

## 統計処理の文学への応用 —ヘミングウェイの場合—

平井 千津子\*・松木 孝幸\*・新井 哲男\*\*

(平成 27 年 1 月 7 日査読受理日)

### An Application of Statistical Analysis of Literary Works: With Special Reference to E. Hemingway

HIRAI, Chizuko MATSUKI, Takayuki and ARAI, Tetsuo\*\*

(Accepted for publication 7 January 2015)

キーワード：TF-IDF, 形態素解析, ヘミングウェイ

Key words: TF-IDF, morphological analysis, Hemingway

#### 1. 背景・目的

日本では書籍の電子化は 1980 年代の後半から考えられ始めたが、欧米ではそれ以前から書籍・文献を電子化しそのデータの公開がおこなわれている。たとえば、アメリカでは 1971 年 12 月に著作権切れの書籍を電子化し、それらをインターネット上で広く公開することを目的にマイケル・S・ハートによる「プロジェクト・グーテンベルク」という電子図書館がつくられた<sup>1)</sup>。このサイトでは現在でも作品数は増加している。また、米国タフツ大学は 1987 年よりアリストテレスやプラトンが書いた古代ギリシャの作品を中心に集めており、それらを CD-ROM におさめて発行した。更に 1995 年には電子図書館「ペルセウス図書館」としてインターネットを通して無償で電子化したデータを公開しはじめた<sup>2)</sup>。

一方、日本でも書籍の電子化より前に博物館、美術館では収蔵してある歴史的文化財の情報化がおこなわれ、東京国立博物館、国立歴史民俗博物館においては収蔵品を画像資料として保存し、データベース化していた。1989 年には情報処理学会の研究会の一つに情報技術をいかして、人文科学分野の情報資源を幅広く収集・記録、提供し、人文科学分野の研究の推進及び発展に寄与することを目的に「人文科学とコンピュータ研究会(じんもんこん)」ができた<sup>3)</sup>。

この動きとは別に、企業では雑誌、書籍などの印刷文書を電子化し、それを利用して新たな文書(資料)を作成したり、ワープロやパソコンを使用して文書を作成することが増えはじめた。これ以前はワープロやパソコンを用い、

再度手入力することで電子化していたが作業効率が悪いとことから、企業によっては印刷文書の読取システムを独自に開発し、正確に資料の文字を認識する精度を高める取り組みがされた。

1994 年頃には東京大学の月尾嘉男名誉教授により、博物館・文書館・図書館が所蔵する資料などを電子化し保存するという意味である「デジタルアーカイブ(digital archives)」という言葉がつくりだされた。そして、日本政府は 2003 年 8 月の『e-Japan 重点計画—2003<sup>4)</sup>』の中で、日本文化の理解向上を図るため、あらゆる情報の電子化、アーカイブ化をおこない、国内外へ積極的に情報を発信するために、2005 年度までに図書館の蔵書などの電子化・アーカイブ化を推進している。これより前の 1997 年より国立国会図書館では、1990 年代後半からインターネットが普及しはじめたことから、情報技術を活用したサービスの展開を目指すため「電子図書館構想」をあげ、著作権法に基づき古典籍、雑誌、新聞などを電子化しそれを利用提供している<sup>5)</sup>。公共図書館では当該地域の情報の活用や発信を促進させるために地域資料を、大学図書館では教育や研究支援のために授業教材や研究資料・文献をそれぞれ対象として電子化がおこなわれている。そして、デジタルマイニングによって、電子化されたテキストデータを単語や語ごとに分解し、それらを分析することで特定の単語の出現頻度や時系列変化などをみることができる。これによって、今まででは得られなかった文書全体の傾向や特徴などをつかむことが可能となる。

当研究室では、6 年前より稀覯本・古文書の電子化システムの研究を始め、適切な機器(Booksnap)と文字化のための OCR ソフト(FineReader Pro)の選定等を数年かけて行い、得られた電子データを統計処理するプログラム

\* 環境教育学科 情報科学研究室

\*\* 英語コミュニケーション学科 第 1 米文学研究室

の開発も行い、文献の電子化データ作成・処理システムの作成に一応の成功を得た。このシステムを英国の Punch という 100 年に及んで刊行された雑誌とフランス語の古文獻に対して適用し、相応の結果が得られている<sup>6)</sup>。

この論文ではこのシステムを使い、アーネスト・ミラー・ヘミングウェイ (Ernest Miller Hemingway, 1899-1961) とフランシス・スコット・フィッツジェラルド (Francis Scott Fitzgerald, 1896-1940) の小説を電子化し、そのデータから作品や著者別に特徴をとらえることを試みる。従来の人文科学の分野における調査・研究方法では、研究対象となる資料中の文章について単語を一つ一つ読みこむことがあげられるが、この試みにより従来とは異なった視点から研究対象資料を調べることで、従来の説を補足し、また新しい事実・解釈を見つけることができることが期待される。このことは、今後の文学作品を用いた英語教育や文学研究の場において大いに役立つものと考えられる。

## 2. 図書の電子化について

書籍を電子化することにより、以下のような効果が期待される。書籍(原資料)が貴重資料・古文獻であった場合でも破損・汚損を恐れることなく多くの利用者に閲覧・提供することができる。また、それらの資料はパソコン・プリンターなどの機械と接続することで利用者は印刷資料として入手することができる。日本では国立国会図書館が 1980 年までに刊行された図書や一部の外国の貴重資料を対象に資料の電子化を進め、「国立国会図書館デジタルコレクション<sup>7)</sup>」として資料をインターネット上で閲覧できるサービスを展開している。

