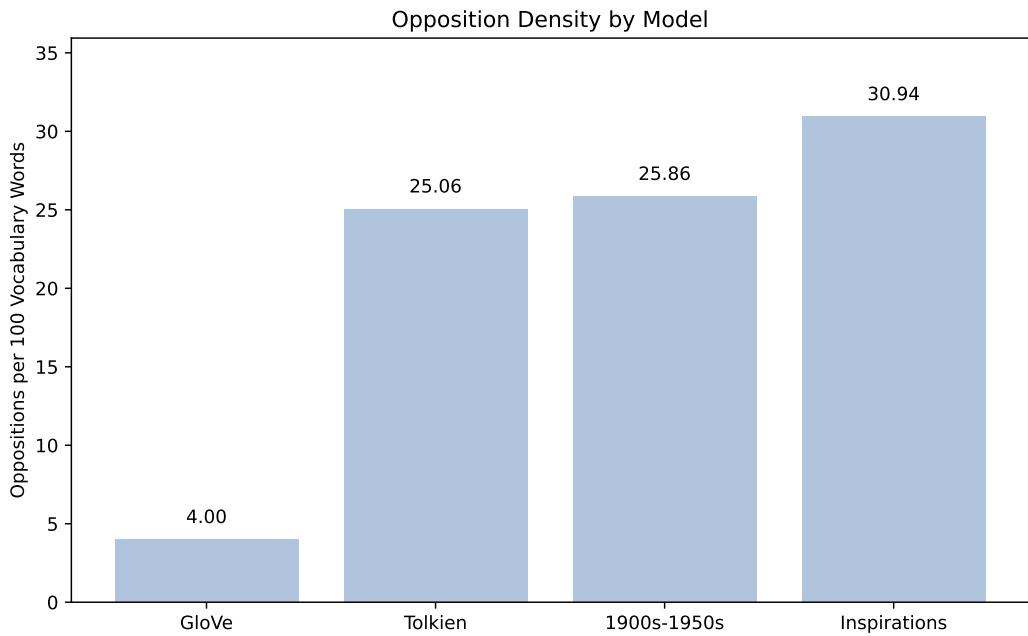


Figure 1: Opposition-to-vocabulary ratios for different word embedding models.

## 2.2. Tolkien Model Implementation in BERTopic

### 2.2.1. Data Preprocessing

As the aim of the methods in this paper is to identify binary oppositions in *the Lord of the Rings* with as little qualitative intervention as possible, I opted to explore topic-modelling as means of identifying opposed words or themes. Preprocessing of the plain text consisted of: stopword removal and document segmentation. *The Lord of the Rings* contains very high frequencies of many proper nouns (see Table 3 for examples) that significantly influence the quality and coherence of the topics, so I included them in the stopword list. To maintain thematic cohesion within all the documents processed by BERTopic I segmented the text on paragraph level and in cases of dialogue I kept the dialogue from start to finish as part of the same segment. There was a need of re-arranging the segments after stopword removal as some segments’ final token count was reduced to very low values such as 3. This resulted to a document length average of 105.13, with

Table 3: Examples of proper nouns and their frequencies in *the Lord of the Rings*

Frodo	1982
Sam	1290
Gandalf	1118
Aragorn	721
Shire	277
Gondor	261

the total tokens of the documents post stopword removal being 142,874—a little over one fourth of the original. While it is the typical fashion in topic-modelling to segment in bigger chunks of text (Misra et al., 2011; Riedl & Biemann, 2012; Tabata, 2020) by setting a set amount of tokens per chunk, I chose to maintain the semantic cohesion of a paragraph—whether long or short—as the work changes places and perspectives often after volume 1 *the Fellowship of the Ring*, where the narrative breaks into three major fronts. Similarly, keeping dialogues intact ensures that the intricacies of character speech (also discussed in sections 1.1. and 3.) remain untouched within each segment.

### 2.2.2. Cosine Similarity in BERTopic and Word2Vec

BERTopic is a neural topic-modelling technique that leverages progress done in model embeddings and clustering techniques to achieve more coherent and representative topics than traditional models like LDA(Latent Dirichlet Allocation) or NMF(Non-Negative Matrix Factorization)(Grootendorst, 2022). For this paper, I utilised the normal BERTopic function with HDBSCAN(Hierarchical Density-Based Spatial Clustering of Applications with Noise) for document clustering and UMAP(Uniform Manifold Approximation and Projection for Dimension Reduction) for dimension reduction. I opted for BERTopic as it comparatively produced more coherent and nuanced results than LDA, but this will not be further discussed in this paper. As BERTopic typically converts documents into embeddings using pre-trained transformer based models such as BERT(Bidirectional Encoder Representations from Transformers) by default, it was also useful in measuring inter-topic distance by applying my Tolkien model as a model for measuring oppositions using cosine distance. For the computation of inter-topic cosine distance, I averaged the Word2Vec embeddings trained on Tolkien’s corpus for each topic. In other words, the mean of the word vectors for each topic. I checked how the word similarities are distributed, and it is a normal distribution (see Fig. 2), so taking the mean instead of the median leads to the same results.

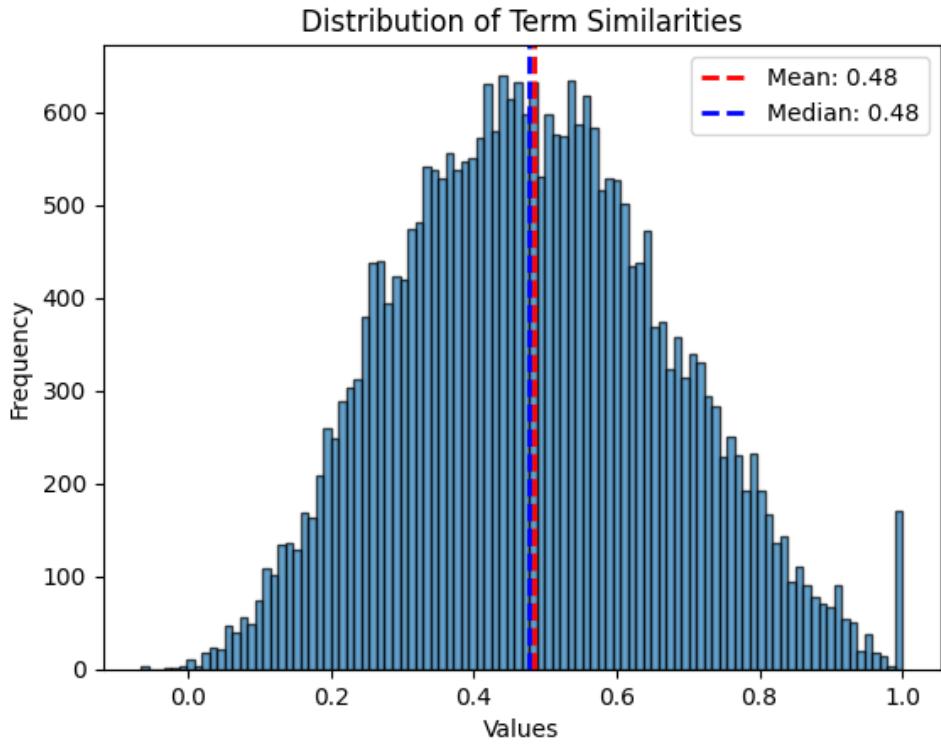


Figure 2: Distribution of term similarities based on the Word2Vec Tolkien model.

Now, BERTopic has its own cosine similarity function, however this similarity is calculated differently. The default function in BERTopic utilises the embedding's model used to extract the topics (this is based on the language set as an argument at the extraction/learning point of the topic-modelling process, in this case the all-MiniLM-L6-v2 (Sentence Transformers all-minilm-L6-V2, n.d.; Wang et al., 2020) is used). While the topics were highly coherent in their contextual nuances, this was not the same for the default inter-topic distances. A clearer representation of the differences between the default inter-topic distance matrix and the Tolkien Word2Vec model can be seen in Figure 3 below.

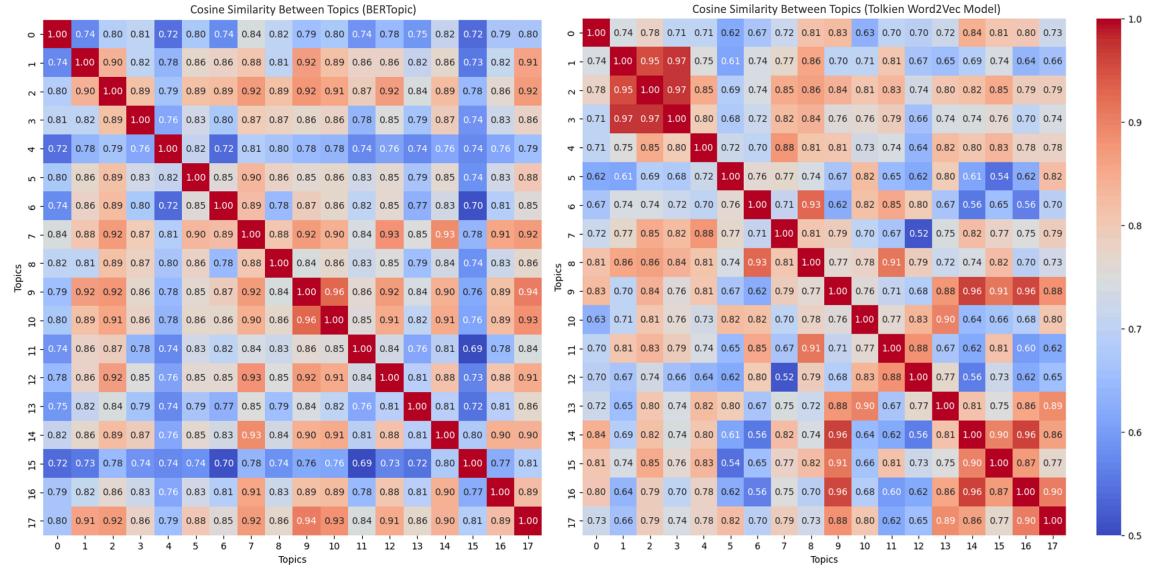


Figure 3: Default BERTopic (left) and Tolkien model (right) cosine similarity matrices

The figure on the left depicts the inter-topic cosine similarity of just the BERTopic embeddings based on the model I mentioned above, while the figure on the right is based on my Tolkien Word2Vec model. As it can be observed, there are distinct differences between the two. The discussion on the topic meanings itself can be found later on in the paper, so I will not go into in depth detail as to what each topic represents, however, I would like to draw attention to a few specific differences that are striking. Topic 12 dealing with war and 7 dealing with trees have a difference of 0.41 between the two graphs, with the Tolkien model showing low-similarity. As Word2Vec embeddings capture the meaning of a word in the context of the corpus that it is trained in, this difference goes to show that to Tolkien trees and war are very far apart in contextual usage. Additionally, while both graphs' cosine threshold was set to 0.5 or higher, the Tolkien graph has more significant fluctuations between high and low values. The Tolkien Word2Vec model overall provides more accurate and educated distances than the default distance.

### 2.2.3. Most Representative Chapters per Topic

As this paper aims to identify oppositions in *the Lord of the Rings*, and as a rhythm as well, I utilised BERTopic's most representative document function to identify which chapters were more representative of which topic (similar to LDA's alpha value, (Blei et al., 2003)). Based on each chapter's narrative note—a chapter's negative or positive note—as well as the terms themselves, I classified the topics into positives and negatives. The argument here is that, since this function aids in understanding in what intervals the topics are appearing throughout the work, by classifying them into these two categories can quantitatively showcase the contextual trochaic rhythm. For the visualisation of the topics throughout the chapters I used ChatGPT and a table containing the top three chapters per topic (see Figure 4). This is further discussed in section 3.

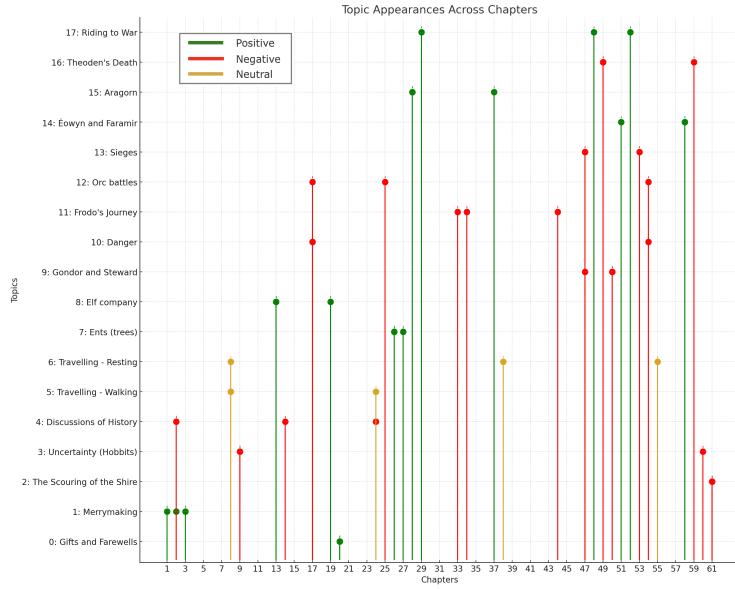


Figure 4: Topic occurrence throughout *the Lord of the Rings*

### 3. Results

The implementation of topic-modelling and semantic vector analysis yielded several notable patterns in the textual structure of *The Lord of the Rings*. This section is divided into three parts. First, I present the extracted topics, including their distribution and content (Table 4). Second, I examine the semantic proximities between topics by applying cosine similarity to averaged Word2Vec embeddings as introduced above. Finally, I focus on identifying recurring thematic binary oppositions throughout the work by analysing the most representative chapters across thematically distinct topics.

