

探討 AI 文字探勘技術應用於日語文本研究之可能性： 開拓創新日語教育之路

落合由治

淡江大學日文系教授

摘要

目前台灣日本語教育面臨一大轉換期。如何應用 AI 第三代發展出的技術來因應經濟社會的變化，是主要的課題之一。此課題若之於創新日語教育的內容、方法、技能等各方面，都是為當務之急。鑑此本論文採用第二、三代 AI 技術最重要之自然言語處理的文字探勘技術，視野不侷限於單字層次，而是跨至句子、文章層次的文本研究。以事例研究方式來進行探究目前人文科學的質化方法與 AI 自然言語處理的量化研究的可能性，進而推展至第三代 AI 技術的日本語學以及日本語教育研究之上。

歷經考察成果顯示：結合量化研究的文字探勘技術與目前質化研究的人文社會領域，是可視為導入文本解讀的新方法，進而窺得應用於日本語教育的無限可能性。

關鍵字：日語教育、文本、AI、量化研究、質化研究

受理日期：2019 年 08 月 19 日

通過日期：2019 年 11 月 15 日

An Applicability of Text Mining Techniques using AI in Japanese Text Research: Exploring New Trajectories for Japanese Language Education

Ochiai Yuji

Professor, Department of Japanese, Tamkang University

Abstract

Taiwanese Japanese language education is at a major turning point. One of the challenges is the response to economic and social changes due to the development of the third AI technology. Taiwanese Japanese language education is under pressure for new responses in the areas of education, methods, and skills. In this paper, therefore, text mining techniques related to natural language processing, which is one of the most important parts of AI technology in the second and third term, are taken up. This paper explored not only the conventional word level, but also a method combining textual and qualitative methods so far with natural language processing by AI for analysis in the text research field at the sentence and text level. This paper attempted to apply the third AI technology to Japanese language studies and Japanese language education as a case study.

As a result of the discussion, we were able to introduce a new method for reading texts by combining text mining technology, which is quantitative research, and qualitative methods of humanities and social studies. From there, the potential for application to Japanese language education has become apparent.

Keywords: Japanese language education, text, AI, quantitative research, qualitative research

日本語テキスト研究における AI による テキストマイニング手法の応用可能性 —日本語教育内容の新軌道を探る—

落合由治

淡江大学日本語文学科教授

要旨

台湾の日本語教育は大きな転換期を迎えているが、その課題のひとつは第三期 AI 技術の発展による、経済的社会的変化への対応である。教育内容、方法、技能の各方面で、新しい対応が迫られている。そこで本論文では第二期に発展し、第三期 AI 技術の最重要部分のひとつ発展してきた自然言語処理に関するテキストマイニング技法を取り上げて、従来の語レベルにとどまらず、文、文章レベルでのテキスト研究分野での分析について、今までの人文科学的な質的方法と AI による自然言語処理とを組み合わせる方法を探求し、第二、三期テキストマイニング技術の日本語学および日本語教育への応用を事例研究の形で試みた。

考察の結果、量的研究である AI 技術と従来の人文社会系研究の質的方法を結びつけることで、テキストの読解に新しい方法を導入でき、そこから日本語教育への応用にも可能性が見えてきた。

キーワード：日本語教育、テキスト、AI、量的研究、質的研究

日本語テキスト研究における AI による テキストマイニング手法の応用可能性 —日本語教育内容の新軌道を探る—

落合由治

淡江大学日本語文学科教授

1. はじめに

現在、台湾の日本語教育は大きな変動期を迎えている。主な要因には三つあり、第一は少子高齢化社会の進行による若年学習者の急減、第二はグローバル化の進展による社会経済的変動、第三は AI 技術の発展による職業と社会構造の変化である。第一の少子高齢化、第二のグローバル化による変化はすでに台湾の日本語教育界でも広く周知され、対応の試みが進んでいるが¹、第三の第三期 AI 技術の発展による職業と社会構造の変化は、現在、急激に進行中の現象で、その対応は摸索の段階にある。²本論文ではこの第三の変動要因に対

¹ こうした社会変動に関する研究、報告は枚挙に暇がないが、人口問題の研究例として、台湾、日本等の東アジア文化圏での少子高齢化の進行には共通性が高く、原因が解明されつつあると、国立社会保障・人口問題研究所(2012)「特集第 16 回厚生政策セミナー「東アジアの少子化のゆくえ—要因と政策対応の共通性と異質性を探る」」『人口問題研究』68-3 は指摘している。台湾における少子化とグローバル化による教育環境の変化については、さまざまな指摘、報告がすでに多数出ているが、概要をまとめたものとして淡江時報(2014)「臺灣高等教育面臨少子高齢化衝擊之因應對策」<http://tkutimes.tku.edu.tw/SAN2016/pdf/948-3.pdf>(2019年8月30日閲覧)、謝青龍(2018)「【大學快報第 174 期】私大如何因應少子化海嘯」台湾高等教育産業工會 <https://www.theunion.org.tw/2018/12/%E3%80%90%E5%A4%A7%E5%AD%B8%E5%BF%AB%E5%A0%B1%E7%AC%AC174%E6%9C%9F%E3%80%91%E7%A7%81%E5%A4%A7%E5%A6%82%E4%BD%95%E5%9B%A0%E6%87%89%E5%B0%91%E5%AD%90%E5%8C%96%E6%B5%B7%E5%98%AF/>(2019年8月30日閲覧)等を参照。

² 台湾の教育における AI 時代への対応については、台湾教育部資訊及科技教育司(2019)「AI 教育 X 教育 AI—人工智慧教育及數位先進個人化、適性化學習時代來臨！」https://www.edu.tw/News_Content.aspx?n=9E7AC85F1954DDA8&s=D4C4CD32CAE3FF5D#で、初中等教育および大学教育での教育改革の方向性が提示されている。世界各国の AI 時代への教育対応については、Tim Dutton(2018)「An Overview of National AI Strategies」<https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd>(2019年8月30日閲覧)を参照。

して、台湾の日本語教育が対応していく方向性を、日本語研究の視点で考えていくことにしたい。

日本語教育が今後の AI 時代に対応するためには、三つの条件が必要になると考えられる。一つ目は AI 時代の社会経済変化に応じた外国語人材のニーズ変化への対応、二つ目は AI 時代の人材を育成する教育方法の革新、三つ目は「読解、書写、聴取、会話、翻訳・通訳」という今までの日本語能力・技能とテキストマイニングなどの自然言語処理関係のスキルを接続したカリキュラムである。³この中で、直接、日本語研究・学習に関する内容に関わるのは三つ目の、従来の日本語運用スキルとテキストマイニング関係のスキルを接続した教育内容の革新の課題である。その点で、現在の最大の問題のひとつは、こうした日本語自体の認識および運用とテキストマイニング関係のスキルを接続させるために、その基礎となる日本語研究が従来の研究を活かしながら AI 技術に関係した自然言語処理やテキストマイニング技術に接続し、新しい分野にどうすれば発展できるかという課題と言えよう。その点で、AI 時代に日本語教育が対応するには、基礎となる日本語研究分野の領域拡大と革新が必要になっているのである。

日本語教育が AI 時代に接続する前提となる日本語研究の現状を考えると、そこには研究分野としての基本的問題が存在していると言える。それは、現在までの日本語に関わる研究において、自然言語処理やテキストマイニング技術に接続している計量言語学的手法

³ AI 時代における教育の方向性は模索の段階であるが、基本的には AI を理解、操作しながら AI にできない部分（認識力、創造性、柔軟性、革新性）を人間が担当するという共通した見解をさまざまな分野の専門家が提案している。経済面では野村直之(2016)『人工知能が変える仕事の未来』日本経済新聞出版社、基本的な学力については新井紀子(2018)『AI vs. 教科書が読めない子どもたち』東洋経済新報社、学校教育では藤原和博(2017)『10年後、君に仕事はあるのか？未来を生きるための「雇われる力」』ダイヤモンド社等を参照。教育方法としては、すでに提唱されてきたアクティブ・ラーニング、反転学習、学習者中心学習など近代教育を超えたモデルがすでに実施されている。一例として佐藤学(2017)『教育の方法(放送大学叢書)』左右社、栗田佳代子、日本教育研究イノベーションセンター(2017)『インタラクティブ・ティーチングーアクティブ・ラーニングを促す授業づくり』河合出版等を参照。