さらに電子化された資料は印刷資料と比較し、音声や動画、別の印刷資料といった他の情報と組み合わせをする、あるいは、情報の抽出・検索が比較的簡単にできるデータベース化を行うといった原資料とは異なる別の新しい資料をつくるのが容易である。例えば、ある統計データ資料のような一般的には通読しない資料においては、利用者はある一部分の統計情報を閲覧・入手し、その情報をもとに調査・研究し、文書などでまとめることが多い。加えてこのような資料の場合、印刷されたものであると目視での検索に時間がかかり、数字の見間違いも出てくる。これを踏まえて、日本においては中央省庁が編集・刊行している白書は印刷資料だけではなく、情報の検索性が高い電子化された資料も作成され無償提供されている<sup>8)</sup>。また、ウェブなどを通してそれを世界中の人に向けて情報提供することができ、利用者は時間を気にすることなく、そして同時に多くの人数が同じ資料を利用することが可能である。さらに洋書の場合は原文に加え和文を付け加えた資料を作成することが簡便となり、教育面での利用の幅も広がると考え

られる。

近年では出版点数が年々増加し、網羅的に図書・雑誌を収集・保存し、利用提供している図書館などでは、印刷された状態の資料を保存するための大きな保管場所が必要であり、紙質の劣化が起こるという課題を抱えている。そのためこれらの問題を解決する方法の一つとして資料の電子化があげられており、所蔵資料を電子化する図書館が増えている。

図書の電子化についての問題点と課題の一つ目は、著作権である。書籍自体は単独の著者であっても、その中の図表、写真などにもそれぞれ著作権者があるため、電子化の際には書籍の著者及び図表の著作権者からも許諾を得る必要がある。二つ目は電子化するために必要な人と費用である。電子化するためには通常は人がスキャナーを使用し、1 ページずつ手でめくりながら読み取っていくため、数百ページの書籍であった場合にはその作業時間を考える必要がある。そのため、書籍を傷つけることなく自動でページをめくり読み取る機械を使用して電子化することもある。2005 年 4 月から Google はアメリカのハーバード大学、ニューヨーク公立図書館の蔵書やそれらの関係書類を電子化し、ウェブ上で電子化された資料を対象とした全文検索システム「グーグル・プリント (Google Print) (現グーグル・ブックス (Google Books))」を設計した。その際 Google 側は広告収入や検索対象の拡大が期待されると考えたことから高速でページをめくる機械を導入し、図書の電子化にかかる費用と情報技術を提供した<sup>9)</sup>。しかし、企業ではない図書館などで図書の電子化を考える際には、通常ある程度の費用が発生すると考えなくてはならない。

## 3. 具体的な方法

最初に、ヘミングウェイの作品の一つである 15 篇の短編小説集 *In Our Time* (1924) を電子化し、その中に現れた単語とその出現頻度をそれぞれの各短編小説について調べた。方法は、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 FineReader Pro によってその撮影された画像を文字化し、1 作品ごとに Microsoft Word のマクロ機能を利用して単語を数えた。その際に、ソフトウェアによる文字誤読の修正が一番時間のかかる手作業である。その後得られた電子データを使い、TF-IDF 法で各作品の特徴的で重要となるような単語を調べ、その結果を視覚的に表現することを試みた。なお、単語の数え方は He's のような短縮形の単語や、ハイフンでつながっている two-week のような単語については、そのまま 1 つの単語として数えている。

さらに前述の 15 編の短編集 *In Our Time* と同時期に出版された長編小説 *The Sun Also Rises* (1926) と晩年に出版された長編小説 *The Old Man and the Sea* (1952) を電

子化し、合わせて17作品を形態素解析によって単語に品詞のタグ付けをおこない各作品の特徴を調べた。なお、比較するために、ヘミングウェイと同時期に活躍したアメリカの小説家フィッツジェラルドによる長編小説 *The Great Gatsby* (1925) も同様に電子化し、形態素解析をして比較検討を行った。

#### 4. TF-IDF 法について

TF-IDF とは、情報検索や文書要約など文書解析分野などで主に利用されている指標であり、文書中の単語に重みをつける方法で求めることができる。この考えはキーワードの抽出、全文検索エンジンの重みづけにも利用されている。さらにこの方法で求められた値は文書をベクトル化することが可能となり、二つの文書間においてどの程度似ているかを示すコサイン類似度を計算するときの特徴量ベクトルの値となることが多い。

TF-IDF 値, TF 値, IDF 値はそれぞれ (1) 式, (2) 式, (3) 式で求めることができる。

$$TF - IDF = TF \cdot IDF \quad (1)$$

$$TF_{(i,j)} = \frac{n_{(i,j)}}{\sum_k n_{(k,j)}} \quad (2)$$

$$IDF_{(i)} = \log \frac{|N|}{|d: d \ni t_{(i,j)}|} \quad (3)$$

TF は Term Frequency の略で、ある単語  $i$  が出現する頻度である。具体的な計算方法は、ある単語  $i$  がある文書  $j$  中に出現頻度数  $n_{(i,j)}$  を文書  $j$  中に出現するあらゆる単語  $k$  の出現頻度数の和で除する。一方、IDF は Inverse Document Frequency の略で、ある単語が出現する文書の数の逆数である。具体的な計算方法は、比較する全ての文書数  $N$  をある単語  $i$  が出現する文書数  $|d: d \ni t_{(i,j)}|$  で除した値の対数である。つまり、TF はある単語  $i$  が文書  $j$  に出現する回数が多いほど大きな値となり、IDF はある単語  $i$  が比較する複数の文書間において、それら全ての文書に出現する場合には 0 となる。そして TF-IDF 値は前述の TF 値と IDF 値を乗ずることから、文書に出現する一般的な単語の値は小さくなる。

実際には図書館において書籍を主題分類する場合を考えると、人間が資料を読んで主題分類をおこなう場合には、同じ資料でも読む人によって異なる主題に分類してしまうおそれがある。そのために、資料の主題分類の自動化が考えられてきた。その方法の一つとして TF-IDF 法を用いて資料の重要語を調べ、それに基づいて主題分類が行われている<sup>10)</sup>。さらに藤田学園医学・保健衛生学図書館の OPAC では検索語の TF-IDF 値が高い順に表示するこ

