airiti

DOI:10.6205/jpllat.202006_(47).0003

探討文本探勘技術的應用與今後發展的可能性
—促進日語教育與日本相關人文社會研究間聯繫之合作—

落合由治
淡江大學日本語文學系教授

摘　要

當今社會隨著 AI 技術快速的發達，促使不同領域野的資訊通訊技術融入生活當中，也使得社會結構產生了變化。目睹當前的日語教育或是日本相關人文社會研究上，利用資訊通訊技術尚未普及。於是因應今後社會結構產生的變化，日文相關科系的課程中，有必要導入 AI 先進的概念或運用的技能。

本論文論述的重點，著重在各領域中被廣泛地運用的 AI 技術中，其中與語言處理關係最為密切的文本探勘技術的適用性。特別是聚焦於文本探勘工具之一、為樋口耕一教授於 2014 年所開發的 R 軟體中的「KHCoder」技術。

就步驟而言，首先介紹可將資訊可視化的文本探勘「KHCoder」技術。其次再以此工具得出的結果，結合語料的質性分析內容進行比對。再者進一步就 AI 技術應用於人文社會研究、教育的成果與日語教育結合來進行闡述。最終期盼開啟 AI 技術與日語教育現場以及開設課程緊密結的一大契機。

關鍵詞：AI 技術、文本探勘、日語教育、人文社會研究、質性分析

airiti

DOI:10.6205/jpllat.202006_(47).0003

Exploring the Application and Development Potential of Text Mining Technology: Aiming for Cooperation and Collaboration with Japanese Language Education and Japanese Humanities and Social Studies

Ochiai, Yuji
Professor, Department of Japanese, Tamkang University, Taiwan

Abstract

In today's society, with the rapid development of AI technology, information and communication technology is entering life in various fields, and it is about to create a great social change. But, until now, there has not been a sufficient link between Japanese language education, Japanese humanities and social studies, and information and communication technology. In response to future social changes, it is necessary to connect new technologies and skills to the existing curriculum.

In this paper, we discussed the possibility of linking text mining techniques related to language processing to humanities and social studies related to Japanese, among various types of AI technologies. In particular, as an example of a text mining tool, this paper mainly introduced "KHCoder", which can visually process text mining programs such as R developed by Koichi Higuchi (2014).

The results obtained by these tools were linked to the qualitative analysis of linguistic materials, and attempts to understand the contents were introduced. It would be best if these tools could be used to link humanities research and education to Japanese language education and be the starting point for applying AI technology in Japanese language education and in curricula.

Keywords: AI technology, text mining, Japanese language education, Humanities and Social Sciences, Qualitative Analysis

airiti

DOI:10.6205/jpllat.202006_(47).0003

テキストマイニング技術の応用と発展可能性の探究
―日本語教育および日本関係人文社会系研究との連繋と協働をめざして―

落合由治
淡江大学日本語文学科教授

要　旨

現在の社会は、AI 技術の急速な発展によって、さまざまな分野で情報通信技術が生活の中に入り込み、大きな社会変化を生み出そうとしている。今まで日本語教育や日本の人文社会系研究と情報通信技術とは十分な結び付があったわけではないが、今後の社会変化に対応して、今までのカリキュラムに新しい技術やスキルを結び付けていく必要が生まれている。

本稿では、さまざまな分野のある AI 技術の中で言語処理に関係したテキストマイニングの技術を日本語に関わる人文社会系研究に結び付ける可能性を論じた。特に、テキストマイニングツールの事例として、樋口耕一(2014)が開発を進めている R などのテキスト・マイニングプログラムを視覚的に処理できる「KHCoder」を中心に紹介する。

手順としては、まず、「KHCoder」を紹介する。次に、こうしたツールで得られた結果を言語資料の質的分析に結び付けて、内容の把握を試みる。そして、こうしたツールを活用した、今後の人文社会系研究との結びつきを論じる。このようにして、台湾の日本語教育現場やカリキュラムへの AI 技術応用の端緒が生まれてゆけば何よりである。

キーワード：AI 技術、テキストマイニング、日本語教育、
人文社会系研究、質的分析

airiti

テキストマイニング技術の応用と発展可能性の探究
―日本語教育および日本関係人文社会系研究との連繋と協働をめざして―

落合由治
淡江大学日本語文学科教授

1. はじめに

1980 年代に初めて人間の言語をプログラミングで扱うことができる処理が可能になり、第二次 AI ブームが起こってから、技術的な壁にぶつかって試行錯誤を続けていた自然言語処理（コンピューター言語に対する人間の言語の処理）は、2000 年代にプログラムの機械学習が試行されるようになって新しい発展が始まった。[1]2010 年代に入って、その成果が実り、深層学習などプログラムの自立的学習を進める技術が発展した結果、物理的データ、数量化データの処理で飛躍的な性能向上が見られ、第三次の AI 技術として様々な分野に社会的応用が拡がるようになり、現在、その発展が急速に進んでいる。[2]第三次 AI の発展は、産業革命の時代とも認識され、プログラミングの応用の広がりと同時に計算機の面でも量子コンピューターの実用化に拍車がかかっている。[3]

---

[1] 自然言語処理の歴史と基本技術については、奥村学(2010)『自然言語処理の基礎』コロナ社参照。

[2] 現在の AI 技術のビジネス応用は野村直之(2016)『人工知能が変える仕事の未来』日本経済新聞出版社、社会との関係は、総務省『情報通信白書』を参照。一例として、AI の発達と今後の社会については『平成 27 年版情報通信白書　特集テーマ　「ICT の過去・現在・未来」』http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h27/pdf/index.html 参照。なお、リンクの確認は 2020 年 2 月 20 日現在で、以下同様。

[3] 量子コンピューターの開発は、各国で進んでいるが、日本の例として週間 BCN＋(2019)「“6 万量子ビットの量子コンピューター”相当で名刺サイズのアニーリングマシンを日立が開発。エネルギー効率の向上で IoT 機器への実装が可能に」https://www.weeklybcn.com/journal/news/detail/20190220_166456.html 参照。

airiti

教育面でも、欧米はもちろん日本、台湾、韓国で AI 技術を初等中等教育から実施する計画が進み、今後の大学進学者の学科選択に大きな影響を与えると考えられる。[4]2020 年 2 月現在、武漢肺炎が世界的に流行を始めているが、第三次 AI 技術によって登校しなくても授業が受けられる会議や授業システムなどが応用され、自律学習や反転学習など新しい教育方法を応用した授業への転換のチャンスにもなっている。[5]

こうした動きは、実は、日本語教育を始めとする台湾の日本語関係の人文社会系学科にとって非常に大きな脅威となっている。その資料の一つは、2018 年の国際交流基金による「2018 年度海外日本語教育機関調査結果」である。資料に拠れば、台湾では 2015 年の調査と比べて 49886 人学習者が減っている。日本台湾交流協会の調査によると、その理由は中等教育では日本語などの外国語よりも、理数系、プログラム系の選択科目を保護者が生徒に選択するように薦めていた結果、学習者が大幅に減ったことが減少の原因と考えられている。世界全体の調査でも学習者が新たに生まれたり、今までより増加したりした国もあるが、今まで日本語教育が普及していた国を中心に 41 ヶ国では 126573 人減少し、従来のように日本語学習者が順調に増えていく状況はすでにないことがわかる。第三次 AI ブー