Apart from the contents of each topic in the table below, we can also see the positive (+) or negative (-) classification, and the most representative chapters for each topic. Under the umbrella of positive I consider topics that deal with Tolkien's forces of good (Elves, Hobbits, kings, nature) as well as positive concepts as seen in Topic 14 that deals with love and healing. Topic 17's positive classification despite its connotations of war is derived from the context of valiant and honourable warriors. For the negative classification, concepts such as death, war, suffering, battles, danger and uncertainty are marked as negative. Topic 9, which at a first glance cannot outright be discerned as negative by looking at its terms, deals with the state of Gondor at the first stages of the work. A city whose Steward was manipulated by evil forces and later on tragically ended his life while trying to kill his unconscious son in a moment of despair.

Table 4: Topic overview with terms

Topic Labels and Classification		Chapters	Top 15 Terms
0	Gifts and Farewells .....(+)	20, 20, 20	silver, gave, set, virtue, gifts, wrought, lady, spoons, parting, gift, garments, gems, fair, leaves, small
1	Merrymaking .....(+)	1, 2, 3	hobbits, good, began, birthday, guests, supper, food, fire, party, small, table, hobbit, people, hobbiton, eat
2	The Scouring of the Shire . (-)	61, 61, 61	ruffians, farmer, hobbits, rosie, lads, village, leader, chief, weapons, score, hobbiton, waggons, gang, hundred, sheds
3	Uncertainty (Hobbits) .....(-)	9, 9, 60	hobbits, inn, chief, folk, travellers, gate, bob, time, good, hob, times, farmer, ponies, big, pony
4	Discussions of History .... (-)	14, 2, 24	ring, time, men, power, orcs, fear, thought, enemy, dark, found, find, elves, ents, hope, night
5	Travelling - Walking .....()	8, 8, 24	trees, road, light, land, dark, mountains, sun, river, side, hills, path, high, looked, sky, night
6	Travelling - Resting .....()	38, 55, 8	sleep, asleep, water, night, light, heard, woke, lay, sat, dream, wind, fell, soft, eyes, awake
7	Ents (trees) .....(+)	27, 27, 26	ents, elves, hobbits, trees, forest, folk, call, heard, days, time, woods, land, speak, thought, men
8	Elf Company .....(+)	19, 19, 13	sleep, thought, looked, face, food, night, end, lady, hear, bright, head, words, deep, fire, elvish
9	Gondor and Steward .....(-)	50, 50, 47	lord, city, door, house, servants, houses, stood, burn, turned, healing, death, kings, steward, men, face
10	Danger .....(-)	54, 54, 17	door, doom, passage, ran, stairs, orc, orcs, light, side, heard, hall, feet, gate, steps, sprang
11	Frodo's Journey .....(-)	33, 33, 34	eyes, precious, master, dark, felt, hands, looked, thought, head, hand, hobbits, lay, face, hobbit, moment
12	Orc Battles .....(-)	54, 34, 17	orcs, orc, hand, legs, head, fell, eyes, black, sword, sprang, cried, ground, cry, knife, red
13	Sieges .....(-)	47, 47, 53	men, wall, retreat, city, walls, gate, cried, enemy, orcs, black, horsemen, army, lord, fire, rock
14	Love and Healing .....(+)	58, 58, 51	lady, warden, lord, house, king, healing, city, care, days, battle, healer, houses, looked, women, war
15	Aragorn .....(+)	37, 37, 28	sword, lord, man, speak, hand, master, son, men, spoke, ring, eyes, thing, words, leave, days
16	Theoden's Death .....(-)	59, 59, 49	king, city, prince, victory, men, knights, mark, lord, arms, bear, death, rode, ride, honour, lay
17	Riding to War .....(+)	52, 29, 48	king, rode, men, ride, riders, lord, night, rider, host, road, turned, white, days, wind, passed

I would like to separately discuss topics 5 and 6 as they deal with travelling. While they can

be marked as negative and positive, walking and resting respectively, the classification would not follow the same principles as above. There are cases in the work when walking is not a perilous thing even though it is tiring, and there are cases when resting does not provide rest. Thus I classify them as neutral and kept the brackets blank.

Consequently, in the above Figure 4, topics 5 and 6 are coloured yellow, positives green and negatives red. While this figure needs refining, we can observe that at a primary stage there are a few areas showcasing a reversal between positive and negative. Quite intriguing is the back and forth between positive and negative topics from chapters 47 to 54, with 7 different topics actively coming in and out of the narrative.

### 3.1. Inter-Topic Distance

In this section I will closely examine the high and low similarities as depicted in the graph below, while investigating the more appropriate cosine threshold for each.

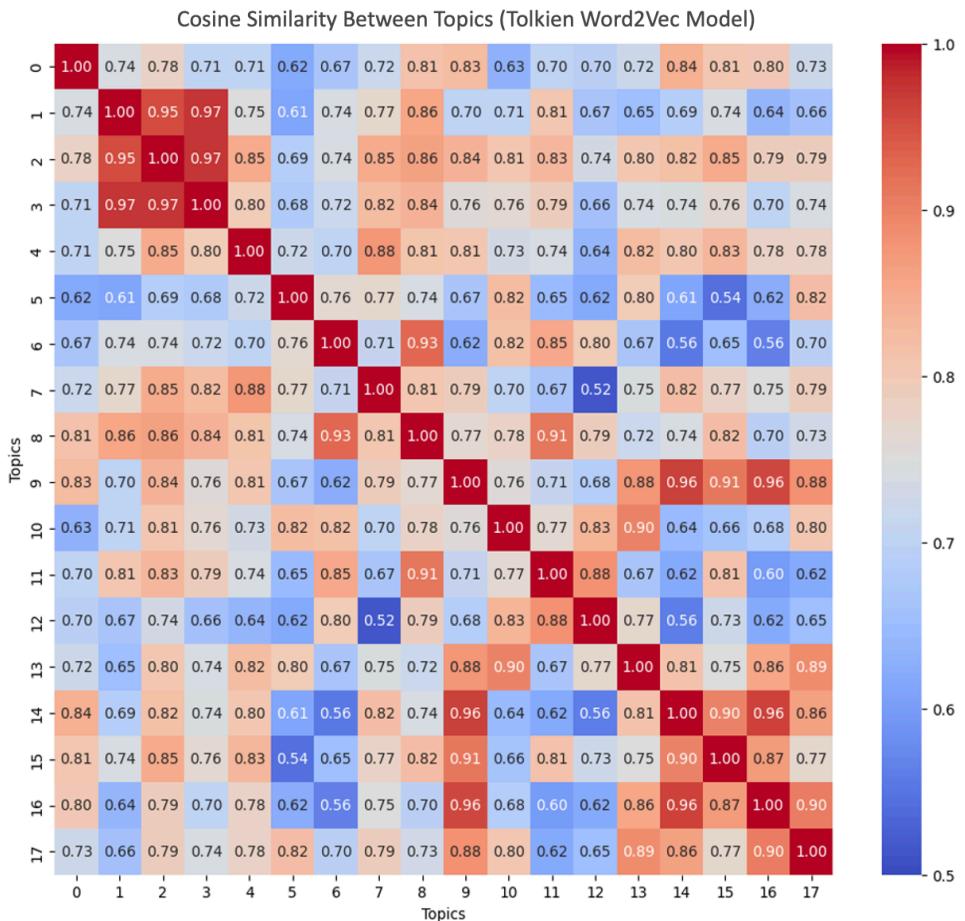


Figure 5: Cosine similarity matrix based on the Tolkien model.

Firstly, when evaluating cosine distance the rule is that the higher the value the closer the distance and the opposite (Sahoo & Maiti, 2025), but what would be the appropriate threshold? What

value or higher and what value or lower works the best in this case? To evaluate this, I firstly crossed out the middle values, from 0.70 to 0.80, with the output range being from 0.5 to 1. I would also like to note that this calculation was also run with range of 0 to 1, but the results were still above 0.5 which did not produce an easily readable graph, hence why in this case it is set to start from 0.5.

Starting with the first 10 decimals on either side—0.60 to 0.70 and 0.80 to 0.90—there are few things that can be easily seen. For example topics scoring at 0.85 like 2 and 15, or 2 and 7, are irrelevant, but not necessarily opposed either. These two pairs deal with the Scouring of the Shire—Hobbits returning home at the end of the work to find it overrun by evil forces and reclaiming it—trees, and Aragorn—the second protagonist reclaiming his position as king. Consequently, anything below 0.85 gives the same results of thematic irrelevance between topics. Now, we can redirect our attention to topic distance of 0.88. Topics that score 0.88 do show some relevance, for example topic 9 and topic 13. Topic 9’s context of Gondor and Steward does occur during the siege of Gondor, but topic 13 that deals with sieges is not limited to Gondor. At the same time, topics 11 and 12 also score 0.88, this score can be attributed to Frodo’s journey taking a negative turn and Frodo suffering for the majority of his travels because of the same forces that mobilise the orcs into battle. However, topics 9 and 17 have much less in common. As previously explained in the above section, topic 9 deals with Gondor’s steward and topic 17 with warriors, these two also do not have thematic connections through the work. As this score is not consistently applicable to function as a threshold for high-similarity, I move one up to 0.89. At 0.89 there are topic pairs like 13 and 17, both dealing with war on the same fronts which confirms their similarity. Interestingly enough, 0.89 appears only in this topic pair. From this threshold and higher the topics along with their scores and similarity characteristics are portrayed in Table 5.

Table 5: High-similarity topic pairs with threshold 0.89 or higher

Similar Themes	Topic Pairs with Scores
Frodo and Resting	8 & 11 ..... 0.91
Elves and Resting	8 & 6 ..... 0.93
Hobbits	1 & 2 ..... 0.95 2 & 3 ..... 0.97 3 & 1 ..... 0.97
Kingdoms of Men	9 & 14 ..... 0.96 9 & 15 ..... 0.91 14 & 15 ..... 0.90 14 & 16 ..... 0.96 16 & 17 ..... 0.90
Sieges and War	9 & 16 ..... 0.96 13 & 17 ..... 0.89

Now focusing on the low-similarity thresholds, the scores are evaluated in a similar fashion. In this scenario however, I am evaluating the opposition between the topics. Topics scoring at 0.68 like 9 and 12, are again irrelevant. Similarly topics 10 and 16 score 0.68, they deal with dangerous situations, and one of the characters' death, respectively. These two are again irrelevant even though they are negative, because topic 16 deals more with a character's death and discussions about his death rather than danger. From here, let us lower the threshold to 0.66. Topics 10 and 15 are up to an extend opposed, as the topic concerning Aragorn deals with heroism and kings in contrast to topic 10 that deals with danger. Yet I cannot call them thematic oppositions with certainty, due to their contextual differences (one deals with a character and the other with a state of things). Topics 3 and 12 are quite irrelevant, as the former deals with uncertain situations around the Shire (where Hobbits live), which could be described as typical British countryside (Curry, 1997; Weidner, 2002). The uncertainty that is occurring around the Shire in the work, and the orc battles in topic 12 have some similarities based on the forces that cause both of them in the narrative, but they are not opposed nor precisely similar. Consequently, we move to the next value of 0.65. Here, topics like 1 and 13 start to show more clear oppositions. Topic 1 deals with merrymaking, especially Hobbit feasts and celebrations, while topic 13 deals with sieges and of course has very strong connotations to war. These two are more thematically opposed than the previously discussed, as this distinctly shows times of peace versus times of war. Similarly, topics 12 and 17 dealing with orc battles and valiant warriors riding to war, are opposed as the two sides of good and evil. From 0.65 and lower the topics along with their scores and opposing characteristics are as found in the table below. Note here that travelling has low-similarity with a variety of different topics and it is difficult to locate the reason. It is possible that the terms in topic 5 are significantly more nature oriented than all the other topic terms which can lead to low-similarity by default. However there are a couple of topics that pair with travelling that show oppositions and are highlighted in Table 6.

Table 6: Low-similarity topic pairs with threshold 0.65 or lower

Opposing Themes	Topic Pairs with Scores
Travelling and Merrymaking	5 & 0 ..... 0.62
	5 & 1 ..... 0.61
	5 & 11 ..... 0.65
	5 & 12 ..... 0.62
Travelling and Healing	5 & 14 ..... 0.61
	5 & 15 ..... 0.54
	5 & 16 ..... 0.62
Resting and War	6 & 9 ..... 0.62
	6 & 14 ..... 0.56
	6 & 15 ..... 0.65
	6 & 16 ..... 0.56
Danger and Healing	0 & 10 ..... 0.63
	10 & 14 ..... 0.64
Frodo(Hobbits) and Men	11 & 14 ..... 0.62
	11 & 16 ..... 0.60
	11 & 17 ..... 0.62
Orcs and Men	7 & 12 ..... 0.52
	12 & 14 ..... 0.56
	12 & 16 ..... 0.62
	12 & 17 ..... 0.65

*Note: Travelling and Merrymaking's respective score and topics are marked with blue. Travelling and Healing's are marked with purple.*

## 4. Discussion

This section evaluates the reliability and limitations of the methods used to extract and measure oppositions in *The Lord of the Rings*. Rather than using the topics as thematic representations, the analysis assesses whether topic-modelling combined with semantic vector similarity can serve as an effective means for identifying binary structures. The subsections that follow are organized around broad thematic clusters—not to showcase oppositions directly, but to examine how well these methods capture meaningful groupings and separations within the text. These reflections aim to determine whether the patterns observed can be trusted as indicators of semantic opposition or merely statistical noise. This critical approach is particularly important given the interpretive risks of applying unsupervised machine learning methods to literary texts, where meaning is deeply context-bound and not always aligned with statistical patterns.

### 4.1. Exploring Topics Dealing with Hobbits and Men

In his legendarium Tolkien has created a variety of different races with their own cultures and intricacies (*History of Middle-Earth: The peoples of Middle-Earth* (Tolkien, 1996)). In this paper I touch upon three of them: Men, Hobbits and Ents. Starting with the case of Hobbits and Men in Table 6, it can be observed that topics that deal with Frodo (the main character) are opposed to

Men. Now, Men in *the Lord of the Rings* and Tolkien's legendarium while they physiologically correspond to the average human, their culture does not correspond in the same way. Tolkien's influence from Anglo-Saxon romances and the Classics is projected onto his depiction of Men—heroic, valiant and kingly, yet easily corrupted (Fimi, 2008; Fernández Camacho, 2023). Hobbits on the other hand represent an idealised and non-technological England (Weidner, 2002; Turner, 2014; *Letters* 230). Turner (2014) further states:

Hobbits are stolidly unheroic, but they have a streak of deep-seated courage which enables them to take on responsibilities greater than their physical stature would suggest, allowing them to lead the modern reader into contact with traditional heroic attitudes and situations which might otherwise seem completely alien.