は計量言語学的課題や統計的視点に限定され、今までのコンピューターの技術的制約から基本的に語単位での統計的処理に限定されたことで、言語事象としてそこから捉えられる現象自体やその理解に大きな制約を受けていたことである。⁴その一方で、他の日本語学の研究領域では、計量言語学が扱っている言語単位の統計的処理によるテキストマイニング技法の応用は限定的なものであり、質的研究と量的研究を接続させる方法は模索の状態と考えられる。⁵しかし現在、第三期 AI 技術の発展で、非常に大きな資料から人間の処理に抛らずに AI 自身で自然言語処理を行う手法が開発されつつあり、さまざまな分野への応用が期待されている。また、AI 技術のために新たな視点で言語事象を捉えなおす必要も生まれている。

そこで本論文では第二期情報処理技術で発展し、第三期 AI 技術の最重要部分のひとつになっている自然言語処理の中で、テキストマイニング技法を取り上げて、語レベルにとどまらず、文、文章レベルでのテキスト研究分野での分析について、今までの人文科学的な質的方法と発展してきた自然言語処理とを組み合わせることで、第三期 AI 技術と密接に関わりを持つテキストマイニング技術の日本語学および日本語教育への応用を試みてみたい。新しい研究課題と分野が広がっていくことで、日本語教育とテキストマイニング等の情報処理および AI 技術との接点を生見出し、拡げてゆく方向性

⁴ 計量言語学に関する研究動向を参照。半沢康(2014)「数理的研究」「特集 2012年・2013年における日本語学界の展望」『日本語の研究』10-3、山口昌也(2016)「数理的研究」「特集 2014年・2015年における日本語学界の展望」『日本語の研究』12-3には、統計的処理による語彙研究が多数紹介されているが、語彙研究としての位置づけは明確ではなく、同じ分野の「語彙(理論・現代)」の動向との関係も考察されていない。課題や領域の設定および研究区分自体に意味があるのか、検討されているとはいいがたい。

⁵ 研究動向の一例として、日本語学会機関誌『日本語の研究』を例に見ると、柏野和佳子(2014)「語彙(理論・現代)」「特集 2012年・2013年における日本語学界の展望」『日本語の研究』10-3、大島中正(2016)「語彙(理論・現代)」「特集 2014年・2015年における日本語学界の展望」『日本語の研究』12-3では、多くの単語の用法に関する研究が紹介されているが、統計的処理による明確な問題設定がされている研究は限られ、多くは用例の分類、数量と%での整理に止まっている。語レベルの言語現象の扱いについても、初歩的な処理の段階で大きな制約が存在している。

が見出せるであろう。同時に語レベルの自然言語処理ばかりでなく、より上位の単位の意味に関わる言語表現理解と自然言語処理との接点を探すテキスト研究として、テキストマイニング技法の応用を文学テキストの分析を事例に試行していきたい。

2. 言語研究における量的研究と質的研究の定位

今回の課題に入る前提として、まず日本語教育と日本語研究がその中に基礎を定位している人文社会系の研究における量的研究と質的研究について考えてみたい。言語に関する量的研究は 1950 年代からの AI 技術の発展と共に発展、普及してきた方法で、計量言語学でも 1950 年代から品詞や語彙使用の量的差異の比較によって、作家の文体特性を見出そうとする研究などが行われてきた。⁶資料間での指標の量的計量と差異の比較、さらに統計的処理での要素の相関性の計量や因子分析などの多変量解析手法による特徴の抽出など、AI 技術の発展と共に、計量言語学の分野も大きく発展してきた。⁷しかし、作者の差異の特定、歴史的変化の素描、ジャンル間の語彙的特徴などのテーマは、研究の草創期から一貫しており、他の課題への発展や応用的課題への拡大の部分では基本的に新しいテーマ、分野に接続することはかなり難しいように見える。そこで、本論文では、質的視点を取り入れて、言語表現に関する量的研究と質的研究のダイナミックスをまず考えてみたい。

量的研究と質的研究は、単に言語形式や語レベルだけを扱うか、

⁶ 語彙の量的分析に関して計量言語学的手法は、現在、最もよく使用される方法である。日本語での計量言語学的手法の研究初期に大きな影響を与えた研究として、樺島忠夫、寿岳章子(1965)『文体の科学』綜芸舎がある。現在でも、研究テーマは基本的には同じであり、一例として、岸本千秋(2017)「ウェブログの計量的文体研究:文とウェブ記号の関係を中心に」『阪大日本語研究』29pp. 71-99、馬場 俊臣(2018)「接続詞の文体差の計量的分析の試み:『BCCWJ 図書館サブコーパスの文体情報』を用いて」『北海道教育大学紀要. 人文科学・社会科学編』69-1pp. 1-14 等参照。

⁷ 研究動向として、統計的手法でテキストの量的特徴を把握しようとする研究は計量文献学と呼ばれ、計量言語学の中では最も有効な分野と言える。一例として、村上征勝(1997)「計量文献学の歴史と課題」『計算機統計学』9-1pp. 65-74、

統計的処理をするかどうかというような手続き的問題ではなく、出発点には主体と客体との関係に関する定位の根本的相違がある。久保田賢一(1997)は教育工学のジャンルで質的研究と量的研究の相違を、客観主義と構成主義という基本的パラダイムの相違として見ている。⁸

表 1 客観主義と構成主義のパラダイム比較⁹

立場 領域	客観主義	構成主義
存在論	「唯一の真実」が研究者の関心とは独立して存在する。それは自然の法則に従って因果的に作用する。	社会的に構成された多様な「真実」が存在する。真実とは、批判にさらされ、人々の合意を作り上げる過程のなかから生み出される。
認識論	研究者は、研究対象に直接関わらずに、研究対象がおかれている場から離れて分析する。	研究者自身が直接、研究対象と相互に関わり合う。そのため、研究者は研究対象と独立して研究することはあり得ない。
方法論	前もって仮説を立て、実験的環境を注意深く作りだし、実証的に検証する。	共通の理解を得るために、研究者は研究対象との絶え間ざる相互作用を通して、真実を構成していく。

客観主義は、ヨーロッパ哲学でのプラトン主義に由来すると言え、「唯一の真実」が人間の主体とは独立にそれ自体で存在しているという前提に立っている。いわゆる科学的思想とも言える。構成主義は相対主義、主観主義とヨーロッパ哲学の主流であるプラトン主義的立場から批判されてきたが、淵源は古代ローマの思想家プロティノスに始まり、20世紀のイギリス、ドイツ、フランス、アメリカ等の質的研究を目指す人文社会系思想の基礎になっている。¹⁰

⁸ 久保田賢一(1997)「質的研究の評価基準に関する一考察:パラダイム論からみた研究評価の視点」『日本教育工学雑誌』21-3pp. 163-173。

⁹ 同注8掲載の表1による。

¹⁰ 客観主義と構成主義は哲学史上、常に対立関係にある思想として相互に作用してきたが、代表的な論としては現象学による客観主義批判がある。高階勝義(1979)「フッサールの科学批判と「生活世界」概念」『鳥取大学教養部紀要』13pp. 1-22、赤松常弘(1982)「フッサールの科学批判と超越論的現象学」『人文科学論集』16pp. 19-30等を始めとして、村上靖彦(2019)「哲学と質的研究:現象学的な質的研究の役割と位置づけについて」『大阪大学大学院人間科学研究科紀要』45pp. 1-18等を参照。また、プロティノスによるプラトン哲学の克服については、美を巡るアイデア概念の検討から分かる。掛下栄一郎(1977)「瞬間の美学への試論-4-プロティノスからアウグスティヌスへ(〔早稲田大学〕社会科学部創設10周年記念号)」『早稲田人文自然科学研究』14pp. 51-68、関村誠(2017)「プ

保田賢一(1997)は教育学での説明のため社会的現実や学習者、教師などの行為を対象にしての説明になっているが、これを言語研究に置き換えて考えて見ると、同じ問題が存在していることがわかる。