---

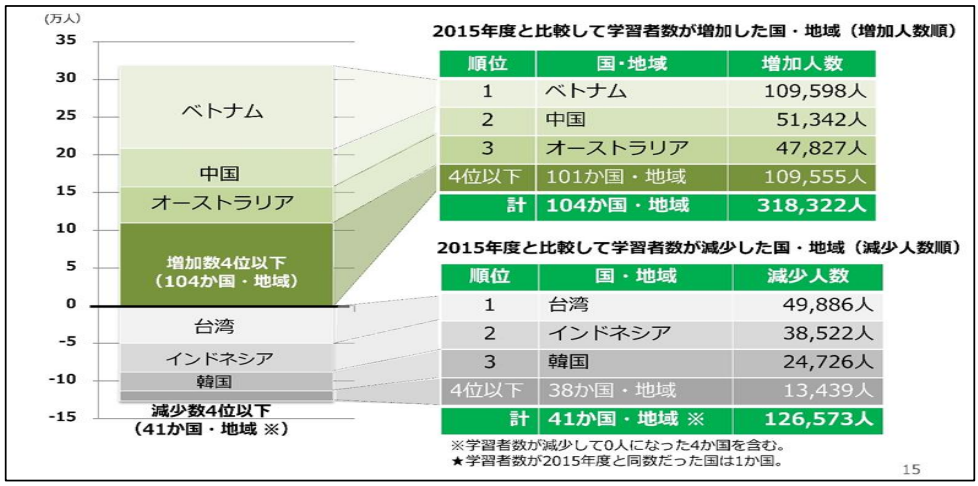
[4] 日本の方針は、文部科学省(2019)「Society 5.0 に向けた人材育成―社会が変わる、学びが変わる」https://www.mext.go.jp/component/a_menu/other/detail/__icsFiles/afieldfile/2018/06/06/1405844_002.pdf 参照。台湾では、教育部(2019)「AI 教育 X 教育 AI－人工智慧教育及數位先進個人化、適性化學習時代來臨！」https://www.edu.tw/News_Content.aspx?n=。9E7AC85F1954DDA8&s=D4C4CD32CAE3FF5D 参照。世界各国の AI 関係教育は、Politcs+AI（2018）「An Overview of National AI Strategies」https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd 参照。

[5] 日本ではオンライン会議システム「Zoom（ズーム）」を応用して休校時にも授業ができる取り組みがされている。EdTheckZin(2020)「新型コロナウイルス休校措置校などに対し、eboard がオンライン教材を無償提供」https://edtechzine.jp/article/detail/3349。台湾でも、オンライン授業で隔離や移動禁止の学生への授業が始まっている。天下雑誌(2020)「無法返校怎麼辦？武漢大學要「教師不停教、學生不停學」https://www.cw.com.tw/article/article.action?id=5098848。

airiti

ムは産業革命であり、日本語学習を巡る環境変化は非常に急激で、特に東アジア圏では時代の変化に即応する教育の革新が求められている。台湾における日本語教育の発展、日本関係の人文社会系研究の普及をすすめるには、AI 技術の発展による産業構造の急激な変化に対応した、教育内容、研究内容の変革が不可避と考えられる。[6]

図 1　国際交流基金「2018 年度海外日本語教育機関調査結果」[7]



台湾における日本語教育および日本関係の人文社会系研究をどのように、今まで関係のなかった AI 技術と結び付けていくかは、未来の方向性を決める上で最も重要な課題のひとつと言える。その点を考える場合、現在の AI 技術発展の中心となっているのは、メディアで注目されているビッグ・データなど数値的データ処理ばかりではなく、人間の言語を処理する自然言語処理技術の大きな発展であることは注目すべき手掛かりと言えよう。2010 年以降の AI 技術の普及は、言語情報処理の大きな進歩によって現在の生活に AI が入

[6] 社会変動、職業変動予想の一例として、総務省(2018)『平成 30 年版情報通信白書:特集　人口減少時代の ICT による持続的成長』https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h30/pdf/index.html 参照。

[7] 国際交流基金(2019)「2018 年度海外日本語教育機関調査結果」速報値 https://www.jpf.go.jp/j/about/press/2019/dl/2019-029-02.pdf

airiti

りこむきっかけとなり、社会全体に大きな影響を与えている基盤になっている。人間の言語を処理する自然言語処理は 1980-90 年代の第二次 AI 技術の時代から本格的な処理が始まったが、第三次 AI 技術の中心のひとつは自然言語処理で、現在、機械学習を中心に質問応答システム、機械翻訳、医学や法律などの専門的情報処理、対話システム、必要な情報を集める統計的潜在意味分析等、人間の言語（自然言語）をプログラムが処理するさまざまな手法が試行錯誤されている。[8]その中で直接人文社会系の研究や教育に応用しやすいのは、2000 年代から利用が広がってきた言語データの量的処理によって目的のデータ特徴を発見するテキストマイニング技術である。[9]

本稿では、台湾での日本語教育および日本関係の人文社会系研究を広く人文社会系研究（語学、文学、歴史、思想、日本語教育学）とする。そして、それらが AI 技術と接続していく今後の可能性について、今までも利用されており、そこからの発展、深化が期待できるテキストマイニング技術との接続、応用を中心に考察を行っていきたい。合わせて、現在、各国で提唱されている AI 技術教育と人文社会系研究・教育との関係も見ていきたい。なお、表の資料はすべて論者によるものである。

2. AI 技術における自然言語処理の発展

まず、現在までの AI 技術の発展について述べる。AI の定義は、

---

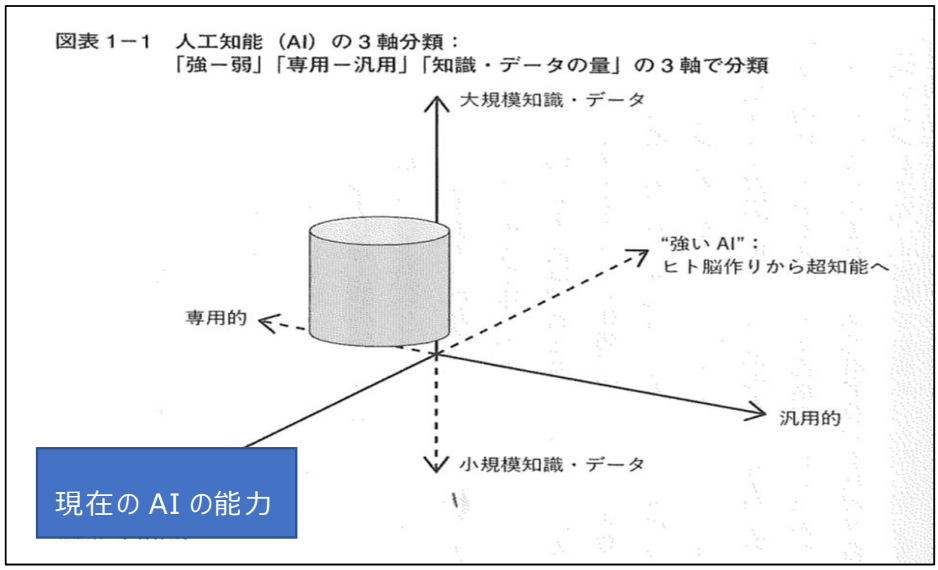
[8] 自然言語処理の最近の処理の全体像はグラム・ニュービッグ、萩原正人、奥野陽編、小町守監修(2016)『自然言語処理の基本と技術』翔泳社、坪井祐太、海野裕也、鈴木潤(2017)『深層学習による自然言語処理』講談社参照。機械学習の基礎については、奥村学監修、高村大也(2010)『言語処理のための機械学習入門』コロナ社参照。最近の自然言語処理技術の動向は、AINOW(2020)「2019 年は BERT と Transformer の年だった」https://ainow.ai/2020/02/25/183082/参照。

