考察應用於各類不同文章之文字探勘技術的效益
―目標朝向適用於日本相關領域―

落合由治

淡江大學日文系特聘教授

摘要

本論文主要是應用自然言語處理當中廣泛被利用於語言資料庫之質化分析的文本探勘技術，來掃描檢視日語相關之日本語學、文學、歷史以及社會文化研究等之人文社會研究或日本語教育等教學活動以及研究之現況與課題。以下面 2 個步驟進行。

(1）文章類別之問題：因自然言語處理技術之發達，掌握到言語表現上的基本差異性。藉由文本探勘技術來探討如何掌握文章類別之差異性，並從中明白文章類別之特徵。

(2)文本分析之日本相關領域之應用：文本探勘解析出言語表達之質性差異，是可以直接導入日語相關之日本語學、文學、歷史以及社會文化研究等之人文社會研究或日本語教育等(總稱為日本相關領域)。其中將活用文字探勘技術之日語教育之教室活動，多加著墨來探討。

關鍵詞：文章類別、文本分析、日本相關領域、日語教育、教室活動

受理日期:2021 年 03 月 09 日

通過日期:2021 年 05 月 14 日

A Study of Text Mining Applications Based on Literary Genres: Toward Application in Japan-related Fields

Yuji Ochiai

Distinguished Professor, Department of Japanese, Tamkang University, Taiwan

Abstract

In this paper, I would like to take up text mining, which has been widely applied in qualitative analysis of linguistic data among the applications of natural language processing, and look at the current status and issues of its application to Japanese language studies, literature, humanities and social sciences such as history and social culture studies, as well as educational activities and research in Japanese language education from the following aspects.

(1) The problem of text genre: With the development of natural language processing, it has become clear that there are fundamentally large differences in linguistic expressions. I will discuss how differences in text genres can be grasped by text mining, and what the characteristics of text genres can be understood from this.

(2) Application of text mining to Japan-related fields: Text mining technology has become capable of extracting qualitative differences in linguistic expressions in a pseudo-realistic manner. Text mining technology can be directly applied to Japanese language studies, literature, humanities and social sciences such as history and socio-cultural studies, as well as to educational activities and research in Japanese language education and learning (collectively, Japan-related fields), where qualitative research is a specialty of the department. In particular, I will discuss the classroom activities of Japanese language education that apply text mining technology.

Keywords：literary genres、text mining、Japan-related fields、Japanese language education、classroom activities

文章ジャンルに基づくテキストマイニング応用の考察
―日本関連分野での活用を目指して―

落合由治
淡江大学日本語文学科　特聘教授

要旨

本論文では、自然言語処理の応用の中で言語データの質的分析での応用が広がっているテキストマイニングを取り上げて、以下の面から日本語関係の日本語学、文学、歴史や社会文化研究などの人文社会系研究、また日本語教育の教育活動と研究への応用について現状と課題を取り上げて見ていきたい。

(1) 文章ジャンルの問題：自然言語処理の発達で、言語表現には基本的に大きな差異があることが明らかになってきている。テキストマイニングによって文章ジャンルによる差異がどのように把握されるか、そこから分かる文章ジャンルの特徴は何かについて述べる。

(2)テキストマイニングの日本関連分野への応用：言語表現の質的差異を擬似的に抽出できるようになってきているテキストマイニング技術は、質的研究が学科の専門性である日本語に関する日本語学、文学、歴史や社会文化研究などの人文社会系研究、また日本語教育の教育活動と研究さらに日本語の学習（以下、まとめて日本関連分野）にも直接、導入可能な技術になってきている。その中で特に、テキストマイニング技術を応用した日本語教育の教室活動について述べる。

キーワード：文章ジャンル、テキストマイニング、日本関連分野、
日本語教育、教室活動

文章ジャンルに基づくテキストマイニング応用の考察
―日本関連分野での活用を目指して―

落合由治
淡江大学日本語文学科　特聘教授

1. はじめに

現在、急速に発展している AI 技術は、自然言語処理において大きな社会的影響を与えてきている。情報処理推進機構(2017,2019)によれば、第二世代までの AI 技術が数理的データ処理中心であったのにたいし、第三世代の AI 技術では、自然言語処理において擬似的に言語の意味を解析する手法が開発され、また分析結果を視覚的に表示する技術が発達したことで、言語に関する研究や教育にも幅広く応用できる可能性が生まれてきている。

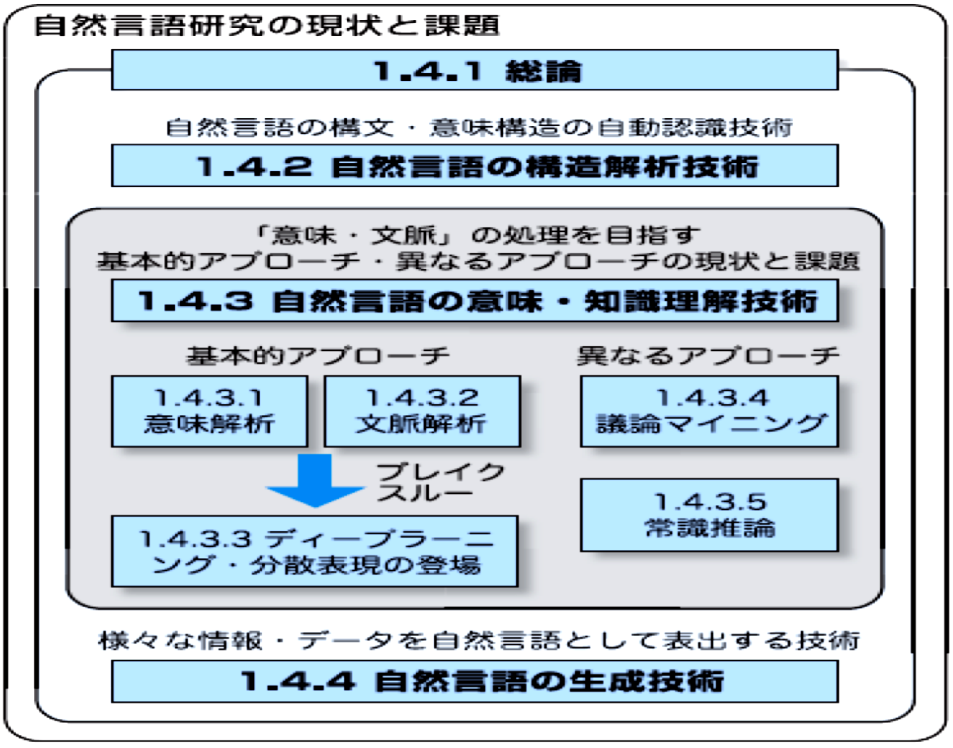


図 1　情報処理推進機構による自然言語処理の現状

情報処理推進機構(2017)によれば、第三世代 AI では、自然言語の構造解析技術が深化し、意味・文脈の処理が現在、開発と応用の中

心になってきている。その中で、現在、特に注目されている技術のひとつは、「感情極性解析」（Sentiment Polarity Analysis）である。これは、ある文、又は文章が与えられたとき、そこに表明されている意見がポジティブなものか､ネガティブなものかを解析する技術で、例えば、レストランのレビュー記事に書かれた「The atmosphere was good. I really liked the cesar salad.」（雰囲気は良かった。シーザーサラダは本当に気に入った）に対して、これをサービスに対する好意的評価として「ポジティブ」と出力するような特徴抽出技術である。現在、すでに多くの解析手法が提案されて、ビジネスで使用されるようになっている。[1]

	第二世代（～2010年）	第三世代（2010年～）
自然言語処理の基本	日本語を機械で処理する 形態素解析	深層学習によって人間の言語を機械言語に変換（word embeding）
自然言語処理の課題	コーパス作成、辞書作成、 形態素解析、情報抽出、構造解析／形態素レベル	意味解析、意味的特徴抽出、構造的特徴抽出／構造レベル／自立学習・分類word2vec/BERT
テキストマイニングの基本的発想	ルールベース（演繹）または統計処理で言語の特徴を決定	深層学習の手法で言語の特徴を帰納 ／×統計処理ではない
テキストマイニングの基本手法	統計的分析（R言語） 言語要素の統計的相関性	機械学習、深層学習（Python等） 機械の学習から有意味な結果を取り出す
人文系の接続	言語の統計的処理の分野でコーパス言語学、計量言語学、計量文献学	コーパス等による計量的研究での特徴量、確率・予測、類似・分類に関する量的研究 淡江日文系で試行：質的研究でのテキストマイニング応用

図 2　第二世代 AI から第三世代 AI への基本的処理方法の革新

今まで、日本語の計量的研究では自然言語処理の形態素解析に基づいた統計的分析などが用いられ成果を上げてきたが、それらは第二世代 AI の技術に依拠していた。[2]こうした試みは、現在も継続されているが、しかし、現在、発展が進んでいる第三世代 AI は第二世代 AI とは異なる方法で発達し、以前は扱うことができなかった言語の

[1] 引用は、情報処理推進機構(2017)『AI 白書 2017』P.68。自然言語処理の現状について、情報処理推進機構(2017)『AI 白書 2017』https://www.ipa.go.jp/about/report/ai/201707.html、情報処理推進機構(2019)『AI 白書 2019』https://www.ipa.go.jp/ikc/info/20181030.html(2020 年 10 月 6 日閲覧)を参照。

[2] 計量言語学での各種処理例については、計量国語学会編(2017)『データで学ぶ日本語学入門』朝倉書店参照。

質的側面を近似的に把握できるようになってきている。図 2 に第二世代 AI から第三世代 AI への基本的処理方法の革新についてまとめた。[3]第二世代の自然言語処理の基本的発想は統計処理などによる一般化を強く志向した量的手法であったが、第三世代の技術によって、深層学習・自立学習により言語の質的特徴を捉える方向に技術は発展してきており、従来の人文系の自然言語処理の基本的発想や研究手法だけでは言語の意味的処理と応用には限界がきている。