This is can be also observed in the language the two races use when they speak. Figure 6 below is taken from the Digital Tolkien Project, an online scholarly source that provides a plethora of different quantitative information on Tolkien's works. This particular figure is a rug plot of where each not-contraction occurs in the three volumes of *the Lord of the Rings*, separated by chapters. Hobbits appear consistently from the start of *the Fellowship of the Ring* (henceforth FotR) until its end, however as the narratives break from there onwards there is a very noticeable difference in the usage of not-contractions whenever Hobbits appear. In *the Two Towers* the contractions appear from chapter 3 to 5 and from 9 to 11 when two of our four Hobbits—Merry and Pippin—appear. The chapters before and in between only deal with Men, and have 0 not-contractions. Following chapter 11, the narrative shifts to only Frodo and Sam—our other two Hobbits—and Sméagol—a creature that used to be a Hobbit. For the remainder of *the Two Towers* (henceforth TT) we follow these three characters' journey. An interesting point of note here is in chapters 16 to 18, as they encounter Men and the frequency of the contractions becomes much lower as it is visible by the amount of occurrences. In *the Return of the King* (henceforth RotK) for the first 10 chapters where Merry and Pippin appear, the contractions are much sparser as the characters are separated and meet again in chapter 8 and the contractions increase. Again Hobbits appear in chapters 11 to 14, and from chapter 16 to the end. This stylistic difference further supports the opposed topics of Table 6, with evidence that is usually part of stopword removal and not included in the topic-modelling process.

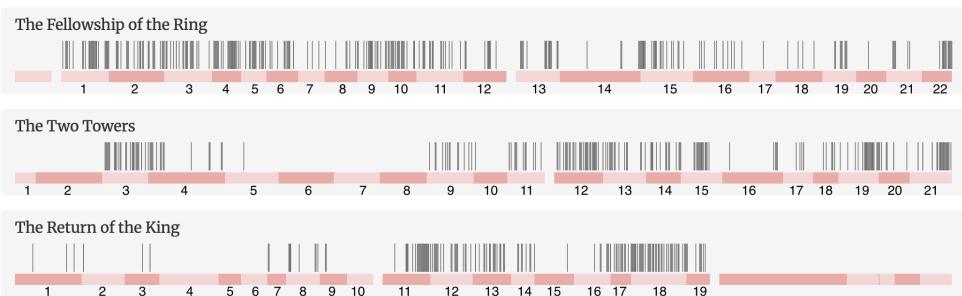


Figure 6: Not-contractions throughout *the Lord of the Rings* as taken from The Digital Tolkien Project's Glossary

To add to this, topic 11 that particularly deals with Frodo's journey (TT chapters: 12-21 and RotK chapters: 11-19), and is opposed to topics that deal with Men, has one more dimension of opposition within it that is irrelevant to style. Frodo's quest objective is to destroy the One Ring. He proves to be resilient against the corruption of the One Ring in contrast to Men that throughout the history of Tolkien's legendarium were easily influenced by it, and failed in its destruction. Now, a knowledgeable reader of the work might argue with me here, that Frodo in the end did not actually succeed in destroying the Ring, but it was rather, a coincidence, or the result of his pity for Sméagol that led to the destruction of the Ring. To which, I agree, yet even though this is indeed the case, Frodo withstanding the will of the Ring up until the last moment, while other characters fail (Boromir), showcase the opposition in resilience between them.

#### 4.2. Lexical Distances in Similar and Opposing Topic Pairs

A part of this overall study is to also identify binary oppositions at a lexical level. In this subsection I examine the relationship of terms in opposed and similar topics. While I cannot examine every single pair in this paper, I select two from each. The goal of this examination is to evaluate if terms semantically follow their topic distances, by showing opposing words.

Starting from high-similarity topic pairs, we have the topic pair of 1 and 2 also seen in Table 5 dealing with Hobbits. Topic 1 that deals with Hobbits and Merrymaking contains terms like: bithday, supper, food, good, party, eat. This topic is mostly respresented by chapters at the start of the work (see Table 4) and is classified as positive. On the other hand, Topic 2 is only represented by chapter 61 (or RotK 18) *the Scouring of the Shire*. This topic as briefly discussed in section 3., is negatively classified. The interesting point between these two topics is that they portray departure (topic 1) and return (topic 2) of the Hobbits from and to the Shire. There is an opposition that appears here, as when the Hobbits leave the Shire, they leave it in a peacefull state, untainted by the evils of the world, yet when they return back the find it in the exact opposite state. The high-similarity appearing here can be attributed to the fact that both topics deal with Hobbits containing relatively similar vocabulary, and as discussed in the previous subsection language and style are changed. Taking a closer look at the terms in Figure 7 below, the term 'weapons' in Topic 2 is opposed or dissimilar to most of the words in Topic 1, especially words like 'hobbits' and 'good'. From an evaluating standpoint this shows that the Word2Vec model functions well in measuring semantic distances. Similarly, other words, like 'village' and 'hobbiton' show high-similarity which is also to be expected. A few key points can be concluded here, high-similarity can mean similar or semantically close themes, but it can also lead to opposing ones which questions the primary argument that low-similarity means oppositions.

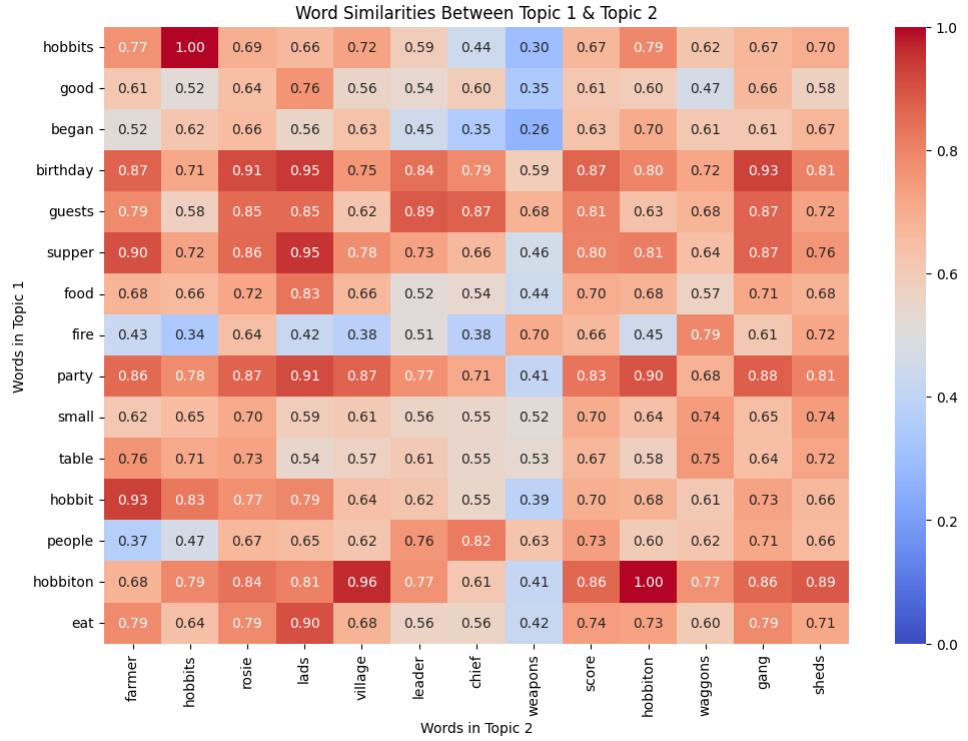


Figure 7: Top 15 term distances of topics 1 and 2 with similarity score of 0.95

Continuing to another high-similarity pair, topic 9 and 14 (Figure 8). Now topics 9 and 14 both deal with Men, but again we have a negative and a positive one. Topic 9 labeled Gondor and Steward, is most represented by chapters 50 (Rotk: 7) *the Pyre of Denethor* and 47 *the Siege of Gondor* (Rotk: 4). As briefly introduced in section 3., topic 9 deals heavily with negative concepts like death, suicide and war. This is visible in terms like: ‘burn’, ‘death’, ‘steward’ and ‘city’. Topic 14 on the other hand labeled Love and Healing, deals precisely with these two concepts and it can be seen in the terms: ‘healing’, ‘lady’, ‘lord’, ‘care’. Topic 14 is mostly represented by chapter 58 (RotK 15) *the Steward and the King* and chapter 51 (RotK 8) *the Houses of Healing*. Both topics connect to the character of Faramir. In the documents included in topic 9 he is nearly killed by his father in a moment of despair before killing himself, and in the documents included in topic 14 he is healed, falls in love and takes his father’s place as steward when the city is healed. This case, is very similar to the one above, high-similarity topics leading to opposing themes. The high-similarity that arises in this topic pair again could be attributed to similar vocabulary and register. Of course, it can also be said that, both topics deal with Faramir and Men in general and thus they are similar. However, the more nuanced analysis would suggest otherwise.

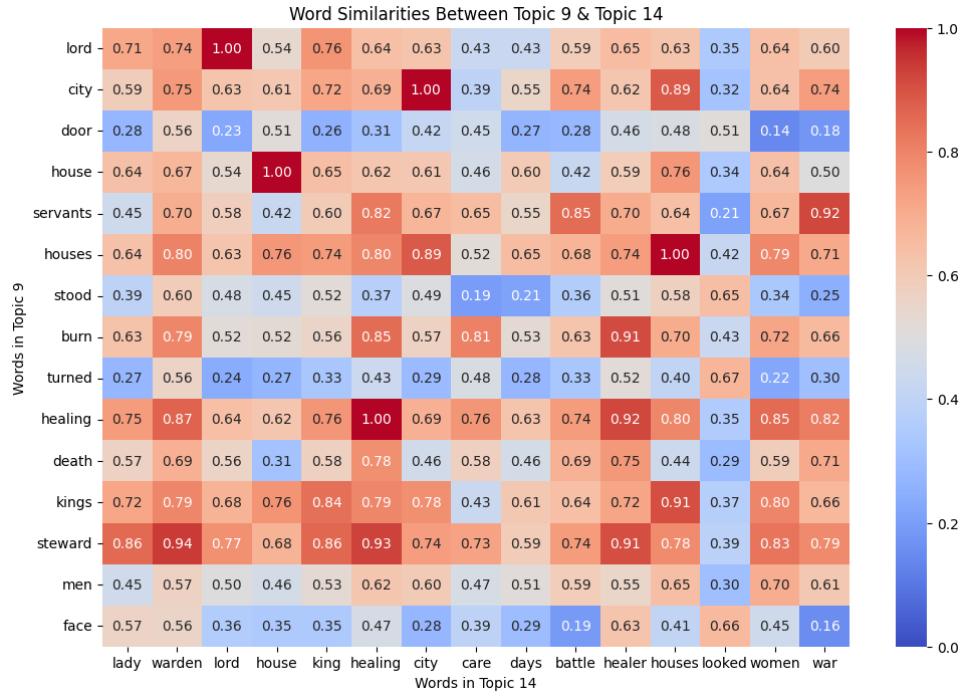


Figure 8: Top 15 term distances of topics 9 and 14 with similarity score of 0.96

Moving on to low-similarity topics, I first examine topics 11 and 14 (Figure 9). Topic 11 deals with Hobbits, but particularly Frodo's quest to destroy the Ring. The most representative chapters are 33 (TT: 12) *the Taming of Sméagol* and 34 (TT: 13) *the Passage of the Marshes*. Frodo's quest is that of suffering, physical and more importantly psychological, and as such this topic is classified as negative. However, the terms do not so clearly express this. The terms 'precious' and 'master' relate to Frodo's relationship with Sméagol, but the rest are not that easily distinguished. Nevertheless, there is a clear opposition between these two topics. Topic 14, as discussed above, deals with healing and characters being able to mend their physical and psychological wounds. Frodo even after the quest was done and the Ring destroyed, could not be psychologically healed. The trauma induced by the Ring was too much, as opposed to Faramir or other characters being healed in the documents that represent topic 14 (Éowyn and Merry). The overall discussion of Frodo or Hobbits being opposed to Men as above, is still applicable in this case, however if we look at isolated instances we can delve deeper into specific oppositions that appear through measuring similarity. Yet, and this is an issue that arises when looking at the relationship between terms, as can be observed by the heatmap below, the terms are not opposed with each other, nor are they similar. They are vastly irrelevant, especially the cases where cosine similarity is the lowest: felt and house, felt and city, head and war. The words 'hobbits' and 'lord' do show the opposition discussed above, but this is not the norm of the relationship between the terms.

