The University of Osaka
Institutional Knowledge Archive

Title	ビッグデータとしてのウェブ小説ー言語特徴およびトレンド解析ー
Author(s)	黄, 晨雯
Citation	大阪大学, 2022, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/89502
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

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

The University of Osaka

様式３

論　文　内　容　の　要　旨

氏　名　（　　黄　晨雯　　）	
論文題名	ビッグデータとしてのウェブ小説 ―言語特徴およびトレンド解析―

論文内容の要旨

インターネットとデジタル化の急速な進展に伴い、デジタル媒体で創作され、消費されるウェブ小説が年々増加の一途をたどり、サブカルチャーの一つとして人気を博している。本論文では、2021年までの日本のウェブ小説を対象として、小説のタイトルとジャンルなどのメタ情報をはじめ、本文と読者のコメントを含む膨大な量のデータを収集して解析を行い、特に大量の投稿作品のうち、読者の好みを反映する人気作品の特徴を解明することを目的とする。

第1章では本論文の背景とウェブ小説に関わる先行研究を概観した上で、リサーチクエスションを設け、研究目的と本論文の構成を示した。

第2章では研究対象とするウェブ小説投稿サイト「小説家になろう」および「アルファポリス」の特徴を述べて、小説のメタデータならびにその取得方法について述べた。また、取得したデータをもとに、各項目情報の分析手順、および本研究が拠り所とするコーパスの構造について詳述した。

第3章では数値データに注目し、主に「小説家になろう」と「アルファポリス」のそれぞれの作品数に着目し、投稿される作品の傾向を時系列的に分析した。これら二つのウェブ小説サイトに共通するのは、投稿作品数が年々上昇し、ウェブ小説への関心の高まりが観察できることである。ジャンルに関してはファンタジーと恋愛が人気を二分している。「小説家になろう」の細かなジャンル分類を行うと、異世界を舞台とするハイファンタジー物語の作品数が圧倒的に多いことが明らかになった。ただし、ここ3年では異世界恋愛ジャンルが新たなトレンドとなり人気を博していることが見てとれる。またVRゲームを題材にした小説の投稿数はまだそれほど多くないが、このジャンルに強い興味を抱くファンダムが存在し、人気度の高いジャンルであることは否定できない。一方で、「アルファポリス」においてSFはそれほど重要視されていないことをデータで示した。

第4章ではウェブ小説のタイトルに焦点を当てた考察を行った。まずはタイトルの長さ、つまり文字数に着目し、時系列的な変化を考察した。その結果、近年だんだんと長いタイトルが用いられる傾向があることを明らかにした。特に人気の高いファンタジージャンルにおいてその趨勢が一層目立つ。その一因として、投稿サイトではタイトルを長くすることで他の作品との差別化を図り、読者の注意をより惹きつけようというストラテジーが取られているという解釈を提示した。次に、タイトルにおける語彙の多様性に焦点を当て解析を行った結果、語彙の多様性は年とともに減少していることが明らかとなった、特に人気度の高い作品ほどその傾向が激しい。ジャンルごとに分類して考察したところ、SFと文芸に比べ、人気ジャンルである恋愛とファンタジーの平均語彙多様性が極めて低いことが判明した。さらに、タイトルにおける語彙の使用頻度を分析対象としてコレスポンデンス分析手法を用いて実行した結果、ジャンル間（異世界と現実世界、恋愛作品と非恋愛作品）における明らかな差異と、投稿年による明白な違いが見受けられた。ファンタジーの方では英雄や勇者など冒険物語から異世界に転生するファンタジーへと変化し、また最近では追放ファンタジーへと流行が推移していることが明らかになった。異世界恋愛の方では婚約をめぐる展開と、悪役令嬢を主人公とするパターンが近年において非常に好まれるようになったことがデータに反映されている。また、人称代名詞が文芸や恋愛ジャンルのタイトルにおいて愛用されている事実に着目した。「君」が初期の現実恋愛作品のタイトルに多く出現しているのに対して、近年の恋愛作品では一人称の「私」が主流である。人気作品では内容自体と関係性がない「書籍」「化」が特徴的に出現している。こうした語句がタイトルにおいて多用されるのは、宣伝効果をあげるための追加説明を付与しやすいためであるという考察を行った。こうした追加情報が挿入されていることで、人気作品のタイトルの語数増加に寄与していることが読み取れる。また、近年の現実恋愛作品では「幼馴染」という設定が基調を成していることも興味深い現象である。

