テキストマイニングの文学への応用
－ヘミングウェイの場合－

An Application of Text Mining to Literary Works
-With Special Reference of E. Hemingway-

平井 千津子, 松木 孝幸, 新井 哲男
Chizuko HIRAI, Takayuki MATSUKI, Tetsuo ARAI
東京家政大学
Tokyo Kasei University
Emails: tk9825@tokyo-kasei.ac.jp, matsuki@tokyo-kasei.ac.jp, t_arai@tokyo-kasei.ac.jp

あらまし：本稿では，アーネスト・ミラー・ヘミングウェイの長編小説 3 篇を中心にテキストマイニングの手法を用いた分析をおこなった．各作品の異なり語および延べ語の数や延べ語数の異なり語数に対する比，一文当たりの単語数や単語の長さ，さらに形態素解析でも作品の特徴をとらえることが可能であることから，今回は助動詞および接続詞を取り上げ，同時代人のフランシス・スコット・フィットジェラルドの長編小説 1 篇と比較しながら特徴をみた．
キーワード：テキストマイニング，文学作品，異なり語，延べ語，形態素解析

1. はじめに

1990 年代から一般でもインターネットが普及されはじめたことに伴いデータが大幅に増えたことから，それらを処理するための方法としてデータマイニングが様々な場面で実施されるようになった．とくに対象とするデータが文字列（テキスト）であった場合にはテキストマイニングとよぶ．このテキストマイニングの方法は対象とするテキストを単語や文節などに分割する自然言語処理をおこない，ある特定の語句や単語の出現頻度などを抽出する．そこから，テキストの重要語の抽出，形態素解析することで品詞構成が明らかとなるなどテキストの特徴や傾向を把握することができる [(1)].

このような背景から，前回研究した内容 [(2)] をさらに深めた統計解析をおこなった. 今回はアーネスト・ミラー・ヘミングウェイ（Ernest Miller Hemingway, 1899-1961）の長編著作 3 篇 *The Sun Also Rises*（1926），*A Farewell to Arms*（1929）と *The Old Man and the Sea*（1952）を電子化しそれぞれの特徴をみた. 加えて，これらの作品と比較するためにヘミングウェイと同時期に活躍したフランシス・スコット・フィットジェラルド（Francis Scott Fitzgerald, 1896-1940）の著作 *The Great Gatsby*（1925）も同様に電子化し，各作品や著者の特性を探った．従来の人文科学分野における研究方法は研究対象となる資料中の文章を細かく読むことであったが，情報技術分野において頻繁に使用されるテキストマイニングの手法を用いることで新しい事実・解釈の発見が期待され，人文科学分野における研究および教育の面で大いに役立つと思われる．

2. 方法

Atiz 社の Book Snap を用いて書籍を 2 ページずつ同時撮影し，ABBYY 社の OCR ソフトウェア Fine Reader Pro によってその撮影された画像を PDF 化し、最後にテキスト文字を採集した．1 作品ごとにコンピュータ言語の一つである C#言語で作成したプログラムを用いて，文字コードを UTF-8 形式で取り扱いながら出現する単語を勘定した．このプログラムにより, テキスト形式で保存された文章を読みこみ，それを UTF-8 形式で保存し，半角スペース（” ”）で単語を区切り，単語はすべて小文字に変換させ，最後に特定の記号を取り除いた．今回は，”.”（ピリオド），”,”（カンマ），”(”（始め丸括弧），”)”（終わり丸括弧），””””（二重引用符），”:”（コロン），”;”（セミコロン），”?”（疑問符），”[”（始め角括弧），”]”（終わり角括弧），”_”（アンダースコア），”!”（感嘆符）を除き，出現した単語をアルファベット順に並べ替え，各単語について出現回数を勘定した．

さらに，ドイツの Stuttgart 大学の Helmut Schmid 氏によって開発されたフリーソフトの Tree Tagger [(3)] を利用して各作品の品詞構成を調べた．加えて，作品や著者の特徴をより詳しくつかむために，今回は前述のプログラムによって単語の出現回数を勘定した結果から，前出のプログラムで出現した各単語の文字数を調べると共に Microsoft Office Excel でその確認をおこなった．

3. 結果および考察

各作品の異なり語数と延べ語数，延べ語数の異なり語数に対する比の結果について調べると，ヘミングウェイは同一単語を何回も用いる傾向がみられた．加えて，”.”（ピリオド），”?”（疑問符）と”!”（感嘆符）の出現回数と延べ語数の結果から，一文当たりの単語数を調べた．その結果，ヘミングウェイの 3 作品の中で *The Old Man and the Sea* が他の 2 作品と比較すると 5 単語ほど一文当たりの単語数が多いことがわかった．文末に用いられる記号の勘定方法に

ついて，”.”（ピリオド）の場合，平叙文の文末に出現することもあれば，男性の敬称をあらわす Mr.のように”.”（ピリオド）がつく単語もある．今回はこのような単語も平叙文の文末を意味する”.”（ピリオド）と同様に勘定したため文章数は正確ではないが，おおよその文章数として勘定した．今後の課題として，単語についている”.”（ピリオド）と平叙文の文末を意味する”.”（ピリオド）を目視以外の方法で判別して勘定することが可能かどうか検討したい．

