小説テクストの計量的分析
アーサー・コナン・ドイルの作品から

黒田　絢香

大阪大学大学院言語文化研究科

〒 560-0043 豊中市待兼山町 1-8

E-mail:　kuroda22@gmail.com

あらまし　本研究では，Arthur Conan Doyle の推理小説と歴史小説を対象とし，その語彙頻度や生起パターンを計量的に分析することで，作品の特徴やジャンル間の違いを考察する。これまで客観的なデータに基づく分析が行われていなかった作家の作品を量的な観点から考察することで，文学研究に新たな視点を提案することが目的である。推理小説と歴史小説を区別する言語的特徴を検討するため，Random Forests を用いて機械的な分類を試み，分類に寄与したキーワードを抽出する。次に MALLET を用いたトピックモデリングを行い，結果をネットワークグラフに表す。どのような語がトピックを構成しているのか，両者がそれぞれどのようなトピックを持っているかグラフをもとに考察し，その差を検討する。以上の結果から，ジャンル間の相違を反映する特徴を明らかにする。

キーワード　ジャンル，ランダムフォレスト，トピックモデル

Quantitative Analysis of Literary Works
Novels of Sir Arthur Conan Doyle

Ayaka Kuroda

Graduate School of Language and Culture, University of Osaka

1-8 Machikaneyama-cho, Toyonaka, Osaka, 560-0043 Japan

Abstruct　This study attempts to provide a new perspective for literary studies through quantitative investigation of words in texts with special reference to word frequency patterns. Two types of machine-learning analyses are conducted to find differences between historical fiction and detective fiction of Sir Arthur Conan Doyle. While Conan Doyle is well-known for the Sherlock Holmes series, his strong inclination for historical fiction has hardly been recognized. A number of studies have carried out to examine personalities of characters or estimate the dates of composition for some of the texts that belong to the Holmes series. Few studies, however, have focused on Doyle's historical fiction. Still less critical attention has been paid to stylistic aspects of his novels and short stories. Machine-learning approaches made it possible to highlight linguistic/stylistic features that distinguish Doyle's historical fiction from his detective fiction. We used Random Forests to show genre-specific 'key words', or words with a high keyness value so as to discriminate between the two categories of texts. MALLET was used in conjunction to build topic models based on Latent Dirichlet allocation (LDA). What emerges from our analyses are linguistic features that differentiate between the two text genres.

Keywords　text genre, Random Forests, topic model

1.　はじめに

1.1.　背景と目的

Arthur Conan Doyle は Sherlock Holmes シリーズの作者として著名なイギリス人作家である。彼は元々医師であったが，診察の間の空いた時間を利用して小説を書き，1887 年に Holmes シリーズの第 1 作となる *A Study in Scarlet* を発表した。当初は注目されなかったが，短編を雑誌に掲載したことから突如人気となり，以降このシリーズは世界的に有名となった。この功績から，推理小説を世に広めた人物と言われている。しかし，Doyle は推理小説以外にも歴史小説や SF，ノンフィクションなども数多く手がけており，どちらかというと歴史小説の方が自身の本分であると考えていた。このことは，Holmes を『殺害』する作品である *The Final Problem* を発表する前に自身の母に宛てた手紙に書かれており，「最後の物語でホームズを殺し，この仕事を打ち切ることを考えています。彼のために私は他のもっと素晴らしいこと (歴史小説) を考える余裕がなくなっているからです」と漏らしている (河村, 1991)。

しかし，そうして発表された *Micah Clarke* を始めとする歴史小説群は，評価こそされているもののこれまで Holmes シリーズの陰に隠れ，あまり批評や研究の対象となっていない。Holmes シリーズは現在でもファンが多く，Sherlockian と呼ばれる熱狂的なファンによって様々な研究が行われている。例えば Baring-Gould は，Holmes やその他のキャラクターたちが実在の人物であるという前提のもと事件の年代や日付の特定を試み，1977 年には架空の人物である Holmes の伝記を発表している。しかし，そういった研究の多くはキャラクターの人物像や事件の詳細に注目しており，物語を実際に形作る言語や文体について注目したものは少ない。

一方で文体解析の分野では，計量的な分析が注目されている。近年，情報技術の発達により，ビッグデータと呼ばれる大規模データを解析する手法が進歩している。その進歩は工学分野のみならず人文系分野にも応用されており，現在では何億語もの言葉の集合を対象とした検索や網羅的な分析，統計的処理も行うことが可能になっている。こういった計量的テクスト分析は，著者推定や文体論の分野はもちろんのこと，文学研究にも従来の手法とは異なる新たな視点をもたらすことができると考えられている。

そこで本研究は，Doyle の作品群を対象として計量的な文体分析を試みることで，新たな研究の視点を提案することを目的とする。

1.2.　先行研究

文学作品の批評的な研究に計量分析を用いた例として Burrows (1987) の Jane Austen 研究が挙げられる。彼は Austen の作品に生起する単語をマッピングし，そのパターンから作中キャラクターの分析などを行った。特徴的であったのは，主成分分析 (Principal Components Analysis, PCA) を用いていた点と，デジタルの手法を用いたことで機能語のような生起頻度の極めて高い単語も分析対象とできた点である。これにより，個人言語 (idiolect) の類型化に関する研究が進んだ。

一方で，Doyle 本人を対象とした研究は，Doyle の死の翌年に刊行された Lamond (1931) の伝記が最初の例として挙げられる。しかしこれはスピリチュアリストとしての Doyle を中心に取り上げており，作家としての Doyle を初めて描いたのは Pearson (1943) の伝記である

と考えられている。この本はDoyleの人生を出生から晩年まで辿り，彼の政治的思想の移り変わりや出版の背景について詳細に述べている。その後も数多くの伝記が出版されているが，Pearsonのものと同様に作者本人の来歴や実績などbiographicalな研究がほとんどで，作品の言語表現を取り扱っているものは少ない。大賀 (1988) はDoyleの文体に関する研究を行っており，作品中に見られる倒置や比喩，逆説，ironyなどの文体技法に関して具体例を幾つか挙げて論じていたが，質的な分析を中心に行っており，客観的な量的データは挙げられていない。

2. 分析対象

本研究で分析対象とするデータは表1に挙げている通り，*Micah Clarke*を始めとする歴史小説9作品 (1889–1906)，Sherlock Holmesシリーズの7作品 (1887–1915) の計16作品である。いずれもProject Gutenbergよりダウンロードしたが，編集時の注釈やタグ，目次などは取り除き，本文のみのプレインテクストとしている。総語数と総異なり語数は，その状態で計量したものである。

表 **1:** 分析対象の作品一覧

標題	作品ラベル	発表年	総異なり語数	総語数
Micah Clarke	HF_1	1889	14,091	177,593
The Firm of Girdlestone	HF_2	1890	11,703	136,665
The White Company	HF_3	1891	11,487	150,252
The Great Shadow	HF_4	1892	5,250	49,599
The Refugees	HF_5	1893	9,564	122,671
Rodney Stone	HF_6	1896	8,515	90,875
Uncle Bernac	HF_7	1897	6,339	57,446
A Desert Drama	HF_8	1898	5,918	46,579
Sir Nigel	HF_9	1906	9,801	130,232
A Study in Scarlet	SH_1	1887	5,910	43,185
The Sign of the Four	SH_2	1890	5,578	43,015
The Adventures of Sherlock Holmes	SH_3ss	1892	8,424	104,361
The Memoirs of Sherlock Holmes	SH_4ss	1894	7,604	87,395
The Hound of the Baskervilles	SH_5	1901–1902	5,812	59,046
The Return of Sherlock Holmes	SH_6ss	1905	8,679	111,968
The Valley of Fear	SH_7	1914–1915	5,977	57,463