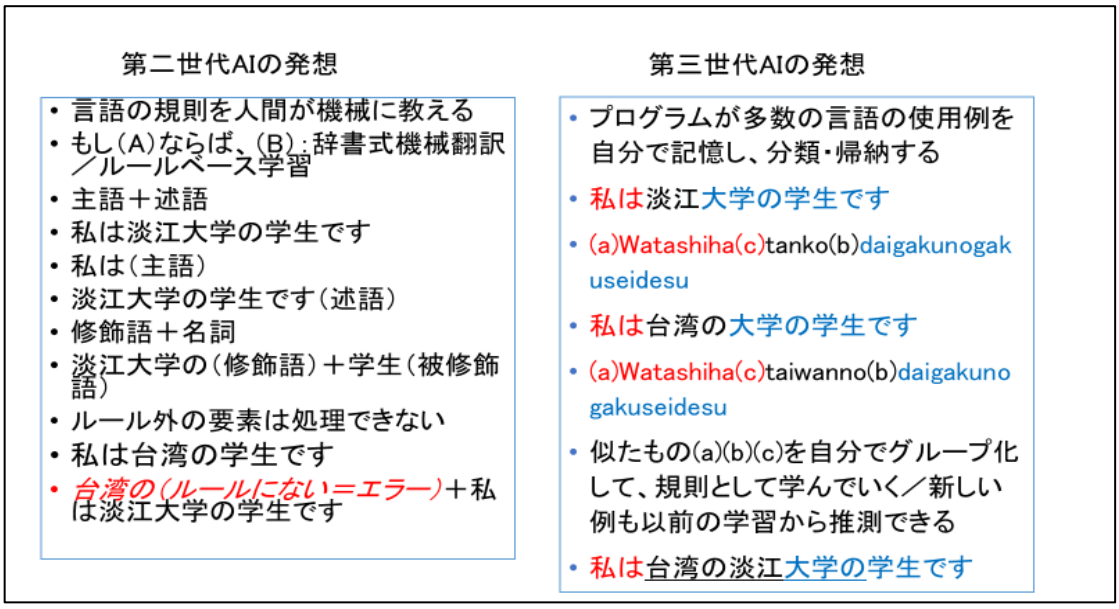


図 3　第二世代 AI と第三世代 AI の質的差異

図 3 のように、第二世代 AI の技術は、人間のルール（品詞分類と文法的結合規則）を演繹的にプログラムに置き換える処理であり、要素と要素の関係を統計処理で規則化していた。しかし、第三世代 AI の深層学習・自立学習では、プログラムが多数の用例を記憶して、語と語の関係をすべて学び、そこから語と語の共起関係、また文中での語の位置、さらに文と文の関係、文の纏まり（段落）の展開などを、自身が規則化して学ぶことで処理をおこなっており、完全に

[3] 第三世代 AI の技術を活かした日本関連分野に関する研究は、現在、進行中の課題である。計量的手法による日本語研究の動向は、間淵洋子（2020）「特集 2018 年・2019 年における日本語学界の展望―数理的研究」『日本語の研究』16-2pp.114-121 参照。主な動向として、確率の説明・予測に関する統計手法を用いた研究、類似度・分類に関わる統計手法を用いた研究（クラスター分析，対応分析等）、特徴語に関する研究などの研究が進んでいる。しかし、質的特徴把握を目指す研究は、まだ試行錯誤の段階と言える。

帰納的学習を元にした処理が可能になってきている。より人間の言語学習に近い方法が近似的に可能になっており、質的な側面を反映させることが可能になってきている。

本論文では、自然言語処理の応用の中で言語データの質的分析での応用が広がっているテキストマイニングを取り上げて、以下の面から日本語関係の日本語学、文学、歴史や社会文化研究などの人文社会系研究、また日本語教育の教育活動と研究への応用について現状と課題を取り上げて見ていきたい。

(1) 文章ジャンルの問題：今まで、自然言語処理は言語データを要素の堆積として数理的に処理してきたため、言語表現の持つ質的差異はまったく扱われてこなかった。しかし、自然言語処理の発達で、言語表現には基本的に大きな差異があることが明らかになってきている。テキストマイニングによって文章ジャンルによる差異がどのように把握されるか、そこから分かる文章ジャンルの特徴は何かについて述べる。

(2)テキストマイニングの日本関連分野への応用：言語表現の質的差異を擬似的に抽出できるようになってきているテキストマイニング技術は、質的研究が学科の専門性である日本語に関する日本語学、文学、歴史や社会文化研究などの人文社会系研究、また日本語教育の教育活動と研究さらに日本語の学習（以下、まとめて日本関連分野）にも直接、導入可能な技術になってきている。その中で特に、テキストマイニング技術を応用した日本語教育の教室活動について述べる。

紙数の関係で、今回は(1)を中心に述べる。AI 技術は、狭義では 2000 年以降の第三期の深層学習等の機械学習の発達に関わる技術や人間的活動ができる強い人工知能の技術開発を指すが、ここでは広義に第一期から第三期までの自然言語処理に関わる情報処理技術の意味で用いることとする。[4]

[4] 情報処理技術、人工知能技術の歴史については、総務省(2016)『平成 28 年版

2.テキストマイニングにおける文章ジャンルの問題

第二世代 AI までの自然言語処理は言語データを要素の堆積、集積として数理的に処理してきたので、言語表現の持つ質的差異は対象外であった。自然言語処理の発達で、言語表現にはジャンルに応じて基本的に大きな差異があると考えられるようになってきた。テキストマイニングによって文章ジャンルによる差異はどのように把握されるか、また、そこから分かる文章ジャンルの特徴は何か、日本関連分野への応用に入る前に、確認しておきたい。

ここでは、文章ジャンルについて、まず書き言葉として文章の基本的類型と言える論説、評論などの文章構成である「話し手の思いを述べる文章」と小説などに典型的に見られる「事件の話をする文章」について特徴を述べる。[5]さらに、話し言葉として、もっとも典型的なジャンルと言える「雑談」について、「友人同士の雑談」を取り上げて、特徴を考察する。[6]テキストマイニングには「KH Coder」の頻度順リスト、共起ネットワークを使用した。[7]

2.1 論説文：話し手の思いを述べる文章の分析

以下、「KH Coder」を使って、各ジャンルの文章をテキストマイニングした結果について述べる。まず、話し手の思いを述べる文章と

情報通信白書』第一部第四章を参照。http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h28/（2021 年 3 月 3 日閲覧）。

[5] 文章構成の分類基準は、形態的指標が明確で分析再現性がある基準として永尾章曹(1992)「第 4 章日本語の文法について」永尾章曹編著『日本語学』和泉書院 pp.103-134 を参照。

[6] 話し言葉のジャンル分類は、目的別の場面で設定されている場合が多い。一例として、宇佐美まゆみ監修(2020)『BTSJ 日本語自然会話コーパス（トランスクリプト・音声）2020 年版』では、話者の性別、年齢、職業と目的に応じて、「親しい同性友人同士雑談（男性、女性）」、「初対面及び友人同士雑談（女性）」「論文指導（日本人教師男女、日本人学生男女）」、「女性同士の断りの電話会話（対先輩、対同級生、対後輩）」など 25 種類の場面が設定されている。國立国語研究所(2020)「BTSJ 日本語自然会話コーパスについて」https://ninjal-usamilab.info/btsj_corpus/（2021 年 2 月 28 日閲覧）。

[7] 「KH Coder」については樋口耕一(2020)『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第二版』ナカニシヤ出版参照。

して、論説文の一種である新聞社説を分析した。[8]

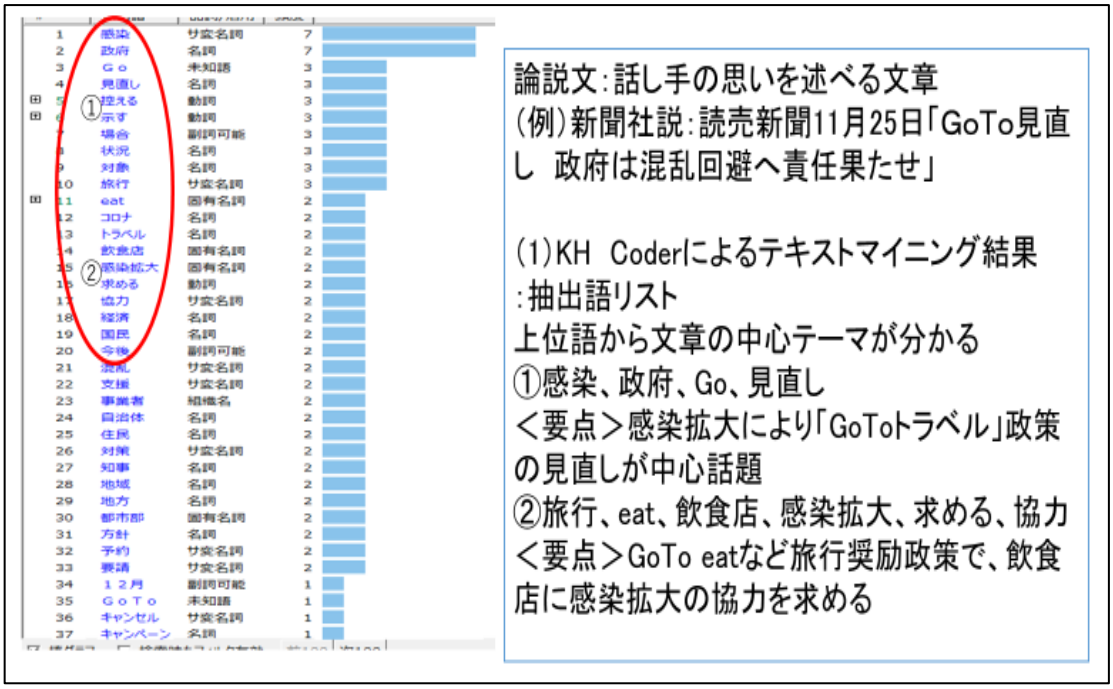


図 4　論説文の抽出語リスト結果

資料を形態素解析し、出てきた主な語彙を頻度順で並べた抽出語リストの結果を見ると、最上位の「感染、政府、Go、見直し」から要点として感染拡大により「GoTo トラベル」政策の見直しが中心話題になっていることが分かる。さらに続く「旅行、eat、飲食店、感染拡大、求める、協力」から GoTo eat など旅行奨励政策で、飲食店に感染拡大の協力を求めることが読み取れる。話し手の思いを述べる文章の場合、上位語から文章の中心テーマを明らかにすることができ、これがテキストマイニングをした場合の、論説文など話し手の思いを述べる文章の特徴と言える。

続いて、出現した語彙の出現距離によって意味的なまとまりを示す共起ネットワーク分析を行ってみると、分析した結果、7 つのクラスターに分かれ、相互に関係の深い要素が抽出された。

[8] 資料は、新聞社説：読売新聞 11 月 25 日「Ｇｏ Ｔｏ 見直し　政府は混乱回避へ責任果たせ」 https://www.yomiuri.co.jp/editorial/20201124-OYT1T50273/（2021 年 3 月 3 日閲覧）。