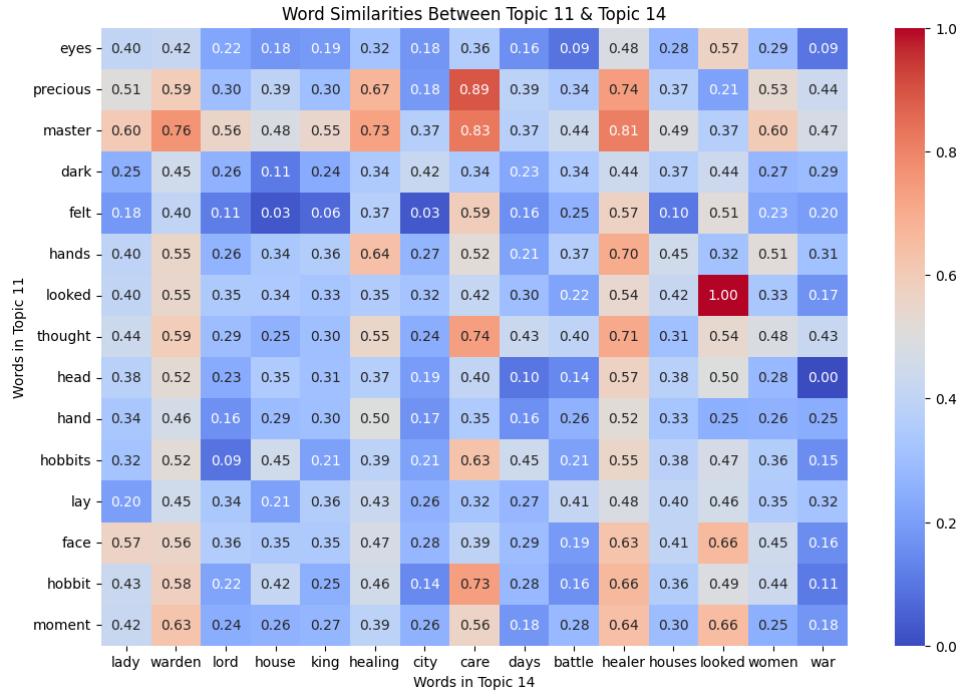


Figure 9: Top 15 term distances of topics 11 and 14 with similarity score of 0.62

Finally, the last pair I examine is topics 7 and 12. Topic 7 deals predominantly with Ents. They are sentient trees that shepherd the forests and guard them. There has been extensive research on Tolkien's portrayal of nature and his personal love for it (Saguaro & Thacker, 2013; Curry, n.d.; Tolkien, 2022), with this in mind as well as the context of the topic itself I classified it as positive. The most representative documents for this topic are in chapters 26 (TT: 4) *Treebeard* and 27 (TT: 5) *the White Rider*. Trees and nature in general in *the Lord of the Rings* suffer and is easily affected by evil forces like orcs, as they cut trees for war which very well opposes topic 12. Topic 12 deals with orc battles, with most representative chapters being 54 (RotK: 11) *the Tower of Cirith Ungol*, 34 (TT: 13) *the Passage of the Marshes* and 17 (FotR: 17) *the Bridge of Khazad-Dûm*. All three of the chapters mentioned deal with battles or armies, orcs and evil forces—Ringwraiths, goblins, Balrogs. Now, taking a closer look at the term distances a similar pattern as the previous topic pair can be observed. Terms with very low-similarity scores are irrelevant, e.g. 'trees' and 'cried', or 'heard' and 'sword'. The terms overall seem to have almost no semantic relevance either, yet nature versus evil is part of Tolkien's major oppositions in *the Lord of the Rings*.

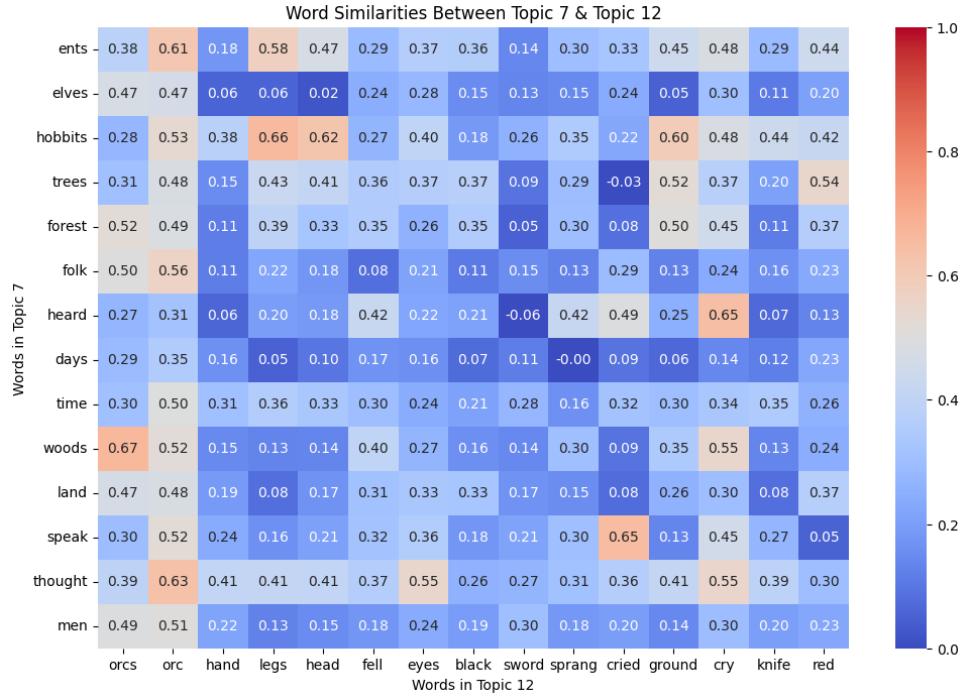


Figure 10: Top 15 term distances of topics 7 and 12 with similarity score of 0.52

There are two major points that can be concluded from the above examinations. First, oppositions can be found in both ends of the spectrum. Highly similar topics (as were grouped in 3.) score high-similarity because they are indeed similar, i.e. Hobbits being in the Shire (topics 1,2 and 3). However, oppositions can be found in high-similarity topic pairs. The difference from the oppositions of low-similarity pairs, is that these oppositions are found within the same concepts—positive Hobbits and negative Hobbits. Whereas in low-similarity pairs, it is about two separate concepts that are opposed by default—trauma and healing. Measuring inter-topic distance through Tolkien’s Word2Vec model seems to be providing significant insight into these conceptual relations. However, distance on a lexical level cannot be measured well with this method as of this moment. Secondly, Tolkien’s Word2Vec model can also catch differences in style even when the style appears more pronounced in words that are by default in stopwords, which further solidifies its usability in identifying thematic oppositions.

## 5. Conclusion

This study presented a primary quantitative method for identifying binary oppositions in *the Lord of the Rings*, exploring how they contribute to the work’s structural and thematic rhythm. By creating a Tolkien-specific Word2Vec model and combining it with BERTopic topic-modelling technique, I was able to identify distinct thematic oppositions, such as healing and trauma, as well as more subtle oppositions within Tolkien’s narrative. The model, trained exclusively on Tolkien’s

texts, enabled a deeper semantic alignment, revealing patterns that non-literary models could not capture.

The integration with BERTopic allowed for a detailed examination of inter-topic distances, showcasing that binary oppositions do not always appear through low-similarity scores. In fact, high-similarity topic pairs often revealed oppositions within the same theme, for instance positive and negative situations concerning Hobbits. Additionally, examining low-similarity pairs revealed more distinct oppositions, such as nature versus war or simple Hobbits against higher Men. These results suggest that Tolkien's use of oppositions is intricately woven into both the structure and semantics of the narrative, and it can be further examined as a form of rhythm.

While the findings offer insight into the possible contextual rhythm of oppositions, this paper also points to areas for further exploration. Analysing the recurrence and distribution of these oppositions across other works like *the Silmarillion* could provide stronger evidence for their role in Tolkien's larger legacy. This would help solidify the idea that these thematic oppositions are not isolated to *the Lord of the Rings* but are a deliberate stylistic choice in his entire legendarium.

## Bibliography

- [1] Albrecht, J., Ramachandran, S. & Winkler C. (2020) *Blueprints for Text Analytics Using Python: Machine Learning Based Solutions for Common Real World Applications*. O'Reilly Media, Incorporated.
- [2] Alden, S. (2021). *Hyphens as Sub-Lexical Morphemes in The Hobbit*. Mallorn: The Journal of the Tolkien Society, 62, 36-40. <https://www.jstor.org/stable/48650605>
- [3] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent dirichlet allocation*. Journal of Machine Learning Research, 3, 993-1022. <https://doi.org/10.5555/944919.944937>
- [4] Curry, P. (1997). *Defending Middle-earth: Tolkien, myth and modernity*. St. Martin's Press.
- [5] Curry, P. (n.d.). *Tolkien and Nature*. The Tolkien Estate. <https://www.tolkienestate.com/writing/patrick-curry-tolkien-and-nature/>
- [6] Ding, L., Wang, L., Liu, X., Wong, D. F., Tao, D., & Tu, Z. (2021). *Progressive Multi-Granularity Training for Non-Autoregressive Translation*. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.findings-acl.247>
- [7] Doyle, M. (2019). *Utopian and Dystopian Themes in Tolkien's Legendarium*. Lexington Books. <https://doi.org/10.5040/9781978738928>
- [8] Fernández Camacho, P. (2023). *Elven-Latin and Semitic Adûnaic: Linguistic, Religious, and Political Strife in Tolkien's Island of Númenor*. Journal of Inklings Studies, 13(1), 67-86. <https://doi.org/10.3366/ink.2023.0176>
- [9] Flieger, V. (2002). *Splintered light: Logos and language in Tolkien's world* (2nd ed.). Kent State University Press.
- [10] Fimi, D. (2022). *Tolkien, Race, and Racism in Middle-earth* (2022) by Robert Stuart. Journal of Tolkien Research, 15(1), 3. Available at: <https://scholar.valpo.edu/journaloftolkienresearch/vol15/iss1/3>
- [11] Goldberg, Y., & Levy, O. (2014). *word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method*. arXiv preprint arXiv:1402.3722.
- [12] Grootendorst, M. (2022). *BERTopic: Neural topic-modelling with a class-based TF-IDF procedure*. arXiv preprint arXiv:2203.05794.
- [13] Holmes J.R. (2014). *The Lord of the Rings*. In Lee, S.D. (Ed.), *A Companion to J. R. R. Tolkien* (2020) (pp. 116-126) John Wiley & Sons
- [14] Le Guin, U. (2001). *Rhythmic pattern in The Lord of the Rings*. In K. Haber (Ed.) *Meditations on Middle-earth*
- [15] Lee, S. D. and John Wiley & Sons (eds). (2022). *A Companion to J.R.R. Tolkien*. Second edition. Hoboken, NJ:John Wiley & Sons.
- [16] Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv preprint arXiv:1301.3781.

[17] Misra, H., Yvon, F., Cappé, O., & Jose, J. (2011). *Text segmentation: A topic-modelling perspective*. Information Processing & Management, 47(4), 528-544. <https://doi.org/10.1016/j.ipm.2010.11.008>

[18] Parker, Robert, Graff, David, Kong, Junbo, Chen, Ke, and Maeda, Kazuaki (2011). English Gigaword Fifth Edition Linguistic Data Consortium doi:10.35111/WK4F-QT80. <https://catalog.ldc.upenn.edu/LDC2011T07> .

[19] Pennington, J., Socher, R. & Manning, C.D. (2014). *GloVe: Global Vectors for Word Representation*. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages1532-1543, Doha, Qatar. Association for Computational Linguistics.

[20] Riedl, M., & Biemann, C. (2012). *Text Segmentation with Topic Models*. Journal for Language Technology and Computational Linguistics, 27(1), 47-69. <https://doi.org/10.21248/jlcl.27.2012.158>

[21] Sahoo, S., & Maiti, J. (2025). *Variance-Adjusted Cosine Distance as Similarity Metric*. arXiv preprint arXiv:2502.02233.

[22] Saguaro, S., & Thacker, D. C. (2013). *Tolkien and Trees*. In J. R. R. Tolkien (pp. 139-155). Macmillan Education UK. [https://doi.org/10.1007/978-1-37-26401-5\\_10](https://doi.org/10.1007/978-1-37-26401-5_10)

[23] Sentence transformers all-minilm-L6-V2 (n.d.) Available at: <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

[24] Tabata, T. (2020) *Latent topics in British classic fiction: using LDA to classify texts into meaningful groups*. Text Mining and Digital Humanities, 2019, 47-58

[25] Tauber, J. (2025a), Digital Tolkien Project: A scholarly project focused on Tolkien from both a corpus linguistic and digital humanities perspective. <https://digitaltolkien.com/>

[26] Tauber, J. (2025b), Tolkien Glossary. <https://glossary.digitaltolkien.com/>

[27] Tolkien, J. R. R. (1936). *Beowulf: The monsters and critics*. Humphrey Milford for the British Academy.

[28] Tolkien, J. R. R. (1937). *The Hobbit* (1st ed.). London: George Allen & Unwin.

[29] Tolkien, J. R. R. (1945). *Leaf by Niggle*. In The Dublin Review, no. 432, Jan., pp. 46-61. Dublin: Burns Oates & Washbourne.

[30] Tolkien, J. R. R. (1949). *Farmer Giles of Ham*. London: George Allen and Unwin.

[31] Tolkien, J. R. R. (1954). *The Fellowship of the Ring*. The Lord of the Rings. London: George Allen & Unwin.

[32] Tolkien, J. R. R. (1954). *The Two Towers*. The Lord of the Rings. London: George Allen & Unwin.

[33] Tolkien, J. R. R. (1955). *The Return of the King*. The Lord of the Rings. London: George Allen & Unwin.

[34] Tolkien, J. R. R. (1962). *The Adventures of Tom Bombadil*. London: George Allen & Unwin.

[35] Tolkien, J. R. R. (1967). *Smith of Wootton Major*. Houghton Mifflin

[36] Tolkien, J. R. R. (1976). *The Father Christmas Letters*. London: George Allen & Unwin.

[37] Tolkien, J. R. R. (1993) *Morgoth's Ring*, in *The History of Middle Earth* (Tolkien C., Ed.). HarperCollins

[38] Tolkien, J. R. R. (1996) *The Peoples of Middle-earth*, in *The History of Middle Earth* (Tolkien C., Ed.). HarperCollins

[39] Tolkien, J. R. R. (1998). *Roverandom* (Scully, C. & Hammond, W. G, Eds.). HarperCollins.

[40] Tolkien, J. R. R. (2015). *The Story of Kullervo*. (C. Tolkien, Ed.). HarperCollins.

[41] Tolkien, J. R. R. (2017). *The Children of Húrin* (Tolkien, C. Ed.). HarperCollins.

[42] Tolkien, J. R. R. (2017). *Beren and Lúthien* (Tolkien, C. Ed.). HarperCollins.

