日本語テキストマイニング技術の文学語学教育分野への応用可能性の検討

落合由治，曾秋桂，王嘉臨，葉麥
淡江大学日本語文学科
{098194, ochiai, 137176, 152790}@mail.tku.edu.tw

1 はじめに

台湾には世界有数の規模の日本語学習者が存在し，日本語文学語学を教える大学の学科も多数存在してきた．しかし，現在，急速な少子高齢化とグローバリズムによる社会変動の進行で，大学の多くは統廃合が避けられない状態となり，同時に職業的可能性の減少で日本語関係学科への進学者も減少しつつあり，今までの日本語習得だけを目的にしたカリキュラムや人文系学科だけを中心にした研究内容から踏み出して，今後の新しい方向性を求める必要に迫られている．[1]

そこで，本学科の有志教員で日本語自然言語処理の技術を取り入れて，新しい教育内容や研究方法の革新ができないかを昨年から勉強会等を開いて，摸索してきた．その中で，社会学，心理学，教育学，経営学などに採り入れられているテキストマイニングの技法が人文系の研究や教育に活かせるのではないかと考え，試行錯誤を行い，また，台湾の日本語教育関係者に呼びかけてシンポジウムを開催し，日本から自然言語処理の専門家を招聘して講演を依頼し，台湾でできる研究発表をおこなって，新しい分野との接続を探ってきた．[2]

その過程で，言語処理学会の存在を知り，また他の分野との交流を広げているという情報を得て，2020年の本大会のテーマセッションでの交流発表を目指すことにした．現在，テキストマイニング技法の人文系研究への応用の中で，中心的課題にしているのは，テキストマイニング技法のテクスト(言語単位としての文章・談話)の質的研究への応用である．今回の発表では，資料の質的読解の結果と，テキストマイニング結果とを比較して，原資料の中でどこに焦点がそれぞれ当たっているのか，また，相互の関係性はあるのかについて検討し，文学・語学・教育など人文系研究へのテキストマイニング技法の応用可能性について考えていきたい．

2 人文系テクスト研究での文章の基本的類型

文学，歴史，文化，社会，人間，思想，宗教などに関する物語，論説，批評，説明，記事などを扱う人文系のテクストを質的統一体として扱う研究が始まったのは，1960年代以降である．いわゆる言語学，国語学的研究が主体と切り離した言語一般を対象として扱い，主に語，文の単位を対象としたのに対して，テクストを対象にした研究は，作品単位，あるいはジャンルを対象とした研究を行ってきた．[3]

言語現象を捉える場合，研究の方向には大きく分けてソシュールのラングの概念が妥当する一般的対象として言語を扱う量的研究と，ソシュールのパロールの概念が適当な個別的対象として言語を扱う質的研究があり，それぞれのパラダイムの相違に応じて，言語の何を対象とし，何を問題にするかは大きく異なってきた．日本語教育の中でも，両者の立場は対立的で，語と文法形式の習得に中心を置く立場と，社会的使用に中心を置く立場が対立している．[4]

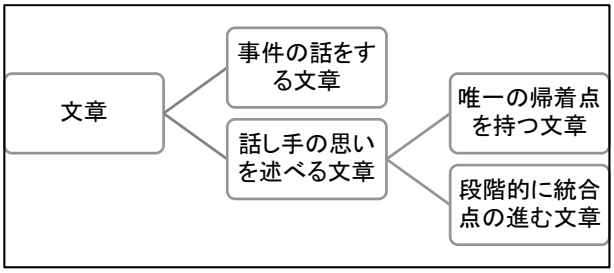
以上のようなパラダイムの対立が人文系での言語の扱いを一層困難にしているが，一般に思われているように，決して言語一般を扱うことが人文系研究の圏域ではない．特定の視点に立つ選択をすれば，研究対

[1] 国際交流基金(2019)「【ご報告】過去最多142の国・地域で日本語教育2018年度「海外日本語教育機関調査」結果（速報）」https://www.jpf.go.jp/j/about/press/2019/029.html 参照(2020年1月14日閲覧).
[2] 台湾日語教育学会ホームページ http://www.taiwanJapanese.url.tw/j-index.htm での相関活動参照.
[3] テクスト研究としての言語研究については，表現学会(2013)『言語表現学叢書』清水書院参照.
[4] 日本語研究と日本語教育のパラダイム問題については，西口光一(2013)『第二言語教育におけるバフチン的視点―第二言語教育学の基礎として』くろしお出版，西口光一(2015)『対話原理と第二言語の習得と教育―第二言語教育におけるバフチン的アプローチ』くろしお出版参照.