とができる仕組みにも応用されている<sup>11)</sup>。また、アメリカのノースカロライナ州立大学 (NCSSU) 図書館でも目録に TF-IDF 法を用いて適合度順にランキング表示させる機能がある<sup>12)</sup>。加えて TF-IDF 法を用いて重要語を抽出し、その重要語が使用されている文章自体が重要であるという考えから、文書の重要文抽出型の自動要約手法の一つにも応用されている<sup>13)14)</sup>。

#### 5. 形態素解析について

形態素解析とは、文章を意味のある単語ごとに分解し事前に用意した辞書や文法の規則にもとづいて単語ごとに品詞を付与することであり、コンピュータを利用する自然言語処理技術の一つである。文章が日本語の場合は、はじめに文章を分かち書きによって語と語の間に区切りを入れてから解析を始めるが、今回は対象とした文章が英語であり、各単語は通常スペースでわかれていることから分かち書きは必要ない。

今回はドイツの Stuttgart 大学の Helmut Schmid 氏によって開発されたフリーソフトの TreeTagger<sup>15)16)17)</sup> を使用して、形態素解析をおこなった。このソフトは無償提供されており、英語以外にもフランス語やスペイン語の文章にも対応しており、比較的簡単に操作できることが特徴である。このソフトでは単語に 68 種類の品詞を表すタグ付けができる。また、isn't のような短縮形の単語の場合は is と n't (not) に分け、Tom's のように Tom is の短縮形か Tom's という所有格をあらわす単語のどちらに当てはまるのか、前後の文章で判別してそれぞれタグ付けがおこなわれる仕組みである。そこで、今回はその結果を、冠詞、名詞、代名詞、形容詞、動詞、助動詞、副詞、接続詞、前置詞、数詞、関係詞、there is における there (不定副詞)、give up のような句動詞における up (Particle)、外国語、記号 (List Marker)、感嘆詞・間投詞の 16 種類にわけて、各作品の特徴を研究した。

#### 6. 分析結果

##### 6-1. 出現する単語の種類数と総単語数について

はじめに、作品別に出現する単語の種類数、総単語数およびそれらの結果から 1 作品あたり一つの単語が出現する平均回数を表 1 に示す。なお、表中の作品名に付けられた 1 から 15 の番号はヘミングウェイの初期の短編集 *In Our Time* に収録された作品順であり、例えば、1 は *In Our Time* の 1 作品目の "Indian Camp" を表し、以下 2 は "The Doctor and the Doctor's Wife", 3 は "The End of Something", 4 は "The Three-Day Blow", 5 は "The Battler", 6 は "A Very Short Story", 7 は "Soldier's Home", 8 は "The Revolutionist", 9 は "Mr. and Mrs. Elliot", 10 は "Cat in the Rain", 11 は "Out of Season",



12 は “Cross-Country Snow”, 13 は “My Old Man”, 14 は “Big Two-Hearted River: Part I”, 15 は “Big Two-Hearted River: Part II” を表している。

表 1 各作品における単語の種類と総単語数

作品名	単語の種類	総単語数	総単語数/単語の種類
In Our Time_1	459	1,453	3.17
In Our Time_2	451	1,408	3.12
In Our Time_3	424	1,434	3.38
In Our Time_4	721	3,138	4.35
In Our Time_5	695	2,939	4.23
In Our Time_6	251	633	2.52
In Our Time_7	671	2,786	4.15
In Our Time_8	188	379	2.02
In Our Time_9	452	1,389	3.07
In Our Time_10	347	1,144	3.30
In Our Time_11	616	2,177	3.53
In Our Time_12	574	1,740	3.03
In Our Time_13	1,080	6,313	5.85
In Our Time_14	851	3,710	4.36
In Our Time_15	867	4,319	4.98
The Sun Also Rises	5,023	67,652	13.47
The Old Man and the Sea	2,543	26,589	10.46
The Great Gatsby	5,674	48,274	8.51

表 1 から, *In Our Time* の 2 作品目の “The Doctor and the Doctor’s Wife” と 9 作品目の “Mr. and Mrs.

Elliot” は出現する単語の種類数はほぼ同じであるが, 前者のほうが総単語数が多いことから, 同じ単語が何回も繰り返して使われていることが分かる. さらに同様のことが長編小説同士で比較した場合にもみられ, ほぼ同じ時期に出版されたヘミングウェイの *The Sun Also Rises* とフィッツジェラルドの *The Great Gatsby* とを比較してみると, 前者の方が出現する単語数は少ないにもかかわらず総単語数が多くなっており, ヘミングウェイは *The Sun Also Rises* においてフィッツジェラルドが *The Great Gatsby* で同じ単語を何度も使うよりも多い頻度で同じ単語を繰り返し用いていることが分かる. このことは, この 2 作品だけを検討して断言するのは危険であるが, ヘミングウェイがフィッツジェラルドと比べ, 同じ単語を繰り返し使うことを好む傾向にある作家であることを示唆していると考えられる.