[43] Tolkien, J. R. R. (2018). *The Fall of Gondolin*. (C. Tolkien, Ed.). HarperCollins.

[44] Tolkien, J. R. R. (2022). *Nature of Middle-Earth*. (Hostetter, C. F., Ed.). HarperCollins Publishers.

[45] Tolkien, J. R. R. & Carpenter, H. (2014). *Letters of J. R. R. Tolkien*. Houghton Mifflin Harcourt Publishing Company.

[46] Tolkien, J. R. R. & Gordon, E. V. (1925). *Sir Gawain & the Green Knight*. Clarendon Press.

[47] Tolkien, J. R. R. & Tolkien C. (1977). *The Silmarillion* (Tolkien, C. Ed.). HarperCollins.

[48] Tolkien, J. R. R., Higgins, A., & Fimi, D. (2020). *Secret Vice*. HarperCollins Publishers Limited.

[49] Turner, A. (2007). Prose Style. In Drout, M. D. C. (Ed.), *J. R. R. Tolkien Encyclopedia* (2013)(pp. 545-546) Routledge

[50] Turner, A. (2014). *Style and Intertextual Echoes*. In Lee, S.D. (Ed.), A Companion to J. R. R. Tolkien (2020) (pp. 374-385) John Wiley & Sons

[51] Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). *Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers*. Advances in neural information processing systems, 33, 5776-5788.

[52] Weidner, Brian N. (2002). *Middle-earth: The Real World of J.R.R. Tolkien*, Mythlore: A Journal of J.R.R. Tolkien, C.S. Lewis, Charles Williams, and Mythopoeic Literature. 23(4), 7. Available at: <https://dc.swosu.edu/mythlore/vol23/iss4/7>

## 6. Appendix

Table 7: Top 10 similar words for example words in Tolkien and GloVe Models

Example	Tolkien Model	GloVe Model
light	gleam ..... (0.77)	air ..... (0.78)
	bright ..... (0.75)	lights ..... (0.76)
	star ..... (0.75)	heavy ..... (0.75)
	dimmed ..... (0.75)	lighter ..... (0.75)
	shine ..... (0.74)	surface ..... (0.75)
	pale ..... (0.74)	display ..... (0.75)
	sunlight ..... (0.73)	bright ..... (0.75)
	shining ..... (0.72)	visible ..... (0.74)
	brighter ..... (0.71)	ground ..... (0.74)
	visible ..... (0.71)	color ..... (0.73)
ring	finger ..... (0.71)	rings ..... (0.89)
	ruling ..... (0.69)	hanging ..... (0.69)
	sauron ..... (0.69)	triangle ..... (0.69)
	burden ..... (0.67)	gate ..... (0.68)
	wield ..... (0.64)	sword ..... (0.68)
	weight ..... (0.62)	wire ..... (0.67)
	chain ..... (0.62)	hidden ..... (0.66)
	stroke ..... (0.62)	mask ..... (0.66)
	dreadful ..... (0.62)	inside ..... (0.65)
	enemy ..... (0.62)	into ..... (0.65)
party	finished ..... (0.96)	democratic ..... (0.88)
	bag-end ..... (0.96)	opposition ..... (0.88)
	arrived ..... (0.96)	parties ..... (0.84)
	attention ..... (0.94)	coalition ..... (0.83)
	polite ..... (0.93)	leader ..... (0.83)
	week ..... (0.93)	election ..... (0.83)
	eaten ..... (0.93)	candidate ..... (0.83)
	tea ..... (0.93)	elections ..... (0.82)
	actually ..... (0.92)	socialist ..... (0.82)
	important ..... (0.92)	liberal ..... (0.81)

Table 8: Top 10 similar words for example words in Tolkien and GloVe Models

Example	Tolkien Model	GloVe Model
<i>destroy</i>	suffer ..... (0.97)	destroying ..... (0.84)
	perish ..... (0.96)	destruction ..... (0.82)
	attempt ..... (0.96)	protect ..... (0.78)
	fulfilment ..... (0.95)	terrorists ..... (0.76)
	abide ..... (0.95)	enemies ..... (0.76)
	restrain ..... (0.95)	rid ..... (0.75)
	vengeance ..... (0.95)	invade ..... (0.75)
	rule ..... (0.95)	kill ..... (0.74)
	fulfilled ..... (0.95)	seize ..... (0.74)
	achieved ..... (0.95)	dismantle ..... (0.73)
<i>quest</i>	betray ..... (0.96)	ultimate ..... (0.77)
	folly ..... (0.96)	dream ..... (0.76)
	release ..... (0.96)	destiny ..... (0.73)
	test ..... (0.96)	dreams ..... (0.71)
	desires ..... (0.96)	survival ..... (0.69)
	choice ..... (0.95)	success ..... (0.69)
	vain ..... (0.95)	pursuit ..... (0.69)
	fulfilled ..... (0.95)	glory ..... (0.69)
	doubtless ..... (0.95)	opportunity ..... (0.68)
	nonetheless ..... (0.95)	struggle ..... (0.68)



# Hardy の Wessex 小説における伝達節と キャラクタライゼーション

曹 芳慧

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: u327503a@ecs.osaka-u.ac.jp

**あらまし** 本稿では、19世紀イギリスの作家 Thomas Hardy の Wessex 小説における「伝達節」に注目し、その機能および人物造形への寄与を明らかにすることを目的とする。小説における会話は物語の進行を助け、登場人物の性格や心理を描写する重要な手法であり、特に長編小説では地の文のみでは伝わりにくい情報を補完し、物語に躍動感を与える役割を果たす。伝達節は、話者を示すだけでなく、発話の口調や感情、態度を伝えることにより、読者が登場人物の発話をより臨場感を持って理解する手助けをする。

本研究では、14冊の Wessex 小説を対象に独自のコーパスを構築し、まず伝達節を抽出した。その後、伝達動詞と副詞の頻度を分析し、さらに動詞と副詞の共起パターンを調査した。また、キャラクター別の分析を行い、ネットワーク可視化することで、主要登場人物における伝達動詞と副詞の使用傾向の違いを明確化し、それによって Hardy のキャラクタライゼーション手法の特徴を浮き彫りにすることを目指した。今後は他の作品との比較分析や感情分類を含むさらなる発展的研究が期待される。

**キーワード** 伝達節、動詞と副詞の共起、ネットワーク可視化、  
キャラクタライゼーション

## Reporting Clauses and Their Role in Characterization

in Hardy's Wessex Novels

Cao Fanghui

Graduate School of Humanities, The University of Osaka

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

**Abstract** This study focuses on the role of reporting clauses in the 19th-century Wessex novels of Thomas Hardy, aiming to clarify how they contribute to characterization. Dialogue in novels not only advances the plot but also serves as a key tool for expressing the personalities, emotions, and relationships of characters. In particular, reporting clauses—indicating the speaker and manner of speech—help convey nuances of tone, attitude, and emotion, allowing readers to grasp the atmosphere of the conversation beyond what is directly stated. By examining how Hardy uses these clauses, this study seeks to shed light on his distinctive techniques of characterization within the fictional world of Wessex.

To achieve this, a custom-built corpus of 14 Wessex novels was created. Reporting clauses were systematically extracted, and the frequencies of reporting verbs and adverbs were analyzed. The study then investigated co-occurrence patterns between verbs and adverbs, followed by a character-based analysis. Additionally, a network visualization was employed to clarify differences in the use of reporting verbs and adverbs across major characters, revealing distinctive stylistic tendencies. Through this multifaceted approach, the study highlights the linguistic strategies underpinning Hardy's character portrayals. Future directions include comparative analysis with other works and advanced sentiment analysis to deepen understanding of speech representation in literary texts.

**Keywords** Reporting Clauses, Verb-Adverb Co-occurrence, Network Visualization, Characterization

## 1. はじめに

小説における会話は、登場人物の性格や関係性を読者に伝える上で重要な役割を果たしている。特に長編小説では、地の文のみに頼った叙述は読者に負担を与える可能性があり、適度な対話の導入によって物語に活力がもたらされる。対話は物語の進行に寄与するだけでなく、登場人物の個性や心理状態を描写するための重要な手法である。Page (1973: 51) は、小説における対話 (dialogue) の重要性について次のように述べている。

The dialogue in a novel is [...] multifunctional: it can serve to further plot, to develop character, to describe setting or atmosphere, to present a moral argument or a discussion on cabbages or kings, or to perform any combination of these purposes.

会話部は小説の成り立ちにおいて重要な役割を果たしているが、会話部を抽出して分析することには技術的な制約が伴う。本稿では、こうした制約を克服し、量的な分析手法を用いることで、19世紀イギリスの作家 Thomas Hardy (1840-1928) の Wessex 小説における会話部に焦点を当て、特に「伝達節」と呼ばれる要素が人物描写に果たす役割を検討する。

### 1.1. 伝達節とは

伝達節 (Reporting Clause) とは、他者 (作家にとっての「他者」、すなわち小説の登場人物) の発言や思考を伝えるために用いられる節のことである。豊田他 (2017: 158) によれば、「文字だけでは話し手たちの表情や声の調子がわからず、朗読を収録した CD などの音源がない場合、会話の雰囲気をつかみにくくことがある。そのような会話に「表情」を与える役割を果たすのが伝達節である」という。伝達節とは、対話文中において話者や語り方を示す節であり、通常、話者、伝達動詞および副詞修飾語が含まれる。例えば、Wessex 小説から次のような例が挙げられる。

“I know hardly any poetry,” he replied mournfully. (*Jude the Obscure*, Part Fourth AT SHASTON, V)

ここで、“I know hardly any poetry” と引用符で囲まれた部分は *he* (*Jude*) の発話である。 *he replied mournfully* が伝達節であり、*replied* が伝達動詞、*mournfully* が動詞を修飾する副詞で

ある。このように、伝達節は単に発話者を示すだけでなく、発話の口調、感情、態度を読者に伝える重要な役割を担う。Hardy が小説を創作する際、この部分でキャラクターの性格づけを行っていたと仮説を立てることができるだろう。本研究では、Wessex 小説における伝達節を対象に、伝達動詞とその動詞を修飾する副詞の結びつき、さらに伝達節とキャラクタライゼーション（人物の特徴づけ）との関連性について掘り下げていく。

## 1.2. Wessex 小説とは

Wessex 小説とは、作家 Hardy が創作した架空の地域「ウェセックス」を舞台とする小説群を指す。この地域は、Hardy の故郷ドーセット州を中心にイングランド南西部の田園地帯をモデルとしており、1871 年から 1897 年にかけて発表された 14 作品がこれに含まれる。Hardy の作品では、農村社会の階級、恋愛、宿命、産業化の影響などがテーマとして描かれ、登場人物の心理や関係性が精緻に描写されることが特徴である。また、叙述と対話が巧みに交錯し、地の文のみならず会話部分においても登場人物の個性や感情が色濃く表現される。特に伝達節においては、発話者の性格や心理が副詞の使用を通じて浮き彫りになることが多く、Hardy のキャラクタライゼーションの手法を考察する上で重要な分析対象となる。このような特徴を持つ Wessex 小説は、伝達節の機能や役割を検討する上で格好の分析対象となる。本研究では、これらの作品における伝達節の具体的な用法を分析し、その結果を通じて明らかになるキャラクタライゼーションの特徴を探究することを目指す。

## 2. 研究背景

### 2.1. 伝達節についての先行研究

伝達節は、物語中における発話や思考の提示において不可欠な構成要素である。典型的には *he said* や *she replied* のように発話の出所を示すが、その役割は単なる中継にとどまらない。Leech and Short (2007: 339–241) は、作家 T. F. Powys の短編物語 *The Bucket And The Rope* を分析し、*the bucket* と *the rope* の伝達節を比較した。その際、直接話法における伝達動詞の頻度を調査し、以下の表 (Table 1) のような結果を得ている (Leech and Short, 2007: 340)。その結果、*bucket* には *exclaimed* や *laughed* といった感情的な動詞が多く用いられているのに対し、*rope* にはそうした動詞は見られないことが指摘された。また、*bucket* は *remarked* や *observed* によって発話の開始者としての役割が強調され、*rope* は *continued* や *answered* を通じて応答的な役割を果たしていることが示された。こうした伝達節の使い分けは、*bucket* の衝動的かつ感情的な性格と、*rope* の慎重で理性的な性格との対比を際立たせ、物語におけるキャラクタライゼーションの形成に重要な役割を果たしている。

Table 1: DS reporting clause verbs (Leech and Short, 2007: 340)

Verb	Bucket	Rope
exclaimed	2	0
laughed	1	0
murmured	1	0
asked	2	1
remarked	2	1
observed	4	2
continued	0	2
said	4	8
answered	0	2
replied	1	2
reasoned	1	0
suggested	1	0
Total	19	18

一方で、Semino and Short (2004) は、コーパスを用いて英語の書き言葉における発話提示パターンを包括的に分析し、伝達動詞の使用傾向を明らかにした。当該コーパスは、約 2,000 語のテキストサンプル 120 件（総語数 258,348 語）から成り、20 世紀後半の書き言葉の英国英語を対象としている。このコーパスは、プローズ・フィクション (Prose fiction: 87,709 語)、新聞報道 (Newspaper news reports: 83,603 語)、伝記・自伝 (Biography and autobiography: 87,036 語) の三つの主要ジャンルに分類され、各ジャンルには 40 件のサンプルが含まれる。また、附録には伝達動詞のアルファベット順リストが収録されており、それぞれの動詞がフィクション、新聞記事、(自) 伝記という三つのジャンルで出現するかどうかが示されている。さらに各ジャンルは「serious (硬派)」と「popular (大衆向け)」に細分され、フィクションおよび (自) 伝記においては一人称・三人称ナレーションの別も考慮されている。Semino and Short (2004) の研究は、これらの分析を通じて、英文学テキストにおける発話提示と伝達節の使用がキャラクタライゼーションや物語構造に及ぼす影響を明らかにし、発話提示研究の基盤を築いた重要な先行研究と位置づけられる。