第5章ではTop2Vecという新たなトピックモデルを用いてあらすじに対するトピック分析を行った。LDAモデルなど従来のトピックモデルと比べ、短文処理の性能が高いことならびに機能語などのストップワーズをあらかじめ分析対象から除外する前処理を必要としない点で従来型のトピックモデルよりも処理効率が高いという利点がある。Top2Vecを

用いてあらすじの時系列的な変化を考察した結果、トピックの構成比率をもとに、2004年~2011年、2012年~2017年、2018年~2021年の3つのグループに分類することができた。2011年以前では、切ない片思いに関するトピックが高い割合を占めており、甘酸っぱい恋、そして動揺する心の描写を基調にした作品が人気を博していた。人生の変わり目、運命の転機を象徴するトピックも多く出現していた。2012年からはvrmmorpg（VRゲーム）のトピックの興隆が顕著となった。この背景としてオンラインゲームの爆発的人気沸騰があると筆者は推察している。また、作者と読者のコミュニケーションも緊密になり、2012年からは作者からの誤字脱字の指摘願いに関わるトピックが高い割合を占めている。2018年前後から急上昇が見られるのは皇族・貴族に関わる婚約破棄と、パーティーから追放されるファンタジーストーリーである。この2つに続くトレンドとして、女主人公が乙女ゲームに悪役令嬢として転生するトピックも高い人気を博している。これらの現象に関しては、タイトルの語彙分析でも類似した特徴語が観察されており、この解釈を裏書きする要素となっている。また人気作品に高い割合を占めるトピックを考察すると、VRゲームと負け組の異世界転生が最初のブームであったが、2016年以降では人気作品の上位トピックが婚約破棄、追放系ファンタジーと悪役令嬢で定着し、トピックの多様性が次第に低下していることが明らかになった。人気作品特有の要素として、書籍化の宣伝もしくは小説を応援したことへの感謝を表すトピックも多く出現しており、人気作品独特の著者―読者間のコミュニケーションが観察される。他にも、各ジャンルに対するミクロ分析を行い、それぞれのジャンルにおける流行トピックの興廃変化を解明し、人気作品とそれ以外の作品のトピックの相違を併せて詳細に検討した。

第6章では分析対象を人気作品に絞り、それぞれ第一話をスクレイピングし、第5章と同じ研究手法であるTop2Vecを用いて、人気作品の冒頭部で頻繁に使用されるトピックを解明する試みを行った。第5章では婚約破棄および主人公がパーティから追放されるプロット展開が読者に好まれている様相を明らかにしたが、これら2つのトピックをテーマとする作品の冒頭部における典型的な書き出しは、それぞれあらずじ自体の内容と高い類似性が認められる。つまり、主人公が婚約破棄やパーティから追放されるなど憂き目に遭う物語では、作品の中心的な内容を最初に述べる特徴があると考えられる。また、皇族をめぐる婚約破棄に起因する主人公の苦悩、懊悩など情緒を中心としたトピックも冒頭部に多く使用されている。凡庸に見える主人公がパーティから追放される物語では、教会の儀式で主人公がスキル鑑定を受け、役立たないと判定される場面も冒頭部に多く用いられることが明らかになった。また、異世界転生に焦点を当て考察した結果、ファンタジー作品に関しては主人公が異世界に足を踏み入れるに至る原因や、元の世界で起きたハプニングなどについて詳細に描かれることが多い。一方で、恋愛ジャンルではあまり転生の理由や経緯は重要ではなく、暗示される物語の結末に対し、主人公が既定の運命と対峙するプロット展開が多く、主要登場人物の関係性に関する説明が冒頭に描写される仕組みとなっている。さらに、性格や外見、出身などキャラクターの設定に関するトピックも第一話で頻繁に使用されることが明らかである。異世界恋愛ジャンルでは、瞳や髪に関わる外見、地位（王子、令嬢、伯爵、公爵など）および家族関係（兄、弟、父、母など）が重点的に描写されるのに対して、ファンタジーの方では身長の高さがより重視され、スキル、武器、職業、種族、所属などファンタジー定番の属性が特徴的に描写されていることを定量的に示した。