## 6-2. TF-IDF の結果

表 2 は, *In Our Time* に収録されている 15 作の短編小説の中で 1 作ずつに対して TF-IDF 値を求め, その値が高い順に上からそれぞれ 20 単語まで並べたものである. この表から全体的な結果として, 人名の TF-IDF 値が高く

表 2 *In Our Time* 中の短編 15 作品別の TF-IDF 値が高い単語 (上位 20 単語)

In Our Time_1 TF-IDF	In Our Time_2 TF-IDF	In Our Time_3 TF-IDF	In Our Time_4 TF-IDF	In Our Time_5 TF-IDF	In Our Time_6 TF-IDF	In Our Time_7 TF-IDF	In Our Time_8 TF-IDF
UNCLE	DICK	MARJORIE	BILL	NEGRO	LUZ	KREBS	MILANO
INDIAN	EDDY	MLL	NICK	MISTER	ARMISTICE	HAROLD	PASS
FATHER	BILLY	BOAT	DRUNK	AD	BATTALION	MOTHER	BOLOGNA
GEORGE	BOULTON	NICK	SAID	BUGS	MAJOR	GIRLS	COMRADES
NICK	TABESHAU	MOON	WEMEDGE	NICK	PADUA	WAR	MANTEGNA
BUNK	DOC	BAY	WHISKY	ADAMS	MARRIED	LIES	SHY
BASIN	DOCTOR	BLANKET	LET’S	TRACK	ABSOLUTELY	FATHER	ITALY
INDIANS	CANT-HOOKS	LINE	FIRE	FIRE	HOSPITAL	YOUR	ADDRESSES
BORN	LOGS	LUMBER	ORCHARD	FRANCIS	MILAN	BEAU	AID
STERN	SAND	PERCH	GLASS	HAM	RAINY	COMPLICATED	AOSTA
BOAT	HENRY	FEEDING	HE’S	BASTARD	GOOD-BYE	HARE	BUDAPEST
DADDY	STOLEN	ROWED	THAT’S	EMBANKMENT	JOB	INDOOR	COMRADE
WOMAN	WIFE	BILL	FALL	CRAZY	HOME	PRAY	DELLA
LADY	COTTAGE	POINT	PRACTICAL	BREAD	AFFAIR	LIKED	EAGER
LAMP	GATE	HORTONS	SOCKS	SNOTTY	AMERICA	CAR	FED
LOGGING	AXES	SAWS	WALPOLE	MAN	ARDITI	GERMAN	FRANCESCA
SHANTY	DEAR	SKINNED	GUY	YOU	BALCONY	LOVE	GIOTTO
UPPER	FATHER	REEL	FISHING	CAP	BANNS	YOU	HEADQUARTERS
BABY	LOG	FIRE	WIND	FIRELIGHT	BIRTH	ARMY	HORTHY’S
CAMP	ROW	TROUT	DAD	MAN’S	BLAB	BACON	HUNGARY
In Our Time_9 TF-IDF	In Our Time_10 TF-IDF	In Our Time_11 TF-IDF	In Our Time_12 TF-IDF	In Our Time_13 TF-IDF	In Our Time_14 TF-IDF	In Our Time_15 TF-IDF	
ELLIOT	CAT	PEDUZZI	GEORGE	KZAR	PACK	CURRENT	
CORNELIA	KITTY	GENTLEMAN	SKIS	HE’D	GROUND	TROUT	
MRS	MAID	YOUNG	SNOW	HORSE	NICK	ROD	
DIJON	GEORGE	MARSALA	NICK	KIRCUBBIN	PINE	NICK	
FRIEND	AMERICAN	PIOMBO	SKI	MY	RIVER	HOOK	
HUBERT	SQUARE	WIFE	WINE	OLD	TENT	STREAM	
TOURNAINE	RAIN	CARO	WALL	JOE	HOPKINS	LINE	
TRIED	WIFE	LIRE	SKING	MAISONS	CANVAS	RIVER	
BOSTON	DESK	RODS	TRAILING	RACE	COFFEE	LOG	
POEMS	MONUMENT	CONCORDIA	FENCE	YD	BURNED	NET	
SOUTHERN	RAINING	SEVEN	SINGING	MAN	FERN	SACK	
PARIS	SI	BANK	STICKS	PARIS	PLAIN	SHALLOW	
BABY	UMBRELLA	BOTTLE	ROAD	RIDING	STREAM	LOGS	
CHATEAU	SIGNORA	US	SHE’S	JOCK	TROUT	GRASSHOPPER	
ELLIOT’S	HAIR	FOLLOW	CROUCHING	I	STRAPS	DEEP	
NUMBER	LIKED	SIGNORINA	KICKING	GEORGE	BLACK	SWAMP	
SHOP	PADRONE	SIGNORA	SLOPE	FUNNY	FIRE	WATER	
FRIENDS	POOR	HOUR	STEEP	HORSES	BRANCHES	HOPPER	
MARRIED	SEA	HOTEL	SWISS	CROWD	CURRENT	SUN	
BOAT	I	FISH	ISNT	GILFORD	POT	BOTTLE	

なっていることがわかるが、この中には作品の主人公の名前が多く含まれており、当然の結果と思われる。

個々の短編について、具体的に検討してみたい。In *Our Time* には“Big Two-Hearted River”という話が1部と2部にわかれて収録されている（14作品目の“Big Two-Hearted River: Part I”と15作品目の“Big Two-Hearted River: Part II”）。この話はニックという青年が川に行き、そこでテントを張って一晩過ごし、翌日バツタを捕まえてそれを餌に川で鱒を釣るという内容である。初めてこの題名を見た人には、題名からだけではその内容が鱒釣りをする青年の話だと判断することは困難であるかもしれない。そこで今回求めたTF-IDF値をみると、“Part I”・“Part II”に共通する語として、NICK（ニック：人名）、RIVER（川）、STREAM（小川）、TROUT（鱒）、CURRENT（水の流れ）があるということに気付く。さらに“Part II”にはROD（釣り竿）やHOOK（釣り針）が上位にあり、“Part I”にはPACK（荷物、リュックサック）、GROUND（地面）、TENT（テント）、CANVAS（厚い帆布）が上位に位置していることに気づく。これらのことに気づくと、題名からだけでは類推しがたい話の内容やあらすじをつかむことが容易となる。また、“Part I”にはBURNED（焼けた）やFIRE（火）、BLACK（黒い）の語もあり、“Part II”には、SHALLOW（浅い）、DEEP（深い）、SWAMP（沼地）の語もある。これらの語は、場面を描写する語ではあるが、同時にこれらの語に注目して作品を読むと、森の中で釣りをするためにこの地を訪れた主人公ニックの心の中に、黒く焼け焦げた大地や自ら進んで入っていきたくない沼地があることが見えてくる。