## 2.2. Wessex 小説のキャラクタライゼーションについての先行研究

Hardy の Wessex 小説群（例：*The Return of the Native*, *Tess of the d'Urbervilles*, *Jude the Obscure*）は、架空の地方社会 Wessex を舞台とし、自然と人間、個人と社会、宿命と自由意思といった対立を描出することで知られる。Windle (1906), Lea (1925), Hawkins (1983) などの研究は、小説に描かれる世界と現実の Wessex 地域との対応関係に主眼を置いている。キャラクタライゼーションの分析においては、心理描写や社会的役割の側面に重点を置いた研究が多く（例：Weber, 1965），叙述技法、とりわけ伝達節のようなミクロなレベルの言語資源が果たす役割は、これまでほとんど考察されてこなかった。従来の Hardy 研究では、物語構造や主題分析が中心となり、対話部分の言語的特徴に体系的に焦点を当てた研究は限定的であった。しかし、登場人物の性格や関係性は、しばしば対話を通じて最も鮮明に示され

る。また、伝達節における動詞や副詞の選択は、読者の人物理解に直接影響を及ぼす可能性がある。

以上のような背景を踏まえ、本研究では従来の質的研究の蓄積を基盤としつつ、伝達節に注目した量的・質的な複合的分析を行うことで、Wessex 小説におけるキャラクター造形および語りの戦略の解明に新たな視座を提供することを目的とする。

### 3. 研究目的

本研究の目的は、Hardy の Wessex 小説における伝達節の使用実態を明らかにし、それが登場人物の性格描写にどのように寄与しているかを分析することである。具体的には、伝達節の有無、使用される伝達動詞や修飾副詞の種類、頻度、機能に着目し、Hardy の人物描写の手法を定量的かつ質的に解明する。

### 4. 研究方法

1871 年から 1897 年にかけて初版された 14 の Wessex 小説のプレーンテキストデータは、Project Gutenberg<sup>1</sup>から入手した。伝達節とキャラクタライゼンションの関係を分析するにあたって、小説の本文そのものだけではデータとして扱いにくいため、各登場人物の発話に対してそれぞれ注釈を加えた。対話部分の注釈付けは、Text Encoding Initiative (TEI) のガイドライン<sup>2</sup>に基づいたものである。具体的には、各発話について「話し手」、「聞き手」を筆者が小説の内容を詳細に読み込んだ上で判別し、TEI タグで記録した。会話データの正確性および判断基準の一貫性を担保するため、注釈は筆者自身が手作業で行い、ChatGPT をはじめとした AI ツールは一切使用しなかった。この過程は英語小説の精読、話者と聞き手の判断、タグの属性記入の繰り返しであり、長期間にわたる作業となった。

TEI タグを使用することで、小説の会話部を本文から明確に区別できるだけでなく、伝達節がどの人物の発話を修飾しているかという情報もデータとして抽出できる。このように、本研究で集めた TEI/XML ファイルは、小説会話部の分析において、価値のあるデータになったと考えられる。データを整えた後、Python スクリプトを用いて XML 解析を行い、データを抽出して量的分析を実施した。具体的には、それぞれの発話について話し手、聞き手、伝達動詞、および修飾副詞の有無を記録したデータとなった。

以下では、本研究のデータ作成および分析の各ステップについて述べる。まずは、14 冊の Wessex 小説のコーパスを構築する（第 5 節）。次に、伝達節の分析に焦点を当てる（第 6 節）。最初に、伝達節を抽出し、伝達動詞と副詞の頻度をそれぞれ分析する（第 6.1 節と第 6.2 節）。続いて、動詞と副詞の共起を調査し（第 6.3 節）、さらに、キャラクター別の分析を行う（第 6.4 節）。

<sup>1</sup> <https://gutenberg.org/>

<sup>2</sup> <https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html>

## 5. コーパス構築

Table 2: 14 冊の Wessex 小説の統計データ

	Year	Wessex Novels	Abbr.	Tokens	Types	TTR	STTR
1	1871	<i>Desperate Remedies</i>	DR	141,388	11,220	0.08	0.46
2	1872	<i>Under the Greenwood Tree</i>	UtGT	57,236	6,981	0.12	0.46
3	1873	<i>A Pair of Blue Eyes</i>	PoBE	130,745	11,328	0.09	0.46
4	1874	<i>Far from the Madding Crowd</i>	FftMC	137,802	12,041	0.09	0.45
5	1876	<i>The Hand of Ethelberta</i>	HoE	141,777	11,525	0.08	0.46
6	1878	<i>The Return of the Native</i>	RotN	142,151	10,769	0.08	0.44
7	1880	<i>The Trumpet-Major</i>	TM	113,626	10,054	0.09	0.46
8	1881	<i>A Laodicean</i>	L	138,846	11,427	0.08	0.46
9	1882	<i>Two on a Tower</i>	ToaT	94,254	9,373	0.10	0.46
10	1886	<i>The Mayor of Casterbridge</i>	MoC	116,321	10,781	0.09	0.45
11	1887	<i>The Woodlanders</i>	W	135,930	11,114	0.08	0.45
12	1891	<i>Tess of the d'Urbervilles</i>	TotD	149,685	12,862	0.09	0.45
13	1895	<i>Jude the Obscure</i>	JtO	145,238	11,355	0.08	0.44
14	1897	<i>The Well-Beloved</i>	WB	62,797	7,373	0.12	0.46

Table 2 は、本研究で構築したコーパスに含まれる Wessex 小説データの概要である。1871 年から 1897 年にかけて刊行された 14 作品を TEI 形式に準拠してタグ付けし、コーパスを構築した。「Year」の列は各小説の初版年を示している。また、「Abbr.」(Abbreviations: 略称) の列で示された小説の略称は、Hardy 研究において長年踏襲されてきたものである。「Tokens」(総語数) の列からわかるように、本コーパスに収録された作品の大半は 11 万語以上を有しており、最多は約 15 万語に達する。一方、最も短い作品は 1872 年刊行の *Under the Greenwood Tree* であり、約 6 万語である。これら 14 作品をすべて精読し、筆者自身が TEI に準拠してタグ付けを行ったものを、本研究のデータベースとした。「TTR」(Type-Token Ratio) と「STTR」(Standardized Type-Token Ratio) の列は小説語彙の多様性を示している。数値が大きければ、使用される語彙が豊富であり、作品の語彙密度が高いことを意味する。「TTR」は Type 数(異なり語数) と Token 数(総語数) の割合を示しており、作品の長さ(総語数) によって大きく左右されるため、作品間で語彙多様性を比較する際には、語数が揃っていない場合、その説得力が低下する。「Tokens」(総語数) で分かるように、長い作品と短い作品では、語数に約 3 倍の開きがあるため、ここでは参考値として提示するにとどめる。右端の「STTR」は、1000 語あたりの異なり語数を示しているが、すべての作品で 0.44~0.46 の範囲に収まっている、作品間で大きな差異は見られなかった。これより、Hardy の文体は語彙密度において一定の傾向を示しており、作品の長さに依存せず語彙使用が安定している可能性が示唆される。

## 6. 伝達節の分析

Table 3: Wessex 小説における登場人物の伝達節に使用される動詞と副詞（例）

Novel	Speaker	Listener	Verb	Adverb
1871_DR	Cytherea	Owen	said	sadly
1871_DR	Cytherea	Owen	—	—
1871_DR	Owen	Cytherea	—	—
1871_DR	Cytherea	Owen	—	—
1871_DR	Owen	Cytherea	added	gloomily
1871_DR	Owen	Cytherea	—	—
1871_DR	Cytherea	Owen	—	—
1871_DR	Owen	Cytherea	continued	—
1871_DR	Owen	Cytherea	—	—
1871_DR	Cytherea	Owen	said	firmly
1872_UtGT	Mrs Dewy	Penny	—	—
1872_UtGT	Penny	Mrs Dewy	—	—
1872_UtGT	Mrs Dewy	Penny	—	—
1872_UtGT	Penny	Mrs Dewy	—	—
1872_UtGT	Mrs Dewy	Jimmy	inquired	—
1873_PoBE	Robert	Stephen	—	—
1873_PoBE	Stephen	Robert	repeated	mechanically
1873_PoBE	Stephen	Robert	—	—
1873_PoBE	Robert	Stephen	—	—
1874_FftMC	Bathsheba	Gabriel	answered	—
1874_FftMC	Bathsheba	Gabriel	—	—
1874_FftMC	Gabriel	Bathsheba	murmured	—
1876_HoE	Christopher	Ethelberta	resumed	cheerfully
1876_HoE	Christopher	Ethelberta	—	—
1876_HoE	Ethelberta	Christopher	—	—
1876_HoE	Lady Petherwin	Ethelberta	said	—
1876_HoE	Lady Petherwin	Ethelberta	—	—
1876_HoE	Ethelberta	Lady Petherwin	exclaimed	innocently
1876_HoE	Ethelberta	Lady Petherwin	—	—

Table 3 は、Hardy の Wessex 小説から実際に抽出した伝達動詞とそれに付随する副詞の使用例を示したものである。表の各行には、話し手（「Speaker」）、聞き手（「Listener」）、使用された伝達動詞（「Verb」）、およびその動詞を修飾する副詞（「Adverb」）が記録されている。「Novel」の列には、それぞれの小説が「初版年\_略称」の形式で記載されている。Wessex 小説は 14 作品あるが、対象とするすべての小説における対話ごとの伝達動詞および副詞の組み合わせを一つの表に網羅的に示すことは現実的ではないため、ここでは各作品からいくつかの例を抜粋し、代表的な事例として提示している。選定した例はできるだけ重複を避け、小説中で連続して現れる対話部分を優先的に取り上げた。このように一部のデータを組み合わせて表を作成することで、本研究におけるデータの構成や記録方法について具体的に示すことができる。なお、Table 3 は、全データの統計的傾向や頻度分布を示すものではなく、あくまで分析に用いるデータの形式や内容を説明するための一例である。

例えば、1871 年の *Desperate Remedies* (1871\_DR) では、Cytherea が Owen に対して *said sadly* (悲しげに言った)、Owen が Cytherea に対して *added gloomily* (陰鬱に付け加えた)

といった表現が用いられている。また、1873年の *A Pair of Blue Eyes* (1873\_PoBE) では、Stephen が Robert に対して *repeated mechanically* (機械的に繰り返した) と描写されている。これらの副詞は、登場人物の感情や会話の雰囲気を補足する役割を果たしており、今後の分析ではこの点についても考察を行う。

「Verb」および「Adverb」欄が「-」で示されている場合は、該当する要素が明示されていない、すなわち、伝達節が使用されていないことを表している。小説の中では、すべての発話に対して逐一、発話者がどのように言ったかを伝達節で説明する必要はなく、しばしば引用符のみで人物 A と人物 B（場合によってはそれ以上）の会話が展開される。そのため、伝達動詞および副詞が明示されないケースはむしろ多い。伝達節が挿入されていない場合でも、改行によって発話者が変わったことが示されるため、必ずしも話者を都度明記する必要がない。このため、「Verb」および「Adverb」が「-」（空欄）となる箇所が小説内に多く存在する。この点については、小説を読んだ経験のある読者にとっては理解しやすいであろう。本研究では、上記のデータを用いて、Hardy の *Wessex* 小説に潜在するパターンを探っていく。

## 6.1. 伝達動詞の頻度分析

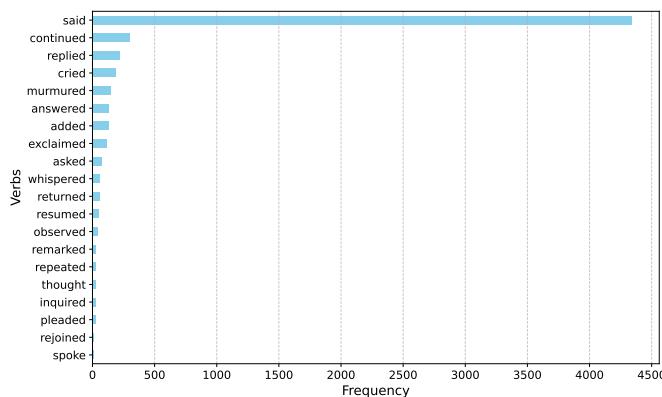


Fig. 1: 動詞の使用頻度ランキング（上位 20 語） Fig. 2: 動詞の頻度ワードクラウド（全語対象）



このセクションでは、Wessex 小説における伝達動詞について、ランキング (Fig. 1) とワードクラウド (Fig. 2) の二種類の視覚化を通じて、頻出する伝達動詞の傾向を分析する。

Fig. 1 に示す棒グラフは、上位 20 の伝達動詞の出現頻度を示している。最も多く使用されているのは *said* (言う<sup>3</sup>) という発話を示す基本的な動詞であり、圧倒的に多いことがわかる。次いで *continued*, *replied*, *cried* などの動詞がよく使われていることが確認できる。*continued* (続ける) は会話が継続する場面で、*replied* (返答する) は受け答えの際、*cried* (叫ぶ・泣き叫ぶ) は感情的な発話を際に多用される。これらの動詞は、登場人物の会話の流れや感情表現において重要な役割を果たしていることがわかる。

<sup>3</sup> 小説中では伝達動詞を過去形（例：*said*）で用いるが、本論文では記述の便宜上、日本語訳は時制を示さない原形（例：言った→言う）で統一する。

Fig. 2 は、伝達動詞の頻度に応じて単語の大きさを変えて可視化したものである。Fig. 2 のワードクラウドも、各単語の出現頻度に基づいて作成されているため、Fig. 1 と同様の傾向が見て取れる。特に、*said* が最も目立つ形で表示されており、頻出していることが視覚的にわかる。また、*murmured* (つぶやく), *whispered* (ささやく), *exclaimed* (感嘆する) など、発話の強弱や感情の込め方を示す動詞が多く含まれていることが特徴である。小説においては、単に「話す」だけでなく、どのように話しているのか（感情・口調）を伝える表現が多様に使用されていることがわかる。Fig. 1 とは異なり、Fig. 2 では、この 14 作品中に登場する全ての伝達動詞が表示されている。上位 20 語は大きめの文字で示されているが、それ以外の動詞もすべて頻度に応じて配置されており、合計 101 語が含まれている。ここから、Wessex 小説における伝達動詞の多様性、ひいては Hardy の語彙の豊富さを窺い知ることができる。