Holmesシリーズには短編小説も多くあり，作品ラベルにssとついているものは全て短編集であるが，短編それぞれを独立したデータとはせず，全体を一冊の本として取り扱っている。長編と短編を一まとめに分析してしまうと，総語数が作品ごとに著しく異なり，ジャンルによる差よりも文書の長さによる差の方が強く出てしまうためである。

3. Random Forests による分類とキーワード抽出

3.1. 特徴語について

特徴語，もしくはキーワードと呼ばれる単語とは，一般に「そのテクストの主題を反映する語」であると考えられている。本論文においては，客観的なデータから計算的に特徴を割り出すことを目的としているため，より明白な定義が必要となる。高見 (2003) は，イギリスの高級紙と大衆紙を分析対象とし，その特徴を統計的手法のもと特定する方法論について述べており，高見 (2004) では，特徴語を「比較対象とした複数のコーパスの中で，特定のコーパスとそれ以外の全てのコーパスとでその出現頻度に統計的に有意な差のある語」(p. 31) と定義している。本研究はこの定義を踏襲し，歴史小説と推理小説を比較した際出現頻度に有意差のある語が何か検討する。

頻度差の有意性を検定するため，対数尤度比 (log-likelihood ratio) やカイ二乗値など様々な統計尺度が考案されている。しかしこれらの指標は，特に文学作品の分析に用いた場合，対象とするテクスト群のうちごく少数の作品にしか出現しない語や固有名詞が特徴語と判断されてしまうという問題がある (Tabata, 2015)。

よって，本論文では Random Forests を用い，分類に寄与した語を特徴語と判断する。判断基準として用いた Mean Decrease in Accuracy は，該当の変数をモデルから取り除いた際にどれくらい分類精度が低下するかをもとに割り出された寄与度の値である。

3.2. 手法

Random Forests (Breiman, 2001) は主に回帰，クラスタリングに用いられるアンサンブル学習のアルゴリズムである。複数の決定木の結果を合わせることで大量のデータを機械的に分類し，各変数が分類にどれだけ寄与したかを出力する。

本研究では高頻度語を変数とし，対象ファイルに対し Random Forests による分類を行った。高頻度語は the や is などの機能語が中心であるが，一般的に機能語は語彙的意味を担っておらず，従って作品の特徴分析においては重要でないと考えられている。しかし，質的な分析では軽視されがちなこれらの項目は，相対的な頻度や生起パターンなどの情報によって計量分析に大いに貢献する。つまり高頻度語は，総語数に占める割合が高く，テクストの情報量の多くを有していると考えらる。実際に，対象の 16 作品では，上位 500 語が総語数のうち 71%から 77%を，上位 100 語の場合でも 54%から 59%を占めている。また，あまりに出現頻度の低い語は統計処理の結果誤差の範囲となってしまい，テクスト間の有意差を証明する証拠として挙げることは難しい。よって本研究では，高頻度語を変数とすることにより語彙頻度の有意な差を抽出し，テクスト間の差を比較する。

分析変数の妥当な項目数を検討するため，上位 100 語から上位 500 語までの間で項目数を変化させて実験を行ったところ，どれも分類精度が 97.26%であった。このため本研究では上位 500 語を変数とした場合の結果について考察する。

また，分析は対象作品を先頭から順に 10,000 語ごとに切り分けて作成した合計 147 のテクストファイルに対して行った。ファイルごとの語数を揃えること，また言語的特徴の作品内での変移を観察することが目的である。例えば HF_1 の *Micah Clarke* のように合計 177,593 語の作品の場合，10,000 語のファイルが 17 個，7,593 語のファイルが 1 個生成され，HF_1.1 から HF_1.18 までのファイル名が順番に付けられている。なお，最後のファイルが 5,000 語未満になる場合は切り捨てとしている。

分析とグラフの描画，コンコーダンスラインの表示には，統計解析ソフトウェアの R，及

びコーパス分析ソフトウェアの CasualConc を並行して用いた。

3.3. 固有名詞に係る考察

出力結果は以下の表 2 の通りである。分類精度は前述の通り 97.26%で，Holmes シリーズのうちの 4 ファイルが歴史小説群と誤分類されていることがわかる。

表 **2: Random Forests** の実行結果

Call:			
randomForest(formula = text.group ∼ .,			
data = tbl, proximity = T, importance = T, ntree = 100000)			
Type of random forest:			classification
Number of trees:			100000
No. of variables tried at each split:			22
OOB estimate of error rate:			2.74%
Confusion matrix:			
	HF	SH	class.error
HF	97	0	0.0000000
SH	4	45	0.08163265

ところが，Mean Decrease in Accuracy の上位項目リストから，固有名詞である Holmes の Mean Decrease in Accuracy の値が圧倒的に高くなっているという問題が発見された。

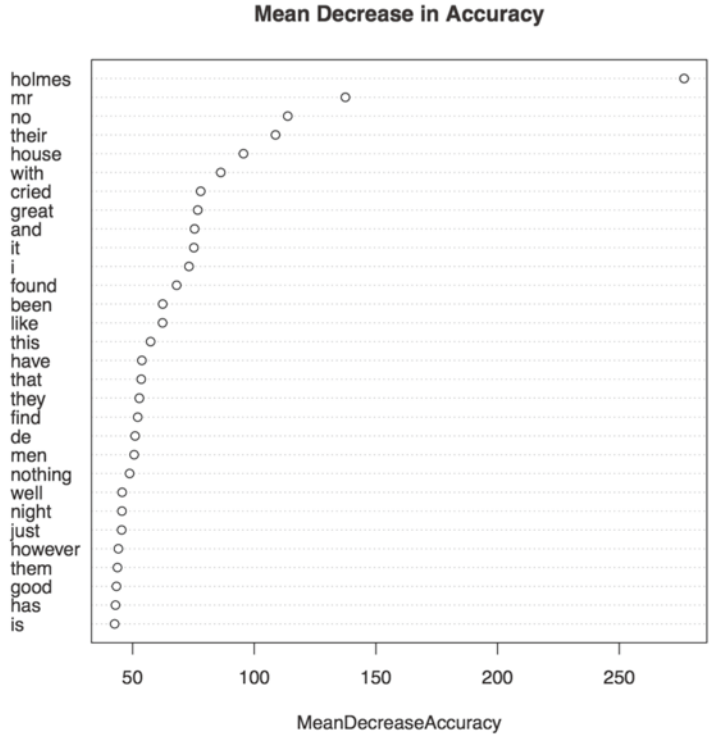

図 **1: Random Forests** の出力結果**: Mean Decrease in Accuracy** の上位項目

頻度の上位 500 語の中には，他にも sherlock や watson など Holmes シリーズの登場人物名が含まれているほか，nigel や ezra など歴史小説側の人物名も含まれていた。それにも関わらず，holmes のみがこのように顕著に高い Mean Decrease in Accuracy の値を取ってしまったのはなぜか。

同じように人物名である ezra は，HF_2 の *The Firm of Girdlestone* にしか出現しない。そのため，歴史小説全体のから見た場合の頻度は相対的に低く，ジャンル間の識別マーカーとしては機能していないと考えられる。一方で holmes は，推理小説群のほぼすべてのファイルに出現してしまっている。グループで一貫して頻度が高いため，分類のマーカーとして十分に寄与してしまうのである。シリーズ作品においてはこのように，複数の作品に共通して出現するキャラクターや用語が存在する。つまり，この問題はシリーズ作品に内在する問題であると解釈できる。