10作品目の“Cat in the Rain”を見てみよう。ここでは、CAT（猫）、KITTY（子猫）が1番目、2番目に置かれ、RAIN（雨、雨が降る）、RAINING（RAINの現在分詞）も上位に位置し、作品のタイトルにもあるようにCATとRAINがキーワードであることがわかる。しかし、作品では、8番目に置かれたWIFE（妻）が主人公として設定され、雨の中でただ一匹、孤独に雨を避けてテーブルの下にうずくまっている猫をじっと見つめている。この作品では、雨の中の孤独な猫の中に自分の姿を重ねて見る妻と、妻に無関心な夫GEORGEとの間に生じている微妙な心のずれが描かれているが、TF-IDF値で見ると、5番目にAMERICANの語が入っていることに注目したい。即ち、この値から見ると、作者は、この夫婦がアメリカ人夫婦であることを強く意識してこの作品を書いていることが示唆されているように思える。作者ヘミングウェイは、この論文では扱わなかったが、1936年に発表した作品“The Short Happy Life of Francis Macomber”において、アメリカ人夫婦であることを強く意識する作品を書いている。作者のアメリカ人夫婦を描く眼は、TF-IDF値に

しっかりと表れているように思える。また、TF-IDF値では、4番目にGEORGEが位置し、8番目にWIFEが位置している。WIFEの対語はHUSBANDであるが、その語はこの表中にはなく、GEORGEの対語となるべき妻の名前もこの表中にはない。実際、作品中においても作品前半でHUSBANDの語は、2度使われているものの、作品後半では、皆無でGEORGEという個人名か代名詞のHEが使われている。これに対し、GEORGEの対となるべき妻の名前は明らかにされていない。妻の存在は、個人としての存在ではなく、THE AMERICAN WIFEとしての存在であることがうかがわれる。

以上2作品の分析により、ヘミングウェイ作品においては、TF-IDFによる語彙分析は、作品の概要把握や作品解釈のヒントを得る上で有効であり、TF-IDF値を上位順に示した表には作品解釈上の多くの示唆が含まれていることが判明した。今後、表化した語彙と実際の作品とを更に詳しく分析し、他の作品に関してもヘミングウェイ作品における語彙と作品解釈との関係を探っていきたい。

今回分析したTF-IDF値をみると、5作品目の“The Battler”の上から17番目および7作品目の“Soldier’s Home”の18番目のYOU、10作品目の“Cat in the Rain”の20番目および13作品目の“My Old Man”の15番目のIのように、どの小説にも使われる可能性がある代名詞も値が高くなっている。ここでは、TF-IDF値の上位に現れた単語をそのまま各作品に特徴的な単語として使用したが、代名詞は排除する処理をおこなった上で分析を行うとより明確な特徴が表れると考えられる。今後はこの方向で改善し、更なる分析を進めていきたい。

### 6-3. TF-IDF法による特徴語の抽出結果の可視化

前述のTF-IDF法によって抽出した特徴語を、jQueryを利用してランキング表示させることとした。表ではなく可視化することでそれらの情報を視覚的に人間に効果的に伝えることができる。その一部として図1にIn *Our Time*

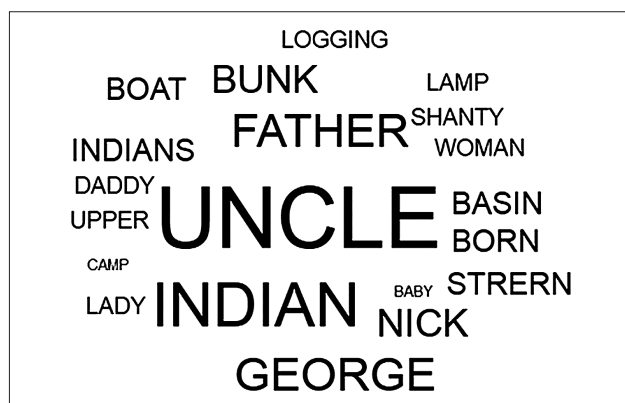


図1 In *Our Time* の1作品目の“Indian Camp”のTF-IDF値が高い単語（上位20単語）

の1作品目の“Indian Camp”の結果を示す。図1では、TF-IDF 値が最も高い UNCLE から続いて INDIAN, FATHER, GEORGE, ……、CAMPと20個の単語の大きさをTF-IDF 値に比例させて、その値が大きいほど文字の大きさを大きくして表示させている。表よりも可視化することで瞬時にどの単語が重要であるかがわかり、利便性が高い。

#### 6-4. 形態素解析の結果

次に、*In Our Time*の15作品と*The Sun Also Rises*, *The Old Man and the Sea*および*The Great Gatsby*のあわせて18作品を形態素解析し、それぞれ品詞の構成比率を調べた。その結果を表3および図2に示す。また、各作品の動詞と副詞および名詞と形容詞をそれぞれ合わせた割合が、前述の全16種類の品詞に対してどの程度あるか調べた。次に代名詞に分類された単語を一人称、二人称、三人称の代名詞およびoneのような不定代名詞の4つにわ

け、それぞれ代名詞全体に対する割合を各作品で求めた。それらの結果を各々図3と図4に示す。

表3および図2で、ほぼ同時期に出版された長編小説*The Sun Also Rises*と*The Great Gatsby*を比較してみると、前者は後者に比べ形容詞の使用度が少なく、動詞の使用度が多いことが分かる。また、このことは、「動詞と副