表 2 言語学における客観主義と構成主義のパラダイム比較¹¹

立場 領域	客観主義	構成主義
存在論	「唯一の真実」である言語が研究者の関心とは独立して存在する。それは独立した言語の法則に従って因果的に作用する。	社会的に構成された多様な「真実」である言語が存在する。社会的言語の真実とは、批判にさらされ、人々の合意を作り上げる過程のなかから生み出される、ジャンルの表現である。
認識論	研究者は、研究対象の言語に直接関わらずに、研究対象がおかれている場から離れて分析する。言語形式はそれ自体で意味を持つ。	研究者自身が直接、研究対象と解釈、分類、操作などの行為を通じて相互作用的に関わり合う。そのため、研究者は研究対象と独立して研究することはあり得ない。
方法論	作例や調査などにより、前もって言語に関する理論的仮説を立て、分析の観点や方法を注意深く作りだし、実証的に検証する。	社会的に共通の理解を得るために、研究者は研究対象である言語表現との絶え間ざる相互作用を通して、用法を認識し、表現の規則性を見出していく。

言語学の研究史を振り返ってみると、ソシュールが 20 世紀の初めに、言語活動を言語(ラング)と発話(パロール)に分けて、以下のように述べたときにこの二つのパラダイムは、すでに併存していたと言えよう。

言語を使うことはあくまで個人の行為で、私たちが発話を認識するのはこの部分においてなのです。(社会的なのは)受容的で連携的な部分です。異なる個人の中に貯蔵されてはいても、すべての個人にかなり同じ存在が形成されるのは、この部分です。

この領域こそ言語の領域を表します。個々人において、無数の言葉のイメージが、対応する同数の概念と結びつけられています。一人の人を取り上げると、そのたった一つのサンプルの中にも社会的総体の中での言語とは何かを見ることができると

ロティノスの美の理論におけるエイドスとミシェル・アンリのカンディンスキー論におけるフォルム」『ミシェル・アンリ研究』7pp. 41-61等を参照。

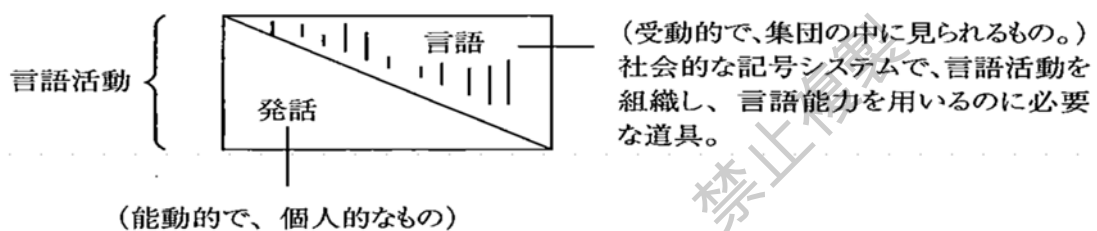
¹¹ 同注 8 掲載の表 1 を元に論者作成。

しょう。もし個人の中に貯蔵され、ある秩序と分類に従って配置されている言葉のイメージを調べることができるならば、そこに言語をなす社会的な結びつきを見ることでしょう。

この社会的な部分は純粹に心の中のもので、純粹に心理的なものです。このようにして、人は言語を保持しているのです。

12

図1 ソシュールが「一般言語学講義」の授業で板書した言語活動の図¹³



客観主義の言語学は、「言語(ラング)」を「異なる個人の中に貯蔵されてはいても、すべての個人にかなり同じ存在が形成される」「受動的で社会集団の中に見られる」「社会的な記号システム」と見て、それを研究対象としている。言語を個々人の差異による個別化の方向ではなく、社会的な共通性に還元できる一般化の方向で定位する。基本的には一般化を試行する量的研究の方法に合致する研究と言える。一方、構成主義的な立場、あるいは相互作用による表現主義的な立場では、「能動的で、個人的なもの」として顕れている言語表現を、相手との共通点として研究者が持っている「受動的で社会集団の中に見られる」「社会的な記号システム」を通じて、「受動的で社会集団の中に見られる」「社会的な記号システム」から差異化する、個別化することを目指すと見えよう。個人の発話から個別的社会的ジャンルまで、その個性を対象とする質的研究の方向を目指している。

¹² フェルディナン・ド・ソシュール/景浦峽・田中久美子訳(2007)『ソシュール一般言語学講義--コンスタンタンのノート』東京大学出版会 pp. 88-89。

¹³ 同注 12。

大事なことは、人文社会系の研究にとって客観主義的量的研究と構成主義的質的研究のパラダイムは相補的であって、両者の関係が調整されて初めて、ソシュールが図1のように「言語活動」と呼んだ人間の言語表現あるいは多様な表現活動を全体的に解明できるということである。「言語(ラング)」だけを扱っても、「発話(パロール)」だけを扱っても、実は言語活動の半分にしか光が当たっていないことになる。言語研究が今まで中心的課題として扱ってきた、文法的形式と意味は、「言語活動」の中で成り立っている現象であり、客観主義的な「言語(ラング)」だけを扱っても、構成主義的な「発話(パロール)」だけを扱っても、実は言語活動・表現の半分しか見えていないのである。この点は、今後、言語研究それ自体として、同時に言語などの表現についてAIを利用したテキストマイニングを進めて行く場合、非常に重要な注意点になる。十全に資料や対象の言語的表現的特徴を理解したい場合、個別的な表現・現象は全体的表現・現象に(語彙を例にすれば、語彙が用いられた作品、ジャンル、社会的表現に)位置づけてみる必要がある、全体的現象は個別的現象に(資料の特徴を帰納する場合は、個々の表現、形式、文、語彙の規則性に)還元してみるという作業が必要となると言える。

今までの日本語研究や言語に関する人文社会系の諸学が研究として行っていた語、文、文章をそれぞれ切り離し、相互に無関係な単位として扱う研究態度や、逆に作品、資料の読み手による受容だけを研究態度には、基本的限界が元々存在しているのである。第三期AIの発展は、言語表現形式しか処理できない人工知能に意味との関係を学習させる方法の模索によって生まれたものであり、こうした言語現象における客観(言語表現形式)と主観(社会的ジャンルの個人的なそれぞれの意味)との関係を明らかにする端緒を初めて開いたと言えるであろう。¹⁴

¹⁴ 第二期から現在の第三期AIの自然言語処理の概要については、小町守監修、奥野陽、グラム・ニュービッグ、萩原正人(2016)『自然言語処理の基本と技術(仕組みが見えるゼロからわかる)』翔泳社、および各年度版総務省『情報通信

3. テキスト分析へのテキストマイニング応用

ここでは、第二期から連続する第三期テキストマイニング技術を応用し量的研究と質的研究を相補的に実施するテキストマイニングの事例として、村上春樹の短編集『女のいない男たち』の中から、文章構成が明確で語彙の機能を特定しやすい「ドライブ・マイ・カー」を事例として取り上げて、量的研究と質的研究の手法を組み合わせ、分析をおこなってみたい。¹⁵