象と方法を確立でき，また，教育内容についても内容と方法を確立できる．自然言語処理と人文系研究が接続する場合，どの視点から接するかで扱う対象と問題は異なってくる．ここでは，文章研究の視点での文章の基本的類型の概念から，自然言語処理の一分野としてテキストマイニングの方法との接続を考察していきたい．

文章の基本的類型は日本でテクスト研究が始まった 1960 年代～80 年代にかけてさまざまな考察が行われた課題の一つであるが，現在，言語研究のパラダイムが混乱しているため，論じられることは稀になってきている．日本で初めてテクスト研究を提唱したのは時枝誠記であり，その時から言語研究の語，文，文章という基本的単位が認定されるようになった．ここでは，文章の基本的類型のモデルとして，文法形式との対応がはっきりしている永尾章曹の二種三類の文章の基本的類型を取り上げる．

図１　文章の基本的類型[5]



永尾章曹は，文章を質的統一体として捉え，以下のように二種三類の文章の基本的類型を提示している．この類型は，他の研究者の類型と異なり，テクスト全体の特徴から対応する言語形式の特徴を見出しており，読解，作文等に応用できる言語モデルである．[6]

「事件の話をする文章」は小説，童話，事件の報告に見られ，「話し手の思いを述べる文章」は，詩の形を取ることが多い「唯一の帰着点を持つ文章」と論説，評論，随筆として実現することが多い「段階的に統合点の進む文章」に分けられる．テクストの読解において，このモデルを適用し，文章構成を明らかにすることで，基本的には語彙の相互関係の数量的分析であるテキストマイニングの結果と対照させ，どの程度，特徴が帰納できるかを確認してみることにする．

3　文章の基本的類型とテキストマイニング

以下では，「事件の話をする文章」の事例として，芥川龍之介「羅生門」，「話し手の思いを述べる文章」の事例として芥川龍之介「大川の水」を取り上げて，文章の基本的構成に基づいた文章構成の分析を行い，次に資料をテキストマイニングで分析した結果と比較して質的分析（意味的な内容理解）との関係から，相互関係を考察する．

3.1　「羅生門」と「大川の水」の文章構成

3.1.1「羅生門」の文章構成

典型的な「事件の話をする文章」である「羅生門」は以下の表１のような文章構成を採っている．

表１　「羅生門」の文章構成

段落	内容	機能
Ⅰ	**羅生門で雨止みを待つ下人の登場**	**描写Ⅰ**
Ⅱ	門の周りに誰もないことの説明	説明
Ⅲ	誰もない理由の説明	説明
Ⅳ	羅生門の様子の説明	説明
Ⅴ	下人が羅生門にいる理由の説明	説明
Ⅵ	**ある日の暮れ方の羅生門の情景**	**描写Ⅱ**
Ⅶ	下人が羅生門にいる理由の説明	説明
Ⅷ	**下人の動き＋羅生門の情景**	**描写Ⅰ＋Ⅱ**
Ⅸ	**下人が羅生門の上にあがる動き**	**描写Ⅰ**
Ⅹ	**羅生門に上がった下人の動き**	**描写Ⅰ**
ⅩⅠ	**羅生門に上がった下人の動き**	**描写Ⅰ**
ⅩⅡ	羅生門の中の様子	説明
ⅩⅢ	**鼻をおおう下人の動き**	**描写Ⅰ**
ⅩⅣ	下人が老婆を見つける	説明
ⅩⅤ	老婆の様子の説明	説明
ⅩⅥ	下人の心理の説明	説明
ⅩⅦ	下人の心理の説明	説明
ⅩⅧ	**下人と老婆のやり取り（以下同じ）**	**描写Ⅰ**
ⅩⅩⅦ	**下人と老婆のやり取り**	**描写Ⅰ**
ⅩⅩⅧ	下人の喪失	説明