最後に第7章では「小説家になろう」サイトから抽出した読者の大量な感想コメントをTop2Vecで解析した。その結果、2012年以前のコメントは全体として「丁寧」を表象するトピックを構成しているのに対して、2013年以降では、より気軽な「面白さ」を評価軸にするトピックが目立ち、かつSFとファンタジーの肝心な要素であるスキルやアイテムなどをめぐる議論と異世界恋愛の重要な登場人物である貴族の諸々の事情に関わるコメントも非常に多く付与されている。気になる点として強めに指摘されているのはスキルやアイテムなどの戦闘系要素と、文章の表現問題である。文章の表現に関する指摘は改行と句読点のような視覚と直接関わることが大部分を占めている。また、読者による問題点の指摘に関するコメントでは、丁寧な態度を取り著者の特権性を尊重するものより、役が立つアドバイスの方が採用されやすい。

本論文は作品のメタデータならびにタイトル、本文からなるビッグデータを自然言語処理の技法を用いて分析することで顕在化するウェブ小説のトレンドをつかむ試みであった。機械学習は大量なデータの中から有用な情報を取り出してくれるため、研究者が遠読の視点から入った上でミクロな分析を結合し、より客観性を踏まえた解釈ができるという利点をもつ。本研究がデジタルヒューマニティズによるサブカルチャー研究へ新視点をもたらす一助になれば幸いである。一方、中国においてもウェブ小説が非常に流行りのサブカルチャーに発展してきていることは注目に値する文化現象であり、本研究で行われた分析結果は今後の中日のウェブ小説の比較にもつながるのではないかと思われる。本研究が定量的な手法により、文化の異同を考察するための一助となることを期待する。

様式 7

論文審査の結果の要旨及び担当者

氏　名　（黄　晨雯）		
論文審査担当者	（職）	氏　名
主 査	教授	田畑 智司
副 査	准教授	ホドシチェク ボル
副 査	教授	岩居 弘樹

本博士学位請求論文は，最新の自然言語処理技術を応用し，ビッグデータとしてのウェブ小説の言語特徴とトレンドの解析を試みた研究である。デジタル媒体で創作され，読まれるウェブ小説は年々増加の傾向にあり，現代サブカルチャーの一つとなっている。本論文は，2021 年までに公開された日本のウェブ小説作品のメタデータをはじめ，本文と読者のコメントを含む膨大な量のデータを収集し，無数にある作品の中からどのような小説が読者の人気を勝ち得ていくのかという問いに，データ駆動型の分析を通して答えた研究である。

本論文は全 8 章からなる。第 1 章では，論文の背景としてウェブ小説に関わる先行研究を概観し，知見を整理した上でリサーチクエスションを設け，研究目的と論文の構成を示している。

第 2 章は，研究対象とするウェブ小説投稿サイト「小説家になろう」および「アルファポリス」の特徴を記述し，メタデータの構成とその取得方法，ウェブ・スクレイピングの技法を説明する一方，取得したメタデータならびに本文と読者コメントからなるウェブ小説コーパスの構造について詳述している。

第 3 章では，二つのウェブ小説サイトそれぞれにおいて掲載される作品の傾向を時系列的に分析している。二つのサイトに共通しているのは，投稿作品数が年々増加し，ファンタジーと恋愛物という 2 つのジャンルが人気を分けていることである。黄氏は詳細なジャンル分類により，異世界を舞台とするハイファンタジー，特に，異世界恋愛をテーマにした作品が新たなトレンドとなり人気を博していることなどをデータによって明示している。

第 4 章は，ウェブ小説のタイトルに焦点を当て，タイトルの文字数を時系列データとして調査した結果，近年，長いタイトルが付けられる傾向があること，中でも人気の高いファンタジーにおいてそれが顕著であることを指摘している。長いタイトルは，前景化効果をもち，競合する他の作品との差別化を図るためのストラテジーとなっているが，その一方で，対照的に，タイトルの語彙多様性は漸次的に減少しているという事実が明らかにされているのは実に興味深い。人気の高い作品ほどその傾向が顕著であり，特に，恋愛物とファンタジーでは，語彙多様性が極めて低いことが指摘されている。さらに，多変量解析のコレスポンデンス分析による多次元マッピングにより，作品ジャンル間の差異を明示するとともに，ファンタジーでは英雄や勇者などを主人公とする冒険物語から，異世界転生をテーマとしたファンタジーや，さらには「追放系」ファンタジーと称されるジャンルへの通時的トレンドの推移が生じていることが可視化されている。