3.1 語彙の重要度による分析

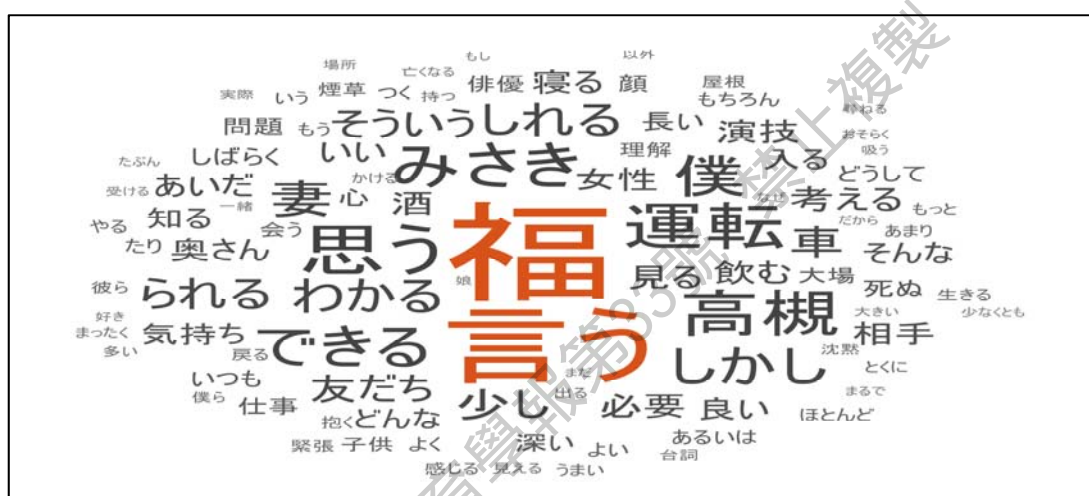
第二期までの自然言語処理技術の中で発展してきた今までの計量言語学的研究やテキストマイニングの基本的処理は、語彙の量的処理、統計的処理で、その処理は資料中での語彙の出現頻度を基礎にして、重要度を算定する方法が第二期から第三期へのテキストマイニング技術発展の中で工夫されている。しかし、客観主義的立場からその語彙が出ている作品、資料を研究者が読んで出現した語彙の意味を考察することは主観主義として退けられ、数理的処理のみに課題を置いていたため、実施できる研究にはさまざまな限界が生

白書』<http://www.soumu.go.jp/johotsusintokei/whitepaper/h30.html>等を参照。自然言語処理の実際については第二期 AI までの技術について奥村学(2010)『自然言語処理の基礎』コロナ社、Steven Bird、Ewan Klein、Edward Loper/萩原正人、中山敬広、水野貴明訳『入門自然言語処理』オライリージャパン等を参照。第三期 AI の自然言語処理技術については坪井祐太、海野裕也、鈴木潤(2017)『深層学習による自然言語処理(機械学習プロフェッショナルシリーズ)』講談社、斎藤康毅(2018)『ゼロから作る Deep Learning②—自然言語処理編』オライリージャパン等を参照。日本語の自然言語処理の先端的研究は、各年度の言語処理学会『自然言語処理』<https://www.anlp.jp/guide/index.html>等を参照。

¹⁵ ワードクラウドは、データの中の単位を頻度順で並べて数値データを可視化する手法で、現在、ビジネス等で広く用いられるようになっている。WEB の大量の言語データからキーワードを抽出することでネット上での利用者の評価、意見を簡単に知ることができるのでマーケティングや世論調査、意識調査などでは有益な情報が得やすい。一例として、小林稔(2017)「ソーシャルメディアのコンテンツから分析したコンビニエンスストアの特徴と分類」『和光経済』50-1pp. 1-17、丸野由希(2017)「卒業論文のテーマに何が選ばれてきたか：現代社会学部の卒業論文題目の語彙分析から」『現代社会研究』19pp. 35-50 等参照。ここでは、MatLab の Analyze Japanese Text Data のツールを使用した。プログラムの内容について、分析の準備(前処理)は <https://jp.mathworks.com/help/textanalytics/text-data-preparation.html>、実際の分析の流れは <https://jp.mathworks.com/help/textanalytics/ug/analyze-japanese-text.html> を参照。

じていた。¹⁶逆に、研究者の作品や資料読解や手作業での分析を中心にした構成主義的質的研究は、量的に処理される言語形式のデータを必ずしも踏まえているわけではないために、いわゆる批評・評論・論説のように根拠となる言語の客観的現象との関係、根拠との関係が曖昧になりがちであった。現在、テキストマイニング関係の自然言語処理技術の発展で、量的分析を質的分析に接続することが容易になっている。その一つは重要度による語彙の類別の視覚化である。

図2 語彙の重要度による「ドライブ・マイ・カー」のワードクラウド



今までの言語の計量的処理の研究では語彙リストとして表形式で出現頻度や比率等で数値表示された内容は、現在の技術では TF、IDF、TF-IDF など語の重要度を判定する方法によって、よりその資料内の語の意味や役割に応じた語彙を選んで、視覚化して表示することができる。¹⁷そのひとつがワードクラウドである。重要度判定の差異により、より大きな字で中央に近い位置に提示される語ほど、単なる出現頻度ではなく、たくさん使われている語彙の中で資料中の重要度が高いことを示している。¹⁸図2では、中心に出ている「福」

¹⁶ 日本での人文社会系の量的研究の基本的方法は、樋口耕一(2014)『社会調査のための計量テキスト分析—内容分析の継承と発展を目指して』ナカニシヤ出版、研究の現状の一例として、日本社会学会(2017)「特集号・テキストマイニングをめぐる方法論とメタ方法論」『社会学評論』68-3 参照。

¹⁷ TF、IDF、TF-IDF の意味と算定法は注 16pp.164-166 を参照。

¹⁸ 文章中に表れる単語の重要性を測る手法は、情報検索でのキーワードを決定する場合などの重要技法で、第二期の自然言語処理で発展し、現在まで開発

は「ドライブ・マイ・カー」の主な登場者「家福」で最も多く、以下、「言う(動詞)」「みさき(名詞)」「思う(動詞)」「高槻(名詞)」「運転(名詞)」「僕(名詞)」「妻(名詞)」「わかる(動詞)」「できる(動詞)」など主な登場者と主要な場面のキーワードが出ており、またよく用いられる形容詞的な機能語として「少し」「しかし」が出ている。作品の頻出語彙を調べる量的処理を作品の内容と関係させてすぐに分かるのは、小説の場合、作品中で大きな部分を占めている構造に関わって見られる内容が量的処理によるワードクラウドで表示されていると考えられるということである。老境に入り視力の落ちた俳優の家福がドライバーを依頼した渡利みさきと車中で、経歴や亡くなった妻と浮気相手のエピソードを語るストーリーの「ドライブ・マイ・カー」では、家福と渡利みさきの会話で成り立つストーリーの流れに関する部分を作品の頻出語彙が話題の中心として表示しているということである。「ドライブ・マイ・カー」では、「運転(名詞)」する場面で「(家)福(名詞)=僕(名詞)」と「みさき(名詞)」が「高槻(名詞)」と「妻(名詞)」について「言(動詞)」い、「思う(動詞)」ことで「僕(名詞)」が「わかる(動詞)」ことが「できる(動詞)」ようになったというように、アウトラインが示されているのである。ワードクラウドの量的処理だけでは、これらの語がいったいどのように言語的にテキスト中で機能しているのかを決めることはまったくできないが、質的処理として作品内容と構成と関係させることで、こうした量的処理で出てくる語の機能を特定することができ、また、逆にこれらの語のテキスト中での用例を元に特徴的な語の用法を調べることで、テキストの構成や重要語の機能を定めることができる。

この方法は、研究はもちろん日本語教育への読解などで応用がすぐにでも可能である。教科書であれ実際の資料であれ新聞、論説、

が続いている。TFIDF と Okapi BM25 が有名である。こうした各種の語彙の重要度判定方法の比較については滝川真弘、山名早人(2016)「特定分野を対象とした単語重要度計算手法の提案と Twitter における専門性推定への適応」『FIT2016 (第 15 回情報科学技術フォーラム) 第二分冊』情報処理学会 pp. 1-7 参照。

小説などの電子化されたテキストを語に句切るなどの前処理をして、プログラムに入れば、こうした重要なキーワードが何かをすぐに知ることが出来、またその語をたどりながら逆にテキストを読むことで、読解を進めることができる。¹⁹技術としては第二期情報処理技術からのテキストマイニング技術であるが、質的分析と連携させることで、日本語研究、日本語教育への応用は極めて容易にできる。その他の、統計的手法との関係は、稿を改めて見ていきたい。

3.2 統計的潜在意味解析による分析

また、第三期 AI 技術の発展で、自然言語に関するテキストマイニングで用いられるようになったトピックモデルによる統計的潜在意味解析の手法は、新しい第三期 AI のテキストマイニング技法の一つとして広く用いられるようになってきている。²⁰LDA を用いると、Bag of Words と呼ばれる単語と出現頻度のペアをモデル化し、「共起しやすいアイテムをモデル化することができ、さらに潜在的な共起性を推定することができるため、データに表層上に現れていないが、共起する可能性があるアイテムを推定することができる」²¹と期待されており、現在、活発に開発が進められている技術である。第二期 AI までの量的分析は、単語と出現頻度のペアを取り出して処理していたが、それら相互の関係を取り出すことは容易ではなかった。LDA など第三期に発展が続いている統計的潜在意味解析の発達は、資料全体での単語と出現頻度のペア相互の関係から、類似している複数のグループをモデルとして抽出できるので、より人間が言語表現を理解しているときの意味の理解に類似した処理ができると考え