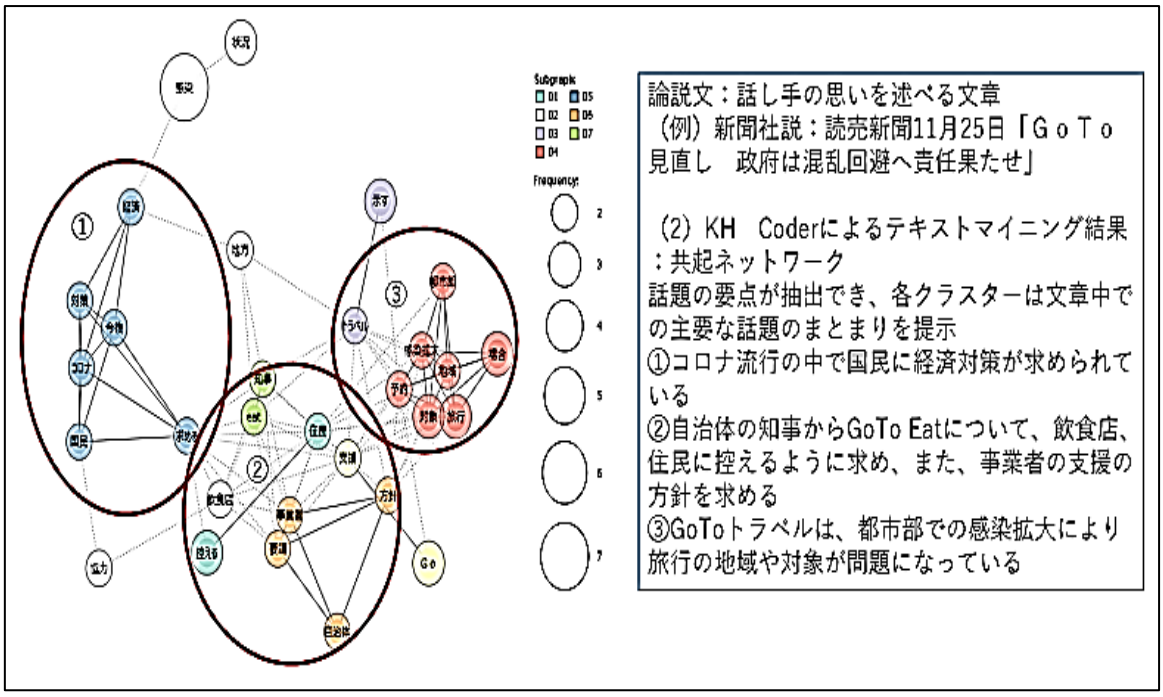


図 5　論説文の共起ネットワーク分析結果

まず①では「経済、対象、コロナ、国民、求める」などの語からコロナ流行の中で国民に経済対策が求められていること、②では「知事、eat、住民、控える、事業者、要請、自治体」などの語から、自治体の知事から GoTo Eat について、飲食店、住民に控えるように求め、また、事業者の支援の方針を求めるなどの具体的な対策が示されている。③でも、「トラベル、都市部、感染拡大、予約、地域」などの語から GoTo トラベルは都市部での感染拡大により旅行の地域や対象が問題になっていることが浮かぶ。テキストマイニングを論説文に適用した場合、結果は、やはり、内容の話題の要点が抽出でき、各クラスターは文章中での主要な話題のまとまりを提示していることが明らかになった。

2.2 小説：事件の話をする文章の分析

次に、小説などによく見られる「事件の話をする文章」でのテキストマイニングの結果を示す。ここでは芥川龍之介「手巾」を資料として分析した。形態素解析し、出てきた主な語彙を頻度順で並べた抽出語リストの結果を見ると、上位語からは「先生、婦人、奥さ

ん」のような登場人物しか分からない。また、「ストリンドベルク、本」は登場者の先生が読んでいた本の作者、「岐阜提燈、椅子」は先生が読書していた部屋の調度で、やはりごく部分的な場面しか分からない。

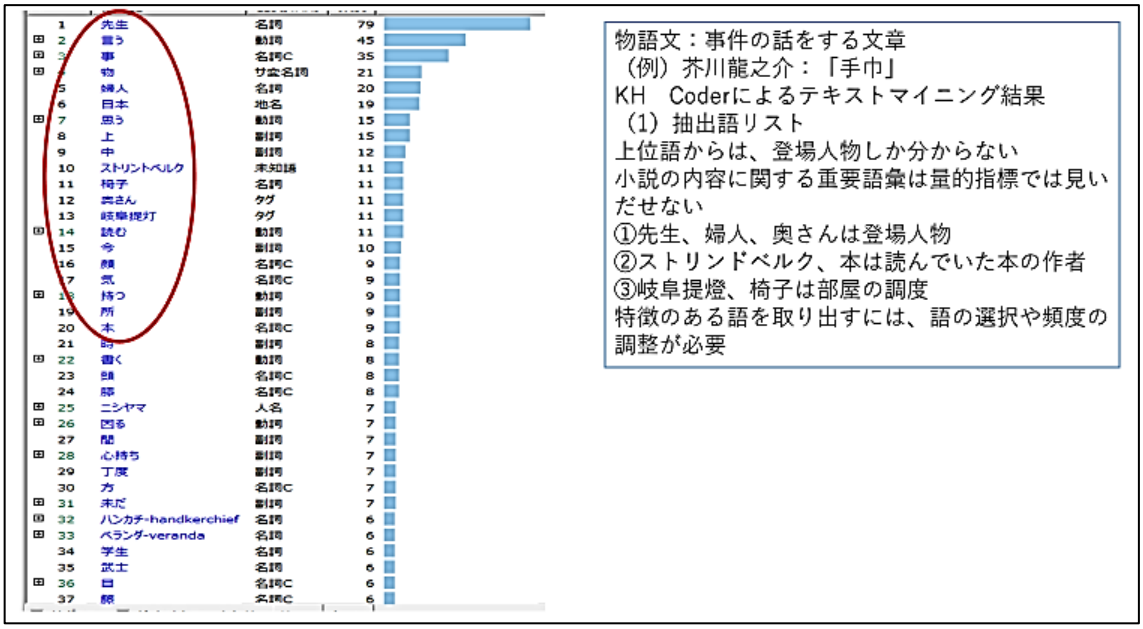


図 6　小説の抽出語リスト結果

小説の内容に関する重要語彙は量的指標だけでは見いだせないため、特徴のある語を取り出すには、語の選択や頻度の調整が必要である。先の話し手の思いを述べる文章の場合、上位語から文章の中心テーマを明らかにすることができたが、小説のような事件の話をする文章の場合は、語彙の頻度は内容理解の手掛りにはできない。これがテキストマイニングをした場合の、小説のような事件の話をする文章の特徴と言える。

続いて、共起ネットワーク分析を行ってみると、やはり、論説文のような話し手の思いを述べる文章とは異なって、各クラスターから頻出する共起語は分かるが、直接、作品理解には役に立たないまとまりが多く、小説の内容に関する重要語彙は量的指標では見いだせないことが分かる。図 7 のように、「ベランダ、読む、ストリンドベルク」は、先生が読書をしている場面、「奥さん、文明、日本、武

士」は、今日の出来事で先生が考えた内容であるが、これだけでは作品のストーリーや人物関係、描かれている主な話題などは分からず、頻出語を基準にすると内容の大半は出てこないままである。

