トピックモデルによるアガサ・クリスティ作品の計量的分析

土村 成美（大阪大学大学院　言語文化研究科）

本研究では，アガサ・クリスティのミステリー作品に対して，潜在的ディリクレ配分法(Latent Dirichlet allocation)を使用したトピックモデルを用いて分析を行うことを目的とする．比較対象として，同時代の女性作家であるドロシー・セイヤーズの作品を使用した．自然言語処理ツールキット MALLET を使用してトピックモデルを行い，その結果を用いてトピックとトピックを構成する語との関係性，及びトピックとトピックを構成するテクストとの関係性のネットワークグラフによる可視化を行なった．この分析を通してクリスティ作品における特徴的なトピック及び語彙を探ることを試みる．

Topic Modeling of Agatha Christie's Works

Narumi Tsuchimura (Graduate School of Language and Culture, Osaka University)

This study applies topic modeling based on Latent Dirichlet allocation (LDA) to the detective works by Agatha Christie. The works by Dorothy Sayers, a female mystery writer contemporary with Christie, are used in this study to compare with Christie's works. MALLET, a Java-based package for statistical natural language processing, is used to build a topic model, and we visualized relationships between topics and words that compose topics, and those between topics and texts. We attempt to investigate characteristic topics and stylistic features in Christie's works.

１．はじめに

本研究では，イギリスのミステリー作家アガサ・クリスティ作品に関して，機械学習の一種であるトピックモデルを用いたアプローチを通して分析を行うことを目的とする．比較対象として，クリスティと同時代に活躍し，1930 年代には彼女と小説の初版部数を競っていた[1], 同じくイギリスの女性ミステリー作家のドロシー・セイヤーズの作品を使用する．日本におけるセイヤーズの知名度はあまり高くないものの，累計出版部数が作家としては世界一のクリスティよりも，セイヤーズを理想の作家として挙げる女性作家も多く(森(1998)[2]), 「ミステリーの 2 大女王」として並んで語られることが多い．そのような 2 人の比較を通して，クリスティの作品の特徴を探ることが本研究の目的である．

クリスティ作品の計量的な分析は，語彙多様性や曖昧語の使用に焦点を当てた Lancashire & Hirst (2009) [3]や，それに加えて統語面にも焦点を当て，アルツハイマー病を患っていた Iris Murdoch と健康に加齢した P. D. James を比較対象として，アルツハイマー病を患っていた可能性があるクリスティの言語能力の低下に着目した Le et al. (2011) [4], スモールコーパスから抽出された特徴語をもとに考察を行なった稲木(2013) [5]などが存在する．しかし使用作品が限定的であり，またその選定理由が述べられていないものが多く，作品選定の恣意性が払拭しきれない．また，分析手法として機械学習の手法は殆ど用いられていないのが現状である．

以上を踏まえて本研究では，クリスティのトピックモデルを用いた文体的特徴の分析に向けた第一段階の研究として，無作為抽出したテクストを用いてトピックモデルを行い，セイヤーズ作品と比較したクリスティ作品の文体の特徴を探ることを試みる．

２．使用データ

本研究で使用したデータは，クリスティの作品 20 作品(1,269,461 語)と，セイヤーズの作品 10 作品(977,001 語)である．データは筆者作成のものであり，本文のみを分析対象としている．両者の作品共に長編作品全作品から無作為に選定した．両作家共に複数ジャンルでの作品執筆を行っているが，ジャンルによる文体の差異を最小限にするため，ミステリー作品のみを分析対象とした．以下の表 1，2 に分析に使用した作品の発表年・本論文内での作品ラベル・総語数を示す．

Jockers & Mimno (2013)[6]において，トピックモデルは語の共起関係からモデル生成を行うため，文書が長過ぎるとそれだけ文書内で共起する語が多くなってしまい，生成されるトピックが漠然としたものになることが指摘されている．そのため本研究ではトピックモデルを行うにあたり，各作品のテクストファイルを作品の先頭から 2000 語単位に分割し，ファイルごとの語数を揃

えた．各作品の最後の 2000 語に満たないファイルは分析対象から除外した．ファイルの分割には統計解析環境 R Ver. 3.4.0 を使用した．この処理を経たクリスティ作品 625 ファイル，セイヤーズ作品 484 ファイルの合計 1,109 ファイルに対して分析を行う．

表 1 クリスティの分析対象作品一覧
Table 1　Christie's works used in this study.