「羅生門」は，全部で 28 段落に分かれているが，特定の時の持続の中で，ある登場者が登場し，その動きを時の経過に従って描写するⅠ，Ⅷ，Ⅸ，Ⅹ，ⅩⅠ，ⅩⅢ，ⅩⅧ～ⅩⅩⅦの描写Ⅰの段落と，第Ⅵ段落，第Ⅷ段落の一部に見られる，特定の時の持続を止めてその時空にある何らかの登場者を取り上げて気分を示す，いわゆる情景描写に当たる描写Ⅱが，「羅生門」の事件の話をしているストーリーである．一方，その他の各段落は，必要に応じてストーリーに様々な説明を

[5] 永尾章曹(1975)『国語表現法研究』三弥井書店，永尾章曹編著(1992)「第四章日本語の文法について」『日本語学』和泉書院参照．

[6] 落合由治(2007)『日本語の文章構成に関する基礎的研究―テクスト論と結合して』致良出版社参照．

加え，プロットを構成している説明の段落である．「事件の話をする文章」は，このように大きくはストーリーの表現とプロットの表現で構成されている．

3.1.2 「大川の水」の文章構成

一方，「話し手の思いを述べる文章」である「大川の水」は，表2のように「羅生門」とはまったく違った文章構成をしている．

表2 「大川の水」の文章構成

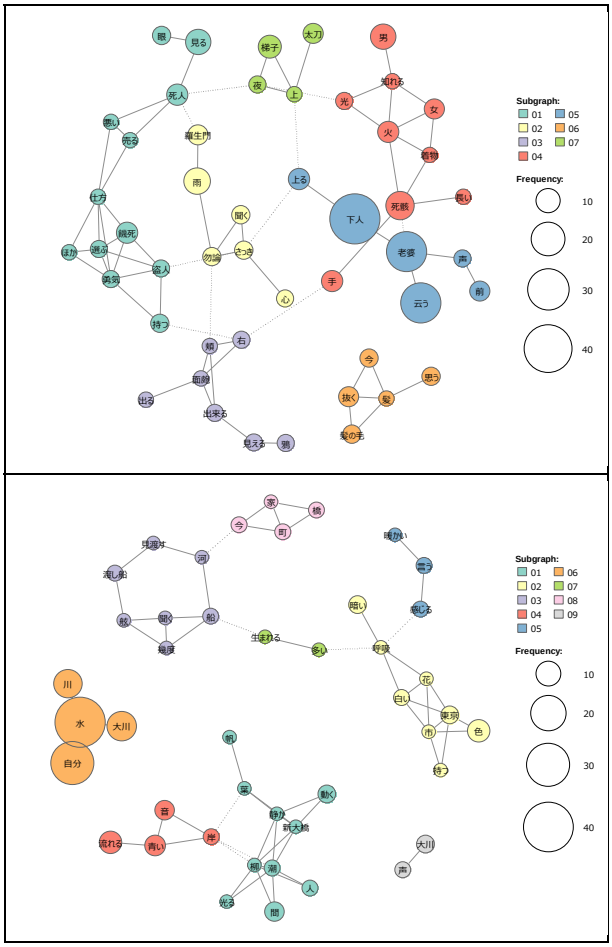
段落	内容	機能
Ⅰ	大川端に近い町に生まれ，「親しく思い出す」ことの説明	話題Ⅰ
Ⅱ	大川を「懐かしい思慕」のゆえに愛する理由の説明	話題Ⅱ
Ⅲ	川の眺めにこころを「おののかせた」思い出の説明	話題Ⅲ
Ⅳ	今でも「さびしい，なつかしい」大川に行く説明	話題Ⅳ
Ⅴ	大川の季節の光景に「さびしさ」を感じる説明	話題Ⅴ
Ⅵ	大川を「慕わしく」思い出す説明	話題Ⅵ
Ⅶ	「なつかしく」思い出す大川の様々な光景の説明	話題Ⅶ
Ⅷ	「なつかしく，さびしさを感じる」大川の渡しの説明	話題Ⅷ
Ⅸ	自分を「魅する」大川の光の説明	話題Ⅸ
Ⅹ	大川の「なつかしい」水の説明	話題Ⅹ
ⅩⅠ	大川の日暮れに「涙を流した」説明	話題ⅩⅠ
ⅩⅡ	大川と東京を「愛する」理由の説明	話題ⅩⅡ
ⅩⅢ	大川の渡しが消えて行く説明	話題ⅩⅢ