しかし，分類はシリーズ作品のキャラクター名だけで決定づけられているわけではなく，実際には他の様々な語が分類に寄与している。このことを示すため，変数とした 500 語のうち前述のシリーズ作品に関わる 3 単語，holmes, sherlock, watson を除いてもう一度同様の解析を行った。結果は以下の表 3 の通りである。

表 **3: Random Forests** の実行結果**:** 固有名詞を除外した場合

Call:			
randomForest(formula = text.group ∼ .,			
data = tbl, proximity = T, importance = T, ntree = 100000)			
Type of random forest:			classification
Number of trees:			100000
No. of variables tried at each split:			22
OOB estimate of error rate:			3.42%
Confusion matrix:			
	HF	SH	class.error
HF	97	0	0.0000000
SH	5	44	0.1020408

分類精度には若干の下降が見られたものの，それでも 96.58%の確率で正しく分類できていることがわかった。つまり，単語の出現頻度をもとにした分類は，holmes のようなシリーズに共通して登場する固有名詞は寄与するものの，それ以外の語だけでも十分に高精度な分類が可能であると判断できる。

以下，考察はこの 3 単語を除いたものに基づいて行う。

3.4. 結果と考察

図 2 はファイルの散布図を三次元空間に表したものである。Holmes シリーズと歴史小説群は概ね左右に分かれているものの，SH_1.3, SH_1.4, SH_2.4, SH_7.4, SH_7.5 の 5 ファイルが歴史小説側に配置されているのがわかる。

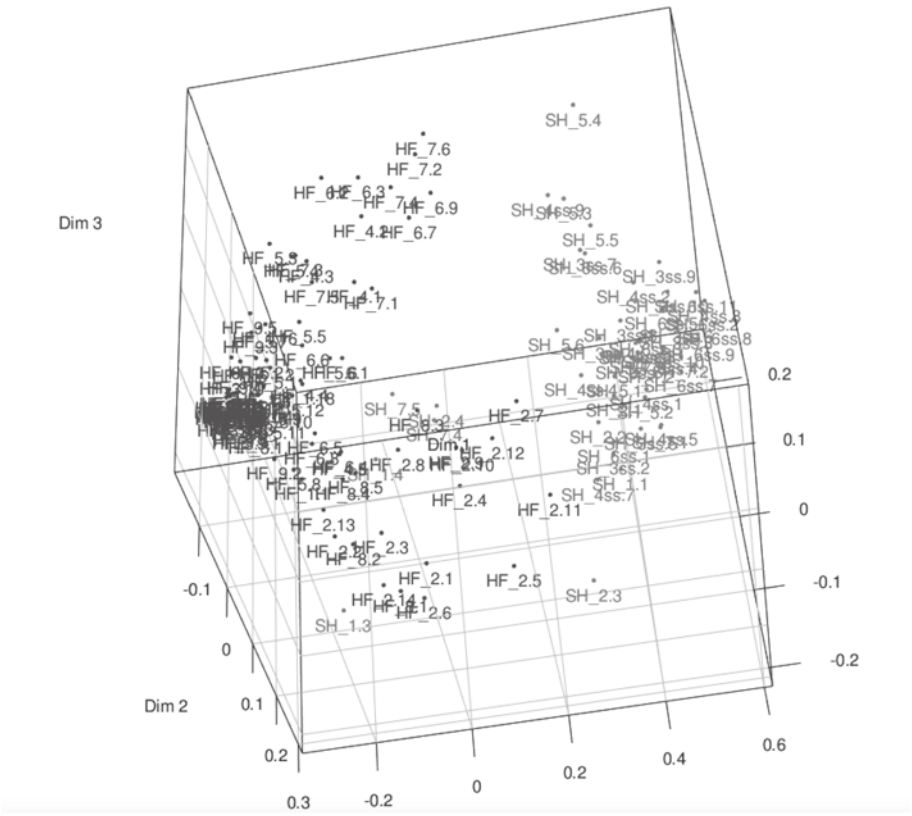

図 **2: Random Forests** の出力結果**:** テクスト分類の三次元散布図

SH_1, *A Study in Scarlet* と SH_7, *The Valley of Fear* には，どちらも二部構成であるという共通点がある。第一部では事件の発生から Holmes の推理と解決までを，第二部では第一部から時代を遡り，事件に至った背景を描写している。SH_1.3, SH_1.4, SH_7.4, SH_7.5 はそれぞれ作品の後半部分であり，この第二部が該当する。

第二部は retrospective narrative，つまり過去に起こった出来事を回顧する語りである。作品全体ではなく，この部分のみが歴史小説に近い部分に位置付けられているという事実によって，第二部が歴史小説と近い語彙生起パターンを持っていると解釈できる。*The Sign of the Four* は前述の 2 作品のように明確に二部構成になっているわけではないが，全 12 章のうちの第 12 章は，犯人の長い独白によって事件の経緯が説明されている章である。SH_2.4 では，この独白の部分が retrospective narrative となっているため，この位置に配置されているのだと考えられる。

推理小説である Holmes シリーズにも，探偵や謎解きの登場しない，歴史小説に近い箇所が部分的に存在している。この 5 ファイルの「誤分類」は，過去の出来事に言及する narrative の言語的特徴が分類に反映されている証拠でもある。

続いて，各ジャンルごとにキーワードとなっている上位項目を調べるため，それぞれのグループを key group として Mean Decrease in Accuracy の値が 20 より大きいのものを取り出した。以下の表 4 と 5 である。

表 **4:** 歴史小説群において
Mean Decrease in Accuracy の値が **20** 以上の語

Key Group: HF	
Words	MeanDecreaseAccuracy
cried	110.78
and	107.86
great	95.61
arms	90.64
de	66.63
high	64.13
fight	61.83
heart	56.07
ere	51.96
good	47.09
hard	41.95
head	41.73
faces	40.72
girdlestone	40.61
as	39.77
ezra	39.36
france	37.75
from	35.66
english	33.18
horse	32.97
eyes	30.91
hath	29.70
horses	29.69
green	27.73
father	25.36
england	23.68
blue	23.53
arm	22.12
ever	21.09
full	20.46
fair	20.13

表 **5: Holmes** シリーズにおいて
Mean Decrease in Accuracy の値が **20** 以上の語

Key Group: SH	
Words	MeanDecreaseAccuracy
case	245.41
house	134.56
certainly	126.32
i	94.27
been	84.42
found	82.11
find	66.61
has	66.43
fellow	60.58
have	56.52
chair	53.64
course	51.71
however	51.02
first	49.31
got	48.71
any	46.41
anything	42.85
evening	39.10
hall	38.45
an	37.29
hardly	36.24
door	34.47
about	34.09
he	29.86
friend	29.81
did	29.72
does	28.39
doubt	28.03
enough	26.56
dear	25.47
clear	25.18
business	24.16
ask	22.95
get	22.28
home	21.77
during	21.58
answer	21.55
here	20.48

歴史小説群のキーワードの最上位は cried という語である。この語は推理小説群も含めすべてのファイルに出現していたが，HF の方が一貫して相対的に多く出現していた。登場人物の激情を表現する際に，said や answered などに代わって用いられている。

また，HF 側のキーワードには，head, faces, arms, eyes など身体的部位に関する語が多く挙がっている。コンコーダンスラインから実際の用例を確認したところ，身体的部位として用いられる例がほとんどであったものの，多義語や，metonimical に用いられている例も多く見つかった。例えば，arms は with her arms folded のような「腕」の意味だけでなく，men-at-arms (重騎兵) や the handling of arms (軍隊の指揮) のように「武器・兵士」の意味でも使われている。また，head も He shook his head のように行動を描写するほか，the head of the

army(軍のトップ) の文脈でも使われていたり，from head to foot のように慣用句の一部として出現している場合もある。以下の図 3 と図 4 にコンコーダンスラインの一部を掲載する。