作品名	出版年	コード	総語数
The Mysterious Affair at Styles	1920	c1	56,938
The Secret Adversary	1922	c2	75,647
The Murder on the Links	1923	c3	59,751
The Man in the Brown Suit	1924	c4	75,558
The Murder of Roger Ackroyd	1926	c5	70,086
The Mystery of the Blue Train	1928	c6	70,365
Three-Act Tragedy	1935	c7	57,430
Death in the Clouds	1935	c8	60,208
The ABC Murders	1936	c9	59,043
Dumb Witness	1937	c10	74,817
Sad Cypress	1940	c11	55,594
The Moving Finger	1942	c12	56,669
They Do It with Mirrors	1952	c13	51,364
Destination Unknown	1954	c14	60,878
Hickory, Dickory, Dock	1955	c15	56,968
4.50 from Paddington	1957	c16	66,063
The Pale Horse	1961	c17	63,099
By the Pricking of My Thumbs	1968	c18	70,619
Passenger to Frankfurt	1970	c19	69,317
Sleeping Murder	1976[a]	c20	59,047

表 2 セイヤーズの分析対象作品一覧
Table 2　Sayers' works used in this study.

作品名	出版年	コード	総語数
Whose Body?	1923	s1	59,363
Clouds of Witness	1926	s2	82,203
Unnatural Death	1927	s3	81,146
The Unpleasantness at the Bellona Club	1928	s4	72,201
Strong Poison	1930	s5	77,691
The Five Red Herrings	1931	s6	111,728
Murder Must Advertise	1933	s7	108,599
The Nine Tailors	1934	s8	107,924
Gaudy Night	1935	s9	159,577
Busman's Honeymoon	1937	s10	116,569

3．分析手法

本研究ではトピックモデルを用いた分析を行う．トピックモデルとは，確率論的アルゴリズムに基づいてトピック(話題，テーマ)を抽出する手法である．本研究では Blei et al. (2003)[7]によって提唱された潜在的ディリクレ配分法 (Latent Dirichlet allocation: LDA)を用いたトピックモデルを実行する．潜在的ディリクレ配分法は，文書は複数のトピックから構成されるということを前提としたモデルである．単語は潜在的にトピックを持ち，同じトピックを持つ語同士は共起しやすい，という考えに基づき，単語のトピックを確率的に求める手法である．

潜在的ディリクレ配分法によるトピックモデルを行うにあたり，本研究ではマサチューセッツ大学で開発された JAVA ベースの自然言語処理ツールキット MALLET(Machine Learning for LanguagE Toolkit)[b] バージョン 2.0.7 を使用した．

MALLET を使用した潜在的ディリクレ配分法の実行においては，あらかじめ用意されているストップワードリストの読み込みを行うことで，そのリストに含まれる語を分析対象から除外することが可能である．そのリストに加え，分析者自身がストップワードを設定することも可能である．今回筆者がストップワードを自身で設定せずにトピックモデルを行なったところ，作中に登場する人名がトピックの抽出に影響を与えることが確認されたため，本研究では既存のストップワードリストに加え，人名をストップワードに設定して分析を行う．

MALLET を用いた潜在的ディリクレ配分法により抽出するトピック数は分析者が設定するが，全ての分析において最適なトピック数は存在しない．設定されたトピックの数が少な過ぎると，多くの文書にまたがる大まかなトピックしか抽出されず，文書ごとの差異を観察することが困難になる．逆にトピック数を多く設定し過ぎると，個々のトピックが細分化され過ぎてしまい，こちらも解釈が複雑で困難となる．本研究では抽出するトピック数を 20 から 50 まで変化させて試行錯誤を試みた．以下ではトピックを 50 に設定した場合の分析結果を提示する．

[a])*Sleeping Murder* は 1940 年代初頭に完成しており，クリスティの死後に出版される予定であったが，1975 年頃には彼女が新作執筆を行うことの出来ない健康状態であったため，出版の許可が下りた．そのため執筆時期と出版時期とが大きく乖離している．