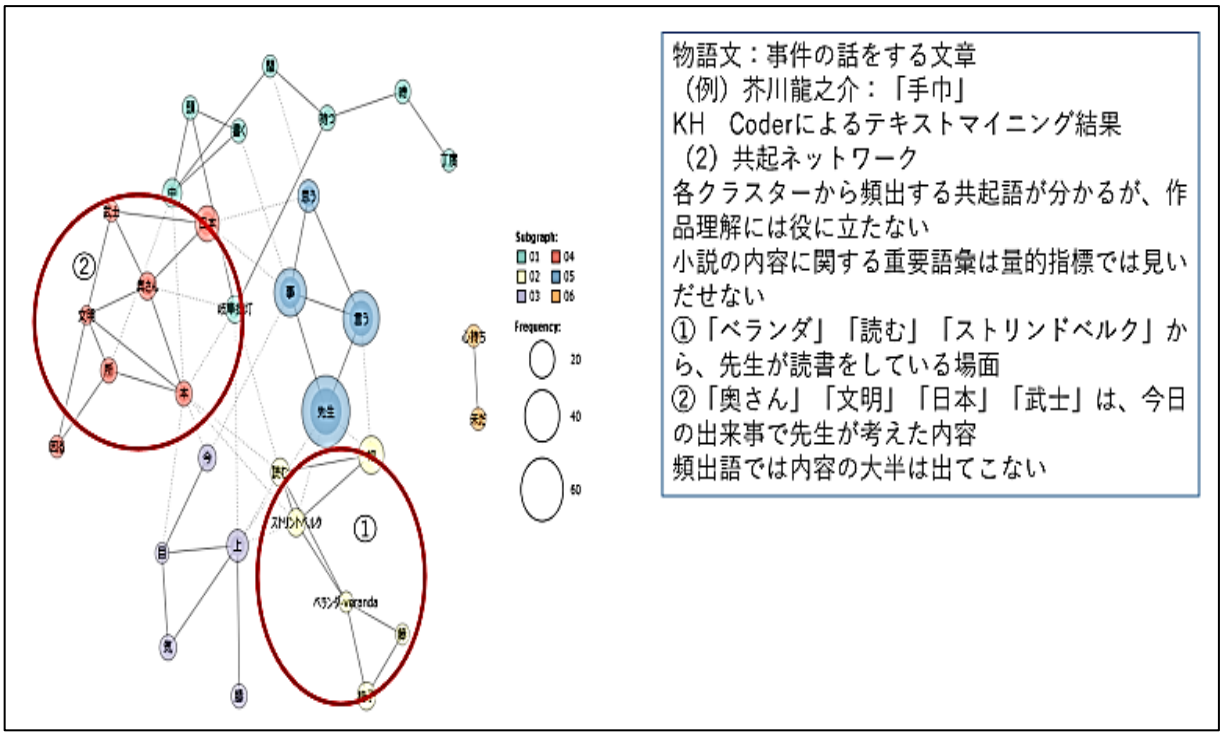


図 7　小説の出現頻度上位語の共起ネットワーク分析結果

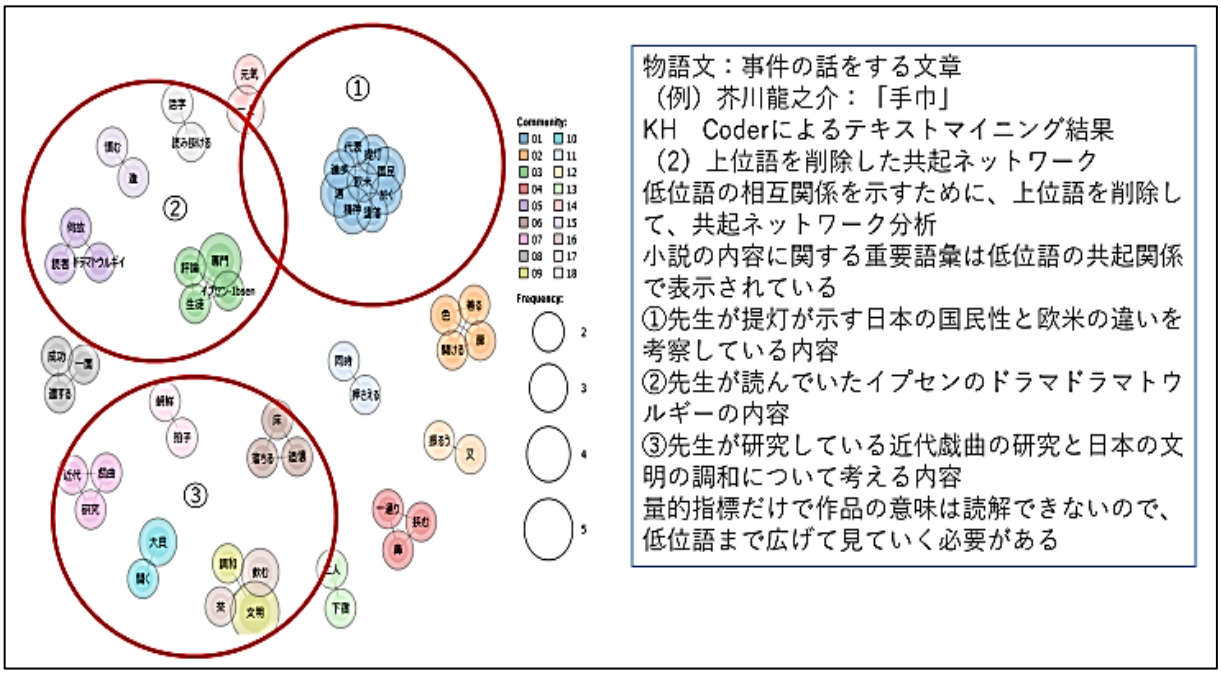


図 7　小説の出現頻度下位語の共起ネットワーク分析結果

そこで、出現語数の頻度を低頻度の語まで下げて、共起ネットワ

ーク分析を行ってみると、以上のようになった。小説の内容に関する重要語彙は低位語の共起関係で表示されていることが分かる。①の語彙は先生が提灯が示す日本の国民性と欧米の違いを考察している内容、②は先生が読んでいたイプセンのドラマドラマトウルギーの内容、③は先生が研究している近代戯曲の研究と日本の文明の調和について考える内容で、作品での登場者の思考が、日本対西洋の対比にあることが分かる。また、ストーリーは、亡くなった学生の母親が学生の死を報告に来て、平然と話すその様子に先生が気丈な日本の母親を見いだし、満足を感じたところ、母がハンカチを握った手が細かく震えていて、悲しみを表に出さない演技をしていたことに気づき、先生が不快を感じるという内容であるが、それらもテキストマイニングでは取り出すことはできない。検出する語彙の頻度数を調整することで分かるのは、登場者の思考や様子などに関する説明の内容である。小説のような事件の話をする文章の場合、量的指標だけで作品の意味は読解できないので、低位語まで広げて見ていく必要があることと、どの部分が出ているかは予め充分に作品を読解していないと意味が理解できないため、テキストマイニングが適用できる範囲は、補助的なものになる。

2.3 友人間の雑談：話し言葉

さらに、話し言葉のひとつのジャンルである友人間の雑談を取り上げてみる。ここでは、『BTSJ 日本語自然会話コーパス』から「同性の友人同士の会話」の資料を用いた。[9]先にみた話し手の思いを述べる文章や事件の話をする文章とはかなり異なった結果が出た。まず、上位語として出てくるのは資料中にでている「地名」「人名」であり、これらについて話していることは推測できるが、具体的な話題は分からない。そのほか、資料に付記されている言語行動として「笑う」「笑い」が多数あり、友人間で笑いを交えながら話が進んで

[9] 国立国語研究所(2020)「BTSJ 日本語自然会話コーパスについて」参照 https://ninjal-usamilab.info/btsj_corpus/（2021 年 3 月 3 日閲覧）。

いることが推測できるが、やはり内容を示す手掛りは得られない。

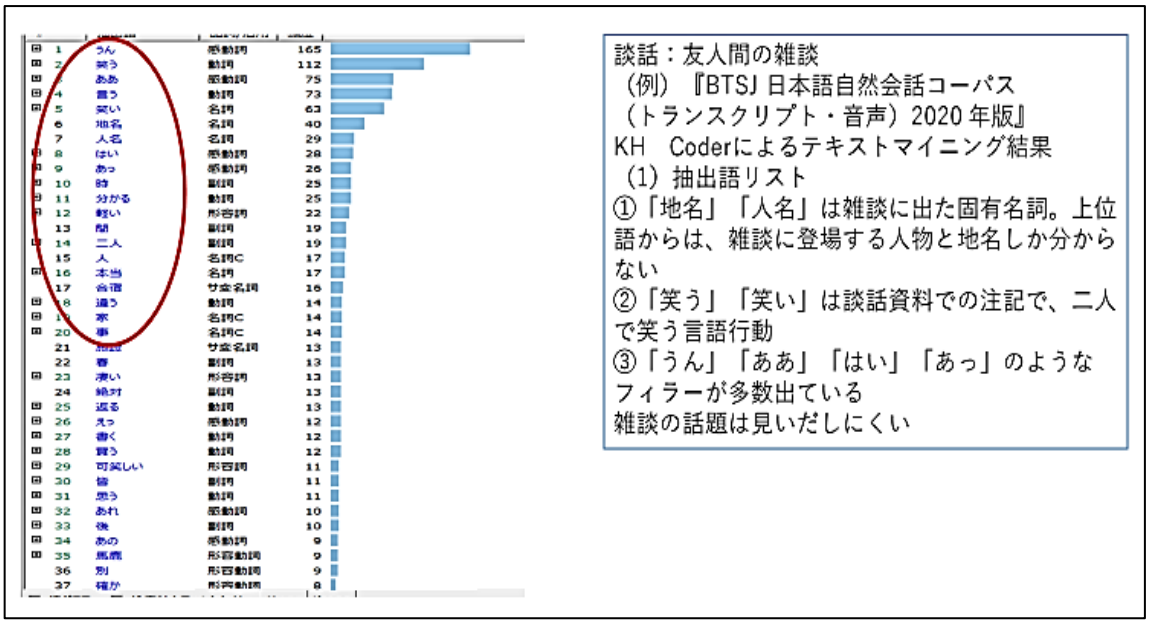


図 8　雑談の抽出語リスト結果

また、上位語では「うん」「ああ」「はい」「あっ」のようなフィラーが多数出ている、その他、「言う」「思う」「書く」「買う」などの動詞で、雑談の話題は見いだしにくい。話し言葉の友人間の雑談のような談話の場合、頻度順で内容の分かる語彙を取り出すことはできず、また、語彙だけでは内容を推測することはほとんどできない。

しかし、語彙の出現の距離を元に関係を描く共起ネットワーク分析をしてみると、内容を知る手掛りが出てくる。共起ネットワークの結果を見ると、雑談をしていた話題と付随する言語外行動のまとまりが推測できる。まずは、①「人名、人」などの語彙から会話に出てくる人物について、噂や出来事を話していたことが分かる。②では「合宿、地名、春」などの語から、春に合宿をした町と合宿の内容について話題になっていたことが推測できる。さらに、③では「大学、書く、引っ越す」などの語から、大学での生活や勉強、また住居について話していたことが分かる。また、こうした話題とは別に付記されている言語活動の付随行動として「大笑い、笑い、軽い笑い」と「うなずき」が出ていて、雑談の進行にこうした付随行動が重要であることがうかがえる。談話の資料でのテキストマイニ

ングでは、共起ネットワーク分析やクラスター分析を使用すれば話題のまとまりと言語外行動で構成されている談話の構造が理解できる。また、談話は、一人が話しかけてもう一人が関連する話題を出す形で複数の異なる話題を出しながら展開する形で成り立っていることも推測できる。

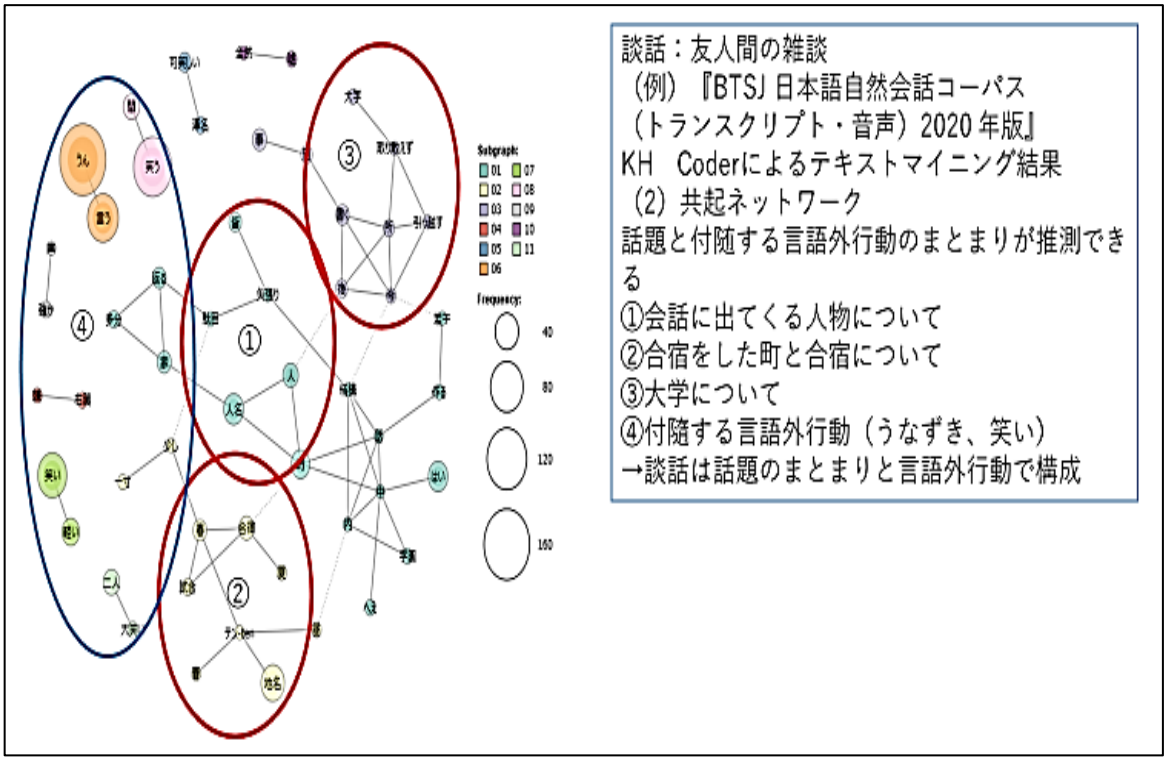


図 7　雑談の共起ネットワーク分析結果

以上、資料とした書き言葉と話し言葉の資料で、それぞれテキストマイニングで理解できる内容を確認して、各ジャンルで差異があることが明らかになった。つまり、テキストマイニングのように表現単位を語彙に還元して特徴を取り出す手法でも、語彙の出現する傾向や意味には、各ジャンルで大きな違いがあり、それは語彙を表現単位に構成しているより上位の規則性があって、そうした出現傾向になっていることを示している。今までの語、文、文章のような言語単位に加えて、社会的表現ジャンルという上位の単位を設ける必要があると考えられる。[10]今回は、これ以上は踏み込まないが、テ

[10] 言語の形式的単位については、時枝誠記(1950)『日本文法 口語篇』岩波書店

キストマイニングを検出方法に使うことで、言語研究の新しい次元を開拓することが可能ではないかと思われる。

3. 日本関係分野へのテキストマイニングの応用―日本語教育中級教科書の文章ジャンル分析

自然言語処理技術の発達によって、AI 技術は社会的に運用されている各種の言語ジャンルの中で直接、関わる分野が拡大している。その分野の一つは、文章作成、要約などに関する部分で、会議 Hack（2020）のような Web 広告や雑誌が広く宣伝しているように、すでにビジネス現場で実用レベルの技術が用いられている。[11]IT トレンド(2020)のように、テキストマイニングも以前の言語の統計処理を中心とした内容から言語資料の質的分析を行うためにビジネス界で利用が広がっている。[12]人間の作業のように正確にはできない等々