「大川の水」は，全部で13段落に分かれているが，それぞれ別々な話題を取り上げているわけではなく，すべて大川に関する自分の思いと思い出に関する内容である．それぞれの話題で取り上げている内容は異なるが，それらに抱く思いは「親しさ」，「懐かしさ」「さびしさ」「慕わしさ」など自分にとって極めて身近な存在に向ける愛情で共通しており，全体として故郷の東京の象徴である失われつつある大川の原風景への郷愁を語っていると言える．このように，「話し手の思いを述べる文章」は，異なる素材であっても共通点のある話題を並べて，それらに共通する焦点を浮かび上がらせる文章構成である．「事件の話をする文章」と「話し手の思いを述べる文章」の文章構成は，大きく異なっている．

3.2 「羅生門」と「大川の水」のテキストマイニング

以上の質的分析に対して，両作品にテキストマイニングを行った結果は以下のようになった．ここでは，共起ネットワーク分析とLDA分析の結果を示す．[7]まず，語彙の共起ネットワーク分析であるが，以下の図2のように，作品中での主要部分が抽出された(以下，ローマ数字は作品の段落，色は図の該当色を示す)．

図2 「羅生門」(上)と「大川の水」(下)の共起ネットワーク分析の結果



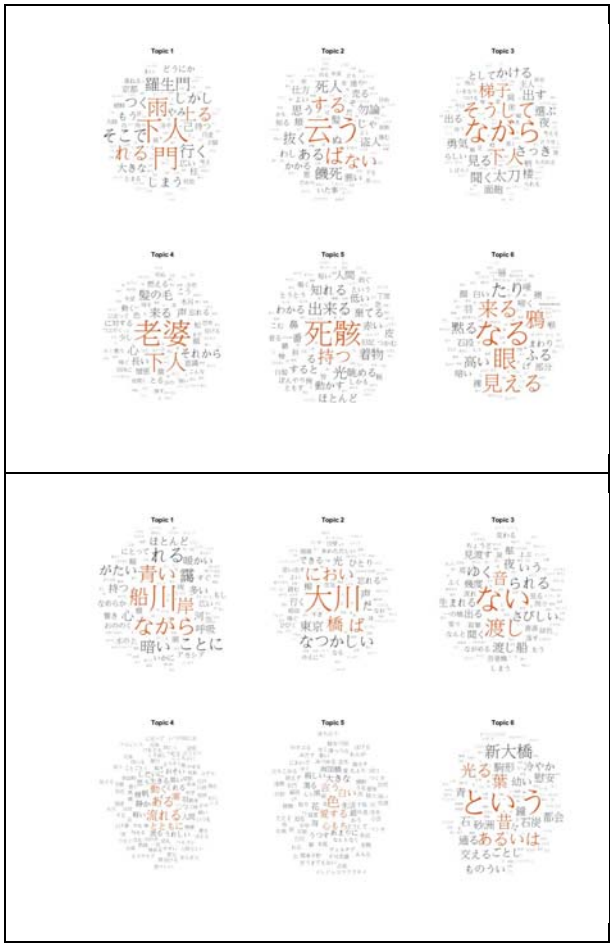
「羅生門」では，7クラスターに分かれ下人の心理(青緑:Ⅶ，ⅩⅥ，ⅩⅦ)，羅生門の情景(黄:Ⅴ)，下人の様子(紫:Ⅹ)，羅生門に上がった様子(赤:ⅩⅡ)，下人と老婆の対話(青:ⅩⅧ～ⅩⅩⅦ)，死人から髪を抜く様子(橙:ⅩⅤ)，階段を上る下人の様子(黄緑:Ⅸ)に主に関わる語彙が7クラスターに分かれ，「大川の水」では，大川と自分が各段落に共通する語彙(橙)として大きなクラスターになり，新大橋などⅧの内容(青緑)，水に関わるⅩの内容(黄)，渡し船に関わるⅩⅠの内容