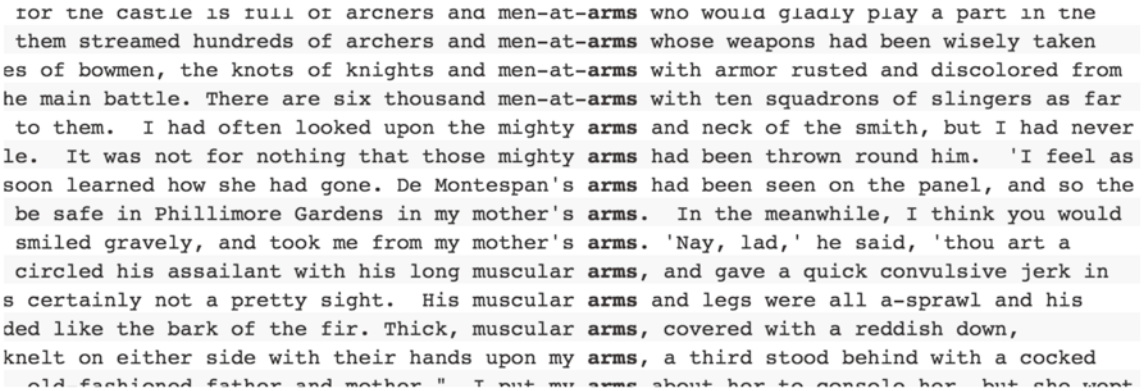

図 **3: 'arms'** のコンコーダンスライン

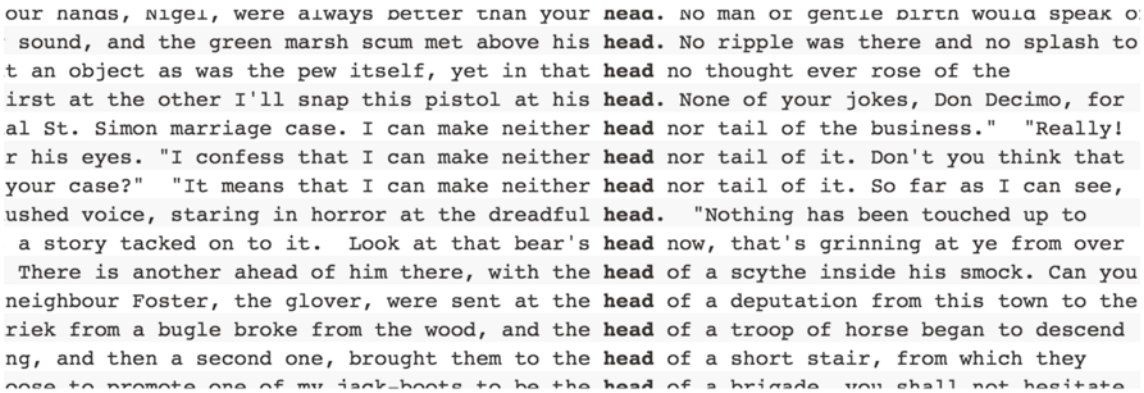

図 **4: 'head'** のコンコーダンスライン

他にも，キーワードとして father という語が挙がっているが，実際の用例を確認すると my father など一般的な用法のほか，Father Matthew や Father Lamberville のように固有名詞の一環として出てきている場合もあった。歴史小説側にこの語が多いのは，血統に関する話題が多いためと考えられる。

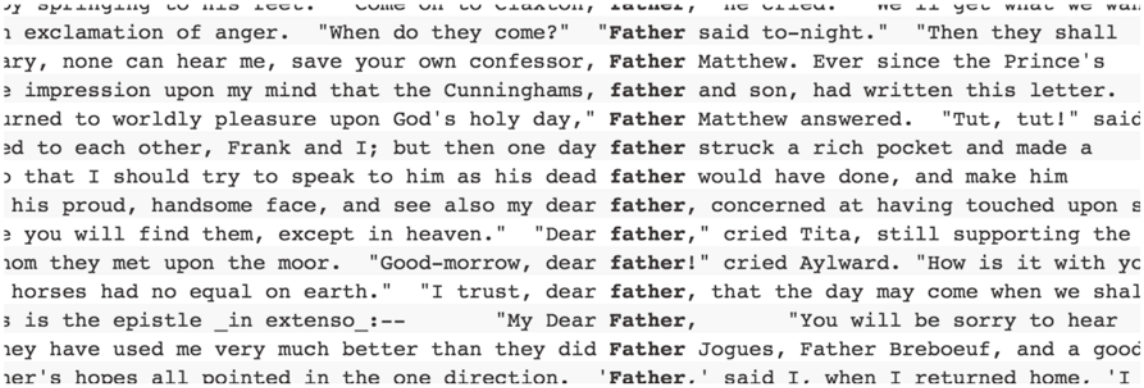

図 **5: 'father'** のコンコーダンスライン

また，キーワードの中に de というものがあるが，これは Thomas de Bray や Monsieur de Laval のようにフランス語由来の人名の中に多く出現している。この語が歴史小説群のキーワードとなったのは，Holmes シリーズが主にイギリスを舞台としており，フランス系の人物の登場頻度が非常に低いためであると考えられる。

一方 SH グループに特徴的なのは，動詞の find，found である。これらの語は推理のプロセスを描写するにあたり，何か手がかりや証拠を見つけた，という文脈で多く用いられている。find out というコロケーションも多く見られた。found がどのような文脈で用いられているのか，実際の用例を図 6 に幾つか示す。

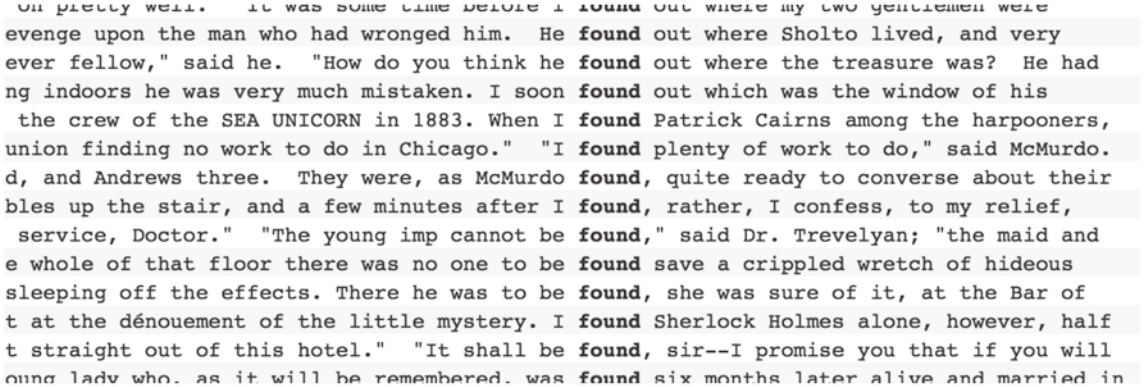

図 **6: 'found'** のコンコーダンスライン

find や has, have, does, get, ask など，現在形の動詞が多く用いられているのも特徴的である。これらの実際の用例を見てみると，会話文の中で用いられていることが多い。小説において会話ではない地の文は過去形で書かれることが多い為，現在形は会話のマーカーとしても機能することが考えられる。一人称代名詞である i が特徴語として挙がっているのも，会話文の多さを裏付ける証拠となる。

また，course という語も主に会話で用いられているが，これは 'of course' という形で用いられていることが多い為である。実際に，Holmes シリーズでは course は 282 回出現しているが，そのうち 228 回，約 80%が 'of course' である。

house や door, chair が特徴語として挙げられているのも注目に値する。歴史小説に比べ，舞台が室内に限定されやすいという傾向を表していると考えられる。