等を参照。本論では、従来の言語の形式的単位のみではなく、言語が実際に社会的に使用される場面に基づいた単位言語を社会的ジャンルと考えている。なお、自然言語誤処理で言語形式を細分化する単位については諸説あり、決着が付かないため、便宜的な区切りをその都度用いている。テキストマイニングでの形態素解析もプログラムでの便宜的な区切りになっている。より上位の単位についても、経験的な概念で論じられることはあるが、言語学的な論証等が行われているわけではない。従来より実態に即した細かい語構成の問題が出ているコーパスでの単位付与については、岡照晃(2018)「『国語研日本語ウェブコーパス』からの新規語彙素獲得の試み」『言語資源活用ワークショップ発表論文集 = Proceedings of Language Resources Workshop』3pp.586-592、西川賢哉、渡邊友香(2019)「『日本語日常会話コーパス』の短単位解析：作業工程を中心に」『言語資源活用ワークショップ発表論文集 = Proceedings of Language Resources Workshop』4pp.238-250 等を参照。

11 ビジネス界では自然言語処理を利用した各種の AI による言語処理作業が実用レベルで導入されている。一例として、Web 会議などの音声会話を議事録化する AI の利用は、2020 年の新型コロナウイルス大流行によって急速に普及している。会議 HACK(2020「自動でテキスト化！会議に役立つ音声議事録 5 選」https://www.kaigishitu.com/meeting-hacks/detail/id=32856 参照。また、文字や会話の資料の内容を要約する要約 AI も普及が加速しており、会社で使われている言語データの要約処理は人間の手から AI に交代しつつある。AIsmaily(2020)「高精度の AI エンジンが対話や議事録を自動で要約・分類」https://ai-products.net/product/voice-recognition-ai-quicksummary/（2021 年 3 月 3 日閲覧）等を参照。

12 テキストマイニングも始まってからすでに 20 年を経過しているが、以前の統計的計量を中心にしていた手法から言語の質的分析に使える擬似的に言語の意味を抽出する各種の技法が開発されて、技術の一般化が進行している。IT トレンド(2020)「テキストマイニングツール 7 つの選定ポイント」https://it-trend.jp/textmining/article/choice（2021 年 3 月 3 日閲覧）。

の批判をすでに超えて、AI 技術による自然言語処理の応用は利点を活かし欠点を人間が補助する形で普及が加速している。学習者の進路に大きな影響を与えるため日本語教育でも自然言語処理の導入や接続は緊急の課題と言える。ここでは、日本語教育の読解教材などで使われている資料の社会的言語ジャンルを調べ、テキストマイニングで特徴抽出がどの程度できるか、比較検討していくことで、日本関係分野でのテキストマイニング技術の応用の手掛りを見つけていきたい。

3.1 日本語教育の読解教材の文章ジャンル

今回は、日本語教育で使用される文章教材のジャンルのサンプルとして、日本事情理解と読解、会話等の技能練習を組み合わせた教材『上級へのとびら』に出ている読解部分を取り上げる。[13]対象となる読解部分を表 1 に示した。

表 1『上級へのとびら』各課の文章教材の文章構成

課数	テーマ	主な内容	文章構成
1	日本の地理	日本の地図を示して地理、名所、名物、行事、祭り、昔話の紹介	話し手の思いを述べる文章
2	日本語のスピーチスタイル	スピーチレベルの使い分け、男女の言葉、言葉の省略・短縮・倒置、書き言葉と話し言葉、	話し手の思いを述べる文章
3	日本のテクノロジー	ロボット、テクノロジーの発達	話し手の思いを述べる文章
4	日本のスポーツ	日本のスポーツ、日本の武道、心・技・体の考え方、	話し手の思いを述べる文章
5	日本の食べ物	「インスタントラーメン発明物語」	話し手の思いを述べる文章
6	日本人と宗教	宗教、宗教的習慣、行事、信仰、神話	話し手の思いを述べる文章
7	日本のポップカルチャー	日本のポップカルチャー、マンガ、手塚マンガ	話し手の思いを述べる文章
8	日本の伝統芸能	「狂言と笑い」	話し手の思いを述べる文章
9	日本の教育	日本の教育制度のいい点と問題点、学歴社会、受験戦争、いじめ、登校拒否	話し手の思いを述べる文章

[13] 岡まゆみ構成(2009)『コンテンツとマルチメディアで学ぶ日本語上級へのとびら』くろしお出版。

10	日本の便利な店	「自動販売機大国ニッポン」、自動販売機の便利さと問題点、	話し手の思いを述べる文章
11	日本の歴史	日本の輸入の歴史、織田信長、豊臣秀吉、徳川家康	話し手の思いを述べる文章
12	日本の伝統工芸	和紙の特性、千羽鶴の話	話し手の思いを述べる文章
13	日本人と自然	日本の自然、自然描写、俳句	話し手の思いを述べる文章
14	日本の政治	日本の政治制度、世襲議員、タレント議員のいい点と問題点	話し手の思いを述べる文章
15	世界と私の国の未来	世界の社会問題、もったいない運動	話し手の思いを述べる文章

全体で 15 課中、各課のテーマにしたがって読みものとして提示されている内容は以上となるが、そこで使われている文章構成はいずれも各種の関連した話題を並べて焦点を示す論説文のような「話し手の思いを述べる文章」であった。[14]

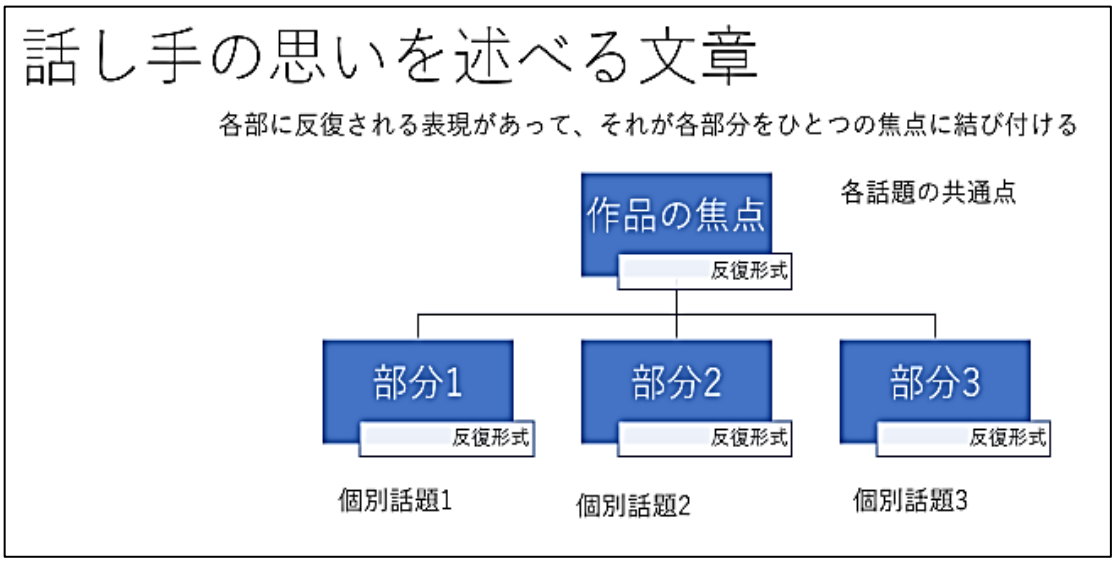


図 8　話し手の思いを述べる文章の文章構成

他の教科書では異なるジャンルの文章が取り上げられている可能性もあるが、しかし、15 課のすべてが同じ文章構成で書かれていることから推測すれば、日本語教育の中級以上のレベルで読解等のために使われている基本的な文章ジャンルはほぼ「話し手の思いを述

[14] 文章の基本的類型および「話し手の思いを述べる文章」については、永尾章曹(1975)『国語表現法研究』三弥井書店 pp.102-129、落合由治(2004)「文章の基本的構成について―基本的構成から次の段階の構成へ―」『台灣日本語文學報 19pp.195-220、台灣日本語文學會参照。

べる文章」に限定される傾向があると見ても問題はないであろう。主に扱われる文章ジャンルが限定されているということは、テキストマイニングを実施したときも同じ結果になることを意味している。

3.2 日本語教育の読解教材へのテキストマイニングの応用

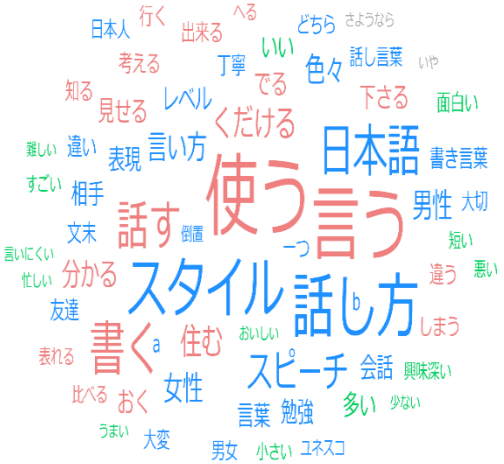
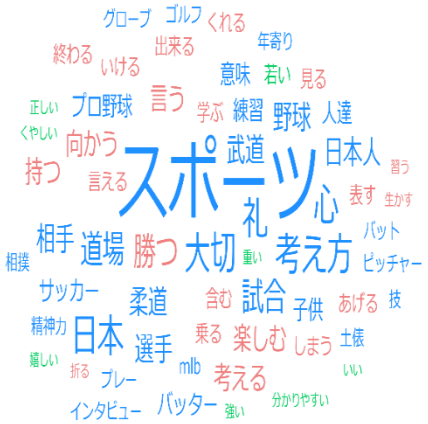
1課　日本の地理	2課　日本語のスピーチスタイル
	
3課　人とロボット	4課　日本のスポーツ
	

図 9　ユーザーローカルを使ったワードクラウド作成

教育現場でそのまま応用できるように、今度はフリーのテキストマイニングツールを使用して結果を読み取る場合の例を示したい。[15]紙数の関係で 15 課すべてを表示できないが、前半の 4 課を事例と

[15] 以下の分析結果は、フリーで使えるテキストマイニングツールとしてユーザ

して分析結果を提示したい。まず、以上の図 9 のように文章中の重要な語を表示するワードクラウド(TF-IDF 法・頻度順)で本文のキーワードを取り出した。[16]話し手の思いを述べる文章の場合、ワードクラウドあるいは頻度順の語彙リストでその文章の中心的話題と重要キーワードを知ることができる。第 1 課を例にすれば、ここでの課の話題は「日本の地理」なので、地理に関する地名（北海道、沖縄等）と各地の景物や文化事象（温泉、城、桜の花、道後温泉等）が抽出されている。動詞・形容詞を見ると、旅行に関する「楽しむ」「泳げる」「楽しい」「おいしい」等と、知識を得ることに関する「調べる」「探す」などがあり、内容の中心が日本各地の地方と景物等の知識を紹介する内容であることが推測できる。中央附近に大きく出ている語ほど、その文書の特徴を示す語彙なので、その部分を手掛かりにして内容のテーマを探すことができ、読む場合も要約する場合も、ポイントを絞る部分を決めることができる。こうした話し手の思いを述べる文章は語彙の相関性や重要度を計算できる AI の処理に適している。