次に各作品に対して形態素解析をおこない品詞構成を調べ，品詞の割合について特徴がみられた品詞について考察した．その結果，名詞率は *The Great Gatsby* が今回比較した他の 3 作品と比較すると高いことから，文章が硬く形式的で要約された文章に近いことがわかった．また，感嘆詞・間投詞の割合は *The Old Man and the Sea* が今回比較した他の 3 作品と比較すると低いことから，会話文が少ないことが示唆された．さらに，助動詞の割合について *The Old Man and the Sea* は can や could のような可能をあらわす助動詞が多く，義務をあらわす助動詞では must のように強制力が大きい単語が比較的多く使用されていることがわかった．Google books Ngram Viewer [(4)] を使用してヘミングウェイの 3 作品と同年代にアメリカで出版された小説における義務をあらわす助動詞の must, should, ought の出現率を調べた結果，should，must，ought の順で高かったが，今回対象としたヘミングウェイの *A Farewell to Arms* を除く 2 作品はいずれも義務をあらわす助動詞について must，should，ought の順に出現率が高いことから，ヘミングウェイは must を好んで使用することが判明した．その一方で，*A Farewell to Arms* は義務をあらわす助動詞は should が最も多く、かつ will や would のような丁寧な意思表現をあらわす助動詞の割合が助動詞全体の約 6 割と他のヘミングウェイの 2 作品と比較すると多く用いられていて、控えめでやわらかい表現が多いという特徴がみられた．そして，接続詞の割合について，一般にヘミングウェイの文体の特徴として接続詞のとりわけ and を非常に多く使用するといわれている．今回の結果から晩年の作品ほど，and のような等位接続詞を”,”（カンマ）を用いて省略せずに接続詞畳用という修辞技法を使う傾向がみられた．*The Sun Also Rises* は接続詞全体に対する and の割合が約 9 割と接続詞のほとんどを and が占める．加えて，*The Old Man and the Sea* は and の使用率が約 4.73%と今回比較した他の 3 作品と比較すると多いことがわかった．

次に単語の長さについて，出現するすべての単語の文字数を各作品勘定したところ，フィッツジェラルドの作品と比較するとヘミングウェイの 3 作品はいずれも平均単語長の値は小さく，内容への依存が比較的小さい動詞に該当した単語のみの平均単語長の結果からも同様の結果が得られたことから，文字数の多い単語より少ない単語をより多く用い，ヘミングウェイの作品はフィッツジェラルドの作品と比較すると平易で単純な傾向にあることが明らかとなった．さらにヘミングウェイの作品について晩年の作品ほど全単語の平均単語長の値が小さくなっていくことも明らかとなったが，動詞に該当した単語のみの平均単語長はこの傾向がみられなかった．この原因として，一般的に動詞において過去形を表す場合は規則動詞では原形＋ed となり，進行形を表す場合は原形＋ing となるなど，原形を表す場合と比較すると文字数が多くなる．このことを踏まえると原形での出現頻度が高いほど文字数は少なくなると思われる．実際，動詞が原形で出現する確率が高いほど，動詞のみの平均単語長の値は小さくなっていて，*The Old Man and the Sea* に出現する動詞の活用形についてみると原形が少なく，同じ単語でも文字数が比較的多くなる進行形および過去分詞形での出現がヘミングウェイの 3 作品の中で最も多かったため，動詞に該当した単語のみの平均単語長の値が大きくなったと思われる．これらのことから，動詞に該当した単語のみの平均単語長について取り扱う際には活用形を原形にもどしてから文字数を勘定することで，より正確に比較することができると推察する．

最後に今回調べた結果をウェブ上で閲覧できるような仕組みをつくった．ホーム画面から，それぞれボタンをクリックすることで作品別に本文（作品の原文）や，品詞構成率のグラフや品詞別に作品比較した結果の表及びグラフを別ウィンドウで表示させることができた．それと同時に php 言語を埋め込むことで利用者が作品名を選択し，単語名を入力することで，その単語の出現回数と単語長（文字数）を同時に表示させることも可能にした．

4. おわりに

現在はヘミングウェイの作品を利用して，あらゆる統計学的手法によって作品（文献）の特徴を数値で表現し，作品の解析や他作品との比較や，文章解析にかんする適切な統計量や自然言語処理手法の検討をおこなっている準備段階である．今後の予定について，語学学習，人文科学分野の研究など多方面で応用できるようにし，とくに教育面においては，大学の授業の課題に取り組む際に学習者自身がオンライン上で膨大な情報（データ）を簡単に解析できるような仕組みを構築したいと考える．

引用文献

(1) JapanKnowledge Lib:“テキストマイニング”
(2) 平井千津子, 松木孝幸, 新井哲男:“統計処理の文学への応用―ヘミングウェイの場合―”，東京家政大学研究紀要，第 55 集, (2)自然科学, pp.39-47 （2015）
(3) Helmut Schmid:“Tree Tagger”，http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/, 2015 年 4 月 30 日 12 時 00 分（最終閲覧）
(4) Google:“Google books Ngram Viewer”, https://books.google.com/ngrams, 2015 年 4 月 30 日 12 時 00 分（最終閲覧）