[9] 計量言語学的手法で発展したテキスト・マイニングの技法は金明哲(2007)『R によるデータサイエンス―データ解析の基礎から最新手法まで』森北出版、具体的な応用方法は樋口耕一(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めざして』ナカニシヤ出版、李在鎬(2017)『文章を科学する』ひつじ書房参照。

airiti

現在、発展中の技術であるため非常に曖昧で、広範囲であり、定義は研究や応用の方向によってそれぞれ大きく異なっている。[10]ここでは、自然言語処理に関わる情報通信技術という暫定的な定義としておきたい。その点で、現在の AI 技術は、特定の目的の処理を行うことに特化した「弱い AI」であり、それぞれ目的によって特性が異なっている。大きく分類すれば、以下の図 2 のように「強-弱」、「専用-汎用」、「知識・データ量」でそれぞれの AI が存在している。

図 2 人工知能(AI)の 3 軸分類[11]



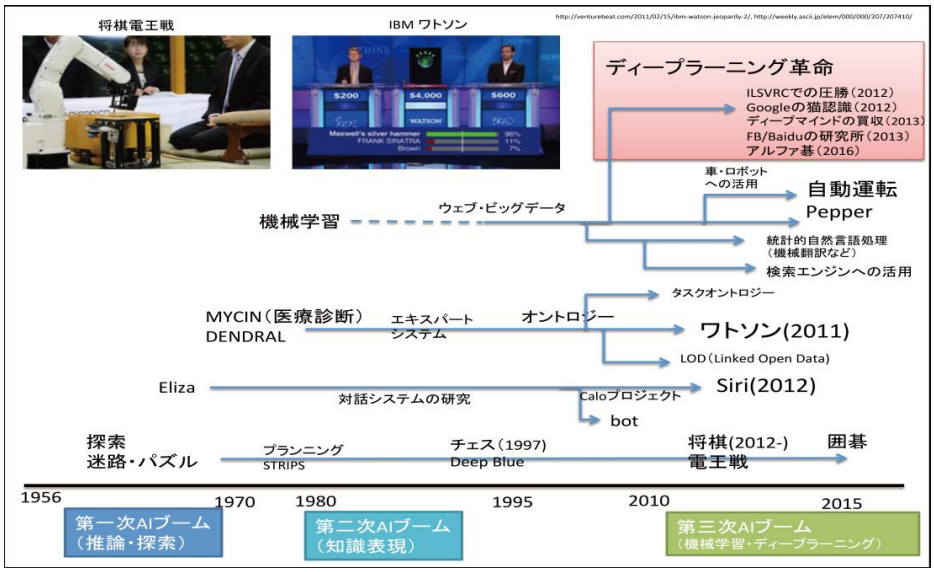
現在までの AI は、すべて特化型 AI の開発で、現在までのところ、メディアで喧伝されている汎用的 AI「強い AI」の開発や、その能力が人間を超えるようなシンギュラリティーを実現する方法、技術は存在していない。[12]

---

[10] 同注 2 野村直之(2016)第 1 章参照。
[11] 同注 2 野村直之(2016)p.43 による。
[12] 2019 年現在、第三次 AI ブームを支えた深層学習に限界があることが指摘されている。まったく新しい方法が生まれない限り、現在のデータ処理の限界を超えることは難しい。AINOW(2019)「ディープラーニングはすでに限界に達しているのではないか？」https://ainow.ai/2019/02/18/161998/等を参照。

airiti

図 3　AI の発展の歴史[13]



現在、AI の社会的利用が拡大する環境では、図 3 のように AI は汎用化ではなく高度に機能的分野的に分化していく可能性が高く、特定の機能に特化した課題、分野では、従来の人間の仕事、作業を容易に代替できるようになっており、実際に現在、事務部門、作業部門等で情報化が急速に進み、職業面での変動が拡がっている。[14] こうした社会的変化の中で、教育機関では訓練すべき能力を再検討、再編成する必要に迫られており、現在の学校で教えている知識や技術が AI 化社会の中で、どう発展し、職業的社会的に応用できるかを考えることは必須の課題と言える。大学教育も、もちろんそれに応

[13] 人工知能の技術的歴史については、人工知能学会「人工知能の歴史」https://www.ai-gakkai.or.jp/whatsai/AIhistory.html および、総務省(2016)「第 4 章第 2 節人工知能（AI）の現状と未来」『情報通信白書平成 28 年度版』http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h28/pdf/n4200000.pdf に拠る。図は各時代に開発された技術を示している。松尾豊（2015）『人工知能は人間を超えるか―ディープラーニングの先にあるもの』KADOKAWA による。

[14] AI に代替される仕事については各種予想が出ているが、一例として、週刊現代(2018)「これから給料が「下がる仕事」「上がる仕事」全 210 職種を公開」https://gendai.ismedia.jp/articles/-/55428 参照。

airiti

じて、情報化 AI 化できない分野の能力を育成し、さらに AI 技術を応用した新しい職業、分野に結合できる教育と研究内容や課程デザインを産み出す必要性に迫られている。[15]以下では、現在の日本語教育および日本関係の人文社会系関係研究の領域を広く人文社会系領域と考え、人文社会系領域と接続しやすい情報科学の分野として自然言語処理の領域で、応用できる技術について述べていきたい。

3. 自然言語処理の人文社会系領域の質的研究への応用

最初に、人文社会系領域の研究に対してすぐに応用が可能な AI 技術として、量的研究について述べたい。すでに第二次 AI 技術が拡がった 1980 年代から計量言語学、社会学、心理学、経営学などでは、言語資料を量的に処理して、特徴を取り出す研究が始まり、応用が進んできた。これらは 2000 年代にデータマイニング、テキストマイニングとして応用がさらに拡がった。[16]2010 年以降の第三次 AI 技術の発展で、言語データから特徴を取り出す手法が開発され、質的研究への接続も可能になってきた。現在までの日本語関係の自然言語処理を応用した研究では、計量言語学的手法(量的研究)が代表的であろう。主なテーマになっているのは語彙の統計的処理による特徴の抽出で、すでに 1950 年代から電算機を使った研究が開始され、語彙(品詞)の定量計測によって、各文章ジャンルでの品詞や記号分布や量的文体特徴の把握が行われ、2000 年代からは各種コーパスの作成で幅広く応用されている。[17]また、計量言語学的手法は作者を