[b])http://mallet.cs.umass.edu/index.php

4．分析結果と考察

潜在的ディリクレ配分法により抽出されたトピックと，それぞれのトピックを構成する語(語の重みが 120 以上のもの)の関係性をネットワークグラフで図示したものが図 1 である．ネットワークグラフの描写にはソフトウェア Gephi[c]を使用した．また，それぞれのトピックを構成する語で，各トピックに対する重み(寄与度)の高い語を表 3 に示した．

ネットワークグラフ内の数字はトピックを表しており，単語と結びついているエッジの太さは，構成するトピックに対する寄与度と対応している． 単語も複数のトピックを持つことも多々あり，単語とトピックとが一対一の関係にあるとは限らない．この図，また表 3 で示された各トピックを構成する主要な語から，それぞれのトピックが何に関するものかを推測することも可能であり，例えばトピック 0 は *room, window, upstairs* などから室内に関するトピック， 42 は *murder, murderer, killed* などから殺人に関するトピックであると考えることが出来る．トピック 0 と 42 を構成する語をワードクラウドで視覚化したものが図 2, 3 である．

表 3 各トピックを構成する主な語
Table 3　List of words that compose topics.

トピック	キーワード
0	door, room, window, bed, back, house, open, opened, looked, thought, light, night, heard, key, stood, left
1	mr, whittaker, lady, good, story, suppose, business, solicitor, family, made, town, act, property, woman
2	church, rector, bells, st, superintendent, bell, mr, great, good, grave, rope, ah, mrs, years, dear, fenchurch
3	miss, dear, i'm, people, it's, afraid, person, young, feel, asked, poor, tea, hope, great, i've, showed, booth
4	young, girl, time, guess, replied, chance, minute, long, boy, hand, heard, suddenly, plan, men, continued
5	dr, family, dog, fact, died, doctor, aunt, ball, left, house, stairs, money, husband, fool, lawyer, woman, manner
6	things, world, people, young, youth, men, power, great, country, thought, money, interesting, music, looked
7	man, people, time, plane, world, end, place, free, dr, american, work, understand, part, cold, french, word
8	seat, norman, woman, madame, lady, blowpipe, business, case, paris, english, shook, murder, noticed
9	sir, mr, man, police, asked, butler, strange, fellow, gentleman, lady, death, put, gentlemen, glass, er, party
10	mrs, mr, friend, asked, room, coffee, village, remember, strychnine, af, dr, cried, husband
11	mr, ain't, that's, mrs, lordship, vicar, put, superintendent, lady, pot, good, gentleman, morning
12	mrs, house, it's, people, don't, woman, time, things, picture, kind, aunt, i'm, years, remember, husband
13	man, mr, inspector, time, girl, people, crime, murder, crome, murderer, police, murders, young, minute
14	voice, face, hand, round, eyes, turned, suddenly, sat, moment, back, hands, head, words, passed, arm, stood
15	man, horse, pale, mark, woman, called, cabin, asked, told, diamonds, secretary, poppy, night, didn't, africa
16	it's, that's, he's, don't, i'm, there's, man, i've, can't, you're, we've, you've, isn't, we'll, we're, won't
17	house, dr, wife, mrs, remember, time, father, years, didn't, it's home, thought, reed, garden, asked
18	train, station, time, minutes, mind, London, hotel, day, looked, hat, town, left, passed, ticket, line, started
19	nurse, mrs, aunt, dr, girl, time, death, patient, made, thought, morphine, make, died, head, left
20	didn't, wasn't, couldn't, told, back, thought, he'd, i'd, time, made, knew, that's, thing, wouldn't bit, hadn't
21	monsieur, madame, mademoiselle, man, ah, comte, paris, train, magistrate, woman, told, moment, de, eh