¹⁹ インターネットで利用できる分析サイトとして UserLocal「AI テキストマイニング」<https://textmining.userlocal.jp/>、使いやすいパソコン用ツールとしては、樋口耕一「KHCoder」<https://khcoder.net/>を参照。

²⁰ 日本でも人文社会系分野で LDA を使った言語データ分析が行われるようになってきている。一例として、佐藤 圭(2017)「マーケティング研究におけるトピックモデルの適用に関する一考察」『経営研究』68-3pp. 125-148、松河秀哉, 大山牧子, 根岸千悠, 新居佳子, 岩崎千晶, 堀田博史(2018)「トピックモデルを用いた授業評価アンケートの自由記述の分析」『日本教育工学会論文誌』41-3pp. 233-244等を参照。

²¹ 奥村学監修、佐藤一誠(2015)『トピックモデルによる統計的潜在意味解析』コロナ社 p. 26。

られている。

以下の図 3 は、「ドライブ・マイ・カー」の LDA 分析による主要トピックである。²²

図 3 「ドライブ・マイ・カー」の LDA 分析によるトピック



「福」は主な登場者の「家福」のことで、トピック 1 は、死んだ妻の浮気相手だった高槻との、作品では後半に位置する何回かの会話場面の内容が取り出されている。家福は、妻の死後に高槻と出逢って、バーで妻に関する会話を交わしているため、「バーテンダー」

²² NTTPC コミュニケーションズ「用語解説」によると、「LDA は Latent Dirichlet Allocation の略で、日本語では「潜在的ディリクレ配分法」と呼ばれている。ディリクレは、19 世紀ドイツの数学者の名前。この数学者が定義した関数が、LDA の中で使われている。(中略) LDA では、各単語や文章には潜在的なトピックがあると考えられる。この場合のトピックは、話題、テーマ、分野、文体といったもの。たとえば人間は、特定の単語や断片的な文章から「あ、あの話だな」と分かることが多い。その要因が、単語や文章が持つ潜在的なトピックということだ。そして、同じトピックを持つ単語は同じ文書に登場しやすい。あるいは、同じトピックを持つ単語が何度も出てくる文書は、それだけ文書のトピックを確定しやすいし、その確率も高い。」
<https://www.nttpc.co.jp/yougo/LDA.html> (2019 年 2 月 20 日閲覧)。統計的潜在意味解析の詳細と手法については、奥村学監修、佐藤一誠(2015)『トピックモデルによる統計的潜在意味解析』コロナ社参照。ここでは、MatLab の Analyze Japanese Text Data のツールを使用し、7 つのトピックを抽出し、その中の上位 4 つの語彙を示した。<https://jp.mathworks.com/help/textanalytics/ug/analyze-japanese-text.html>。

など場所に関する語彙、俳優である家福の仕事の内容に関する「台詞」などの語彙も関連して出ている。トピック 2 は、視力が落ちて運転できなくなった家福から依頼されたドライバーのみさきとの会話場面に関する内容で、車に関する「シフトチェンジ」や、車中でカセットテープによって音楽を聞いていることから「音楽」、話題で出た「結婚」などもドライブ中での話であることから関係して出ている。トピック 3 と 4 は、家福の作中での動きに関わるストーリーと心理状態の説明に伴って出ていると考えられる語彙である。LDA の方法を小説のテキスト資料に適用した場合、先の語彙の重要度を示すワードクラウドよりもさらに、作品中で大きな部分を占めている話題の基本的内容が、グループごとに分類されてより明確に示されていると言える。

テキストの LDA 分析の結果は、テキスト内に語彙の共起関係によって分類できる複数のトピックの存在を示しており、研究はもちろん日本語教育の読解に応用できる。LDA のトピックに出てきた主要語をテキスト中で、たどりながら、その出てきた部分を集めることで、語彙的に関係が深い部分を作品中での重要な話題として抽出できる。例としてトピック 3 で出てきた部分を作品中から集めてみると以下のような内容になる。「ば」「あいだ」「ながら」「思う」「できる」「考える」「知る」等のような頻出する機能語、動詞は取り敢えず除いて、大きな文字で表示されている重要度の高い、以下の名詞 3 語を例として取り出してみる。

表 3 「ドライブ・マイ・カー」の LDA 分析トピック 3 の語彙の作品中での出現部分

(注) 網掛けは対象語

トピック 3 の 重要語彙例	出現部分の用例サンプル	作品中での役割
車	① これまで女性が運転する車に何度も乗ったが、家福の目からすれば、彼女たちの運転ぶりはおおむね二種類に	車は普通名詞ばかりでなく、家福が愛車(サーブ 900 コンバーティブル)を修理に出し、運転手に渡利みさ

	<p>分けられた。p. 15</p> <p>②その車^車を新車で購入したとき、妻はまだ存命だった。(中略)「古い車^車だから、ナビもついていないけど」p. 20-22</p> <p>③「ひとつ質問していいですか?」とみさきが言った。(中略)二ヶ月ほど一緒に長く車^車に乗っていて、みさきが自分の方から口をきくことはきわめて希だったからだ。p. 34-35</p>	<p>きを採用して、仕事の行き帰りに、車中で自分の結婚生活や妻との関係を話す場面での家福の車をさしている。</p>
理解	<p>①みさきが運転手を務めるようになって以来、なぜか頻繁に妻のことを思い出すようになった。(中略)しかし妻の方は時折、彼以外の男と寝ていた。(中略)どうして彼女が他の男たちと寝なくてはならないのか、家福にはよく理解^{理解}できなかった。そして今でも理解^{理解}できていない。p. 30-31</p> <p>②「それで家福さんには理解^{理解}できたんですか。どうして奥さんがその人と寝たのか?」/家福は首を振った。「いや、理解^{理解}はできなかったな。p. 41</p>	<p>渡利みさきが運転手をするようになって、家福が死んだ妻のことをたびたび思い出すようになり、妻の浮気の原因が理解できないことに気がつき、みさきと原因について話すようになった場面で、妻の行為が理解できない家福のそれまで隠蔽されていた思いに使われている。</p>
生活	<p>①そういうのが彼の生活^{生活}スタイルの欠かせない一部になっていた。p. 20</p> <p>②十代半ばから車を運転しています。車がなければ生活^{生活}できないようなところです。p. 24</p> <p>③二人は結婚して以来、夫婦としてまた生活^{生活}のパートナーとして、良好な関係を常に保っていたからだ。p. 31</p>	<p>家福の現在の日常生活と渡利みさきの履歴、家福と妻との結婚生活に関する概要を示す場面に出て、日常性を示す指標になっている。</p>

以上の表1のように、LDA分析のトピック3に出ている名詞類は、家福とみさきが仕事場に向かう「車」のなかで、会話した内容と、みさきが運転手になってから家福が妻との「生活」を見直すようになって、主に話された妻の浮気の原因が「理解」できない部分が集まっており、作品の主要テーマに関わる部分であることが分かる。

「知る」「思う」など心理に関する動詞語彙が付随して出ているのも、家福が妻との関係を見直している中で、妻の行為が理解できない場面で使われているキーワードであることを示していると言えよう。LDA 分析を使うと、以上のように作品中の重要テーマに関わる部分を複数抽出でき、そこに出ているキーワードにそれぞれ注目することで、読解の方向性を複線化して決めることができる。用例を集めて読むという質的方法を合わせて使うことで、それぞれの語彙グループごとに多面的に作品の主要内容を理解することができる。質的方法だけでは、注目点を多数決めることは難しく、研究者、教師が提供した読解ポイントだけが重要とは言えないが、語彙の共起関係によるグルーピングを行う LDA などの統計的潜在意味解析の手法は、語彙の共起関係による複数のポイントを出せる点で、研究や教育での読解を進める上で、言語的事実に基づく確かな手掛かりを提供している。特に日本語教育では、学習者が自分で読みを深め、相互に討論して内容を確認していく、自律的学習のパートナーになりえる方法と言えよう。