次に、出現した語彙の相関性を表示する共起ネットワーク分析の結果であるが、話し手の思いを述べる文章の場合、出現したキーワードの相互関係をかなり正確に抽出できる。第 1 課の「日本の地理」で見ると、例えば「四つ-島」「大きい-九州」という纏まりは日本列島の地形の紹介、「一道-一都」「広島県-静岡県」のグループは都道府県の紹介であることが分かる。文章全体は開始から終了まで連続した線条性で結束されていると想定されているが、文章構成の特徴から言えば、話し手の思いを述べる文章の場合、個々の話題の並列に

ーローカル「AI テキストマイニング」https://textmining.userlocal.jp/（2021 年 3 月 3 日閲覧）を使用。

[16] TF-IDF 法は、「テキスト中に高頻度で出現する内容語はテキストの主題を表す傾向がある」という仮定による TF 値と、「出現する文書数が少ない索引語は、その索引語により該当テキストをより小さく絞り込める」IDF 値を掛けた方法で、資料検索などの場合によく使われる、そのテキストの特徴的な語彙を捉える数値のひとつである。奥村学(2010)『自然言語処理の基礎』コロナ社 pp.117-119 参照。

よって一定の共通性が認識され、内容のまとまりが認知される。

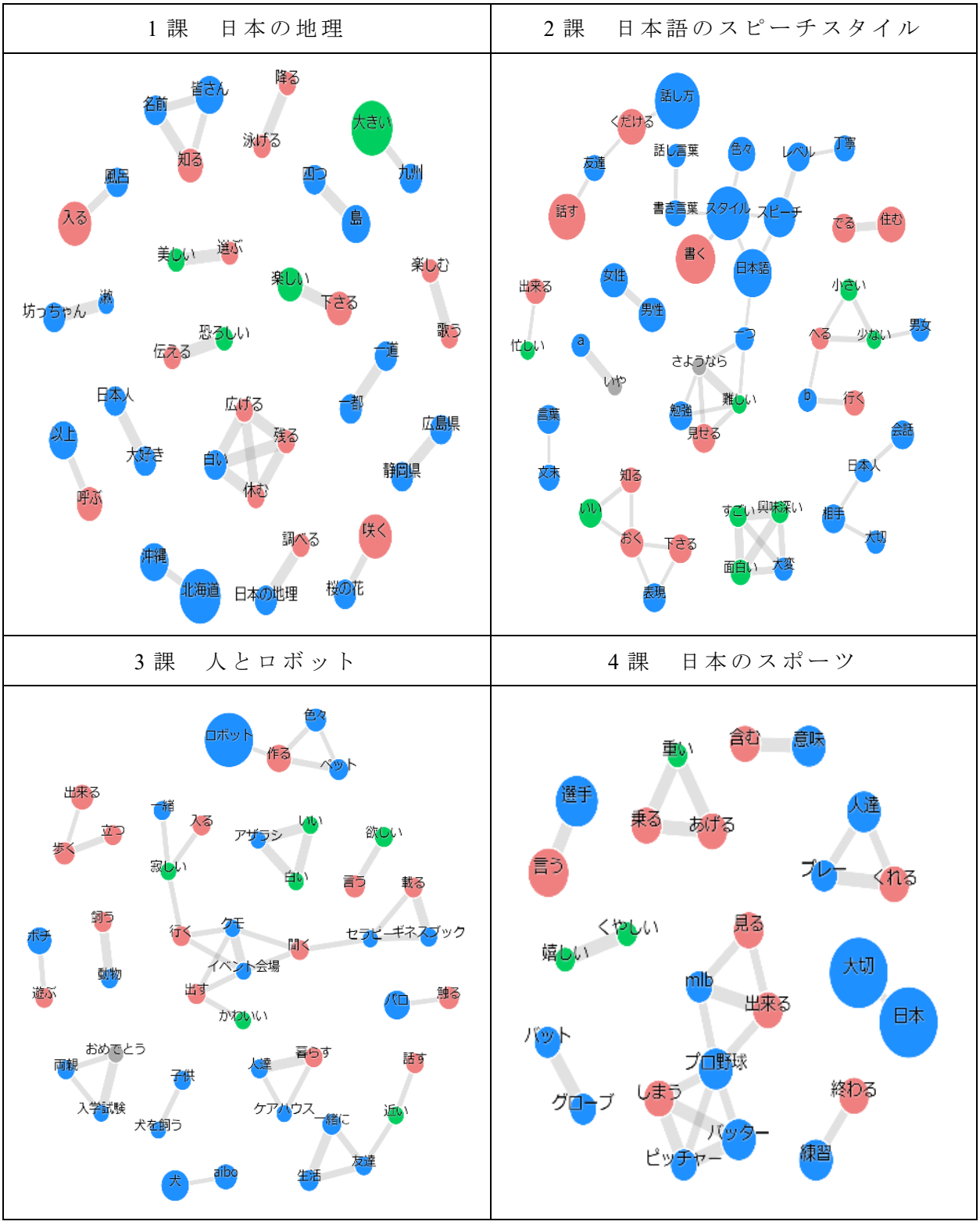


図 10　ユーザーローカルを使った共起ネットワーク分析[17]

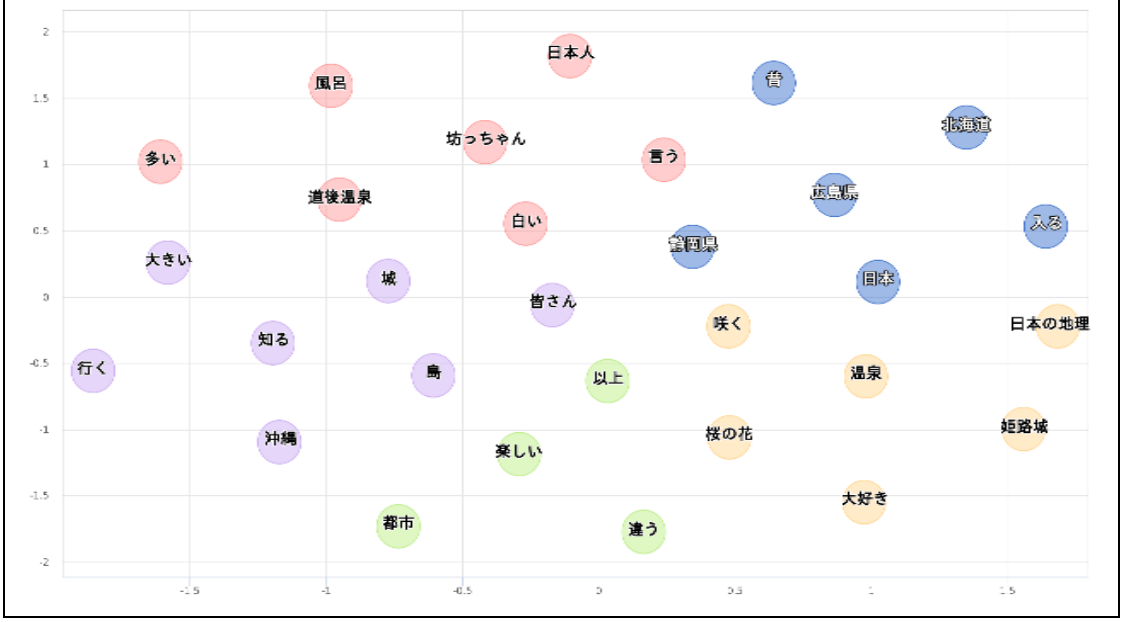
[17] 共起ネットワーク分析の特徴については、樋口耕一(2020)『社会調査のための計量テキスト分析―内容分析の継承と発展を目指して第二版』ナカニシヤ出版 pp.182-189 参照。

説明文や論説文のような話し手の思いを述べる文章では、言語的事象としては実は個々別々な話題の並列が現象として見られる。テキストマイニングは、そうした話題の並列を語彙の共起ネットワークとして擬似的に意味の纏まりとして表示できるようになっている。読解において、共起ネットワーク分析結果を基にして、本文と対照しながら学習者は内容の要点を捉える練習ができ、また討論の手掛りにできる。

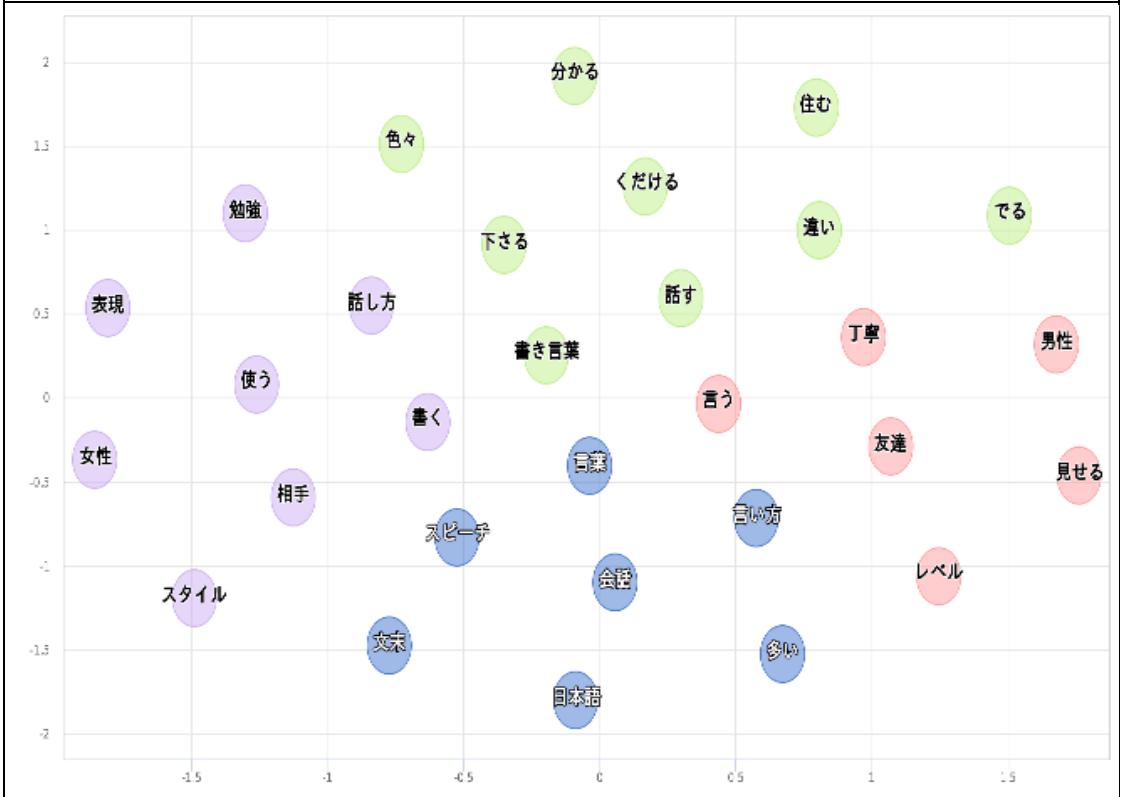
続いて、高次元データを 2 次元又は 3 次元に変換して可視化するための次元削減アルゴリズムである t-SNE 法を使った 2 次元マップの結果を示す。これは、近くにある単語同士は同じ場所で出てくる傾向が強いことを利用した要素の抽出方法で、やはり話題のまとまりを抽出できる。