---

[15] 現在、欧米圏で進んでいる教育改革の様子の一例は、EdTecZine（2019）「日本教育にイノベーションを ～AI 時代に本当に必要な教育とは～」https://edtechzine.jp/article/corner/35 参照。

[16] データマイニングとは、データサイエンスとも呼ばれ「データの収集、加工、蓄積、流通、解析などに関する科学の総称」で科学的な発見の方法として「実験、観察、記録、調査などにより得られたデータから規則、パターン、知識を見つけ出す統計学」的手法であり、言語データの場合は、テキストマイニングと呼ばれる。定義は、同注 9 金明哲(2007)p.2 参照。

[17] 現在まで日本語コーパスの開発は進み、各種にコーパスが公開されるようになっている。ビッグ・データから人間の処理に拠らずに AI 自身で自然言語処理

airiti

特定する計量文献学の方法としても発展してきた。一方、日本語学や日本文学など言語表現の質的理解や特徴把握の研究を中心にしている学科と、こうした量的研究のテーマとは距離が大きかった。日本語学や日本文学など言語を資料としてきた人文社会系研究での手法は質的研究を中心にしており、人間の解釈により語や文の用例分析、解釈、分類によって表現の機能を解明しようとするため、計量的研究が使っていたデータマイニングやテキストマイニング技法の応用は限定的であり、人文社会系研究の中では、今まで量的研究と質的研究が分離していた。[18]

だが、第二次 AI 技術の発展の中で、2000 年代以降、日本語学や日本文学等の人文社会系研究に活かせる各種コーパスの整備が進み、パーソナルコンピューターの性能と自然言語処理技術の飛躍的進歩もあって、従来は量的にしか扱えなかった言語データを質的に捉える自然言語処理の応用が可能になっている。さらに、第三次 AI 技術の発展では、ビッグ・データから人間の処理を経ずに AI 自身で自然言語処理が可能な機械学習手法が発達し、言語の質的内容を反映した自然言語処理が実施できるようになってきた。[19]

データマイニング中で、自然言語処理の応用として言語データから有意義な規則、パターン、知識を見つけ出すテキストマイニングは、現在、幅広い研究分野で活用されるようになっている。最も応用が普及しているのは経営学関係で、市場調査、製品評価、顧客分析など消費者の意識と行動を調査する基本的手法として活用され、専門の調査会社もコンサルティング業務の基本的方法として活用し

を行う手法として自律学習等でもこうしたデータをデータセットして応用する研究が進んでいる。コーパス例として、国立国語研究所コーパス開発センター https://pj.ninjal.ac.jp/corpus_center/goihyo.html 参照。

[18] 日本語学、日本文学、日本語教育の分野では量的研究と質的研究の相違に関する議論は今まで十分とは言えなかった。社会科学分野での議論を参照すれば、問題の本質が分かりやすい。一例として、大谷尚(2017)「質的研究とは何か」『Yakugaku Zasshi』137-6pp.653-658 参照。

[19] 自然言語処理入門の歴史は、黒橋禎夫(2015)『自然言語処理 (放送大学教材)』放送大学教育振興会参照。

airiti

ている。会社などの組織研究での応用も進んでいる。[20]社会科学系研究ではテキストマイニングの利用が一般化し、樋口耕一(2017)は、社会系分野でのテキストマイニングの質的分析への応用と発展可能性を述べている。[21]今まで、テキストマイニングは日本語学、計量文献学、計量国語学等での量的分析で使用されてきたが、研究テーマとして資料間の差異の検出や語彙、品詞の計量など特定の課題にしか適用できず、言語あるいは文学、歴史等人文社会系で問題となっている内容的特徴や意味的特徴の抽出とは距離が大きかった。だが、社会科学の一定の分野では量的方法と質的方法の特性をそれぞれ活かす方法論的検討が近年、進んできており、その理論的検討は質的研究が中心の人文社会系研究全体にも非常に示唆的である。久保田賢一(1997)は、二つの研究のパラダイムの相違を整理し、量的研究と質的研究の違いは価値の優劣の問題ではなく、パラダイムの相違であり、研究目的と対象の相違であると述べている。[22]今まで人文社会系研究では数量化を客観性の基準に置く量的研究と資料の意味的解釈を重視する質的研究は対立的立場と見なされ、研究の評価において優劣善悪の視点で相互の研究を批判する傾向が強かった。しかし、実際には研究方法と目的の相違という客観的問題であり、研究目的と対象が変われば使用すべき研究方法が変わるという方法的問題であることは、改めて認識し直す必要がある。AI 技術の発展と社会変動の中で社会的存在意義が疑問視される現在の人文社会系研究は、従来の自分の受けた教育や訓練をただ反復したり、絶対化した

---

[20] 経営学での応用動向として、疋田眞也、萩原克幸、鶴岡信治(2012)「組織研究におけるテキストマイニングを用いた系統的分析法」『日本情報経営学会誌』32-3pp.97-109、高木修一、竹岡志朗(2018)「経営学におけるテキストマイニングの可能性 ：仮説構築志向の利用方法 (古川勝教授退職記念号)」『富山大学紀要』64-2pp.241-260 参照。

[21] 社会科学分野でのテキストマイニングの応用については、樋口耕一(2017)「計量テキスト分析および KH Coder の利用状況と展望」『社会学評論』68-3pp.334-350 等を参照。

[22] 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。

airiti

りするのではなく、研究方法の問題を意識化し、意識的方法選択によって、新しい価値を生み出していく必要があると言える。

表 1 量的研究と質的研究の比較[23]

比較する点	量的研究	質的研究
焦点量	量(どれくらい)、発見	質(性質、本質)、意味
哲学的前提	客観主義、論理実証主義	現象学、解釈学、象徴的相互作用
関連する用語	実験、実証、統計、客観、中立的	エスノグラフィック、自然主義、データ対話型、主観的、中立的ではあり得ない
目標	予測、制御、確証	理解、描写、協同的構築
研究デザイン	事前に決定、構造的	柔軟、変化していく、次第に明らかになっていく
状況	人工的、操作的、実験室、状況から独立している	自然的、日常的、状況に依存する
標本	大きい、無作為、代表	小さい、意識的
データ収集	データ収集テスト、アンケート、サーベイ調査など	インタビュー、参与観察、日誌など
理論	理論の検証	理論の生成
期　間	(一般に)短い	長い
知見	正確、狭い、還元主義的	理解、全体的、広がり

こうした研究におけるパラダイムと方法論的差異は、教育にも深く関係している。現在、台湾でも導入が進んでいる新しい教授法であるアクティブラーニング、反転学習、自律学習などの淵源は、質的研究を形成してきた構成主義的教育観にあり、従来の知識の伝達を中心にした客観主義的教育観を包含する思潮である。[24]教育面で