以上から、データマイニングの中で発展してきた自然言語処理によるテキストマイニングの代表的な方法である、ワードクラウド、統計的潜在意味解析を用いることで、多数者が形成している SNS やブログなどの大量の社会言語的データから一般的傾向を見出す際に応用する従来のテキストマイニングでの使用ばかりではなく、質的研究と併用することで新聞、小説、論説などの作品、資料でもテキスト中で語彙的に大きな位置を占めている基本的内容を、多面的に知る手掛かりになることが分かる。多数の作品を比較しながら読む場合も、一定の基準で各作品の中心的キーワードを取り出したり、主要部分の内容のキーワードを複数のトピックで取り出したりすることができるので、比較対照しながら読解する横断的あるいは縦断的な日本語研究や日本語教育の読解での手掛かりとして非常に有効であると言えよう。

4. 量的研究と質的研究の相補性の確保

しかし、今までの計量的研究がそうであったように、取り出した単位の数値データを研究目的に合わせて、どう活用するかは大きな問題と言える。通常は、こうしたデータは、作品間、作家間の比較や時代比較、文体比較などに用いられるが、それは抽出された指標による比較であって、それだけでは作品それぞれの内容理解や意義を了解し、評価する手掛かりにはならない。そして、「ドライブ・マイ・カー」の作品を十分に読んでいれば、量的指標で抽出されたこれらの名詞、動詞、形容詞（副詞）が主な登場者・家福と、そのドライバーを担当するみさきとの会話、妻の浮気相手だった高槻とのやり取りに関して、そのストーリー展開の中で出てくる語彙であることはすぐに分かることで、量的指標だけでは実際に読んで、解釈や整理をするという相互行為的な質的研究に有益なデータを提供することは難しい。先に述べたように、言語研究あるいは人文社会系研究全般にも言えることであるが、研究の基本的パラダイムによって見えるものがまったく変わってしまうため、量的研究であれば客観的で、質的研究は全部、主観的というような、単純素朴な二者択一の問題ではないことを人文社会系研究者が理解していく必要がある。人文社会的研究の対象は、もともとすべて人間の意識が生み出した記号体系であり、そうした意識による象徴的操作のための意味的構造は決して物理現象として、人間とは独立に存在する客観的現前ではない。どうすれば、量的研究と質的研究の相補性を確保できるか、絶えず摸索していく必要がある。

そうしたとき、言語作品にはジャンルによって固有の表現的構成があることを見ていくことは重要である。「ドライブ・マイ・カー」の場合、以下のように多元的重層的な意味の複合が為されている。紙数に限りがあるので、冒頭から暫くの部分を取り上げて見ていくことにする。

例 1 「ドライブ・マイ・カー」

I ① これまで女性が運転する車に何度も乗ったが、家福の目

からすれば、彼女たちの運転ぶりはおおむね二種類に分けられた。②いささか乱暴すぎるか、いささか慎重すぎるか、どちらかだ。③後者の方が前者より一我々はそのことに感謝するべきなのだろう一ずっと多かった。④一般的に言えば、女性ドライバーたちは男性よりも丁寧な、慎重な運転をする。⑤もちろん丁寧で慎重な運転に苦情を申し立てる筋合いはない。⑥それでもその運転ぶりは時として、周囲のドライバーを苛立たせるかもしれない。(Ⅱ、Ⅲ、Ⅳ段落省略)

V ①彼が男性と女性を区別して考えることは、日常的なレベルではあまりない。②男女の能力差を感じることもほとんどない。③家福は職業柄、男女ほぼ同数の相手と仕事をするし、女性と仕事をしているときの方がむしろ落ち着けるくらいだ。④彼女たちはおおむね細部に注意深く、また耳がよい。⑤しかし車の運転に限って言えば、女性が運転する車に乗ると、隣でハンドルを握っているのが女性であるという事実を彼は常に意識させられた。⑥しかしそのような意見を誰かに語ったことはない。⑦それは人前で口にするには不適切な話題であるように思えたからだ。

VI ①だから彼が専属の運転手を捜しているという話をして、修理工場の経営者である大場が若い女性ドライバーを推薦してくれたとき、家福はそれほど楽しげな表情を顔に浮かべることができなかった。②大場はそれを見て微笑した。③気持ちはわかりますよ、と言わんばかりに。

「でもね、家福さん、この子の運転の腕は確かですよ。そいつは私が間違いなく保証します。よかったら会うだけでも一度会ってやってくれませんか？」

④「いいよ。あなたがそう言うなら」と家福は言った。⑤彼は一日でも早く運転手を必要としていたし、大場は信頼のできる男だった。⑥もう十五年のつきあいになる。⑦針金のような硬

い髪をした、小鬼を思わせる風貌の男だが、こと車に関しては彼の意見に従ってまず間違いはない。

⑧「念のためにアラインメントを見ておきたいんですが、そちらに問題がなければ、あさつての二時には完全な状態で車をお渡しできると思います。そのときに本人にここに来させますから、試しに近所をちょっと運転させてみたらいかがでしょう？もし気に入らなければ、そう言って下さい。私に気を遣ったりする必要はまったくありません」

⑨「年はいくつくらいなんだ？」

⑩「たぶん二十代の半ばだと思います。あらためて訊いたことはありませんが」と大場は言った。⑪それから少し顔をしかめた。「ただ、さっきも言ったように運転の腕にはまったく問題はないんですがね……………」(以下、会話文 10 文省略)

VII ①飲酒の話題は家福の顔を曇らせた。②右手の指が自然に唇に伸びた。

③「あさつての二時に会ってみよう」と家福は言った。④ぶっきらぼうで無口でかわいげがないというところが彼の興味を惹いた。

VIII ①二日後の午後二時には、黄色のサーブ 900 コンバーティブルは修理を終えられていた。②右前面のへこんだ部分は元通りに修復され、塗装も継ぎ目がほとんどわからないように丁寧に仕上がっていた。③エンジンは点検整備され、ギアは再調整され、ブレーキパッドもワイパーブレードも新しいものに交換された。④洗車され、ホイールを磨かれ、ワックスをかけられていた。⑤いつもどおり大場の仕事にはそつがなかった。⑥家福はもう十二年そのサーブに乗り続け、走行距離も十万キロを超えている。⑦キャンバスの屋根もだんだんくたびれてきた。⑧強い雨の降る日には隙間の水漏れを気にする必要がある。⑨しかし今のところ新車に買い換えるつもりはない。⑩これまで

大きなトラブルは皆無だったし、何よりも彼はその車に個人的な愛着を持っていた。⑪冬でも夏でも、車の屋根を開けて運転するのが好きだった。⑫冬には分厚いコートを着てマフラーを首に巻き、夏には帽子をかぶって濃いサングラスをかけ、ハンドルを握った。⑬シフトの上げ下げを楽しみながら都内の道路を移動し、信号待ちのあいだにのんびり空を眺めた。⑭流れる雲や、電線にとまった鳥たちを観察した。⑮そういうのが彼の生活スタイルの欠かせない一部になっていた。⑯家福はゆっくりとサーブのまわりを一周し、レース前の馬の体調を確かめる人のように、あちこち細かい部分を点検した。

IX ①その車を新車で購入したとき、妻はまだ存命だった。②ボディーカラーの黄色は彼女が選んだものだ。③最初の数年間はよく二人でドライブをした。④妻は運転をしなかったので、ハンドルを握るのはいつも家福の役だった。⑤遠出も何度かした。⑥伊豆や箱根や那須に出かけた。⑦しかしそのあとの十年近くはほとんど常に彼一人で乗っていた。⑧妻の死後、何人かの女性と交際したが、彼女たちを助手席に座らせる機会はなぜか一度もなかった。⑨都内から外に足を伸ばすことも、仕事でそうする必要がある場合を別にして、まったくなくなってしまった。(一文省略)

X ①家福が必要書類にサインし、請求書の詳細について説明を受けているときに、その娘がやってきた。(10文省略)