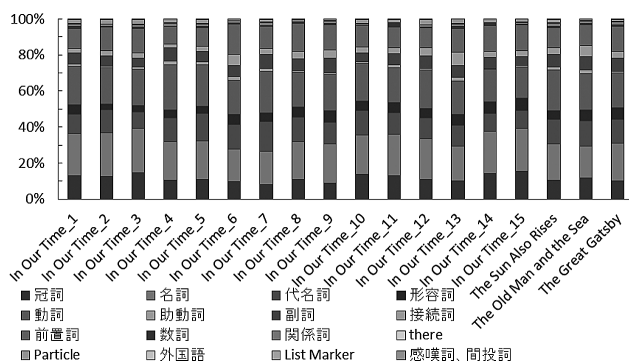


図2 各作品の品詞構成率

表3 各作品の品詞構成率

品詞	In Our Time_1	In Our Time_2	In Our Time_3	In Our Time_4	In Our Time_5	In Our Time_6	In Our Time_7	In Our Time_8	In Our Time_9
冠詞	13.1%	12.5%	14.6%	10.3%	11.0%	9.6%	8.1%	10.8%	8.9%
名詞	23.0%	24.3%	24.6%	21.3%	21.2%	18.0%	18.3%	21.1%	21.8%
代名詞	10.8%	12.7%	9.1%	13.4%	15.1%	13.6%	16.6%	13.7%	11.9%
形容詞	5.2%	3.3%	3.7%	4.3%	4.4%	5.7%	4.7%	5.5%	6.5%
動詞	21.6%	20.1%	20.3%	25.0%	23.0%	19.0%	23.1%	19.5%	20.3%
助動詞	1.1%	1.0%	1.0%	2.1%	1.4%	2.1%	1.7%	0.5%	0.8%
副詞	6.1%	5.5%	5.1%	7.4%	6.0%	6.3%	7.8%	6.6%	8.0%
接続詞	2.4%	3.1%	2.7%	2.3%	2.8%	6.0%	3.0%	4.2%	4.4%
前置詞	11.3%	12.8%	13.6%	9.6%	10.4%	16.9%	12.5%	15.8%	14.3%
数詞	0.9%	0.6%	0.7%	0.2%	0.6%	0.5%	0.4%	0.8%	0.5%
関係詞	1.5%	1.0%	1.4%	1.4%	1.3%	0.8%	1.6%	1.1%	1.5%
there	0.3%	0.1%	0.5%	0.5%	0.3%	0.8%	0.3%	0.0%	0.4%
Particle	1.9%	2.3%	2.3%	1.9%	2.1%	0.8%	1.4%	0.3%	0.8%
外国語	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
List Marker	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
感嘆詞、間投詞	0.5%	0.6%	0.5%	0.2%	0.6%	0.0%	0.4%	0.3%	0.0%
品詞	In Our Time_10	In Our Time_11	In Our Time_12	In Our Time_13	In Our Time_14	In Our Time_15	The Sun Also Rises	The Old Man and the Sea	The Great Gatsby
冠詞	13.7%	13.0%	11.1%	10.1%	14.4%	15.6%	10.5%	11.7%	9.9%
名詞	22.0%	23.2%	22.6%	19.1%	23.3%	23.8%	20.0%	17.7%	21.1%
代名詞	13.3%	11.8%	11.2%	11.9%	9.7%	9.8%	13.8%	13.9%	13.2%
形容詞	5.3%	5.7%	5.5%	5.8%	6.5%	6.8%	5.0%	6.2%	6.4%
動詞	20.9%	19.4%	21.3%	18.7%	18.1%	17.3%	22.4%	19.8%	19.2%
助動詞	1.0%	1.6%	1.0%	1.8%	0.6%	0.8%	1.4%	2.1%	1.0%
副詞	5.0%	6.1%	6.9%	6.6%	6.1%	5.0%	7.0%	7.6%	7.2%
接続詞	3.2%	3.2%	4.3%	7.3%	3.3%	3.2%	3.7%	5.9%	3.9%
前置詞	11.7%	11.5%	11.3%	13.5%	14.4%	14.4%	11.4%	11.7%	13.9%
数詞	0.3%	1.6%	0.4%	1.0%	0.5%	0.6%	0.8%	0.6%	0.9%
関係詞	0.5%	0.6%	0.9%	1.3%	0.7%	0.6%	1.5%	1.4%	1.5%
there	0.3%	0.1%	0.3%	0.2%	0.2%	0.3%	0.3%	0.3%	0.3%
Particle	2.2%	2.0%	2.6%	2.6%	2.2%	1.8%	1.7%	0.9%	1.2%
外国語	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
List Marker	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
感嘆詞、間投詞	0.4%	0.3%	0.6%	0.0%	0.0%	0.0%	0.6%	0.1%	0.3%

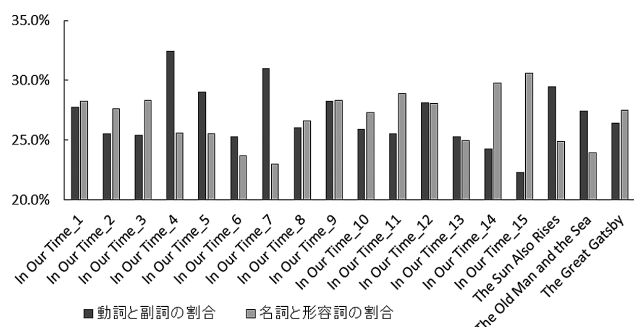


図3 各作品の動詞と副詞の割合  
および名詞と形容詞の割合

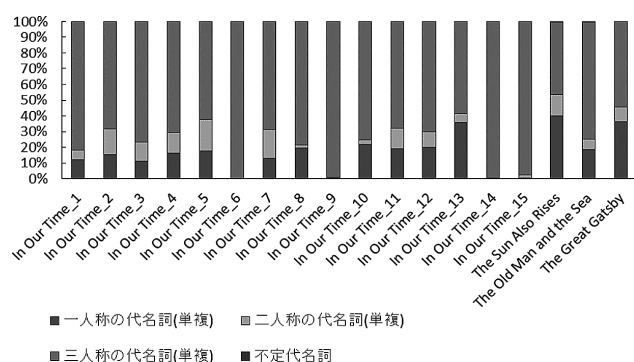


図4 各作品の人称別の代名詞の割合

詞を合わせた割合」と「名詞と形容詞を合わせた割合」を採った図3で見るとさらに明確になる。短編では作品によりややばらつきがみられるが、*The Old Man and the Sea*を含め、「動詞と副詞を合わせた割合」が高いことは、抽象語を嫌い、簡潔で生き生きとした文章を好んだヘミングウェイの文体の特徴を数的に示すものといえる。