[23] 同注 22、久保田賢一(1997)p.166 の表 2 に拠る。原典は、MERRIAM, S.B.(1988) *Case Study Research in Education: A Qualitative Approach.* Jossey-Bass Publishers, San Francisco

[24] 研究パラダイムと教育観の転換については、久保田賢一(2003)「構成主義が投げかける新しい教育」『コンピュータ＆エデュケーション』15p.12-18、久保田真弓、末田清子、浅井亜紀子、小池浩子、小柳志津(2010)「異文化コミュニケーション研究と教育の歩みと展望 ―実証主義から構成主義へのパラダイムシフトの視点から―」『青山国際政経論集』81pp.97-118、久保田賢一(2012)「構成主義パラダイムの学習理論」『情報研究:関西大学総合情報学部紀要』36pp.43-55 参照。

airiti

の転換を進めていくためにも、研究パラダイムの再認識と意識化は、相即的関係にある。本稿では、台湾の日本語の人文社会系研究にテキストマイニングを活用する事例研究として、量的手法を質的研究に活かす社会学での内容分析で実績をあげてきた樋口耕一(2014)の開発した「KHCoder」を取り上げ、量的手法の人文社会系研究の質的応用への可能性を探ってみたい。

4. 質的研究としての人文社会系研究へのテキストマイニングの連繋と協働の可能性

「KHCoder」は樋口耕一により開発された、Perl、R などのテキストマイニングで常用されるプログラミング言語をメニュー形式で操作し、結果を視覚的に表示できるプログラムで、日本語、英語、中国語、韓国語等 13 カ国語が処理できるようになっている。テキストマイニングに使われるプログラミング言語は、プログラミングを全部自分でひとつずつ書く必要があるが、「KHCoder」では、よく分析に使われる分析手法がメニューで選択でき、テキストマイニングに必要な前処理(言語資料を語に句切って分析できるようにする処理)も行えるので、定型的作業で資料の分析を行える非常に優れたテキストマイニングツールである。特に、初心者にも操作しやすいインターフェースなので、情報処理に不慣れな教員でも学生でも自分でテキストマイニングができるように練習しやすい。人文社会系の学生に AI 技術の教育をする場合も、自然言語処理の応用例として、自然言語処理とはどのような手順で行い、結果をどう扱えばよいかを容易に示すことができ、応用方法を簡単に示すことができる。[25]以

[25] KHCoder は https://khcoder.net/で公開されているフリーソフトである。KHCoder では、資料の前処理と抽出語について、抽出語リスト、出現度なので記述統計、語例の KWIC コンコーダンスと、テキストマイニングとして対応分析、多次元尺度構成法、階層的クラスター分析、共起ネットワーク、自己組織化マップが使用できる。サイトは使い方と応用例の論文が公開されている。日本では、すでに KHCoder を使った、社会学、心理学、看護学、メディア研究など 3000 本以上の論文が発表されており、量的手法を質的研究に活かす方向性の模

airiti

下に応用例を示した。分析例は、最近、日本で話題になっている武漢肺炎についての話題を取り上げ、資料として文章量がほぼ同じになるように朝日新聞(3 本)と読売新聞（3 本）の 2020 年 2 月末の社説を取り上げた。[26]

まず、「KHCoder」では、テキストマイニングの前処理により簡単に資料を分析単位である単語に分割できるため、名詞、動詞、形容詞などの頻度順リストを簡単に作成でき、EXCEL 出力できる。その結果を表 2 の頻度順リストに示した。頻度順リストによって、記事中でのキーワードの重要性を比較推測できるようになる。また、教育現場で利用する際、学習者に簡単に語彙表を提供することが可能なため、JLPT の級別で語彙の難易度を提示するなどの方法で主な語について読解の前に注意を促し、読解の際の学習者の語彙理解を助けることができる。読解に際しては、共通する語彙は両社で同様の話題が出ている注目点を示し、それぞれに特有な語彙はその社でのみ取り上げた話題であり、内容理解の手掛かりになる。朝日新聞では、「人 7」「国民 5」「説明 5」「医療 5」などが特異で、読売新聞の場合は、「定年 8」「延長 7」「拡大 6」「企業 6」が特異な語である。各社の意図を推測すると、朝日新聞は武漢肺炎の国民への説明や医療への不安を取り上げ、読売新聞は武漢肺炎に関わる企業システム

索が続いている。応用論文はサイトでリスト化されている。「KH Coder を用いた研究事例」https://khcoder.net/bib.html?year=2019&auth=all&key=。

[26] 使用した資料は以下である。朝日新聞「（社説）新型肺炎対策　きめ細かな現場支援を」2020 年 2 月 27 日 5 時 00 分 https://www.asahi.com/articles/DA3S14380981.html?iref=pc_rensai_long_16_article、朝日新聞「（社説）全国一斉休校　影響の軽減に全力注げ」2020 年 2 月 28 日 5 時 00 分 https://www.asahi.com/articles/DA3S14382477.html?iref=pc_rensai_long_16_article、朝日新聞「（社説）休校の決断　重みに見合う説明を」2020 年 2 月 29 日 5 時 00 分 https://www.asahi.com/articles/DA3S14384067.html?iref=pc_rensai_long_16_article、読売新聞「衆院集中審議　新型肺炎対策を掘り下げよ」2020/02/27 05:00 https://www.yomiuri.co.jp/editorial/20200226-OYT1T50376/、読売新聞「全国臨時休校へ　混乱抑え感染防止に全力を」2020/02/28 05:00 https://www.yomiuri.co.jp/editorial/20200227-OYT1T50338/、読売新聞「新型肺炎　医療と経済に全力で取り組め」2020/02/29 05:00 https://www.yomiuri.co.jp/editorial/20200228-OYT1T50387/。

airiti

の問題を主な論点にしていると考えられる。テキストマイニングは、このように、複数の資料グループでの内容差や論点差を容易に推測する手掛かりを与えている。今回の新聞記事の場合、日本では基本的な編集方針が大きく異なる新聞社二社を選ぶことで、こうした手掛かりから同じ問題でも新聞社の立場で見解が大きく異なることを学習者に気づかせることができ、ただ日本語として記事の意味を理解するだけではなく、実は新聞などの情報には常に一定の編集目的による偏差があるというメディアリテラシー的な読解に結びつけることが可能になる。

表 2　朝日新聞と読売新聞の記事の使用単語頻度順リスト

(左：朝日新聞/右：読売新聞　以下同様)

1	感染	サ変名詞	15
2	休校	サ変名詞	12
3	政府	名詞	10
4	首相	名詞	9
5	人	名詞C	7
6	全国	名詞	7
7	対策	サ変名詞	7
8	措置	サ変名詞	6
9	対応	サ変名詞	6
10	方針	名詞	6
11	医療	名詞	5
12	一斉	副詞可能	5
13	国民	名詞	5
14	説明	サ変名詞	5
15	専門	名詞	5