22	don't, it's, i'm, that's, i've, she's, people, you're, things, isn't, can't, i'd, girl, dear, doesn't, you've
23	hand, paper, pocket, find, small, found, left, made, back, place, carefully, picked, piece, dropped, papers
24	mrs, inspector, police things, girl, young, ring, ink, students, road, room, mr, lane
25	inspector, general, man, asked, major, things, money, ten, question, club, minute, girl, call, pounds
26	man, body, found, make, police, dead, doctor, place, case, doubt, find, chief, put, men, yard, suppose, end
27	morning, time, time, night, telephone, evening, o'clock, give, home, rang, heard, round, yesterday
28	hair, long, eyes, black, kind, small, thought, tall, face, dark, made, sat, side, touch, white, short, large
29	mr, woman, family, good, arsenic, time, inspector, bacon, tea, told, boys, made, hall, asked, making
30	de, french, cat, mon, la, war, English, le, pas, brought, c'est, en, je, mais, bon, bien, rich, suppose, husband
31	head, water, coming, lay, heavy, dead, feet, broken, struck, god, light, full, arms, caught, hands, ran
32	car, back, house, road, turned, place, side, gate, drive, half, country, path, long, close, wall, big, past, garden
33	tae, bicycle, car, inspector, ye, sergeant, constable, ay, wi, gatehouse, morning, road, night, glasgow, tuesday
34	inspector, curry, mrs, christian, mr, shot, looked, dr, hall, study, back, great, door, idea
35	evidence, heard, coroner, witness, deceased, court, called, jury, case, prisoner, put, murder, judge, impey
36	dean, college, de, vine, warden, quad, work, room, oxford, library, students, time, common, term, kind
37	made, mind, felt, life, great, love, feeling, day, feel, knew, world, end, thing, find, set, beginning, moment
38	letter, letters, read, write, written, writing, paper, book, wrote, day, address, good, words, received, morning
39	man, street, business, taxi, day, square, shop, flat, office, corner, morning, paid, moment, information
40	mr, man, night, duke, duchess, denver, lady, loadship, brother, mother, shot, morning, bath, till, body
41	mr, dean, office, armstrong, room, copy, desk, staircase, advertising, victor, headline, replied
42	murder, case, crime, thing, idea, fact, truth, friend, point, story, murderer, person, knew, killed, motive
43	things, thought, put, don't, coming, kind, thing, back, remember, wanted, mind, suppose, thinking, left
44	lady, dear, good, make, hope, poor, matter, bad, days, give, mine, great, ah, interest, excellent, speak, trouble
45	young, woman, money, years, married, life, girl, mother, men, women, father, child, ago, live, wife
46	death, made. matter, people, time, mind, point, case, person, question, fact, opinion, kind, present, subject
47	looked, head, face, asked, nodded, eyes, shook, good, smiled, slowly, chapter, manner, murmured, rose
48	don't, i'm, make, can't, it's, god, i'll, good, thought, won't, afraid, damned, stand, you're, long, word
49	good, round, time, i'll, put, bit, damn, thing, sort, gentleman, fellow, i'd, show, give, fact, what's