石川（2012）によれば、文章内に動詞や副詞が多い場合は、くだけていて動的な文章、名詞や形容詞が多い場合は、かたく説明的であり、描写的な文章だといわれている<sup>18)</sup>。また、代名詞については、回顧録、日記、自叙伝などで一人称の代名詞が多くみられ、小説において多用されている場合は、登場人物と同じ目線で書かれていることから読者は感情移入しやすく、臨場感が得られ躍動感のある文章だと推察される。一方、三人称の代名詞が多くみられるものとしては神話、昔話、伝説などがあげられ、小説において多用されている場合は客観的な文章であると考えられる。さらに、二人称の代名詞が多くみられるものは手紙であり、小説で二人称の代名詞を多く使用する作品は前述の一人称や三人称のそれと比較すると少ない。小説において多用されていた場合、読者は登場人物に語りかけられることから、徐々に何かが迫りくる印象を与えられ、一人称の代名詞が多く出てくる小説とは異なる臨場感を得ることができると考えられる。

このように文章の品詞の構成比率を知ることは、文章の特性の発見につながる。

図3を見ると、*In Our Time*に収録されている小説の中でも、4作品目の“The Three Day Below”と5作品目の“The Battler”と7作品目の“Soldier’s Home”の「動詞と副詞を合わせた割合」は、「名詞と形容詞を合わせた割合」と比較すると多い。一方、14作品目の“Big Two-Hearted River Part I”と15作品目の“Big Two-Hearted River Part II”は「名詞と形容詞を合わせた割合」のほうが、「動詞と副詞を合わせた割合」より多い。長編小説では、先にも述べたように、*The Sun Also Rises*と*The Old Man and the Sea*の両作品とも「動詞と副詞を合わせた割合」の方が、「名詞と形容詞を合わせた割合」と比較すると多い。このことは、登場人物たちの心理や行動様式、それを表現するための語りの方針とも大いに関係するものと思われるが、今後の課題としたい。

また、図4の代名詞の比較では、*In Our Time*の6作品目の“A Very Short Story”と9作品目の“Mr. and Mrs. Elliot”と14作品目の“Big Two-Hearted River Part I”と15作品目の“Big Two-Hearted River Part II”の4作品は、全てあるいはほぼ全ての代名詞が三人称の代名詞であり、一般にヘミングウェイの短編小説においては、代名詞全体に占める三人称の代名詞の割合が高いことが判明した。このことは、できるだけ主観を排し、客観的な描写に徹することを旨とした作家自身の著作態度の反映であるともいえよう。これに比べ長編小説*The Sun Also Rises*では代名詞全体に占める一人称の代名詞の割合が40%となり、ヘミングウェイ作品の中では際立って高率となっている。この作品では、長編小説のため登場人物が多いなどの影響もあると思われるが、ここでは事実の指摘のみにとどめ、詳しい分析は今後の課題としたい。

## 7. まとめ

今回はヘミングウェイとフィッツジェラルドの小説18作品をAtiz社のBook Snapによって撮影し、ABBY社製のOCRソフトFineReader Proを使って画像データから文字を読み取り、書籍を電子化した。撮影で読み取った画像を正確な文字に編集し直す作業は人の手でおこなう必要があることから、かなりの時間を要することがわかった。次に電子化した小説を1作品ずつMicrosoft Office Wordのマクロ機能を利用して、単語数を調べた。ここの課題としては、長い文書の場合は単語を数える時間が非常に長くなることから、今よりも早くそして正確に単語を数える方法を見だし、改善することがあげられる。次に得られた電子データを利用して*In Our Time*の15作品についてTF-IDF値を求めた。前述の結果より、TF-IDF値は人名や地名が比較的高くなる傾向がみられた。その一方でIやYOUなどの代名詞もその値が高くなっていたことから、代名詞や前置詞や冠詞に該当する単語については



TF-IDF 値をみる上で除外する必要があると考えられる。

続いて、Helmut Schmid 氏が開発した文章中の各単語に品詞をタグ付けするソフトの TreeTagger を使用して形態素解析し、18 作品の品詞の構成比率を調べた。今回は各作品の動詞と副詞や名詞と形容詞の割合、代名詞について人称別に割合を求めそれぞれ比較した。今後は、他の品詞に着目してより詳しくその割合を調べ、作品についての特徴を探りたいと考える。たとえば、動詞について現在形、現在進行形、過去形のように時制ごとにそれぞれ割合をみることがあげられる。

これら以外にも特定の単語の出現度や一つの作品を Chapter (章) ごとにわけてそれぞれ出現回数を調べることで、作品のどの位置に多くみられるのかを調べ、結果をグラフなどで可視化することも検討している。