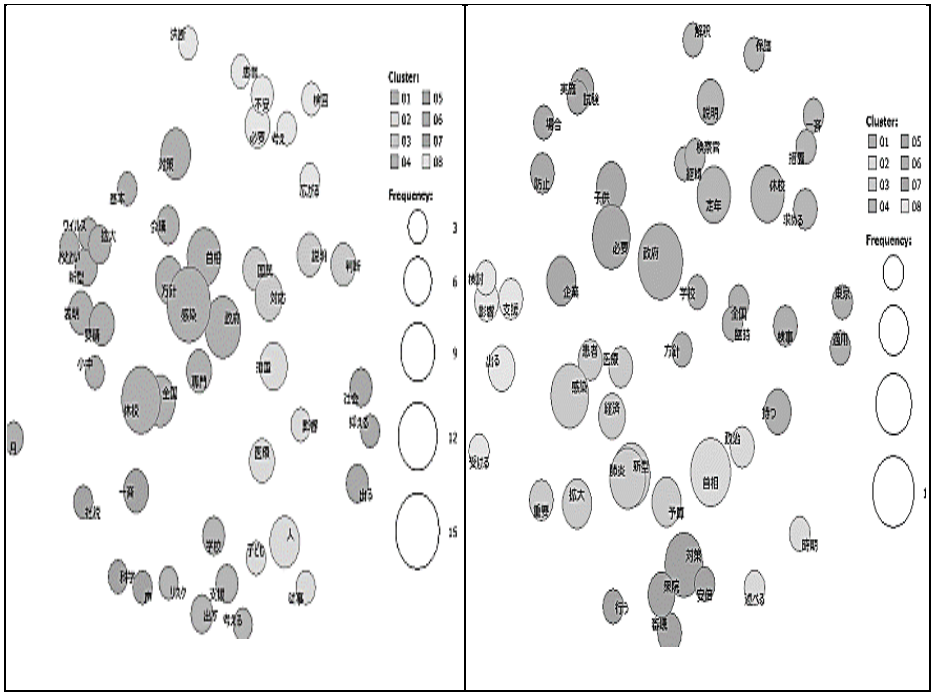
1	政府	名詞	14
2	首相	名詞	11
3	感染	サ変名詞	10
4	新型	名詞	10
5	対策	サ変名詞	10
6	必要	形容動詞	10
7	肺炎	名詞	9
8	休校	サ変名詞	8
9	定年	名詞	8
10	延長	サ変名詞	7
11	拡大	サ変名詞	6
12	企業	名詞	6
13	子供	名詞	6
14	予算	名詞	6
15	経済	名詞	5

今まで人文社会科学分野で応用されてきたテキストマイニングは大きな資料やコーパスから特徴となる傾向や語を見つけ出す目的で使われ、集団的特徴を持ったデータを扱い、一般化を目指してきたため小範囲の資料の質的分析にテキストマイニングが有効かどうかは主な検討課題にはなっていなかった。一方、質的研究を行う人文社会系研究の場合は、比較的限定された個別的資料の内容的意味的特徴を解明するのが基本であるが、今回、例とした少量の言語資料でも、キーになる単語を容易に抽出でき、自分で読解したり、手作業で分析したりするのに比べて、はるかに迅速、容易に質的特徴

airiti

となる部分を見出すことができることは新しい応用可能性を開く点と言える。内容を読みとる点を基礎にすることで、テキストマイニングは質的分析と読み取りに活用できるのである。

表 3　朝日新聞と読売新聞の
記事使用単語の多次元尺度構成法の結果

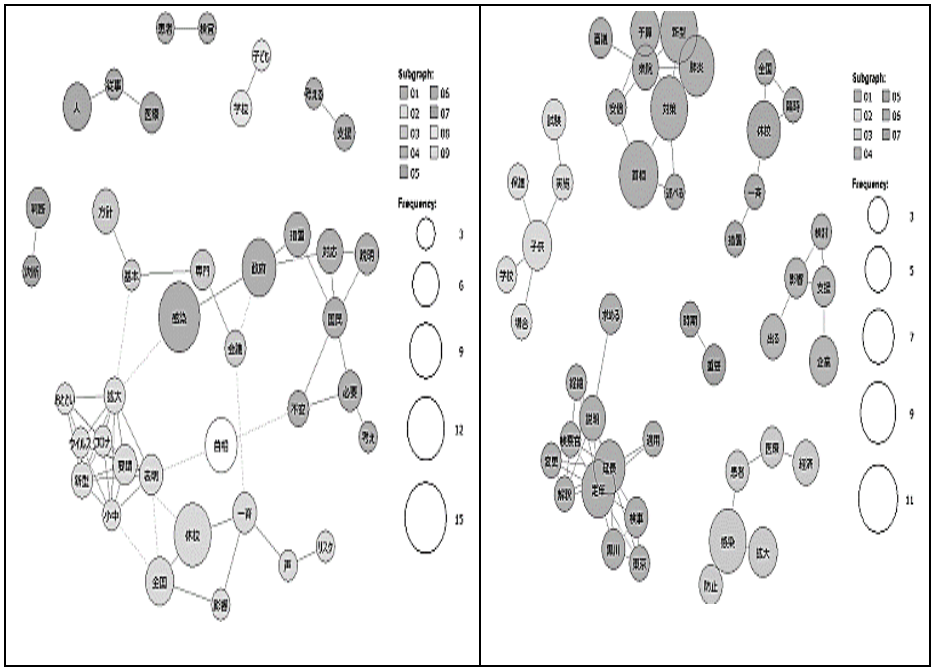


さらに、その他の分析手法を見ていきたい。表 3 の多次元尺度構成法は、共起関係の強い単語を近距離に配置し、ゼロ座標附近にその資料での常用的な語が集中して、周辺には特異な語が分布する語彙の分布状況を視覚化できる。KHCoder では、語彙の共起関係の強さによるグループわけであるクラスター分析を併用して共起関係の強い語をクラスター別に表示できるため、頻度順リストよりも資料中の内容の重点をさらに容易に読み取ることができる。従来は、数値的に両者を比較し、類似度を検出する方法が計量言語学では用いられていたが、質的分析としてテキストの読み取りに重点を置けば、

airiti

出現語のクラスター分類は内容に含まれる潜在的な複数のテーマを示していると考えられる。朝日新聞では、「患者、不安、検査、必要」（右上）、「措置、医療、影響」（右）、「国民、対応、説明」（右中）など中央のグループが示している政府の措置に対する批判や説明を求めるグループが目立ち、読売新聞では「検察官、定年、休校、措置」（右上）、「企業、子供、必要、防止」（左上）、「患者、経済、支援、医療」（左下）、「衆院、審議、安倍」（中央下）等の政府に具体的な政策や対策を求める内容に関わるテーマが多く抽出され、両社の社説に含まれる主な内容の差異が明確化できる。「KHCoder」では、各語をクリックすると資料での用例が表示できるため、原文での意味の確認も容易にでき、用例の収集も簡単におこなえる。