[c]) https://gephi.org/

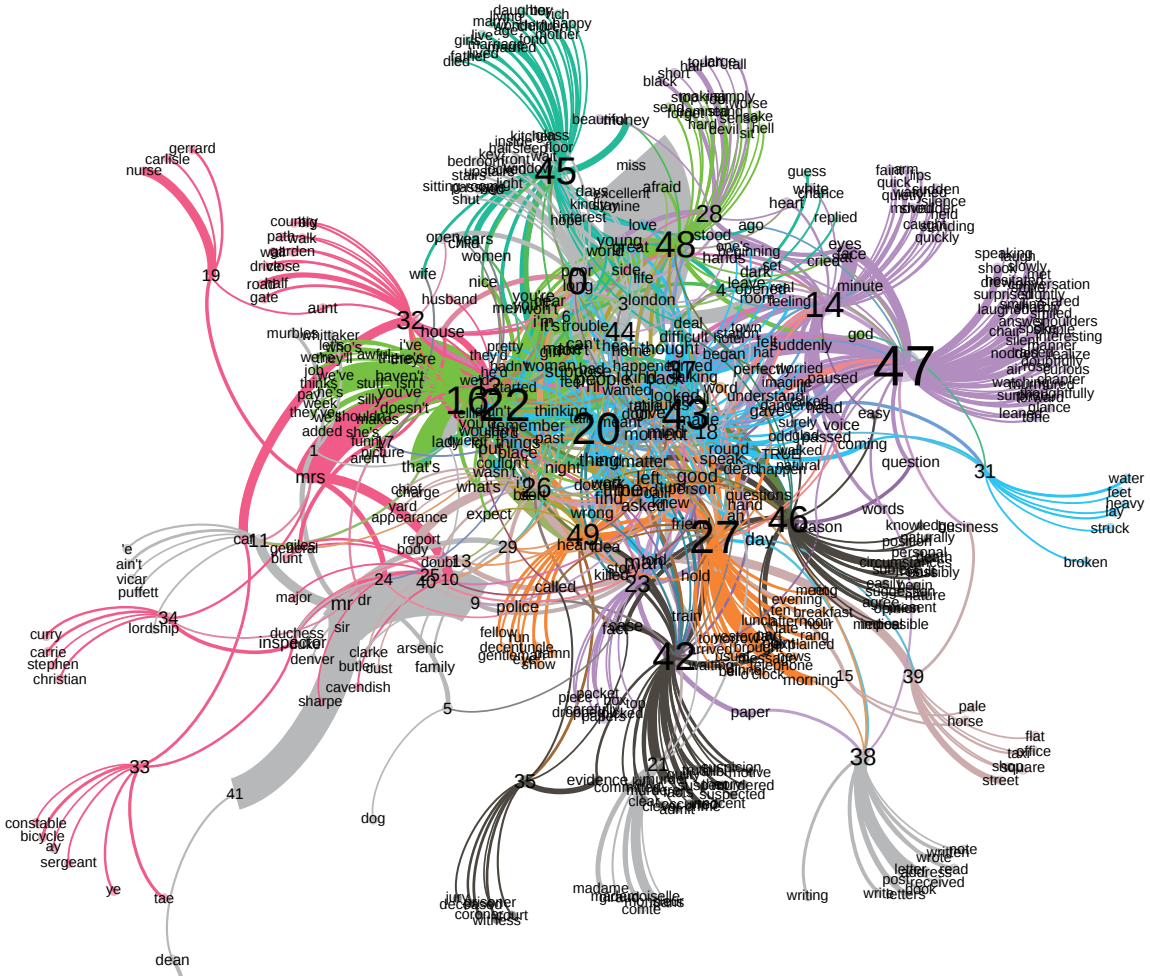

図 1　トピックと単語との関係性のネットワークグラフ
Figure 1　A network graph of relationship between topics and words.

図 2　トピック 0 のワードクラウド
Figure 2　A wordcloud of Topic 0.

図 3　トピック 42 のワードクラウド
Figure 3　A wordcloud of Topic 42.