4. MALLET によるトピックモデリング

4.1. 手法

本研究では，トピックモデルとして潜在的ディリクレ配分法 (Latent Dirichlet Allocation, LDA) を採用する。Blei, Ng, Jordan (2003) によって提案されたこのモデルは，「各単語は潜在的にトピック (話題，カテゴリー) を持ち，同じトピックを持つ単語は同一文書に出現しやすい」という想定を前提とし，文書集合からトピックを確率的に算出するものである。このモデルでは，トピックは数多くの単語の集合であり，また文書は複数のトピックの集合であると考える。LDA は主に文書のクラスタリングに用いられているが，インターネット記事の自動分類やラベリングだけでなく，画像解析やユーザープロファイリングなど，幅広い分野に応用されている。

本分析においては LDA に基づくトピックモデリングを実装する MALLET (MAchine Learn-

ing for LanguagE Toolkit) を用いた。MALLET とは，機械学習に則り統計的自然言語処理，クラスタリングなどを行う Java ベースのツールキットである。

また，前章では 10,000 語ごとに区切ったファイルを分析対象としたが，本章ではさらに細かく切り分け，2,000 語ごとに区切った合計 730 ファイルを対象とした。例えば HF_1 は，HF_1.1 から HF_1.89 までの 89 ファイルに分かれている。トピックモデリングにおいては，対象ファイル数が多いほどより統計的に確かな結論が得られると考えられるためである。また，細かく分けることにより作品内のより細かなトピック変遷を観察することができる。

LDA を実行する際のトピック数はユーザーによって決定されるが，トピック数を少なく設定してしまうと，多くの文書に共通する幅広いトピックしか出力されず，文書間の差異が見つけられない。多く設定した場合，各文書に対応したより具体的トピックが数多く出力されるが，トピック同士の関係性もより複雑なものとなり観察することが難しくなってしまう。そのため対象とする文書の数や規模，分析の目的によって最適解が異なり，実験を行って数値を細かく調整する必要がある。本研究では，対象となる作品数が 16 であることも考慮し，トピック数を 10 個から 50 個まで変化させて実験を行った。10 個の場合は本文の内容に対して具体的なトピックがほとんど発見できなかったため，本論文では特に 20 個の場合と 50 個の場合を取り上げて論じる。

モデリングの結果，それぞれのトピックと文書ファイルとの関連度を数値化するデータ，分類に有用であった主要単語とその重みのデータが得られた。これらのデータをもとに，統計解析ソフトウェアの R，及びネットワークグラフの可視化を行うソフトウェアの Gephi を用いてグラフを作成した。

4.2. 結果と考察

初めに，トピックを 20 個に設定した際の結果について考察する。以下の表 6 は，0 から 19 までの各トピックがどのような語で構成されているかを表す表である。表ではトピックに対する寄与度が特に高いものから順に表記している。一部のトピックはこれらの主要単語から，トピックが何に関するものかを推測することができる。

例えば，トピック 1 は Holmes シリーズに関連するトピックであることが holmes, watson, lestrade などの固有名詞から特定できる。またこのトピックはそれ以外にも case, police, inspector, evidence など，犯罪捜査に関わる語が多く見られており，find, found, doubt などの動詞もそれを裏付けている。このことから，トピック 0 は Holmes シリーズの登場人物，及び犯罪捜査に関するトピックとわかる。

またそれ以外にも，5 は door, room, house, window, bed などで構成されており，家や家具に関するトピックであると考えられる。7 は army, fight, battle など戦争，闘争を表現する語や hundred, thousand, crowd など大多数，群衆を表す語，horse, ground, camp などの語から構成されていることから，陸地を舞台とした戦争に関するトピックと判断できる。8 は knight, lord, archer, sword, castle など，中世の騎士や貴族に関する語が多い。9 にも lord や sword のほか king や soldier など 8 と似たカテゴリーの語が含まれているが，こちらは ye, hath, nay, quoth など古い時代の英語が特徴的なトピックとなっている。16 は monsiur, france, paris など，舞台がフランスであることを示唆しており，de や du という語も，3.2 節で述べたようにフランス語由来の人名を表しているものと考えられる。17 は captain, water, ship, sea, abroad など，航海に関するトピックである。

表 **6:** トピックごとの主要単語 上位項目

トピック	キーワード
0	man, men, business, work, night, find, hand, give, read, hundred, mcmurdo, good, house, matter, money, asked, time, things, make, put, table, ready, brother, worth, pay, hands, paid, word, news, ...
1	holmes, mr, watson, case, sherlock, street, man, found, friend, paper, inspector, police, house, london, lestrade, dear, fellow, crime, chair, young, left, papers, facts, clear, interest, remarkable, morning, ...
2	sir, uncle, man, jim, cried, charles, harrison, lord, fight, face, ring, wilson, ll, young, time, good, london, heard, round, boy, nephew, crowd, belcher, stone, men, prince, mother, avon, champion, lothian, ...
3	ezra, major, young, girdlestone, good, tom, john, kate, son, girl, money, merchant, time, answered, firm, father, business, great, head, office, mind, continued, hope, mr, pounds, thousand, dimsdale, ...
4	king, sire, eyes, brother, court, face, father, king's, hand, church, madame, young, woman, heart, anger, holy, soul, voice, lady, abbot, world, service, turned, power, love, hold, spirit, order, read, noble, ...
5	door, room, house, light, open, night, window, hand, face, round, heard, side, opened, voice, floor, front, fire, suddenly, back, hands, gave, stood, turned, half, steps, bed, corner, sound, instant, ...
6	long, man, road, left, found, lay, white, side, end, body, blood, dead, centre, stone, led, path, small, passed, showed, foot, ran, brought, chance, set, track, feet, show, narrow, water, stranger, reach, ...
7	men, horse, horses, great, side, english, line, hundred, front, army, left, fight, round, rode, thousand, battle, leader, time, morning, crowd, ride, ground, riding, blue, lines, saddle, gallant, camp, field, ...
8	sir, nigel, alleyne, fair, john, cried, nay, aylward, knight, lord, squire, lady, ere, archers, good, prince, hath, castle, chandos, hand, master, saint, pray, young, honor, simon, man, archer, great, sword, ...
9	ye, saxon, hath, answered, good, cried, lord, nay, reuben, sword, monmouth, master, town, scarce, king, gervas, make, sir, friend, soldier, time, set, hand, friends, faith, grey, find, sergeant, quoth, ...
10	cried, head, lay, back, instant, forward, moment, fell, feet, air, stood, looked, round, great, arms, front, struck, slowly, high, turned, eyes, beneath, wild, lost, cry, yellow, blow, sprang, deep, hold, ground, ...
11	small, time, made, asked, answered, remarked, round, companion, appearance, large, make, part, thought, put, knowledge, day, question, place, pair, air, observed, times, sat, thing, sort, fairly, ...
12	great, country, father, good, years, days, day, england, mother, life, hard, village, heard, high, year, set, folk, world, black, glad, part, head, call, things, red, land, rest, coming, talk, full, age, town, sight, ...
13	sir, letter, man, night, moor, matter, told, days, henry, mind, heard, hall, family, letters, day, death, dr, position, doubt, case, london, people, answer, friend, point, dear, end, charles, understand, fear, ...
14	colonel, long, miss, brown, belmont, cried, women, black, prisoners, death, men, desert, white, looked, cochrane, time, mr, hand, sadie, adams, stephens, hope, party, began, figures, frenchman, line, good, ...
15	room, morning, matter, time, knew, made, lady, mrs, heard, find, wife, evening, miss, returned, years, occurred, doubt, gentleman, moment, life, hour, late, surprised, st, husband, character, impossible, ...
16	de, catinat, monsieur, amos, men, emperor, river, france, green, ah, instant, du, side, woods, women, paris, good, lhut, great, french, fire, st, ephraim, young, laval, people, la, white, adele, hat, captain, ...
17	captain, water, ship, sea, boat, long, deck, side, wind, cried, lay, ships, great, seaman, coast, black, mate, round, aboard, vessel, half, put, hour, night, ten, sail, line, channel, waves, land, men, seamen, ...
18	man, eyes, head, back, face, long, hands, held, strange, hear, passed, fear, sudden, god, raised, spoke, eye, words, leave, heavy, quick, stood, made, shoulders, dark, save, hope, carried, life, mouth, ...
19	thought, man, cried, face, eyes, time, back, heart, asked, looked, word, mind, make, hand, made, turned, thing, woman, friend, good, life, give, put, knew, ah, love, lips, mine, hard, things, found, ...