1 課の「日本の地理」で見ると、「日本人-坊っちゃん-道後温泉-風呂」という纏まりは日本人と温泉・風呂文化の紹介、「北海道-広島県-静岡県-昔」の纏まりは都道府県の事例紹介である。「桜の花-温泉-姫路城-日本の地理」は、日本について人気のある季節の風物や有名な景色を取り上げている。内容の主な話題をこうした結果から読み取ることは容易であろう。

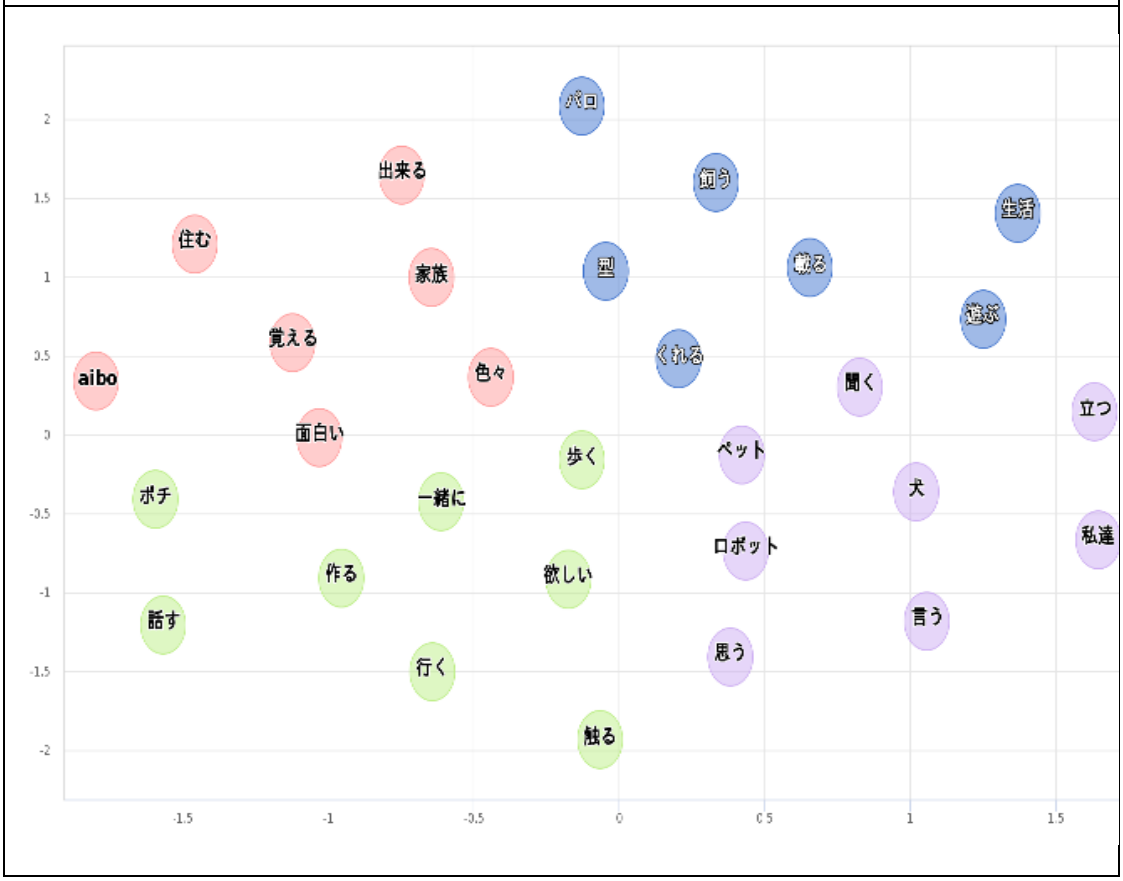
1 課　日本の地理



2課　日本語のスピーチスタイル



3課　人とロボット



4 課　日本のスポーツ

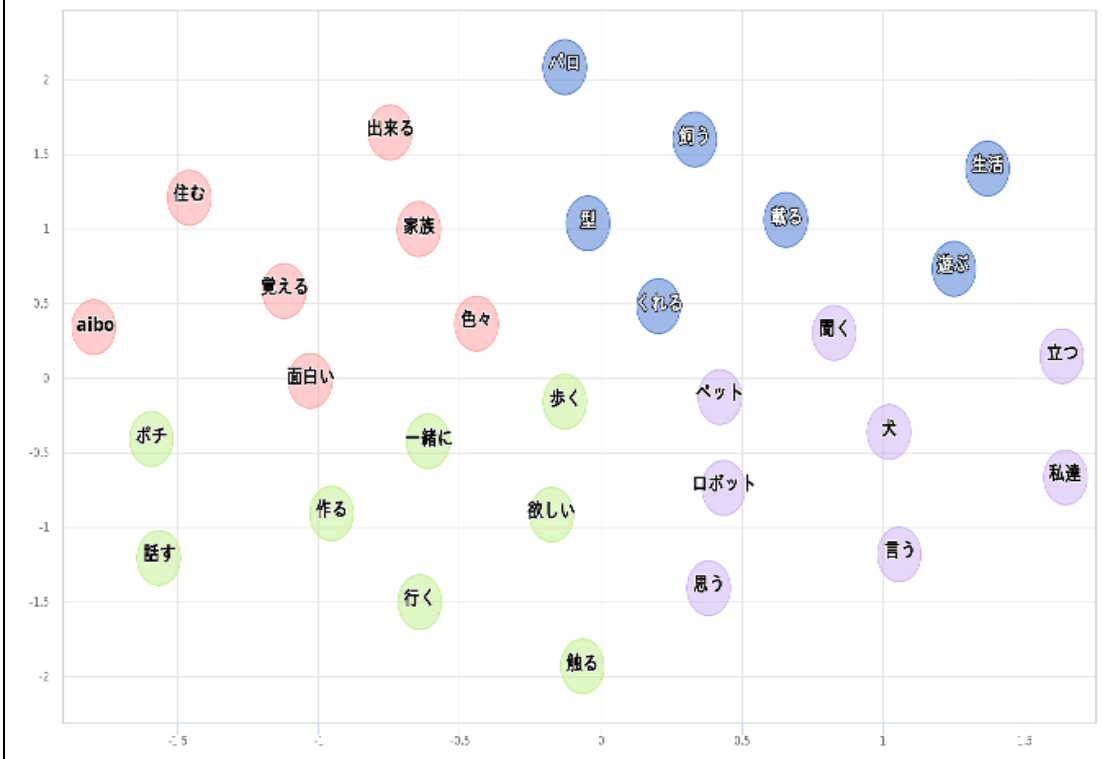


図 11　ユーザーローカルを使った 2 次元マップ（t-SNE 法）[18]

第 2 課も同じく、「勉強-表現-話し方-女性」のまとまりは、相手によって話し方を変える必要がある日本語の特徴、「書き言葉-話す-くだける」のまとまりは、書き言葉と話し言葉のスタイルの違いを説明していると言える。このように話し手の思いを述べる文章の場合、共起ネットワーク分析と同じく、2 次元マップ（t-SNE 法）も話題の並列を語彙の共起関係によって擬似的に意味の纏まりとして表示できると言える。同じように本文と対照させながら、学習者が読解や討論の手掛りに使うことができる。

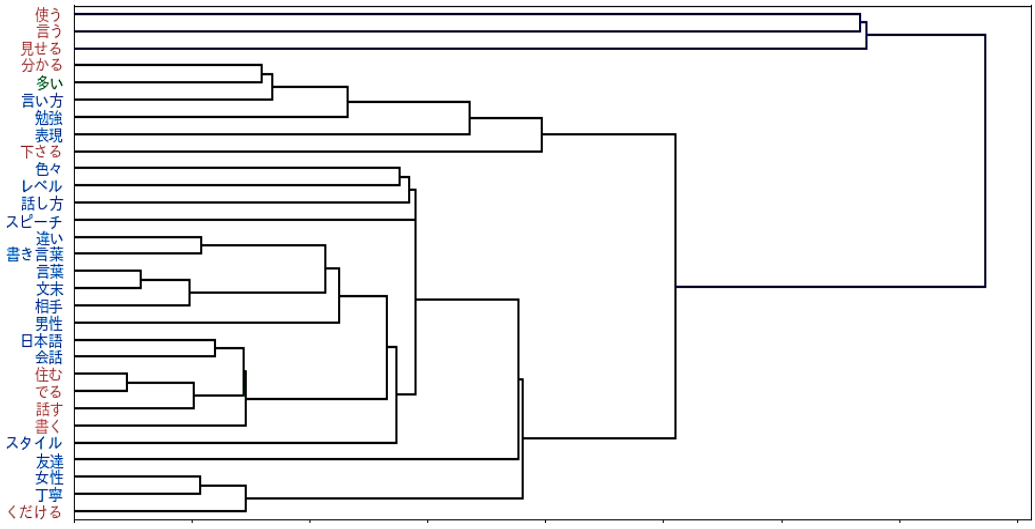
さらに、階層的クラスタリングを使用すると、文章中での出現傾向が似た単語をまとまりとして樹形図で表示できる。階層的クラスタリングを文章に当てはめると、文章全体での語彙出現の相互関係

[18] t-SNE 法については、MARTIN WATTENBERG,FERNANDA VIÉGAS,IAN JOHNSON (2016)*How to Use t-SNE Effectively* https://distill.pub/2016/misread-tsne/（2021 年 3 月 3 日閲覧）参照。