X I ①大場が家福を紹介した。②彼女の名前は渡利といった。③渡利みさき。④「みさきは平仮名です。もし必要なら履歴書を用意しますが」、彼女は挑戦的に聞こえなくもない口調でそう言った。

X II ①家福は首を振った。②「今のところ履歴書までは必要ない。マニュアル・シフトは運転できるよね?」③「マニュアル・シフトは好きです」と彼女は冷ややかな声で言った。④まるで筋金入りの菜食主義者がレタスは食べられる

かと質問されたときのように。

⑤「古い車だから、ナビもついていないけど」(六文省略)

X III ④「わかりました」と彼女は言った。⑤いちいち道順の確認もしなかった。⑥そして大場から車のキーを受け取ると、座席の位置とミラーを手早く調整した。⑦どこにどんなスイッチがあるのか、彼女はすべて承知しているようだった。⑧クラッチを踏み、ギアをひとつおとり試した。⑨ジャケットの胸のポケットからレイバンの緑色のサングラスを出してかけた。⑩それから家福に向かって小さく肯いた。⑪用意は整ったということだ。

⑫「カセットテープ」と彼女はオーディオを見て独り言のように言った。⑬「カセットテープが好きなんだ」と家福は言った。

⑭「CDなんかより扱いやすい。台詞の練習もできる」

⑮「ひさしぶりに見ました」

⑯「僕が運転を始めた頃はエイトトラックだった」と家福は言った。

X IV ①みさきは何も言わなかったが、どうやら表情からするとエイトトラックがどんなものかも知らないらしかった。

(注) 下線部分：話の筋の展開(ストーリーの展開)、枠囲い：特定の時の提示、波線部分：説明(プロット)、網掛け：社会文化記号

「ドライブ・マイ・カー」は冒頭から5段落がストーリーの背景や理由を説明するプロットで主な登場者の家福の女性と男性のドライバーに関する考えが述べられ、ストーリーの展開は第VI段落文②から時の経過に従って下線部及び会話文のように、特に第VIII段落文①の「二日後の午後二時」以降、明確に展開している。その他の部分はすべてプロットに当たる説明で、登場者の背景、登場する景物などの説明である。作品は、以下の表2のように基本的に時の持続の中での時の経過によって展開する登場者の動きを描く事件の話をする文章と、話し手の思いに焦点を当てる話し手の思いを述べる文章という二種類の異なる機能を持った表現で各段落が構成されてい

る。²³

表4 「ドライブ・マイ・カー」の文章構成概要

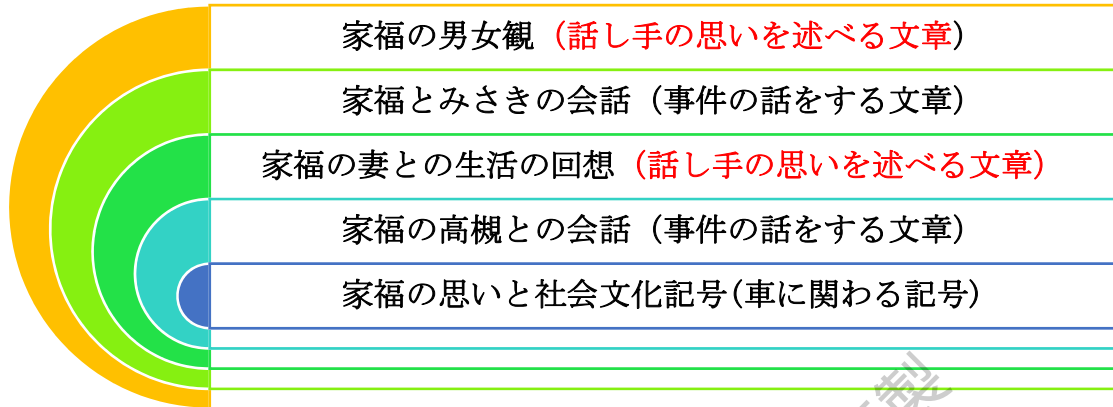
内容区分	段落	文章構成の種類	内容
1	1～5	話し手の思いを述べる文章	男女のドライバーの差＝プロット
2	6・7	事件の話をする文章	時の経過の持続に従う登場者の動き＋説明（舞台、背景、心理、過去の出来事などでストーリーの理由を説明）／（車の修理とドライバーの依頼）
3	8～23	事件の話をする文章	家福とみさきの最初のドライブと会話＝ストーリー
4	24～38	話し手の思いを述べる文章	みさきがドライバーになってからの経過の話＝プロット
5	39～43	話し手の思いを述べる文章	家福夫妻の子供の死の回想＝プロット
6	44～69	事件の話をする文章	みさきと家福の間での、妻の浮気についての会話＝ストーリー
7	70～75	事件の話をする文章	家福が妻の浮気相手の高槻にコンタクトした話＝ストーリー
8	76～86	話し手の思いを述べる文章	高槻の印象と最初の会話の回想＝プロット
9	87～101	事件の話をする文章	家福と高槻と間での、家福の妻に関する会話＝ストーリー
10	102～104	事件の話をする文章	家福と高槻との間での、家福の妻に関する二回目の会話＝ストーリー
11	105～121	事件の話をする文章	家福とみさきとの間での、家福の妻と高槻の関係についての会話＝ストーリー

先にみた図3のLDAの分析は、事件の話をする文章によるストーリーとそこに伴う話し手の思いを述べる文章のプロットの内容の違いをある程度識別して、家福対高槻、家福対みさき、家福対妻というそれぞれの関係のトピックを取り出したと考えられる。作品中での主なストーリーとプロットの識別に有効に使えると言えよう。「ド

²³ 文章の基本的類型については、永尾章曹(1992)「描写と説明について」『小林芳規博士退官記念国語学論集』汲古書院 pp.717-738を参照。

「ドライブ・マイ・カー」の作品の構成を整理すると以下の図4のようになるであろう。

図4「ドライブ・マイ・カー」の作品の構成



しかし、冒頭部分で見ていく必要のあるところは、それだけではない。網掛けで示した「黄色のサブ 900 コンバーティブル」「キャンバスの屋根」「キャンバスの屋根」「カセットテープ」「エイトトラック」など1回しか出ていない語彙は、実はストーリー展開の背後にある隠れた作品テーマのモチーフである。

図5「ドライブ・マイ・カー」における社会文化記号



これは図5のように時間的關係で家福と妻と高槻の關係を示す大事な手掛かりで、同時に家福が生きてきた経歴、時代を示している。こうした社会文化記号的表現といえる内容は、実際に作品を読み、

またその社会文化的意味を探ることではか読者としての研究者、学習者には明示されない。また、主なストーリーである家福とみさきの会話の背後に提示される日本社会全体を背景にした内容で、顕在化していない作品の構造でもある。量的方法では分からない点を明らかにしていくことが、質的手法の大事な点である。

以上、概観してきたように量的指標を手掛かりに作品の文章構成を基にして、各表現の意味構造を明らかにしていくことで量的指標と質的研究の相補性を確保できると言えよう。また、量的に少ない語彙や表現であっても、テキスト全体から見れば重要な意味を与えられている場合は文学作品や思想作品では少なくない。1回しか出ていない語彙や一文しかない説明でも、極めて大きな意味を持っていることは、小説、思想、宗教などの人文社会系テキストの特徴でもある。量的指標を出す作業と同時に、作品自体に即して各語や表現をたどる作業は質的テキストや資料を扱う場合、欠かせないポイントだと言えよう。この点に、人文社会系研究の意義があり、また日本語学や日本語教育の立脚点になる価値が産まれると言える。

5. おわりに

AIの利用によるテキストマイニング技法は、量的指標と質的研究の相補性を確保することで、新しい展開が期待できる。量的指標と質的研究の相補性の確保は、ケースバイケースで、汎用的一般的に何かをすれば必ずこういう結果に成るとは言えない。人文社会系でのテキストマイニング技法の利用は、その結果の利用方法自体が検討、確認すべき大きな課題であり、同時に新しい課題への挑戦でもある。事例に則して可能性の探究を続けることで、AI情報処理技術の日本語教育、日本語学や表現研究への応用を探っていきたい。