前述の通り，トピック 8, 9 における lord や sword など，複数のトピックにまたがって出現している語も幾つか存在する。このように共通する単語によってトピック同士も関係性を持っていると言える。そこで，トピックと単語との関係性を可視化するネットワーク図を作成した。図 7 は，20 個のトピックとそれらと関連性の強い単語を上位 1,000 語まで配置したものである。

数字のノードは 0 から 19 までのトピックを表しており，単語のノードと結びついているエッジの太さはそれぞれの語の重みと対応している。例えば，トピック 5 において最も重みの大きな語，つまり一番太いエッジで 5 のノードと繋がっている語は door である。その右にある suddenly や window はトピック 19 とも細いエッジで繋がっており，トピック 5 と 19 の共通項目であることが図から分かる。

図 **7:** トピックとそれを構成する単語の関係性を表すネットワーク図 **(**トピック **0** から **19)**

トピック 8, 9 のように，共通する項目の多いトピック同士は近くに配置され，多くのノードを介して繋がっている。また，トピック 0 や 4，13 など比較的中心部に位置したトピックは，様々なトピックと共通要素を持っており，man や sir, king など多くの作品に共通して出現する語で構成されている。一方で外縁部に位置する単語のノードは，一つのトピックにしか出現しない語が多い。例えば，下部にあるトピック 1 において最も重みの大きな語である holmes は，繋がっているトピックノードが 1 のみであるため，図全体でも外縁部にある。しかし重みが二番目に大きい mr はトピック 3 とも繋がっているので，比較的内側に配置されている。

次に，各文書がどのようなトピックで構成されているかを分析する。前述の通り LDA においては，トピックと文書は必ずしも一対一関係ではなく，一つの文書に複数のトピックが潜在的に存在していると仮定している。MALLET により，各文書において 0 から 19 までのトピックがそれぞれどれくらいの割合で存在しているかを表すデータが出力されている。図 8 はそのデータをもとに，トピックとテクストファイルの関係を図示したものである。各ノードはトピックとファイル，エッジの太さは生起確率の高さと比例している。

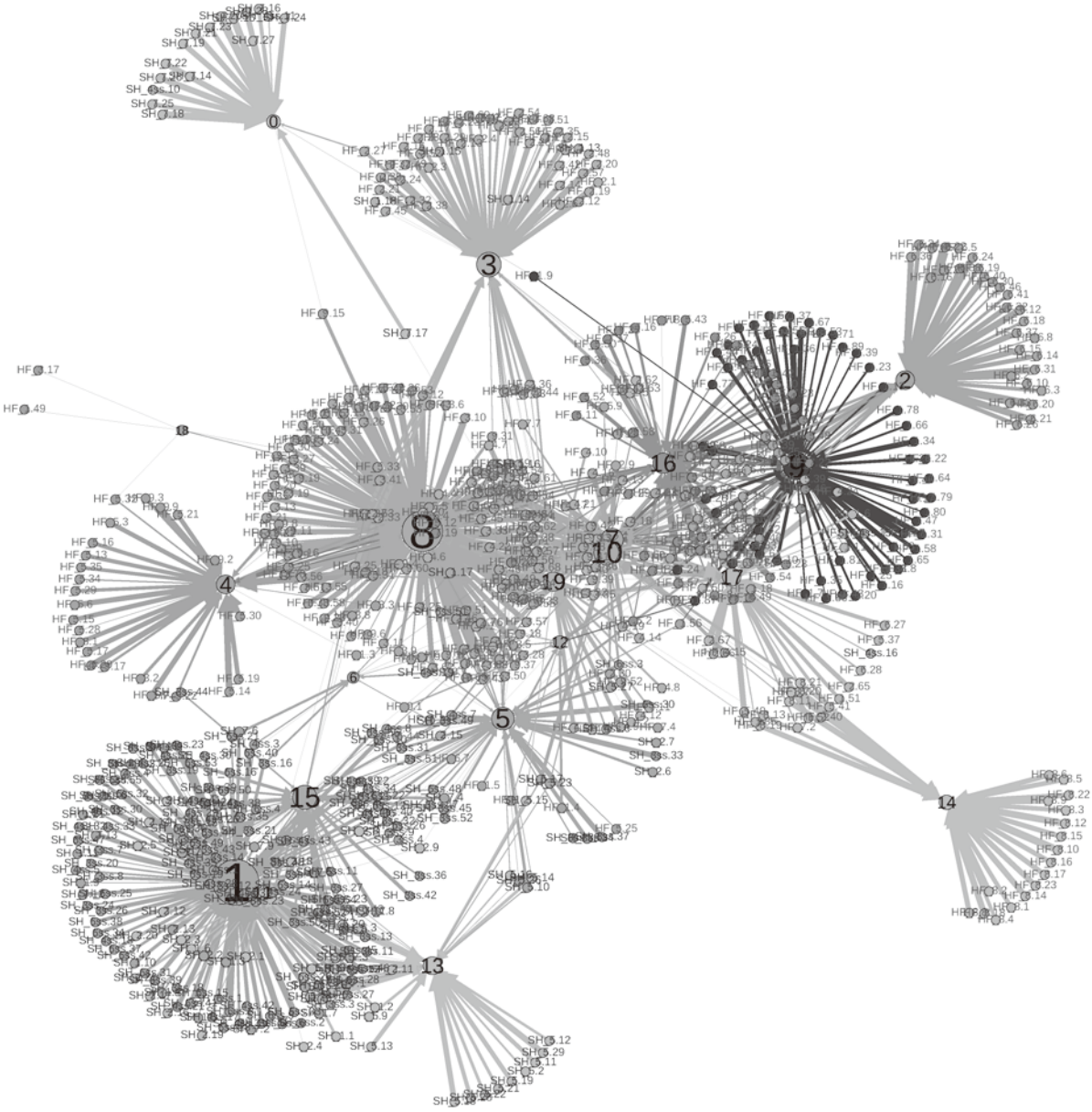

図 **8:** トピックとテクストファイルの関係を表すネットワーク図 **(**トピック **0** から **19)**

図の左下と左上を見ると，トピック 0, 1, 5, 13, 15 は主に Holmes シリーズに頻出する話題であると判断できる。表 6 から考察したように，トピック 1 は犯罪捜査に関係するトピックであり，実際に推理小説群の多くがこのトピックを有していることから，このトピックが Holmes シリーズを特徴づけるトピックであることは明らかである。また特徴的であるのは，先ほど家・家具に関するトピックだと判断したトピック 5 が Holmes シリーズの方に多く出現しているという点である。一方で歴史小説群にはこのトピックはほとんど出現していない。このことから，これら二つの話題は，歴史小説と比較した際 Holmes シリーズを特徴づけるトピックだと言える。図 9，図 10 は，それぞれトピック 1 と 5 の生起確率をジャンル別に分けてプロットしたものであるが，このボックスプロットもまた，各トピックが推理小説側に平均して多く出現していることを裏付けている。