## 6.2. 伝達動詞を修飾する副詞の頻度分析

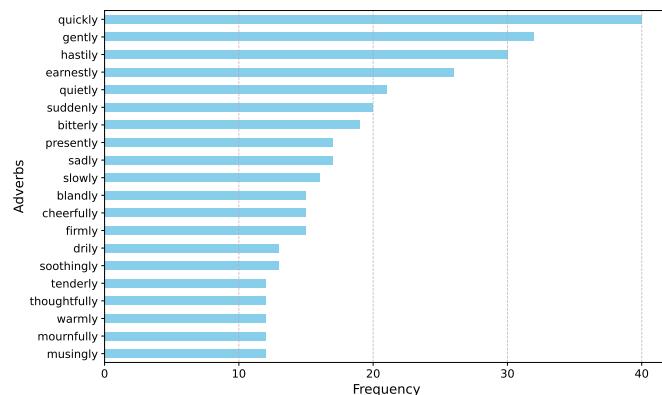


Fig. 3: 副詞の使用頻度ランキング（上位 20 語）

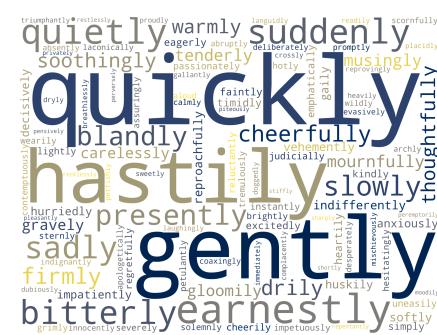


Fig. 4: 副詞の頻度ワードクラウド（全語対象）

次に、伝達節における副詞修飾の使用について検討する。このセクションでは、発話に用いられる伝達動詞を修飾する副詞の使用頻度を分析する。Fig. 3 は、最も頻繁に使用される副詞の上位 20 語を示したもので、Fig. 4 は、全体の副詞の使用傾向を視覚的に表したワードクラウドである。

まず、Fig. 3 の「副詞の使用頻度ランキング（上位 20 語）」を見てみよう。横軸（Frequency）から分かるように、頻度が 40 回を超える副詞は存在せず、Fig. 1 に示される動詞に比べ、伝達節において同じ副詞が繰り返し使用される回数は少ないことがわかる。なお、副詞が実際に伝達動詞を修飾しているかどうかは、本文中の各用例を一つひとつ確認した上で集計を行ったため、結果の精度は高いと言える。副詞の種類は合計 308 語に及び、そのうち頻度 1 のものが 165 語と半数以上を占めていた。頻度上位の副詞としては、*quickly* (素早く) や *gently* (優しく), *hastily* (急いで), *earnestly* (真剣に), *quietly* (静かに), *suddenly* (突然) などが挙げられる。これらは、登場人物の発話の速さや口調、話し手の感情やトーンを繊細に表現する役割を果たしていると考えられる。

次に, Fig. 4 のワードクラウドに注目する。頻出する副詞ほど大きく表示されており, 全体的な使用傾向を視覚的に把握できる。ここでも, *quickly*, *gently*, *hastily* といった副詞が目立ち, さらに, *warmly* (温かく), *blandly* (淡々と), *cheerfully* (陽気に) といった感情を強調する副詞の多さが確認できる。このワードクラウドには, 合計 308 語の副詞がすべて含まれており, 特に *-ly* で終わる副詞や比較的長い語, 難解で文学的な語彙が多く含まれている点は注目に値する。

この分析から, 登場人物の発話には, 伝達動詞とともに様々な副詞が組み合わされことで, 感情や話し方のニュアンスが豊かに表現されていることが明らかとなった。特に, 話し手の感情や発話の速度・強さを表す副詞が頻繁に用いられていることが顕著である。

### 6.3. 副詞修飾の共起分析

このセクションでは, 伝達動詞と副詞修飾の共起分析 (Collocation Analysis) について説明する。こここの「共起」は, Wessex 小説において登場人物が発話する際に使用される伝達動詞と, それを修飾する副詞の「組み合わせ」のことである。小説内の発話伝達節において, 特定の動詞と副詞がどのように共起するのかを明らかにすることで, 登場人物の発話パターンの特徴を把握することを目的としている。

Fig. 5 は, Wessex 小説の伝達節における動詞-副詞の共起を示したバブルチャート (動詞と副詞はそれぞれ上位 20 語<sup>4</sup>) である。バブルチャート (Bubble Chart) とは, データの関係性を視覚的に表現する散布図 (Scatter Plot) の一種である。通常の散布図 (点を使ってデータを示す) とは異なり, データの頻度や重要度をバブルの大きさと色で表現する。このチャートでは, 縦軸に伝達動詞を修飾する副詞 (Adverb), 横軸に登場人物の発話に使われる伝達動詞 (Verb) を配置し, それぞれの組み合わせの頻度をバブル (円) の大きさと色で表している。円が大きいほど, その動詞と副詞の組み合わせが頻繁に登場することを意味する。色も頻度を示しており, 頻度が高いほど赤く, 低いほど青く表示される。

バブルチャートを用いることで, 登場人物の発話がどのような動詞と副詞の組み合わせで表現されるかを一目で把握できる。特定の動詞がどのような感情や話し方のニュアンスとともに使われているかを視覚的に捉えるのに適した分析方法である。

<sup>4</sup> 本研究で使用したデータには, 伝達動詞が合計 101 語, 副詞が 308 語含まれている。すべてのペアの共起関係を視覚化すると視認性が低下するため, 図の可読性を確保する目的で, 頻度の高い順に, 伝達節に現れる伝達動詞上位 20 語と, それを修飾する副詞上位 20 語を可視化の対象とした。

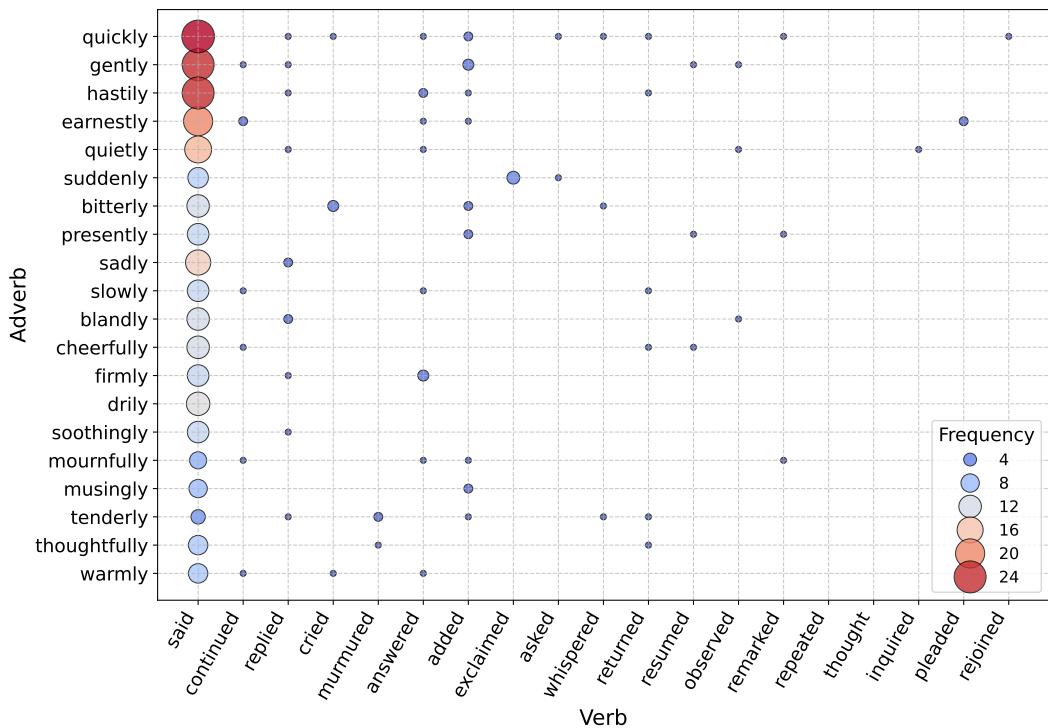


Fig. 5: 動詞-副詞共起バブルチャート（上位 20 語）

バブルチャート (Fig. 5) から、いくつかの興味深いパターンが確認できる。まず、*said* (言う) という動詞は、最も多くの副詞 (頻度上位 20 の副詞全て) と共に起しておらず、特に *quickly* (素早く) : 25 回, *gently* (優しく) : 24 回, *hastily* (急いで) : 24 回、などが頻繁に使われている。これは、一般的な発話動詞として *said* が多様な発話状況を表現するために広く使用されていることが要因と考えられる。

一方で、*continued* (続ける), *replied* (返答する), *answered* (答える), *added* (付け加える) などの発話の流れを示す動詞は、*earnestly* (真剣に), *saidly* (悲しげに), *blandly* (淡々と), *firmly* (しっかりと), *gently* (優しく) といった態度や感情を補足する副詞と共に起しており、会話のトーンや意図を明確にするために使われている。

また、*cried* (叫ぶ・泣き叫ぶ), *exclaimed* (感嘆する) など感情表出の強い動詞は、*bitterly* (激しく・苦々しく), *suddenly* (突然) といった感情の激しさや突発性を表す副詞とよく共起する。これは、感情の強度を強調する意図があると考えられる。

さらに、*murmured* (つぶやく), *whispered* (ささやく) など似たようなイメージを持つ静かな発話動詞は、このチャートから、それぞれ使い方のパターンが見られる。前者は *thoughtfully* (思慮深く) と共に起し、控えめな感情や内省的な態度を表現する役割がある。後者は *bitterly* (激しく・苦々しく), *quickly* (素早く) と共に起しており、抑えきれない感情や緊張感を伴う場面を表している。

他にも、伝達動詞自体の特徴が窺える。例えば、*observed* (指摘する) は、*gently* (優しく), *quietly* (静かに), *blandly* (淡々と) と共に起し、客観的または冷静な態度を伝えること

が多い。このように、動詞と副詞の共起パターンから、各動詞がどのようなトーンや感情とともに用いられているのかが読み取れる。これにより、小説の中で特定の伝達動詞がどのようなニュアンスを持って使われているのかが浮かび上がった。

この共起分析から、登場人物の発話には、特定の動詞と副詞の組み合わせがある程度決まっていることが分かった。特に、話し手の感情や発話の強弱を表現するために、副詞が効果的に使われていることが確認された。

#### 6.4. 人物別の伝達節分析（事例分析）— *Tess of the d'Urbervilles* を例に—

次に、主要登場人物と伝達動詞の関係について説明する。このセクションでは、代表作 *Tess of the d'Urbervilles* を例に、小説に登場する主要人物の発話に用いられる伝達動詞の種類を分析する。登場人物ごとに、どのような伝達動詞が使われているかを視覚化することで、それぞれのキャラクターの発話スタイルや特徴を明らかにすることを目的とする。

Fig. 6 は、主要登場人物である「Tess」、「Angel」、「Alec」の3人を対象に、発話時に使用される伝達動詞のネットワークを示している。この図では、3人の発話に使われる伝達動詞を、それぞれ異なる色で分類している。Tess は赤、Angel は青、Alec は黄色で示され、複数の人物に共通する動詞は混合色（グレーなど）で表示される。また、ノードの大きさは、各キャラクターがその動詞をどの程度使用したかを示している。ネットワークの中心部にある動詞を除き、周辺に配置されている動詞のノードサイズはほぼ同じである。これは、多くの動詞が一度しか使用されていないためである。このネットワークは頻度差よりも、登場人物と伝達動詞の結びつきを可視化するのに役立つ。

この3人が発話の際に実際に使用した伝達動詞の種類とその頻度は、Table 4 にまとめてある。頻度から見ると、女性主人公である Tess の発話には、作家 Hardy がより多様な伝達動詞を用いていることがわかる。女性主人公の心理を男性主人公よりも繊細に描写していることは、Hardy 作品における「フェミニズム」の主題が反映されていると言える。