第 5 章は，新たなトピックモデル Top2Vec (Angelov, 2020)を適用し，作品の「あらすじ」に生起するトピックを扱った章である。Top2Vec は，短文処理に適している(Angelov; 黄, 2021)のに加え，機能語など，ノイズとなりうる項目を分析対象から除外する事前処理を必要としないため，処理コストが低いという利点がある。Top2Vec を用いたあらすじのトピックモデリングを行った結果，黄氏はあらすじのトレンドが，2004 年~2011 年，2012 年~2017 年，2018 年~2021 年の 3 つの時系列グループに分類できることを指摘している。その上で，人気作品の特徴として，書籍化の宣伝や読者による作品支援に対する著者からの感謝を表象するトピックが生起している事実を掘り起こし，人気作品では，あらすじにおいて，著者と読者間のコミュニケーションが形成されていると論じている。その他，各ジャンルに焦点を当てたマイクロトピックの分析を行い，ジャンル内でのトピックの興廃変化に，人気作品と他の一般作品との相異が反映していることを明らかにしている。

第 5 章で示された知見をさらに深めるべく，第 6 章では対象を人気作品に限定し，Top2Vec を用いて，人気作品の冒頭部で頻繁に使用されるトピックの解明を試みている。著者によると，作品の冒頭部のトピック構成は，作品のあらすじと高い類似性が認められるという。その理由として，人気ウェブ小説では，登場人物の出自や社会的地位，外見や性格など，まず主人公の「キャラクタライゼイション」が周到に行われ，読者とテクスト世界の共有が図

られていることをヒートマップやワードクラウド，頻度グラフなど客観的なデータによって裏付けており，説得力のある議論が展開されている。

第 7 章は，「小説家になろう」サイトからスクレイピングにより取得した膨大な量の読者コメントを Top2Vec で解析し，作品に対する reader-response（読者の反応）を論じた章である。Top2Vec の分析結果から読み取れることは，2012–2013 年を境として，読者コメントの傾向が異なることであり，2012 年以前は全体として「丁寧さ」と関わるトピック，すなわち，作者に対して一定の距離を保ったコメントが支配的であるのに対し，2013 年以降になると，読者側の評価軸と関わるトピックが顕著になっていること，換言すれば，読者から作者へのアドバイス，リクエストを表すコメントが増えるなど，読者がより積極的に制作側に関与しようという姿勢が強くなるなど，読者が作品の受動的な消費者から，ウェブ小説というサブカルチャーの積極的な参与者になってきていることが浮き彫りにされており，非常に興味深い。

本論文が用いた機械学習という手法は，大量のデータから有用な情報を掘り起こすことを支援し，分析者が巨視的視点で発見した現象を微視的に考察する糸口を与えてくれる。この手法を文化データの分析に効果的に活かすには，ズームアウトしてデータの全体像を把握する一方，個々のテクストにズームインすることで，マクロな分析で得た仮説をミクロな分析やアウトプットの質的解釈により検証し，それをマクロな分析にフィードバックし，さらに仮説を精緻化していくという仮説発見と検証の循環を築くことが不可欠である。黄 晨雯氏の論文は，その点において，定量的手法とアウトプットの質的な読みをうまく組み合わせて，ウェブ小説の特徴を手堅く記述しており，客観性，再現性や反証可能性を担保するための科学的な手順を踏まえた研究となっている。本博士学位請求論文は，自然言語処理と機械学習を適切に組み合わせ，ウェブ小説というビッグデータを分析し，デジタルヒューマニティーズの射程を一歩進めた研究であると評価できる。

一方で，論文の表現において，助詞の使用が自然ではない箇所があること，そして，口語的な表現やウェブ小説コミュニティで共有される「業界用語」的表現が一部で用いられているために，文意を汲むのが必ずしも容易ではない箇所が若干あることも指摘せねばならない。しかし，これらは本論文の学術的価値を損なうほどの瑕疵ではなく，全体として本論文は，自然言語処理の解析技法と，テクスト分析を有機的に組み合わせてウェブ小説の特徴を多角的に論じた意義深い研究であり，その学術的貢献は明確である。

以上により，本論文を博士（言語文化学）の学位論文として価値のあるものと認める。

なお，チェックツール iThenticate を使用し，剽窃，引用漏れ，二重投稿等のチェックを終えていることを申し添える。