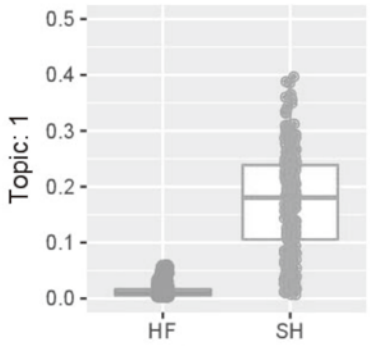

図 **9:** トピック **1** の生起確率を表すプロット

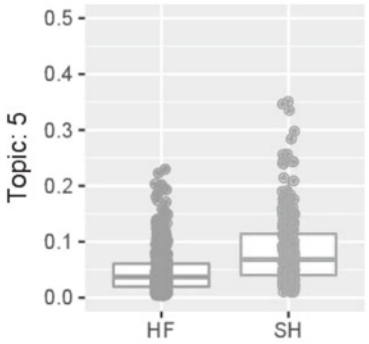

図 **10:** トピック **5** の生起確率を表すプロット

その他多くのトピックは，歴史小説群の方に頻出する話題である。これらにおいて特徴的であるのは，特に外縁部に位置しているものに関して，トピックと各作品が対応関係になっているものが多いことである。例として，トピック 2 と HF_6，3 と HF_2，4 と HF_5，14 と HF_8 などが挙げられる。一方で前述の推理小説群に関しては，トピック 0 と SH_7，13 と SH_5 はほとんど一対一関係であると言えるが，1, 5, 15 は作品を問わず SH_1 から SH_7 までの様々なテクストファイルで出現している。この差は，Holmes シリーズが基本的に共通した登場人物，時代背景，舞台の下で展開される物語であるのに対し，歴史小説群は多様な時代背景のもと様々な出来事を描写していることが原因だと考えられる。

以上の結果がトピック数を 20 個に設定した場合であるが，続いてトピック数を 50 にした場合において作成した同様のグラフについて考察する。

まず，先ほどと同様にトピックとそれを構成する語彙との関係を示すネットワーク図を作成したが，項目数が増えたため見づらいグラフとなってしまった (図 11)。ノード数を増やし過ぎればより複雑になりすぎてしまうため，一トピックあたりの語数は制限されてしまう。

図 **11:** トピックとそれを構成する単語の関係性を表すネットワーク図 **(**トピック **0** から **49)**

一方で，トピックを細かくすることへのメリットもある。次の図 12 は，トピックと各文書ファイルの関係性を示すものであるが，トピックが各ファイルに対してより具体的になるため，20 個の場合では発見できなかった特徴が明らかとなっている。

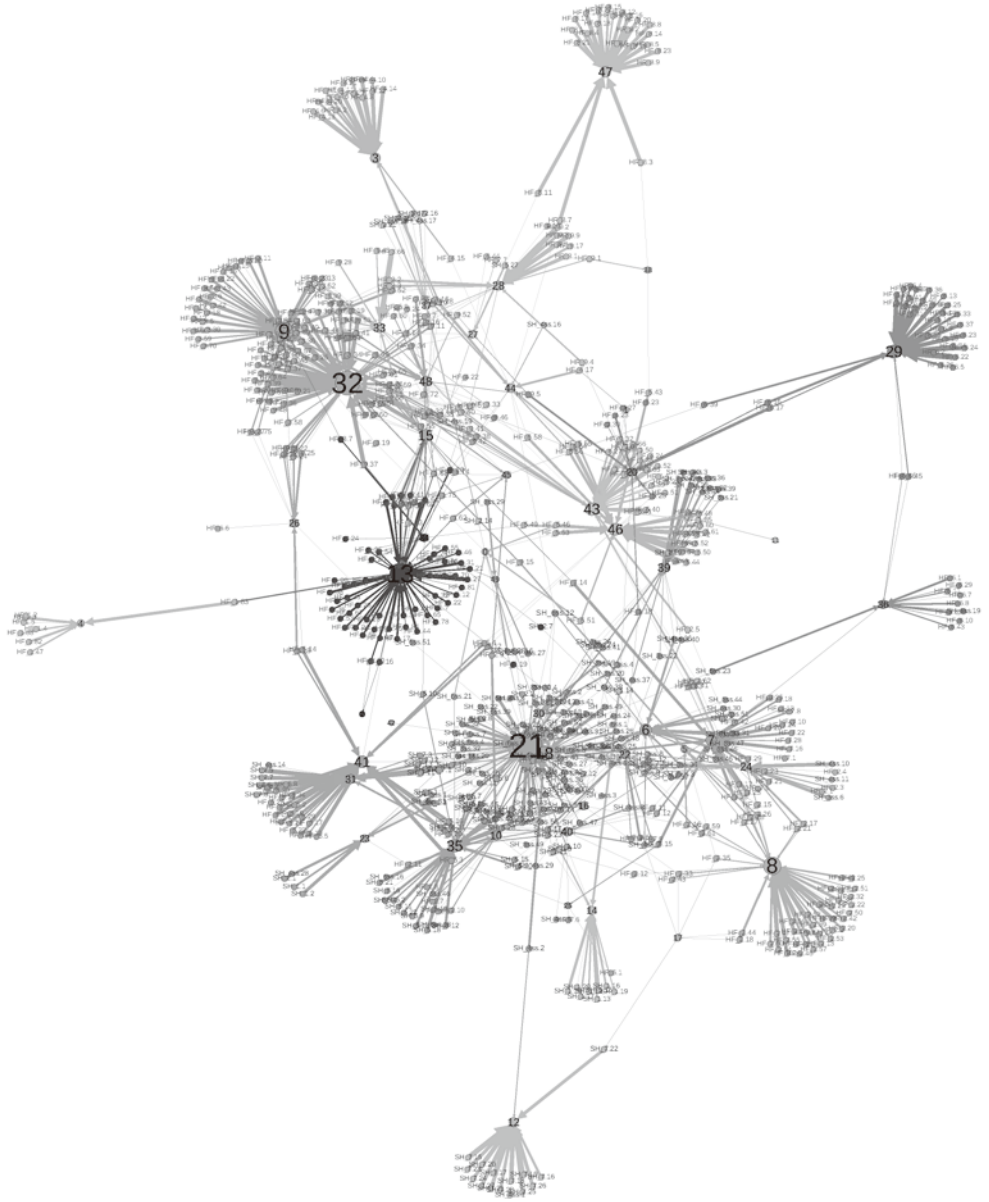

図 **12:** トピックとテクストファイルの関係を表すネットワーク図 (トピック **0** から **49**)

例えば，先ほど推理小説に特徴的なトピックは作品を問わず様々なファイルで出現していたが，トピックが 50 個の場合は 12, 14, 31, 35, 37 など，作品と対応関係にあるトピックが数多く見られる。特に，14 は SH_1 の中でも 1.11 から 1.21 まで，つまり *A Study in Scarlet* の後半に特徴的なトピックとなっている。第 3 章でも論じた通り，この作品は前半と後半とで大きく言語的特徴が異なっている。トピック 14 を構成する語を精査することで，後半部の言語的特徴を見ることができる。図 13 を見ると，girl や daughter など少女を表す語が多いほか，asked と answerd がセットで出現しているのも特徴的である。

