様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月 ２６ 日現在

機関番号：１６１０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ０２７３２

研究課題名（和文）文章の時系列変化に関する研究

研究課題名（英文）Time Series Analysis of Sentence Length

研究代表者
石田　基広（ISHIDA, Motohiro）

徳島大学・大学院社会産業理工学研究部（社会総合科学域）・教授

研究者番号：４０２３２３１８

交付決定額（研究期間全体）：（直接経費）　　1,300,000円

研究成果の概要（和文）：本研究は、日本語の文章の流れを予測することを目指したものである。日本の学校教育では、作文に「起承転結」を求めることが多い。これは、文章の進行が起、承、転、結とおおまかに４つの段階を経て進むことを意味している。本研究では、小説や研究書などの日本語の文章を対象に、文章の流れに規則性があるか、また規則性があるとして、そこに個人差、あるいは執筆ジャンルの差が存在するかどうかを確認しようとしたものである。

研究成果の学術的意義や社会的意義
今回対象とした文章データから、文の時間的流れに従来の統計分析を適用するのは妥当でないことは明らかになった。最新の状態空間モデルという方法をあてはめることで、文の流れにトレンド（文章のメリハリ）の存在が確認できた。これが本研究の学術的な意義である。社会的な意義として、トレンドを複数の作家から抽出することで、日本語書き手のいわゆる「作風」を数値的に定義できるようになる。今回の研究では数値的な定義には至らなかったが、本研究を発展させることで、「文体」と漠然と表現されている文章の特徴をルール化し、学校教育における作文指導に貢献しうると考える。

研究成果の概要（英文）：The purpose of this study is to predict the flow of Japanese sentences. In Japanese school education, the composition is often called "invocation transfer". This means that the text progresses through roughly four stages: start, acceptance, turn, and conclusion. In this research, time series of sentences of novels and research books are analyzed and checked whether there are individual differences or writing-genres differences.

研究分野：テキストマイニング

キーワード：データサイエンス　テキストマイニング　統計学

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

自然言語の文章では、その構成要素である単語数や文字数に確率分布をあてはめられることが、たとえば代表者による 基盤研究(C)(一般)「文長にみる言語の確率分布」課題番号20520389(2008-2010)などを通して明らかになっていた。この分布に、書き手の個人差、あるいはテキストのジャンル差があるかどうかについては明らかにされていなかった。一方で、文の分布と、時間系列の変化を同時に考察することで、文章の個人差、あるいはジャンル差を明らかにする可能性があるという知見が一部に存在した(文献①)。

２．研究の目的

(1) 文章の計量研究では、テキストがデジタル形式で保存されていることが前提となる。そのため、日本語に関する文章研究では、著作権が切れた作品をボランティアが入力したデータが公開されている青空文庫などが使われることがほとんどであるが、その性質上、比較的古いテキストに限定される。最新の文章についていえば、たとえばツィッターなどから収集したデータなども公開されているが、青空文庫以外のデジタルデータの場合、文章が短い上に総数も少ない。時系列として分析するためには、一部抽出したデータではなく、テキストの総体が望ましい。そこで、テキストそのものの整備を研究課題として行った。

(2) デジタル化したテキストを文章ごとに分割し、単語数や文字数をカウントしたテーブルデータを作成し、時系列モデルをあてはめることで、文章にトレンドや季節性に相当するものが存在するかどうかを明らかにする。

(3) また、トレンドや季節性に、書き手個人、あるいはジャンルにもとづく違いがあるかどうかを確認する。もし確認できれば、「文体」と一言でくくられている概念は、こうした差を指しているとも考えられる。

３．研究の方法

(1) 対象データの整備については、青空文庫から（サイトの利用規約に反しない範囲で）機械的にデータを収集するスクリプトを作成した。また、個人、ジャンルごとの違いを確認することを目的とし、小説家であり、また文学研究者でもある柴田翔氏の小説と研究書を OCR 化してデータ化した。さらに、最近の日本語文章のサンプルとして、比較的文章が「単調」であろうと（研究者が想定した）ラノベというジャンルからも数作品をデータ化した。

(2) テキストごとに、文章から抽出した数値の時系列に対して、伝統的な時系列解析手法を適用した。なお、抽出する単位として「単語」あるいは「文字」が適切であるかは再検討が必要であると考えられたが、Denner S. Vieira *, Sergio Picoli, Renio S. Mendes らの研究(文献②)から、「単語」あるいは「文字」は妥当な計量単位であり、また一般化しやすいと判断した。

(3) 従来の時系列解析はデータの時間的性質に関する制約が強く、テキストによってはこの仮定を満たさない（逆にいえば、文章の時系列は、この点で個別的であり、一般性に欠けている）。そこで、制約のゆるい状態空間モデルを適用し、テキストごとに時間変化を予測することを試みた。

４．研究成果

(1) データベースの作成について、青空文庫などすでに公開されているデータを除き、独自に小説、研究書、ラノベのデータベースを作成した。
(2) データとして選定した個々のテキストごとに、まず伝統的な時系列モデルを適用することを検討した。一般に、時系列分析では、データの原系列が分析の仮定を満たすことはまれである。今回、対象としたテキストにおいても、多くが原系列のままでは仮定を満たすことはできなかった。この場合、テキストごとに差分を取るなどの処理を行うことになるが、その調整について、一貫性のある手順を定めることはできなかった。そのため、テキストごとに時系列分析を適用した結果から、一般化した知見を得ることは難しかった。
(3) 従来の時系列分析に比べ、状態空間モデルでは仮定の縛りがゆるい。そこで、文章から抽出した単位について、正規分布を仮定したモデルを適用し、予測の精度を確認した。説明変数をおかないモデルにおいても、文長の流れについて、相応の予測精度を得ることはできた。そこで、ここに説明変数を追加することを試みた。具体的には、書き手自身が設定したと思われる時間区間（すなわち章番号や節番号）を加えて、予測精度が上がるかどうかを確認したが、大きな改善が見られることはなかった。
(4) 総じて、一般化できる知見は得られなかったが、データとして収集したラノベについては、時間の推移に依存すると思われるトレンドの存在があるように思われたが、この段階では、単なる偶然とも解釈できた。
(5) 共通のトレンドが認められるとして、その変曲点の相対的位置（文の冒頭からの位置）を確認するために、機械学習における異常検知の手法を適用することが考えられる。ただし、本研究の範囲では、この分析にまで到達することはできなかった。

＜引用文献＞

① Dee L. Clayman : 'Time Series Analysis of Word Length in Oedipus the King', Favonius, Suppl. Vol1, pp. 65--79, 1987.

② Denner S. Vieira, Sergio Picoli, Renio S. Mendes: 'Robustness of sentence length measures in written texts', Physica A, pp.749 -- 754, 2018.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計0件

〔図書〕　計1件

１．著者名 石田 基広	４．発行年 2020年
２．出版社 森北出版	５．総ページ数 160
３．書名 実践 Rによるテキストマイニング	

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考