表 4　朝日新聞と読売新聞の記事使用単語の共起ネットワーク



続いて、表 4 のように記事に出ている語の共起ネットワーク図を作成すると、出現パターンの似た語を線で結ぶ形式で語彙の共起関係の種類分けをクラスターごとに表示できる。朝日新聞では、中央

airiti

下「首相」の語を中心に、「感染、政府、対応、国民、不安、説明」（右）、「休校、全国、一斉、リスク」（下）、「新型、ウイルス、拡大、表明」（左）が出ており、首相に対して、感染対策や休校措置について、国民に説明を求める内容が中心になっている。逆に、読売新聞では、「首相、衆院、対策、予算、肺炎」（上）という国会審議を巡る点が中心で、「全国、休校、臨時」（右上）、「企業、支援、影響、検討」（右中）、「患者、医療、拡大、防止」（右下）、「検察官、定年、延長、説明」（左下）というように、国会での審議内容への論究がポイントであることが分かる。共起ネットワークの繋がりが強く、また、関係図のネットワークの多いものは、それぞれ重要な内容を表示していると考えられる。こうしたテキストマイニングにおける語彙の共起ネットワークは、日本語学での概念的階層性に基づく語彙のネットワークのようなラング的で一般的な意味による語彙ネットワークではなく、そのテキストの中で使用されている語彙の共起関係に基づく、パロール的で固有な用法による語彙ネットワークで、そのテキストの中での主題に関係した特徴的語彙の分布を示している。こうした共起ネットワークを手掛かりにして、その記事の要点を抽出したり記事の主題にあたる内容を特定したりすることができ、従来の日本語としての語彙的理解から一歩踏み込んだ内容理解に学習者を導くことができる。

以上、3 つのテキストマイニング手法で、新聞社説を例にして、質的分析に結び付ける読み取りについて述べてきたが、新聞記事のような文章の基本的類型（話題を並列させて焦点を産み出す文章構成）では、テキストマイニングは、文章の基本的話題をかなり的確に取り出すことができ、質的読み取りに非常に有益なアシスタント機能を提供していると言える。

5. 人文社会系研究とテキストマイニングとの連繋と協働の探究

本論では、事例を示す形で R 等によるテキストマイニングをメニューで操作し、視覚的に結果を出力できる「KHCoder」を使って、

airiti

テキストマイニングを実施し、質的読解に結び付ける結果を示した。これまでも計量的言語研究や社会学、心理学等の分析でもこうしたテキストマイニングは行われていたが、その目指す対象と目的は、個別的な対象の特性を見出す人文社会系の質的研究とは異なっていた。今回、台湾での日本語学、日本文学、歴史学、日本語教育研究、社会学質的調査分析等の質的研究に活かす目的で、質的データの読み取りに重点を置いて、テキストマイニングを実施した。その結果、各資料の持つ意味のまとまった表現をテキストマイニングの手法で取り出すことができ、それによって内容のテーマを容易に抽出できることが明らかになった。こうした手法は、研究はもちろん、教育現場でも授業や学習者の読解のサポート等に幅広く使用できる。キストマイニングは、今回取り上げたような文章の基本的類型（話題を並列させて焦点を産み出す文章構成）で書かれる通常は読解困難な抽象的な思想、宗教、評論などの文章で、範囲を句切ってテキストマイニングを実施することで、資料の基本的キーワードとそのまとまりが形成している要点を抽出できるため、研究、教育、学習におい、資料の読解、意味の理解に極めて有益な手掛かりとなり、人文社会系分野に広く応用できる可能性を持っていると言えよう。

テキストマイニングは元来、中立的技術で、対象と目的により出てくる結果の読み取り方を変更すれば、今まで利用して来なかった人文社会系の研究や教育に幅広く応用できる技術になっていると考えられる。また、従来は数値でしか表示できなかった結果が現在では視覚的に表示することができ、意味的把握が容易になり、語彙の共起関係も明確化できるので人文系の研究や教育に応用しやすい環境も生まれている。以下に、テキストマイニングに関わってきた従来の分野と、新しく応用を始める人文社会系の分野を比較して、整理した。

元々、言語データの量的処理は 1950 年代に第一次 AI ブームでコンピューターとプログラムが実用化されたときにコーパス作成や電子的データの統計的解析などを行う計量言語学の分野が誕生したと

airiti

ころから始まり、以後、その手法が人文社会科学の各分野に拡がって、データの量的処理が発展してきた。しかし、1980 年代から量的研究だけでは解明できない各種の質の異なる問題が明らかになったことで質的方法が提唱され、以降は量的研究と質的研究が相互に対立する立場の中で各分野において展開されるようになっている。[27]

表 5　各分野でのテキストマイニングの応用

項目	社会学、心理学、教育学、福祉、医療	経営学	計量言語学 計量文献学	人文社会系（語学、文学、歴史、思想、日本語教育学）
対象	社会現象、人間行動、心理的現象の規則性、量的質的特徴	ビジネスに関する社会現象、心理要因、消費行動、システム	言語単位の量的処理、資料間の差異	資料となる文献、データの質的特徴、内容的理解
目的	言語データの分析による社会現象、心理現象の一般的規則性あるいは個別的特性の解明	言語データの分析によるビジネス関係の現象の量的質的特性の解明	言語データの量的分析による言語の一般的特徴の解明	言語データの質的な読解、整理、解釈による資料の一般的あるいは個別的特徴の解明
方法	量的方法と質的方法が対立	量的方法と質的方法が併用	量的方法	量的方法と質的方法が対立
応用	内容分析、インタビューやアンケートの記述内容の質的分析、SNS などの言語データの質的分析	インタビューやアンケートの記述内容の質的分析、SNS などの言語データの質的分析	言語データの量的分析と整理、統計的処理	言語データの量的分析を手掛かりにした質的分析手法の開発 教育での質的分析導入
成果	質的研究で広く応用	ビッグ・データ分析で広く応用、商業化	データベース作成、資料間の差異の検定	開拓途上

[27] 量的研究と質的研究の特徴については注 9 樋口耕一(2014)参照。また、同注 22 久保田賢一(1997)等を参照。

airiti

今まで、いわゆる人文系と言われた語学、文学、歴史、思想、日本語教育などの言語データの個別的特徴を捉え、意味的内容を分析、整理し資料との相互作用的解釈を行ってきた分野は、特に量的方法との関係づけが困難で、第二次 AI ブーム時代の自然言語処理技術を言語データの質的分析に応用する方法も難しかった。だが、現在の第三世次 AI ブームの技術革新により、自然言語処理にとって言語の意味的特徴を捉える技術、手法が発達し、同時に結果を視覚的に提示できるようになったため、飛躍的に応用範囲と効果が拡大し、従来、質的研究を行ってきた人文社会系分野との接続が容易になっている。いかにして量的方法を活用し人文社会系分野の特徴である質的方法の利点を発揮するかは今からの大きな課題と言える。数値データの示す意味を元の資料に即して解釈するのは人間の質的活動そのもので、まさに質的研究としての人文社会系分野の中心となる研究方法である。テキストマイニングは日常の日本語教育内容である読解、社会文化理解また、アクティブラーニング等の新しい教育法に活用できる方法である。