図 **13:** トピック **14** のワードクラウド

他にも考えられるメリットは，トピックの中身が細分化されることにより，共起関係が分かりやすくなる点である。20 個に分けた場合のトピック 1 は，Holmes シリーズの登場人物と事件捜査に関する語が両方とも出現していたトピックであった。しかし，50 個に分けた場合を見てみると，Holmes シリーズの固有名詞が出現しているのはトピック 21 だが，case, found, inspector など捜査に関する語が含まれているのはトピック 18 である (図 14，図 15)。つまり，トピック 1 の内容が 18 と 21 にうまく分かれた形となり，Holmes の登場人物名とよく共起する語や，犯罪捜査の話題において使われる語をそれぞれ分析することが可能となる。

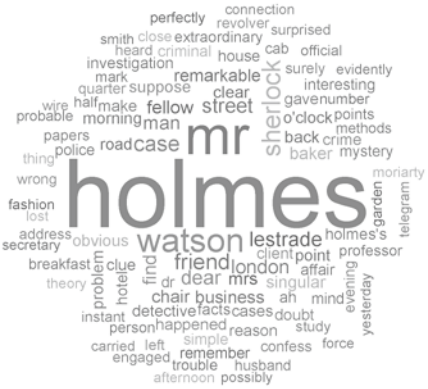

図 **14:** トピック **21** のワードクラウド

図 **15:** トピック **18** のワードクラウド

以上のように，トピックを構成する語とそれらの関係性，及び文書とトピックの関係性を考察することで，文書ごとの特徴を明らかにすることができる。また，少ないトピック数は全体の関係性を包括的に把握した分析，多いトピック数は個々のトピックの特性に特化した分析にそれぞれ有用であることがわかった。値を変化させて実験を行うことにより，それぞ

れ別の観点から各ジャンルの特徴を明らかにすることができる。

4.3. 固有名詞の問題

本研究において行った分析では固有名詞に対して処理を行っておらず，そのため前述の通り，結果に人名や地名が含まれてしまっている。3.1 節や 3.3 節でも述べたように，特に小説テクストを分析対象とする際，固有名詞の取り扱いに関しては多くの問題が考えられる。

たとえば，King James や Mr. Watson，Sir Nigel などの場合，king や mr，sir なども固有名詞の一部と捉えるべきだろうか。4.2 節で行った考察において，captain という単語は海に関するトピックのマーカーとして有効であったため，Captain Ephraim Savage という語全体を固有名詞としてしまうのは早計に感じられる。しかし一方で，敬称部分を残して Captain <NNP>や Mr. <NNP>のような形にしてしまうと，captain や mr の出現頻度が相対的に高くなってしまう。また，地名をどの程度まで取り除くかも判断が分かれる。例えば London や France などは，固有名詞ではあるものの，トピックとしても判断材料になり得る。

人名を除去し，地名を残す，といった選択をする場合でも，新たな問題が生じる。特定の品詞をのみを抽出する際には，品詞タグ付け (tagging) の情報が参考にされる。Jockers (2013) では，分析対象を絞るため，Standford POS tagger を用いて名詞のみを抽出した。しかし，固有名詞を抽出する場合，その精度が普通名詞やその他の品詞に比べて大きく下がってしまう。人名や地名，商品名などの固有名詞は無数に存在し，多くの場合辞書に載っておらず，大半が未知語と判断されるためである。地名か人名かの判断を行うのも，現段階では困難な問題とされている。

自然言語処理の分野では，この問題は情報抽出の共有タスクの一つとされ，固有表現抽出と呼ばれる技術が研究されている。現在，隠れマルコフモデル (Hidden Markov Model) や CRF (Conditional Random Fields) などをベースにした固有名詞認識が検討されている (坪井，鹿島，工藤，2006)。しかし，いずれも多くの外部辞書や学習データを必要とする複雑な手法である。

5. おわりに

本研究では Arthur Conan Doyle の 16 作品を対象に，Random Forests や MALLET を用いた分析を行い，Holmes シリーズの作品群と歴史小説の作品群との比較を行った。まず，Random Forests による自動分類とそれに寄与した単語の抽出を行い，各グループの特徴語をリストアップし考察した。次に MALLET を用いた LDA に基づくトピックモデリングでは，作品内に潜在するトピックを単語の出現分布をもとに算出し，各作品とのトピックの関係性，及びトピックを構成する単語との関係性を観察し，ジャンル間の差異を検討した。Doyle の作品研究は，これまで主に質的な観点から行われており，また Holmes シリーズにのみ注目されている傾向にあったが，複数ジャンルを計量的に分析することにより，幾つかの新たな事実を明らかにした。

今後の課題として，他の推理小説作家や歴史小説作家の作品と比較を行った上で，ジャンル間の違いと作家固有の言語的特徴を明らかにすることが挙げられる。本研究では，Holmes シリーズの特徴，歴史小説それぞれの特徴は得られたため，他の作家でもこの言語的特徴が一般化できるかどうかを検討したい。

また，固有名詞に関する問題も今後の課題となる。固有名詞を加えたまま分類するのが適切なのかどうか，また取り除いて分析する場合にはどこまでを固有名詞と判断すべきなの

か，タグ付けが信用に足る精度で働いているのかどうかなど検討する項目は多く，今後稿を改めて論じる必要がある。

文　献

[1] 揚石 亮平・三浦 孝夫. 2008.「固有名詞の認識を含む HMM による英文形態素解析」『電子情報通信学会 第 19 回データ工学ワークショップ論文集』E7–6.

[2] Blei, M., Ng, A. and Jordan, M. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993–1022

[3] Breiman, L. 2001. Random forests. *Machine Learning* 45: 5–23.

[4] Breiman, L. and Cutler, A. *Random Forests*. Available at: http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm [Accessed 5 March 2017]

[5] Burrows, J. 1987. *Computation into Criticism: A Study of Jane Austen's Novels*. Oxford: Oxford University Press.

[6] Ji, H. and Grishman, R. 2006. Analysis and Repair of Name Tagger Errors. *Proceedings of the COLING/ACL* 420–427.

[7] Jockers, M. and Mimno, D. 2013. Significant themes in 19th-century literature. *Poetics* 41: 750–769

[8] 河村 幹夫. 1991.『コナン・ドイル–ホームズ・SF・心霊主義』東京: 講談社.

[9] 小林 雄一郎・田中 省作・冨浦 洋一. 2011.「ランダムフォレストを用いた英語科学論文の分類と評価」『情報処理学会研究報告 IPSJ SIG Technical Report』2011-CH-90: 53–68.

[10] Lamond, J. 1931. *Arthur Conan Doyle: A Memoir*. London: John Murray.

[11] 大賀 信孝. 1988.「Conan Doyle の文体的特徴について」『九州産業大学教養部紀要』25(1): 61–70.

[12] Pearson, H. 1943. *Conan Doyle, His Life and Art*. London: Methuen & Co., Ltd.

[13] Tabata, T. 2015. Stylometry of Dickens's Language: An Experiment with Random Forests, in P. L Arthur and K. Bode (eds.) *Advancing Digital Humanities: Research, Methods, Theories*. Basingstoke, Hampshire: Palgrave Macmillan, 28–53.

[14] 高見 敏子. 2003.「『高級紙語』と『大衆紙語』の corpus-driven な特定法」『大学院国際広報メディア研究科・言語文化部紀要』No.44. 73–105.

[15] 高見 敏子. 2004.「特徴語の特定法 –英・米・豪の新聞英語における語彙比較への応用–」『The Northern Review』No.32. 31-66.

[16] 坪井 祐太・鹿島 久嗣・工藤 拓. 2006.「言語処理における識別モデルの発展 – HMM から CRF まで」言語処理学会第 12 回年次大会 (NLP2006) チュートリアル.