[7] テキストマイニングは，既成のプログラムを用い，共起ネットワークと自己組織化マップはKHCoder，ワードクラウドとLDAはMathlabのJapanese Text Analysisを使用した．

(紫)，青いなど川の水に関わる各段落の内容(赤)，暖かさを感じる各段落の内容(青)，大川に近い町で生まれたことに関わる町，橋，声などの各語彙(黄緑，薄紅，灰)が，9 クラスターに分類された．他の多くの作品と比べる必要があるが，「事件の話をする文章」の場合は，段落単位で見られる比較的内容の大きなまとまりが抽出され，「話し手の思いを述べる文章」の場合は，語彙の共通性がクラスターの中心になって取り出されやすいように思われる．

次に，LDA 分析で 6 トピックに分けた結果は，以下のようになった．

図 3　「羅生門」(上)と「大川の水」(下)の LDA 分析の結果



「羅生門」では，下人と羅生門の様子(Ⅰ，Ⅱ)，羅生門に誰もいない理由と老婆の話(Ⅲ，Ⅳ，ⅩⅧ～ⅩⅩⅧ)，下人の心理と羅生門を上がる様子(Ⅴ，Ⅵ，Ⅶ，Ⅷ，Ⅸ)，老婆が髪を抜く様子と下人の心理(ⅩⅤ，ⅩⅥ，ⅩⅦ)，死骸に関する説明(Ⅳ，ⅩⅡ，ⅩⅤ)，鴉と老婆の様子(Ⅳ，ⅩⅤ)を中心にした内容がそれぞれトピックとして抽出され，「大川の水」では，川に関係した船，色などの語彙，大川に関する橋やにおいなどの語彙，渡し舟に関する語彙，流れに関する語彙，色に関する語彙，新大橋に関する語彙がトピックとして周出された．LDA 分析の場合，事件の話をする文章では，比較的段落ごとでのまとまりがトピックとして取り上げられやすいと同時に，共通性のある語彙のまとまりが段落を越えて取り出されているようである．「話し手の思いを述べる文章」では，頻出するキーワードを中心に共通性のある語彙がとしてそれぞれトピックとして取り出され，中心になるキーワードが見分けやすいと考えられる．「事件の話をする文章」では，頻度は質的分析にはほとんど意味を持たないので，話題のまとまりを捉える手掛かりとしてテキストマイニングが使える可能性がある．「話し手の思いを述べる文章」では，キーワードは話題の中心を示し，また段落を越えて似た語彙が展開していくので，頻度と共起関係は内容の理解に大きな手掛かりを与えると考えられる．

今回の考察結果から，テキストマイニングで取り出される情報は，文章構成により文章中での機能に大きな違いがあると考えられ，語彙的特徴といってもその意味するところには潜在的質的構造の相違が反映している可能性が考えられる．

4　おわりに

以上，人文系研究，教育に自然言語処理の内容を接続する一つの試みとして，質的分析とテキストマイニングの方法との結果の関係を考察した．質的分析に応用する場合，文章構成によってテキストマイニングの結果の意味するところは大きく異なっており，読解の手掛かりになる部分も相違している．何がどう取り出されているのか，それぞれの言語表現ジャンルの資料で比較考察し，人文系研究，教育に自然言語処理の内容を接続する方途を確立していきたい．

参考文献(注以外)

石田基広(2017)『R によるテキストマイニング入門第 2 版』森北出版
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
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
金明哲(2017)『R によるデータサイエンス第 2 版―データ解析の基礎から最新手法まで』森北出版
斎藤康毅(2018)『ゼロから作る DeepLearning2 自然言語処理編』オーム社
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
李在鎬編(2017)『文章を科学する』ひつじ書房