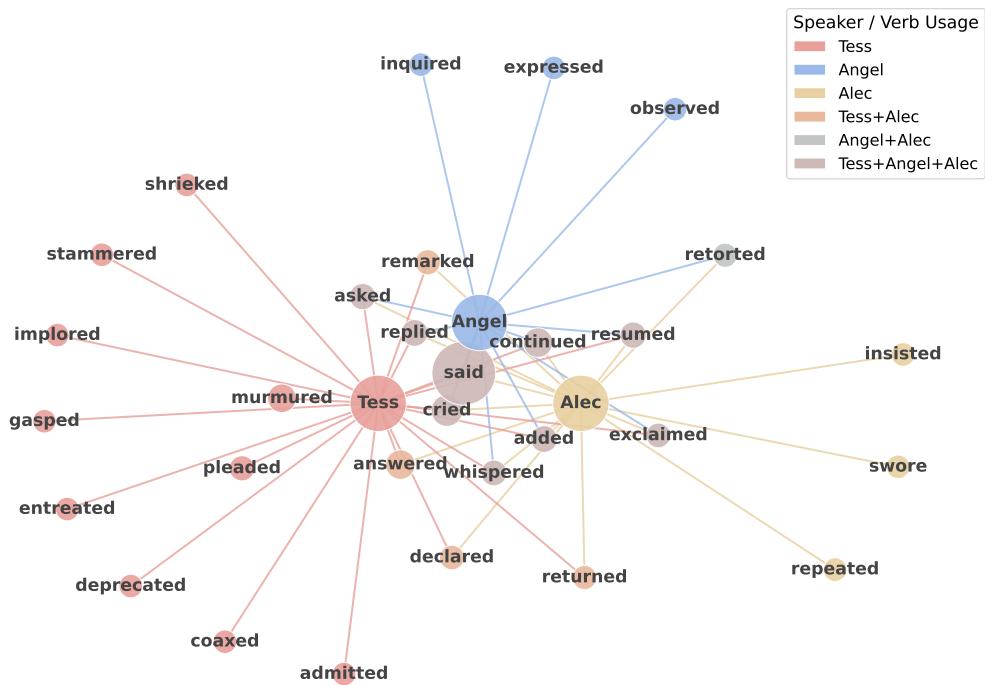


Fig. 6: 主要登場人物と伝達動詞のネットワーク可視化

Table 4: Tess, Angel, Alec と伝達動詞

Tess			Angel			Alec		
No.	Verb	Japanese	No.	Verb	Japanese	No.	Verb	Japanese
1	said	言う	1	said	言う	1	said	言う
2	cried	叫ぶ・泣き叫ぶ	2	continued	続ける	2	continued	続ける
3	answered	答える	3	cried	叫ぶ・泣き叫ぶ	3	resumed	再開する
4	murmured	つぶやく	4	added	付け加える	4	added	付け加える
5	pleaded	懇願する	5	asked	尋ねる	5	cried	叫ぶ・泣き叫ぶ
6	replied	返答する	6	whispered	ささやく	6	answered	答える
7	asked	尋ねる	7	exclaimed	感嘆する	7	replied	返答する
8	continued	続ける	8	expressed	表現する	8	asked	尋ねる
9	remarked	述べる	9	inquired	尋ねる	9	declared	宣言する
10	added	付け加える	10	observed	指摘する	10	exclaimed	感嘆する
11	declared	宣言する	11	replied	返答する	11	insisted	強く主張する
12	whispered	ささやく	12	resumed	再開する	12	remarked	述べる
13	admitted	認める	13	retorted	言い返す	13	repeated	繰り返す
14	coaxed	なだめる				14	retorted	言い返す
15	deprecated	非難する				15	returned	返す
16	entreated	懇願する				16	swore	ののしる
17	exclaimed	感嘆する				17	whispered	ささやく
18	gasped	はっと息を呑む						
19	implored	嘆願する						
20	resumed	再開する						
21	returned	返す						
22	shrieked	金切り声を上げる						
23	stammered	口ごもりながら言う						

ここからは、ネットワーク図 (Fig. 6) と頻度表 (Table 4) を参照しながら、人物別の伝

達節分析をさらに詳細に行い、各登場人物の発話における伝達動詞の使用傾向について説明する。ただし、*said*（言う）、*cried*（叫ぶ・泣き叫ぶ）、*answered*（答える）、*replied*（返答する）、*asked*（尋ねる）、*continued*（続ける）などの基本的な伝達動詞は、主要人物・脇役を問わず、キャラクター全員に共通して使われているため、ここではこれらを特別に検討しないこととする。その代わりに、3人それぞれに特徴的な動詞に着目する。

まず、女性主人公Tessが独自に使用する動詞には、「懇願」や「感情的な発話」に関するものが多く含まれている。具体的には、*pleaded*（懇願する）、*entreated*（懇願する）、*implored*（嘆願する）などの切実な訴えを表す動詞、*gasped*（はっと息を呑む）、*shrieked*（金切り声を上げる）、*stammered*（口ごもりながら言う）などの強い感情や驚きを表す動詞、さらに*murmured*（つぶやく）といった静かな発話を表す動詞や、*coaxed*（なだめる）、*deprecated*（非難する）といった感情的な発話に関連する動詞が多く確認される。これにより、Tessが物語の中で困難な状況に直面し、しばしば強い感情を伴って発話し、苦境に立たされていることが示唆される。

一方、男性主人公の一人であるAngelは、「知的で冷静な発話」が多い。*expressed*（表現する）、*observed*（観察する）、*inquired*（尋ねる）といった、客観的・理性的な発話を示す動詞が多く使用されている。これは、Angelの発話が比較的冷静かつ理知的であることを反映していると考えられる。しかし、彼も時には感情を露わにする場面がある。Tessに比べるとその頻度は低いが、*cried*（叫ぶ・泣き叫ぶ）といった感情的な発話も確認できる。また、*retorted*（言い返す）も1回使用され、議論や対立的な場面が一部含まれていることが分かる。このように、Angelは普段は冷静に発話するが、特定の状況では感情を表出する一面も持つことが読み取れる。

もう一人の男性主人公であるAlecは、「自己主張的で強制的・支配的な発話」が多い傾向にある。*insisted*（強く主張する）、*repeated*（繰り返す）などの動詞が使われており、Alecが相手に対して自己の意見を強く押し付ける傾向がうかがえる。また、*swore*（罵る<sup>5</sup>）が含まれており、攻撃的または支配的な発話をを行うことが示されている。さらに、*retorted*（言い返す）もAlecの発話に使われており、対立的な場面があることがわかる。このように、Alecの発話における伝達動詞の分析から、彼のキャラクターが自己主張が強く、相手に対して自分の意見を押し付けたり、時には感情的かつ支配的な対話を行うことが明らかとなった。

分析の結果をより明確に整理すると、以下のような結論が導き出される。Tessは感情的な発話が多く、懇願や嘆願を表す動詞が伝達節に頻繁に用いられている。Angelは冷静かつ知的な発話が中心で、観察や質問に関する伝達動詞が多く使用される。一方、Alecは自己主張が強く、攻撃的・支配的な発話が特徴的である。実際に小説を読んだ読者にとって、これら三人の性格は、使用される伝達動詞から浮かび上がるイメージと一致するのではなかろうか。このことから、伝達動詞が小説におけるキャラクター描写において、いかに重要な役割を果たしているかが、ここから読み取れる。

<sup>5</sup> “Now, damn it—I'll break both our necks!” *swore* her capriciously passionate companion. “So you can go from your word like that, you young witch, can you?” (*Tess of the d'Urbervilles*, Phase the First: The Maiden, VIII) これは、Alec (*her capriciously passionate companion*) が Tess に対する発話で彼女を罵った場面である。*swore* という語は「誓う」という意味もあるが、ここでは「罵る」の意味に当てはまる。この伝達動詞によって、Alec の性格の悪さが読者に印象づけられる。

## 7. まとめと今後の課題

本研究では、Wessex 小説における登場人物の発話に用いられる伝達動詞と副詞の分析を行い、以下のような重要な知見を得た。

まず、Wessex 小説における伝達動詞および副詞の種類（多様性）が明らかとなり、作家 Hardy がキャラクターの発話時における心理描写において、繊細かつ力強い表現力、ならびに豊富な語彙力を有していることが示された。また、伝達動詞と副詞の組み合わせを分析することで、登場人物の感情や話し方のニュアンスがどのように表現されているかを明らかにした。例えば、*cried* は *bitterly*, *exclaimed* は *suddenly* といった副詞と共に起ることが多く、感情表出の強い動詞は感情の激しさや突発性を示す副詞と頻繁に組み合わされることが確認された。一方で、感情的により中立的な *observed* は、*gently*, *quietly*, *blandly* など、客観的または冷静な態度を表す副詞と共に起する傾向が強い。これにより、登場人物の発話の特徴が、伝達節における副詞によって効果的に反映されていることも確認された。さらに、登場人物ごとの伝達動詞の使用傾向には明確な違いがあり、キャラクターの性格や関係性が発話スタイルに反映されていることが明らかとなった。特に、Tess は感情的な発話が多く、Angel は知的・冷静な発話が中心であり、Alec は強制的・支配的な発話が多いことが分かった。このように、伝達動詞の選択を通じて、登場人物の性格や発話スタイルが的確に描写されていることが示された。

本研究では、伝達動詞と副詞の分析を通じて登場人物の発話パターンを明らかにしたが、今後のさらなる発展に向けて、以下の課題が考えられる。(1) 他の小説との比較分析：本研究では、代表作 *Tess of the d'Urbervilles* を中心に分析を行ったが、他の Wessex 小説にも同様の傾向が見られるかどうかを比較分析することで、作家の文体や作品ごとの発話パターンの違いを明らかにする必要がある。(2) より詳細な感情分析：伝達動詞と副詞の組み合わせだけでなく、文脈を考慮した感情分析（ポジティブ・ネガティブの分類など）を行うことで、登場人物の発話における感情的な変化をより詳細に捉えることができる。(3) 登場人物の関係性と発話の関連性：登場人物同士の関係（敵対・恋愛・友情など）が発話スタイルにどのように影響を与えていているのかを分析することで、物語の構造をより深く理解することが可能となる。(4) 時間的変化の分析：物語の進行に伴い、登場人物の発話スタイルがどのように変化するのかを調査することで、キャラクターの成長や心理的変化をより明確に把握できる。

以上のように、本研究では、小説における発話表現を定量的に分析し、登場人物の性格や物語の展開がどのように言語的に表現されているのかを明らかにした。今後は、他作品との比較や、より高度な感情分析を進めることで、文学作品の発話表現に関するさらなる洞察を得ることが期待される。

## Bibliography

[1] Hardy, T. (1992). *Far from the Madding Crowd*. Project Gutenberg. Online resource. (Last accessed 13 March 2023.) Available online at <https://www.gutenberg.org/cache/epub/27/pg27-images.html>

- [2] Hardy, T. (1994). *Jude the Obscure*. Project Gutenberg. Online resource. (Last accessed 13 March 2023.) Available online at <https://www.gutenberg.org/cache/epub/153/pg153-images.html>
- [3] Hardy, T. (1994). *Tess of the d'Urbervilles: A Pure Woman*. Project Gutenberg. Online resource. (Last accessed 26 May 2022.) Available online at <https://www.gutenberg.org/cache/epub/110/pg110-images.html>
- [4] Hardy, T. (1996). *The Woodlanders*. Project Gutenberg. Online resource. (Last accessed 13 March 2023.) Available online at <https://www.gutenberg.org/cache/epub/482/pg482-images.html>
- [5] Hardy, T. (2001). *The Trumpet-Major*. Project Gutenberg. Online resource. (Last accessed 13 March 2023.) Available online at <https://www.gutenberg.org/cache/epub/2864/pg2864-images.html>
- [6] Hardy, T. (2001). *Under the Greenwood Tree*. Project Gutenberg. Online resource. (Last accessed 13 March 2023.) Available online at <https://www.gutenberg.org/cache/epub/2662/pg2662-images.html>
- [7] Hardy, T. (2002). *The Hand of Ethelberta*. Project Gutenberg. Online resource. (Last accessed 13 March 2023.) Available online at <https://www.gutenberg.org/cache/epub/3469/pg3469-images.html>
- [8] Hardy, T. (2002). *Two on a Tower*. Project Gutenberg. Online resource. (Last accessed 13 March 2023.) Available online at <https://www.gutenberg.org/cache/epub/3146/pg3146-images.html>
- [9] Hardy, T. (2006). *The Mayor of Casterbridge*. Project Gutenberg. Online resource. (Last accessed 13 March 2023.) Available online at <https://www.gutenberg.org/cache/epub/143/pg143-images.html>
- [10] Hardy, T. (2006). *The Return of the Native*. Project Gutenberg. Online resource. (Last accessed 13 March 2023.) Available online at <https://www.gutenberg.org/cache/epub/122/pg122-images.html>
- [11] Hardy, T. (2008). *A Pair of Blue Eyes*. Project Gutenberg. Online resource. (Last accessed 13 March 2023.) Available online at <https://www.gutenberg.org/cache/epub/224/pg224-images.html>
- [12] Hardy, T. (2009). *A Laodicean*. Project Gutenberg. Online resource. (Last accessed 13 March 2023.) Available online at <https://www.gutenberg.org/cache/epub/3258/pg3258-images.html>
- [13] Hardy, T. (2009). *Desperate Remedies*. Project Gutenberg. Online resource. (Last accessed 13 March 2023.) Available online at <https://www.gutenberg.org/cache/epub/3044/pg3044-images.html>
- [14] Hardy, T. (2009). *The Well-Beloved*. Project Gutenberg. Online resource. (Last accessed 13 March 2023.) Available online at <https://www.gutenberg.org/cache/epub/3326/pg3326-images.html>
- [15] Hawkins, D. (1983). *Hardy's Wessex*. Macmillan.
- [16] Lea, H. (1925). *Thomas Hardy's Wessex* (Pocket ed.). Macmillan.
- [17] Leech, G. and Short, M. (2007). *Style in Fiction: A Linguistic Introduction to English Fictional Prose*, Second edition. Pearson Longman.
- [18] Page, N. (1973). *Speech in the English Novel*. London: Longman.
- [19] Semino, E., and Short, M. (2004). *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. Routledge.
- [20] 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, pp. 79 – 106.
- [21] Weber, C. J. (1965). *Hardy of Wessex: His Life and Literary Career*. New York: Columbia University Press.
- [22] Windle, B. C. A. S. (1906). *The Wessex of Thomas Hardy*. J. Lane, Bodley Head.
- [23] 曹芳慧. (2023). *Tess of the d'Urbervilles* の会話部によるキャラクタライゼーション. 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』(大阪大学大学院人文学研究科言語文化学専攻), 2023, pp. 59–78.
- [24] 豊田昌倫, 堀正広, 今林修. (2017). 『英語のスタイル—教えるための文体論入門』. 東京: 研究社.



## 執筆者紹介

田畠 智司 (たばた ともじ)	大阪大学大学院人文学研究科教授
黄 晨雯 (こう しんぶん)	大阪大学大学院人文学研究科助教
藤田 郁 (ふじた いく)	比治山大学 現代文化学部言語文化学科講師
曹 芳慧 (そう ほうけい)	大阪大学 大学院人文学研究科博士後期課程
Dimitra Vogatza (ディミトラ ヴォガジヤ)	大阪大学 大学院人文学研究科博士後期課程

言語文化共同研究プロジェクト 2024

テクストマイニングとデジタルヒューマニティーズ2024

2025年5月31日発行

編集発行者

大阪大学大学院人文学研究科言語文化学専攻