大学における英語教育について『英語指導方法等改善の推進に関する懇談会 報告<sup>19)</sup>』には、大学英語教育の現状に情報検索技術を身につけると同時に情報を得てそれを発信し議論する英語力が必要で、様々な大学が工夫をしているとある。また、ここでは、中学校、高等学校においても情報通信機器の活用と関連して英語力を育成させることが重要であると述べている。たとえば、千葉大学の外国語科目でもこのための取り組みをおこなっており、一部の英語科目では英文雑誌や教科書の要約文の作成、読解した内容の要約を英語で書くなど書く力を養成している<sup>20)</sup>。このように読解した文章から要点をつかみ、まとめるためには話の中心となるキーワードを見つけることが重要であると推察される。今後の研究目標としては文書中の重要単語を調べ、形態素解析だけではなく、ある特定の単語に着目してその出現間隔をグラフで表すことなどを可能にし、文書の電子化によって従来の紙媒体の文書では探し出すことが困難な値、量を算出することを置いている。そして今回は 18 作品だけであったが、さらに多くの小説を電子化することで作品だけではなく、著者ごとの特性や傾向も明らかにすることが可能となるのではないかと考えている。その結果を語学学習、人文科学分野の研究、図書館など多方面で応用できるようにしたい。

### 分析作品

Fitzgerald, Francis Scott, *The Great Gatsby*, Charles Scribner's Sons, 1925

Hemingway, Ernest Miller, *In Our Time*, Charles Scribner's Sons, 1924

Hemingway, Ernest Miller, *The Sun Also Rises*, Charles Scribner's Sons, 1926

Hemingway, Ernest Miller, *The Old Man and the Sea*, Charles Scribner's Sons, 1952

### 引用文献

- 1) プロジェクト・グーテンベルク : <http://www.gutenberg.org/>  
2014 年 8 月 30 日 14:00 最終アクセス
- 2) ペルセウス図書館 : <http://www.perseus.tufts.edu/hopper/>  
2014 年 8 月 30 日 14:00 最終アクセス
- 3) 人文科学とコンピュータ研究会 : <http://www.jinmoncom.jp/>  
2014 年 8 月 30 日 14:05 最終アクセス
- 4) 経済産業省 IT 戦略本部 : e-Japan 重点計画 -2003-, 2003, pp. 5 ~ 69
- 5) 国立国会図書館 電子図書館事業の概要 : <http://www.ndl.go.jp/jp/aboutus/elib-project.html>  
2014 年 8 月 30 日 14:15 最終アクセス
- 6) 海和夏希 : 2012 年度 修士論文要旨集, 東京家政大学大学院 家政学研究科 (東京), 2013, pp. 17 ~ 20
- 7) 国立国会図書館 国立国会図書館デジタルコレクション : <http://dl.ndl.go.jp/>  
2014 年 8 月 30 日 14:20 最終アクセス
- 8) 総務省 政府統計の総合窓口 e-Stat : <http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do>  
2014 年 8 月 30 日 14:25 最終アクセス
- 9) Harvard University Library Harvard-Google Project : <http://hul.harvard.edu/hgproject/index.html>  
2014 年 8 月 30 日 14:30 最終アクセス
- 10) 石田栄美 : 三田図書館・情報学会, **39**, 31 (1998)
- 11) 藤田学園医学・保健衛生学図書館 書誌蔵書検索 検索条件入力 (高機能検索) : <http://library.fujita-hu.ac.jp/scripts/mgwms32.dll?MGWLPN=CARIN&wlap p=CARIN&WEBOPAC=1&i=1409377612589>  
2014 年 8 月 30 日 14:45 最終アクセス
- 12) K. Antelman, E. Lynema and A. K. Pace : *Toward a 21st Century Library Catalog. Information Technology and Libraries* **25**(3), 128 (2006)
- 13) G. Salton and C. S. Yang : *Journal of Documentation* **29**(4), 351 (1973)
- 14) T. Mori, M. Kikuchi, K. Yoshida : *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, 5 (2001)
- 15) Tree Tagger : <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>  
2014 年 8 月 30 日 14:35 最終アクセス
- 16) H. Schmid : *Proceedings of International Conference on New Methods in Language Processing* **12**(4), 44 (1994)



- |  |   |
|--|---|
| 17) H. Schmid : <i>Natural Language Processing Using Very Large Corpora Text, Speech and Language Technology</i> 11, 13 (1999) | <a href="http://www.mext.go.jp/b_menu/shingi/chousa/shotou/018/toushin/010110b.htm">http://www.mext.go.jp/b_menu/shingi/chousa/shotou/018/toushin/010110b.htm</a><br>2014 年 8 月 30 日 14:35 最終アクセス |
| 18) 石川慎一郎：ベーシック コーパス言語学，ひつじ書房（東京），2012, pp. 151 ~ 156  | 20) 国立大学法人 千葉大学普遍教育 英語科目その 2 英語授業形態： <a href="http://www.fuhen-chiba-u.jp/pub/fuhen/1051.html">http://www.fuhen-chiba-u.jp/pub/fuhen/1051.html</a><br>2014 年 8 月 30 日 14:40 最終アクセス                |
| 19) 文部科学省：英語指導方法等改善の推進に関する懇談会報告，2001   |   |

### Abstract

An examination of whether statistical methods can be effectively applied to literary works is made in this paper. First, the TF-IDF (Term Frequency - Inverse Document Frequency) method is applied to selected literary works by E. Hemingway and F. S. Fitzgerald to find the frequency of each word in use. Then, a morphological analysis is undertaken of the words in each work. The results show that the vocabulary of *The Sun Also Rises* by E. Hemingway is smaller than that of *The Great Gatsby* by F. S. Fitzgerald, whilst the number of words used in *The Sun Also Rises* is larger than that in *The Great Gatsby*, which suggests the words in *The Sun Also Rises* might have a tendency to be repeated more than those in *The Great Gatsby*, and that directing our notice to high-frequency words in Hemingway's short stories could lead to a deeper understanding of the works. Furthermore, verbs plus adverbs are more frequently used than adjectives plus nouns in *The Sun Also Rises* while the latter are more frequently used than the former in *The Great Gatsby*. These results show that statistical analysis is an effective way to clarify the author's writing style and can lead to a deeper understanding of literary works.