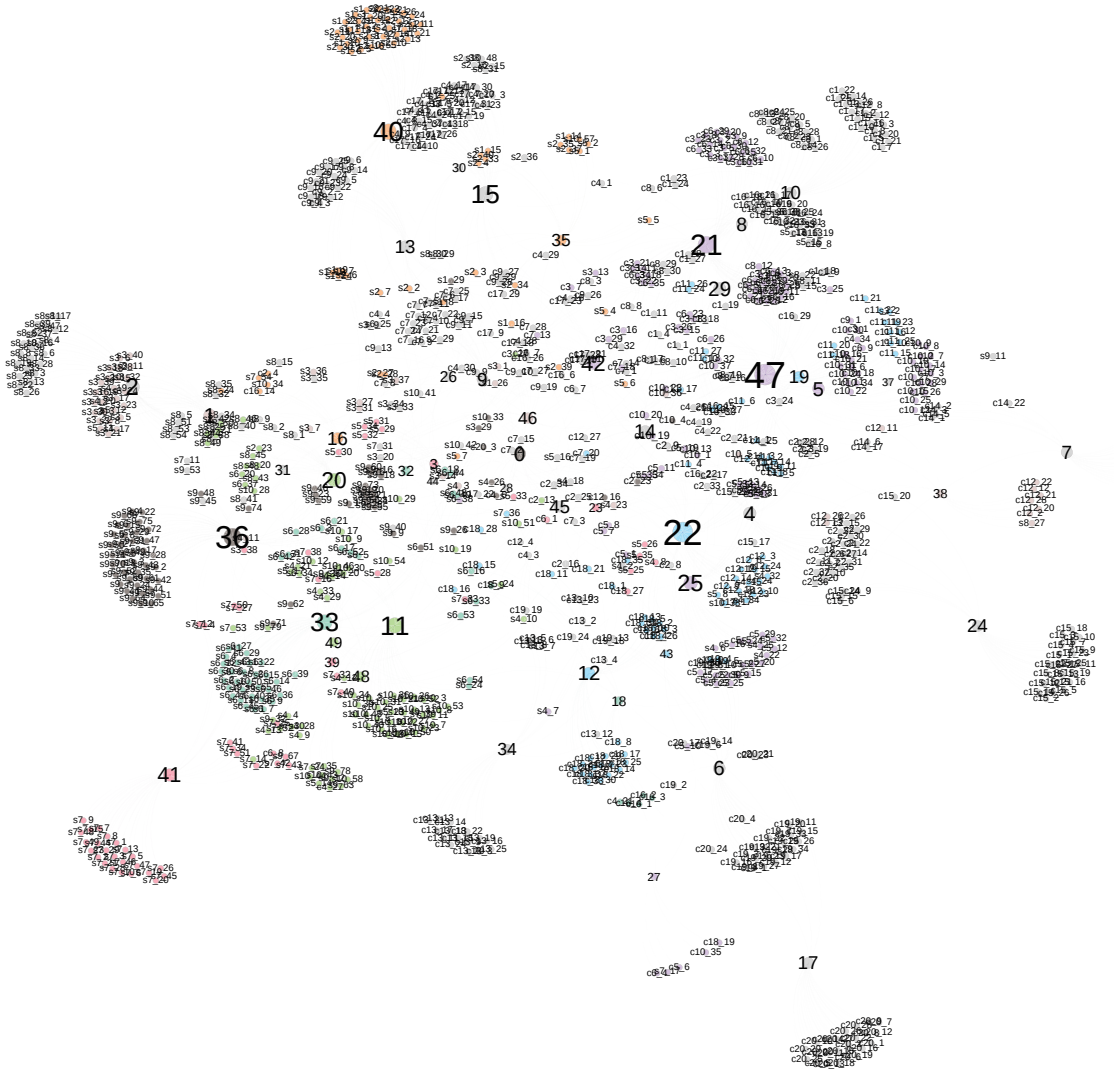

図 4　トピックとテクストとの関係性のネットワークグラフ
Figure 4　A networkgraph of relationship between topics and texts.

次に，トピックとトピックを構成するテクストとの関係性を描写したものを図 4 に示す．

クリスティの作品を表す c から始まるテクストが多く繋がっているトピックを確認すると，比較的大きなトピックを形成しているトピック 21 と 47 に多くの作品が繋がっている．トピック 21 には *monsier*, *madame*, *mademoiselle* などフランス語が多く含まれており，フランス語を話すポアロにまつわる，フランス語に関するトピックであることが推測される．トピック 47 は動詞の過去形や-ly 副詞が多い(図 5)．図 7 にトピック 47 で最も重みの大きい語である looked のコンコーダンスラインを示しているが，この図からも明らかであるように，これら動詞の過去形は会話文よりも地の文で用いられていることが多く，小説におけるナレーションを表すトピックであると考えられる．トピック 47 を構成している語としては他に，身体部位である *head* や *eyes* も含まれている．例として *head* の 3 語クラスターを図 8 に示したが，*shook his head, shook her head* のような表現が使用頻度上位に挙がっている．このトピックに-ly 副詞が多く含まれていることからも，地の文において身体部位や-ly 副詞を用いて登場人物の行動を描写する表現が，クリスティ作品において特徴的であることがうかがえる．

他にもトピック 6, 7, 17, 24 などにクリスティ作品が集中しているが，同一作品のラベルが同じトピックに固まる傾向が見られる．

図 5 トピック 47 のワードクラウド
Figure 5 A wordcloud of Topic 47.