6. おわりに

現在、人文社会系の教育と研究は、社会的環境変化の中で、大きな試練に立たされている。しかし、社会的に注目される AI 技術や自然言語処理など、今まで直接関係していなかった分野で応用できるものは何かを明確に見極めることで、技術変動に連動した今後の社会変動にも人文系の研究の進路を見つけることができる。今回、テキストマイニングを質的研究に活かす事例を提示したが、現在の自然言語処理は人間の質的読解や解釈に有意義な示唆を提供できるところに接近しており、自然言語処理に関わることで人文社会系の従来の知見を活かせる可能性も大きい。今後、分野間の連繋を図っていくことで、相互作用的に新しい可能性を展開できるにちがいない。テキストマイニングを質的研究と結び付けて、人文社会系分野の自然言語処理への接続の入り口に利用し、そこから可能性を広げてい

airiti

くことで、これからの時代への対応の道が開けるにちがいない。

使用データ

①朝日新聞「（社説）新型肺炎対策　きめ細かな現場支援を」2020 年 2 月 27 日 5 時 00 分 https://www.asahi.com/articles/DA3S14380981.html?iref=pc_rensai_long_16_article、②朝日新聞「（社説）全国一斉休校　影響の軽減に全力注げ」2020 年 2 月 28 日 5 時 00 分 https://www.asahi.com/articles/DA3S14382477.html?iref=pc_rensai_long_16_article、③朝日新聞「（社説）休校の決断　重みに見合う説明を」2020 年 2 月 29 日 5 時 00 分 https://www.asahi.com/articles/DA3S14384067.html?iref=pc_rensai_long_16_article、④読売新聞「衆院集中審議　新型肺炎対策を掘り下げよ」2020/02/27 05:00 https://www.yomiuri.co.jp/editorial/20200226-OYT1T50376/、⑤読売新聞「全国臨時休校へ　混乱抑え感染防止に全力を」2020/02/28 05:00 https://www.yomiuri.co.jp/editorial/20200227-OYT1T50338/、⑥読売新聞「新型肺炎　医療と経済に全力で取り組め」2020/02/29 05:00 https://www.yomiuri.co.jp/editorial/20200228-OYT1T50387/。

参考文献

1）インターネット資料の確認はすべて 2020 年 2 月 20 日現在。

2）図書、雑誌の出版地はすべて日本。

3）図書は部分引用ではなく全体を研究の参考にしている。

AINOW(2019)「ディープラーニングはすでに限界に達しているのではないか？」https://ainow.ai/2019/02/18/161998/

AINOW(2020)「2019 年は BERT と Transformer の年だった」https://ainow.ai/2020/02/25/183082/

EdTecZine（2019）「日本教育にイノベーションを ～AI 時代に本当に必要な教育とは～」https://edtechzine.jp/article/corner/35

EdTheckZin(2020)「新型コロナウイルス休校措置校などに対し、eboard がオンライン教材を無償提供」https://edtechzine.jp/article/detail/3349

大谷尚(2017)「質的研究とは何か」『Yakugaku Zasshi』137-6、pp.653-658

奥村学(2010)『自然言語処理の基礎』コロナ社

奥村学監修、高村大也(2010)『言語処理のための機械学習入門』コロナ社

教育部(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

airiti

金明哲(2007)『R によるデータサイエンス―データ解析の基礎から最新手法まで』森北出版
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』21-3、pp.163-173
久保田賢一(2003)「構成主義が投げかける新しい教育」『コンピュータ＆エデュケーション』15、p.12-18
久保田賢一(2012)「構成主義パラダイムの学習理論」『情報研究:関西大学総合情報学部紀要』36、pp.43-55
久保田真弓、末田清子、浅井亜紀子、小池浩子、小柳志津(2010)「異文化コミュニケーション研究と教育の歩みと展望 ―実証主義から構成主義へのパラダイムシフトの視点から―」『青山国際政経論集』81、pp.97-118
黒橋禎夫(2015)『自然言語処理 (放送大学教材)』放送大学教育振興会全体
グラム・ニュービッグ、萩原正人、奥野陽編、小町守監修(2016)『自然言語処理の基本と技術』翔泳社
KHCoder　https://khcoder.net/
「KH Coder を用いた研究事例」https://khcoder.net/bib.html?year=2019&auth=all&key=
国際交流基金(2019)「2018 年度海外日本語教育機関調査結果」速報値 https://www.jpf.go.jp/j/about/press/2019/dl/2019-029-02.pdf
国立国語研究所コーパス開発センター https://pj.ninjal.ac.jp/corpus_center/goihyo.html
週刊現代(2018)「これから給料が「下がる仕事」「上がる仕事」全 210 職種を公開」https://gendai.ismedia.jp/articles/-/55428
週間 BCN＋(2019)「“6 万量子ビットの量子コンピューター”相当で名刺サイズのアニーリングマシンを日立が開発。エネルギー効率の向上で IoT 機器への実装が可能に」https://www.weeklybcn.com/journal/news/detail/20190220_166456.html
人工知能学会「人工知能の歴史」https://www.ai-gakkai.or.jp/whatsai/AIhistory.html
総務省(2016)「第 4 章第 2 節人工知能（AI）の現状と未来」『情報通信白書平成 28 年度版』http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h28/pdf/n4200000.pdf
総務省(2018)『平成 30 年版情報通信白書:特集　人口減少時代の ICT による持続的成長』https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h30/pdf/index.html
総務省『平成 27 年版情報通信白書　特集テーマ　「ICT の過去・現在・未来」』http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h27/pdf/index.html
高木修一、竹岡志朗(2018)「経営学におけるテキストマイニングの可能性 ：仮説構築志向の利用方法 (古川勝教授退職記念号)」『富山大学紀要』

airiti

64-2、pp.241-260
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』講談社
天下雜誌(2020)「無法返校怎麼辦？武漢大學要「教師不停教、學生不停學」https://www.cw.com.tw/article/article.action?id=5098848。
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
疋田眞也、萩原克幸、鶴岡信治(2012)「組織研究におけるテキストマイニングを用いた系統的分析法」『日本情報経営学会誌』32-3、pp.97-109
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めざして』ナカニシヤ出版
樋口耕一(2017)「計量テキスト分析および KH Coder の利用状況と展望」『社会学評論』68-3、pp.334-350
Politcs+AI（2018）「An Overview of National AI Strategies」https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd
松尾豊（2015)『人工知能は人間を超えるか―ディープラーニングの先にあるもの』KADOKAWA
MERRIAM, S.B.(1988) Case Study Research in Education: A Qualitative Approach. Jossey-Bass Publishers,　San Francisco
文部科学省(2019)「Society 5.0 に向けた人材育成―社会が変わる、学びが変わる」https://www.mext.go.jp/component/a_menu/other/detail/__icsFiles/afieldfile/2018/06/06/1405844_002.pdf
李在鎬(2017)『文章を科学する』ひつじ書房

付記

本論文は、2019 年 12 月の 2019 韓国日本語教育学会大会での招待講演発表を元に編集、加筆、訂正をおこなったものである。また、科技部研究案 MOST 107-2410-H-032 -030 -MY2 の研究成果の一部である。研究へのご支援に心からの感謝の意を表するものである。

※2020 年 3 月 2 日原稿受理　2020 年 4 月 20 日審査通過