テキスト

村上春樹 (2014) 「ドライブ・マイ・カー」『女のいない男たち』文藝春秋

プログラム

MatLab2018b 「Analyze Japanese Text Data」

<https://jp.mathworks.com/help/textanalytics/text-data-preparation.html> / <https://jp.mathworks.com/help/textanalytics/ug/analyze-japanese-text.html> (2019年8月30日閲覧)

参考文献

NTTPC コミュニケーションズ 「用語解説」

<https://www.nttpc.co.jp/yougo/LDA.html> (2019年2月20日閲覧)

赤松常弘(1982)「フッサールの科学批判と超越論的現象学」『人文科学論集』16pp.19-30

新井紀子(2018)『AI vs. 教科書が読めない子どもたち』東洋経済新報社

大島中正(2016)「語彙(理論・現代)」 「特集 2014年・2015年における日本語学界の展望」『日本語の研究』12-3

奥村学(2010)『自然言語処理の基礎』コロナ社

奥村学監修、佐藤一誠(2015)『トピックモデルによる統計的潜在意味解析』コロナ社

小町守監修、奥野陽、グラム・ニュービッグ、萩原正人(2016)『自然言語処理の基本と技術(仕組みが見えるゼロからわかる)』翔泳社

掛下栄一郎(1977)「瞬間の美学への試論-4-プロティノスからアウグスティヌスへ(〔早稲田大学〕社会科学部創設10周年記念号)」『早稲田人文自然科学研究』14pp.51-68

柏野和佳子(2014)「語彙(理論・現代)」 「特集 2012年・2013年における日本語学界の展望」『日本語の研究』10-3

樺島忠夫、寿岳章子(1965)『文体の科学』綜芸舎

岸本千秋(2017)「ウェブログの計量的文体研究:文とウェブ記号の関係を中心に」『阪大日本語研究』29pp.71-99

久保田賢一(1997)「質的研究の評価基準に関する一考察:パラダイム論からみた研究評価の視点」『日本教育工学雑誌』21-3pp.163-173

栗田佳代子、日本教育研究イノベーションセンター(2017)『インタラクティブ・ティーチングーアクティブ・ラーニングを促す授業づくり』河合出版

言語処理学会『自然言語処理』 <https://www.anlp.jp/guide/index.html> (2019年8月30日閲覧)

国立社会保障・人口問題研究所(2012)「特集第16回厚生政策セミナー

- ー「東アジアの少子化のゆくえ—要因と政策対応の共通性と異質性を探る」『人口問題研究』68-3
- 小林稔(2017)「ソーシャルメディアのコンテンツから分析したコンビニエンスストアの特徴と分類」『和光経済』50-1pp.1-17
- 斎藤康毅(2018)『ゼロから作る Deep Learning②—自然言語処理編』オライリージャパン
- 佐藤圭(2017)「マーケティング研究におけるトピックモデルの適用に関する一考察」『経営研究』68-3pp.125-148
- 謝青龍(2018)「【大學快報第174期】私大如何因應少子化海嘯」台灣高等教育產業工會 [https://www.theunion.org.tw/2018/12/%E3%80%90%E5%A4%A7%E5%AD%B8%E5%BF%AB%E5%A0%B1%E7%AC%AC174%E6%9C%9F%E3%80%91%E7%A7%81%E5%A4%A7%E5%A6%82%E4%BD%95%E5%9B%A0%E6%87%89%E5%B0%91%E5%AD%90%E5%8C%96%E6%B5%B7%E5%98%AF/\(2019年8月30日閱覽\)](https://www.theunion.org.tw/2018/12/%E3%80%90%E5%A4%A7%E5%AD%B8%E5%BF%AB%E5%A0%B1%E7%AC%AC174%E6%9C%9F%E3%80%91%E7%A7%81%E5%A4%A7%E5%A6%82%E4%BD%95%E5%9B%A0%E6%87%89%E5%B0%91%E5%AD%90%E5%8C%96%E6%B5%B7%E5%98%AF/(2019年8月30日閱覽))
- 関村誠(2017)「プロティノスの美の理論におけるエイドスとミシェル・アンリのカンディンスキー論におけるフォルム」『ミシェル・アンリ研究』7pp.41-61
- 総務省『情報通信白書』<http://www.soumu.go.jp/johotsusintokei/whitepaper/h30.html>
- フェルディナン・ド・ソシュール/景浦峽・田中久美子訳(2007)『ソシュール—一般言語学講義——コンスタンタンのノート』東京大学出版会 pp.88-89。
- 台湾教育部資訊及科技教育司(2019)「AI教育 X 教育 AI—人工智慧教育及數位先進個人化、適性化學習時代來臨！」
https://www.edu.tw/News_Content.aspx?n=9E7AC85F1954DDA8&s=D4C4CD32CAE3FF5D#
- 高階勝義(1979)「フッサールの科学批判と「生活世界」概念」『鳥取大学教養部紀要』13pp.1-22
- 滝川真弘、山名早人(2016)「特定分野を対象とした単語重要度計算手法の提案と Twitter における専門性推定への適応」『FIT2016 (第15回情報科学技術フォーラム) 第二分冊』情報処理学会 pp.1-7
- 淡江時報(2014)「臺灣高等教育面臨少子高齡化衝擊之因應對策」
<http://tkutimes.tku.edu.tw/SAN2016/pdf/948-3.pdf>(2019年8月30日閱覽)
- Tim Dutton(2018)「An Overview of National AI Strategies」
<https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd>(2019年8月30日閱覽)

- 坪井祐太、海野裕也、鈴木潤(2017)『深層学習による自然言語処理
(機械学習プロフェッショナルシリーズ)』講談社
- 永尾章曹(1992)「描写と説明について」『小林芳規博士退官記念国語
学論集』汲古書院 pp. 717-738
- 日本社会学会(2017)「特集号・テキストマイニングをめぐる方法論
とメタ方法論」『社会学評論』68-3
- 野村直之(2016)『人工知能が変える仕事の未来』日本経済新聞出版
社
- 半沢康(2014)「数理的研究」特集 2012年・2013年における日本
語学界の展望」『日本語の研究』10-3
- Steven Bird、Ewan Klein、Edward Loper/萩原正人、中山敬広、水
野貴明訳『入門自然言語処理』オライリージャパン
- 馬場 俊臣(2018)「接続詞の文体差の計量的分析の試み：『BCCWJ 図
書館サブコーパスの文体情報』を用いて」『北海道教育大学紀要。
人文科学・社会科学編』69-1pp. 1-14
- 樋口耕一(2014)『社会調査のための計量テキスト分析—内容分析の
継承と発展を目指して』ナカニシヤ出版
- 樋口耕一「KHCoder」<https://khcoder.net/>(2019年8月30日閲覧)
- 藤原和博(2017)『10年後、君に仕事はあるのか？未来を生きるため
の「雇われ力」』さとう
- 佐藤学(2017)『教育の方法(放送大学叢書)』左右社
- 松河秀哉, 大山牧子, 根岸千悠, 新居住子, 岩崎千晶, 堀田博史 (2018)
「トピックモデルを用いた授業評価アンケートの自由記述の分
析」『日本教育工学会論文誌』41-3pp. 233-244
- 丸野由希(2017)「卒業論文のテーマに何が選ばれてきたか : 現代
社会学部の卒業論文題目の語彙分析から」『現代社会研究』
19pp. 35-50
- 村上征勝(1997)「計量文献学の歴史と課題」『計算機統計学』9-
1pp. 65-74
- 村上靖彦(2019)「哲学と質的研究:現象学的な質的研究の役割と位置
づけについて」『大阪大学大学院人間科学研究科紀要』45pp. 1-18
- 山口昌也(2016)「数理的研究」特集 2014年・2015年における日
本語学界の展望」『日本語の研究』12-3
- UserLocal「AI テキストマイニング」[https://textmining.
userlocal.jp/](https://textmining.userlocal.jp/)(2019年8月30日閲覧)
- (注記) 本論文は2019年3月の淡江大学日文系「AIと日本語教
育」国際シンポジウムで発表した内容を加筆・訂正したもので、
科技部研究案 107-2410-H-032 -030 -MY2 の成果の一部である。