から文章全体の中での各話題の大きな纏まりを階層で表示できる。

1課　日本の地理

2課　日本語のスピーチスタイル



3課　人とロボット

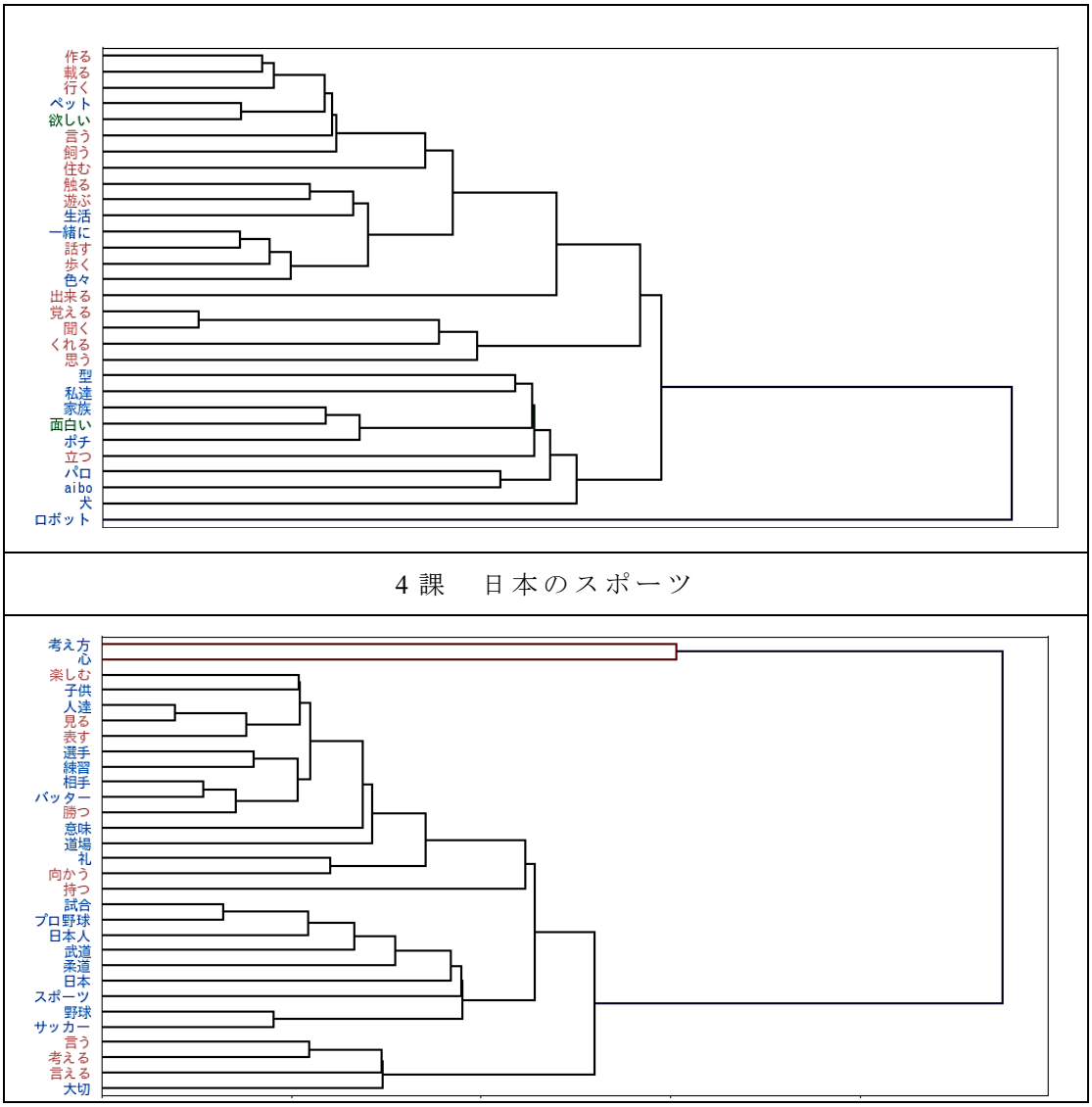


図 12　ユーザーローカルを使った階層的クラスタリング

右に階層が伸びている語ほど文章全体との共通度が高く、左に寄っている語ほど各部分で細かくまとまった話題を示している。1 課の「日本の地理」では、繰り返し登場する全体に関わる語彙として「日本」がある。クラスターから「日本」について「楽し」く「言う」「知る」ことが課の主なポイントであることがまず、分かる。次に、小部分の各話題として日本の「大きさ」、「温泉」や「風呂」、「広島県」等の各県や地方、「桜の花」等の話題があることが分かる。全体に関わる語彙は、右に側伸びた上位のクラスターとして線で示されている。各話題の纏まりが複雑になるほど右側への階層が高く伸び、各話題との纏まりが強くないグループほど左側に集中する。そ

こから言えば、第 1 課の内容は各話題の相互関係は薄く、第 2 課から第 4 課までの内容は、各話題に関する説明がそれぞれ大きな纏まり（段落）を幾つか作っていることが分かる。こうした纏まりは各段落を形成しているので、各段落のキーワードをここから見つけることができる。

こうした中級レベルの読解教材は今まで文法や語彙の解説と内容理解として翻訳を中心にし、内容を教師が解説して、習得知識を筆記試験でテストするという講義型の授業の流れを中心に教育活動がデザインされてきたと考えられる。しかし、今後はテキストマイニングを導入することで、学習者が教材の中で取り出された語彙の文中での使用をたどり、また関連する資料を探していく活動をすることで、学習者中心の教室活動をデザインすることができる。テキストマイニングを導入することによって、学習者が内容の要点を知る手掛りが得られ、読解授業をアクティブラーニング化することができ、同時に AI 技術との接点を導入することも可能になる。

4．おわりに

以上、現在、日本語の中級教科書で使われている読解教材の文章構成は、ほぼ話し手の思いを述べる文章に限定されており、テキストマイニングで内容のポイントを抽出するのに適していることを述べた。台湾の高等教育における日本語教育が AI 時代に対応するには、二種類の方向性が必要と言える。一つは、従来の日本語教育の内容を活かし、縦軸で多義的な社会文化的文脈に対応できるより精緻な日本語の運用力と理解力を高めていくことである。もう一つは、横軸として今まで中心的なスキルではなかった新しいスキルを採り入れて、組み合わせることで人文社会科学の基礎である「リベラル・アーツ」を応用したスキルの訓練を行うことである。AI の応用もその重要な一分野になり、テキストマイニングはその入り口に使える。

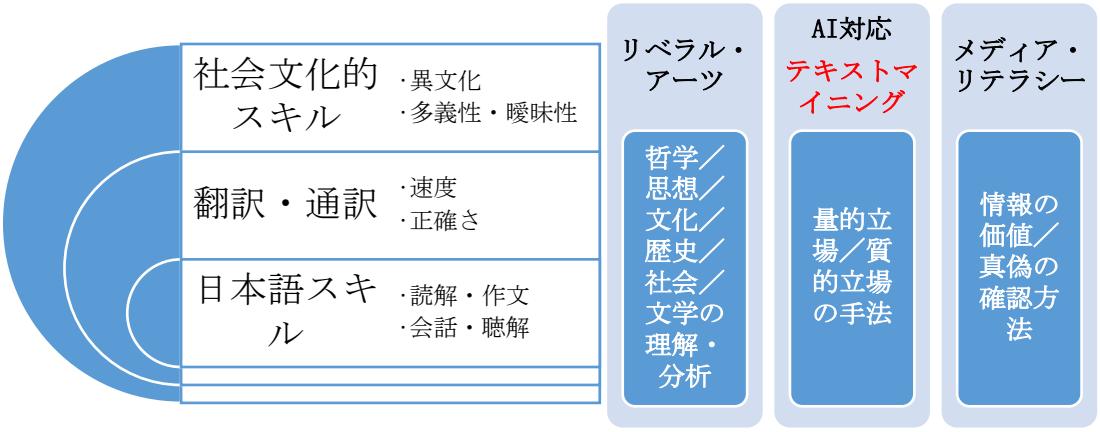


図 13　日本語能力の訓練モデル

テキストマイニングを読解授業に取り入れることで、従来の文法解説と翻訳を中心にした読解から、日本語の表現に即して語彙の用法を理解し、教材以外の資料と対応させながら、内容を理解する手順を学習者中心の活動で形成していくことができる。新しい授業の基本的デザインについて、今後も探究を続けていく必要があると言えよう。

【付記】本論文は、科技部研究案 MOST 109-2410-H-032 -061 -MY3 の研究成果である。また、2020 年 11 月の台湾日語教育学会「2020 年台湾日本語教育研究国際シンポジウムークリエイティブ・ラーニングを目指す日本語教育」で発表した内容に加筆、訂正をおこなったものである。査読委員から丁寧なご意見をいただき、改めて御礼申し上げたい。

テキスト

読売新聞 11 月 25 日「ＧｏＴｏ見直し　政府は混乱回避へ責任果たせ」https://www.yomiuri.co.jp/editorial/20201124-OYT1T50273/（2021 年 3 月 3 日閲覧）

国立国語研究所(2020)「BTSJ 日本語自然会話コーパスについて」参照 https://ninjal-usamilab.info/btsj_corpus/（2021 年 3 月 3 日閲覧）

岡まゆみ構成(2009)『コンテンツとマルチメディアで学ぶ日本語上

級へのとびら』くろしお出版

参考文献

IT トレンド(2020)「テキストマイニングツール 7 つの選定ポイント」https://it-trend.jp/textmining/article/choice（2021 年 2 月 28 日閲覧）

宇佐美まゆみ監修(2020)『BTSJ 日本語自然会話コーパス（トランスクリプト・音声）2020 年版』國立国語研究所「BTSJ 日本語自然会話コーパスについて」https://ninjal-usamilab.info/btsj_corpus/（2021 年 2 月 28 日閲覧）

AIsmaily(2020)「高精度の AI エンジンが対話や議事録を自動で要約・分類」https://ai-products.net/product/voice-recognition-ai-quicksummary/（2021 年 2 月 28 日閲覧）

岡照晃(2018)「『国語研日本語ウェブコーパス』からの新規語彙素獲得の試み」『言語資源活用ワークショップ発表論文集 = Proceedings of Language Resources Workshop』3pp.586-592

奥村学(2010)『自然言語処理の基礎』コロナ社 pp.117-119

落合由治(2004)「文章の基本的構成について―基本的構成から次の段階の構成へ―」『台灣日本語文學報 19pp.195-220

会議 HACK(2020「自動でテキスト化！会議に役立つ音声議事録 5 選」https://www.kaigishitu.com/meeting-hacks/detail/id=32856（2021 年 2 月 28 日閲覧）

計量国語学会編(2017)『データで学ぶ日本語学入門』朝倉書店

報処理推進機構(2017)『AI 白書 2017』https://www.ipa.go.jp/about/report/ai/201707.html（2021 年 2 月 28 日閲覧）

情報処理推進機構(2019)『AI 白書 2019』https://www.ipa.go.jp/ikc/info/20181030.html(2020 年 10 月 6 日閲覧)

総務省(2016)『平成 28 年版情報通信白書』第一部第四章を参照。

http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h28/ （2021年2月28日閲覧）
時枝誠記(1950)『日本文法 口語篇』岩波書店
永尾章曹(1975)『国語表現法研究』三弥井書店 pp.102-129
永尾章曹(1992)「第4章日本語の文法について」永尾章曹編著『日本語学』和泉書院 pp.103-134
西川賢哉、渡邊友香(2019)「『日本語日常会話コーパス』の短単位解析：作業工程を中心に」『言語資源活用ワークショップ発表論文集 = Proceedings of Language Resources Workshop』4pp.238-250
樋口耕一(2020)『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第二版』ナカニシヤ出版
MARTIN WATTENBERG,FERNANDA VIÉGAS,IAN JOHNSON(2016)How to Use t-SNE Effectively https://distill.pub/2016/misread-tsne/
間淵洋子（2020）「特集　2018年・2019年における日本語学界の展望―数理的研究」『日本語の研究』16-2pp.114-121
ユーザーローカル「AIテキストマイニング」 https://textmining.userlocal.jp/（2021年3月3日閲覧）