he asked. ``Henden." Miss Brewis **looked** a little astonished. Poirot recoll
ised and assured." Hercule Poirot **looked** a little astonished by the outburst
ke, smiling, adding hastily as she **looked** a little bewildered, ``My name is L
here will be repercussions." Mary **looked** a little bewildered. ``You mean th
His eyes were amused. Carrington **looked** a little confused. ``Not at all no
explanation can be!" I thought he **looked** a little crestfallen, and rather wi
ed." ``Oh, I see." Rose Humbleby **looked** a little crestfallen. But she went
o a close, Poirot joined them. He **looked** a little dejected. ``Well, mon che
est, Poirot?" The chief constable **looked** a little disappointed rather like a
``Yes, yes…I wonder too." Clayton **looked** a little distrait. ``Well, I must
pe broke off. His nice kindly face **looked** a little disturbed and uncomfortabl
``Such as?" ``Well " Tim Kendal **looked** a little doubtful 'Bread and butter
s and drove away. Miss Vansittart **looked** a little doubtful for a moment or t
perintendent." The superintendent **looked** a little doubtful at the compliment
was goaded into it." Miss Marple **looked** a little doubtful. She could not qu
ne day that he had one." Tuppence **looked** a little dubious. ``You don't thin
ere! Not at all apparently!" Tony **looked** a little embarrassed. ``Rather odd
ed his throat rather nervously. He **looked** a little embarrassed. ``That's
ce long rest." Colonel Clapperton **looked** a little irresolute. Evidently the
rightfully dangerous." Merrowdene **looked** a little irritated. ``Mary's no
nd look at the pictures. A man who **looked** a little like the model for the
as dressed in glittering green and **looked** a little like a mermaid. She was

図 7 *looked* のコンコーダンスラインの一例
Figure 7 An example of concordance lines of *looked*.

	3-word cluster	Freq	In Files
1	shook his head	1,287	158
2	shook her head	704	141
3	nodded his head	234	87
4	his head and	175	79
5	her head and	163	75
6	the head of	144	64
7	his head he	140	69
8	his head i	136	71
9	his head no	126	65
10	turned her head	123	55
11	on the head	119	49
11	turned his head	119	62
13	her head i	112	61
14	nodded her head	108	51
15	head of the	105	57
16	of the head	104	62
17	shook my head	102	26
18	his head in	101	60
19	her head no	95	61
20	her head she	87	50

図 8 *head* の 3 語クラスター
Figure 8 A list of three-words cluster of *head*.

５．まとめ.

本研究では MALLET を用いて潜在的ディリクレ配分法によるトピックモデルを実行し，同時代女性作家であるセイヤーズ作品と比較してのクリスティ作品の文体の特徴の分析を行なった．その結果，クリスティ作品ではポアロが用いるフランス語にまつわるトピックや，地の文におけるナレーション，身体部位や-ly 副詞を用いることで登場人物の動作を表す表現が特徴的であることが明らかになった．今後の課題として，クリスティの作品は分析に使用していない作品が多く残っているためより多くの作品を用いた分析を行い，また抽出されたトピックの精査を行なっていきたい．

参考文献

1) Malling, S. and Peters, B.: AZ Murder Goes: Classic, p.247−269, Poisoned Pen Press, Scottsdale (1998).
2) 森英俊：世界ミステリ作家事典：本格派篇，p.219−222，図書刊行会(1998).
3) Lancashire, I. and Hirst, G.: Vocaburary Changes in Agatha Christie's Mysteries as an Indication of Dementia: A Case Study (online), available from 〈http://ftp.cs.toronto.edu/pub/gh/Lancashire+Hirst-extabs-2009.pdf〉 (accessed 2017-09-14).
4) Le, X, Lancashire, I, Hirst, G. and Jokel, R.: Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists, Literary and Linguistic Computing, Vol.26(4), pp.435−461 (2011).
5) 稲木昭子：謎解きのことば学：アガサ・クリスティの英語を楽しむ，p.65-117，英宝社 (2013).
6) Jockers, M. L. and Mimno, D.: Significant themes in 19th-century literature, Poetics, Vol.41(6), pp.750−769 (2013).
7) Blei, D., Ng, A. and Jordan, M.: Latent Dirichlet Allocation, Journal of Machine Learning Research, Vol.3, pp.993−1022 (2003).