中国のミステリー小説におけるトピック解析の試み*

黄　晨雯**

キーワード：トピックモデル、ミステリー、作家

作为大数据处理的手段之一，主题模型近年来渐渐作为文学作品的计量分析手法之一被广泛地应用。在中国文学作品分析中，主题模型的应用非常少，因此作者希望通过主题模型手法来考察不同悬疑作家的作品之间，在主题选取上有什么不同。

本文选取了中国 4 位著名的悬疑小说作家程小青，鬼马星，雷米，蜘蛛共 26 个作品进行主题模型分析。设定了 200 个主题来运行模型，并将得到的数据可视化以方便进行深入的考察。

从主题的密度分布图上可以分析出，时代因素和主人公背景设定影响着主题的选取。像有关电子监控的主题在程小青的作品中几乎不会出现，而在其他作家，尤其是 2010 年以后出版作品的雷米和蜘蛛的作品中出现的非常多。而像报纸的主题就几乎只出现在程小青的作品中。作者认为这是因为时代的变迁导致了新媒体的传播方式发生了改变。另外遣词上也可以看出时代的影响。同样是有关车的主题，与由汽车，车夫等用语构成的主题几乎只出现在程小青的作品中，而由车，司机，出租车等用语构成的主题几乎都出现在了鬼马星，雷米和蜘蛛的作品中。有趣的是，尽管现如今交通工具的选择越来越多，一般车辆（如出租车）依旧是近年的悬疑小说中的高频元素。

另外主人公的背景设定也对潜在的主题选取起到了很大影响。譬如程小青的作品是以侦探作为主角，所以有关寝室痕迹搜索的主题几乎全部出现在了程小青的作品中。而有关死亡现场的搜查一类的主题，在以警方人员作为主角的雷米和蜘蛛的作品中占比很高。而用作家，记者之类普通人作为主角，并且以言情推理著称的鬼马星的作品中，明显可以看到对话主题出现的非常多。一般认为对话元素在任何虚构作品中都是会频繁出现，但是通过密度分布图可以看出鬼马星对会话，特别是对问答句式的钟爱。通过对恋爱主题的观察可以看出对话元素受到恋爱元素的影响。

总之，通过主题模型手法，可以从一个全新的角度来审视文学作品，用数据客观地考察大量作品中的主题变化和作者在主题选取上的差别。

---

* 中国悬疑小说的主题模型应用研究（黄 晨雯（HUANG Chenwen））

** 大阪大学大学院言語文化研究科博士後期課程

1 はじめに

近年、情報処理技術の急速な進歩により、大量のデータから有益な情報を抽出し処理することが可能になってきた。統計的テキスト解析の研究において、単語の頻度計算や特徴語抽出など様々なテキストマイニングの手法が用いられている。その中に、単語それぞれの特徴分析ではなく多数の文書に潜在するトピックを推定するトピックモデルと呼ばれる方法が存在する。

Blei et al.（2003）によって提案されたLDA（Latent Dirichlet Allocation）、つまり潜在ディリクレ配分法は、「各文書は潜在的にトピックを多数持ち、同じトピックに属する単語は同一文書に出現しやすい」[1]と想定し、大規模な文書を構成する各々の文書に内在するトピックを推定するアルゴリズムである。例えば、オリンピック受賞の記事を書く場合、「年」「月」「日」「時」など日付や時間に関する用語を使いながら、「金メダル」「1位」「優勝」など表彰のことについて言及すると思われる。LDAは文章にそのような潜在的なトピックがあるということを前提として、研究者により設定されたトピックの数をもとに各トピックを構成する単語を確率的に計算する。LDAは代表的なトピックモデルとして、関連アイテム推薦システム、文書分類、プロファイリングなど様々な領域で利用されている。

文学研究におけるトピックモデルの応用研究に関しては、Jockers and Mimno（2013）、黒田（2017）などがある。Jockers and Mimno（2013）は19世紀の英語フィクション作品を大量に収集しトピックを分析した結果、25個のトピックにおいて作家のジェンダーによる相違が見られることを解明した。例えば「Tears and sorrow」というトピックは女性作家に好まれている一方、男性作家の作品にはあまり出ていないことを明らかにした。Jockersの研究により、フィクション創作における潜在的なトピックの選択に関しては、19世紀の作家では男女別に好みがあることが証明された。

黒田（2017）は、コナンドイルのミステリー小説と歴史小説を対象とし、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のか、視覚化したデータをもとに考察し、ジャンル間の相違を反映する特徴を明らかにした。

テキスト解析の重要な手法としてトピックモデルは定着しつつあるが、本研究で対象にしている中国のミステリー小説におけるトピックモデルの応用研究は筆者の知る限りではまだ限られている。中国のミステリー小説の研究については、程小青に関する質的研究が非常に多い。程小青は中国のコナンドイルと呼ばれ、ホームズシリーズの翻訳者

[1] Blei, D. M., Ng, A., & Jordan, M.（2003）Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022. p.996

として名をあげ、さらにホームズの登場人物設定を模倣して霍桑シリーズを創作した。程小青の小説における都市要素の探索（周 2015）、また中国を舞台とする事件の設定をめぐる研究（程 2015）は見られるが、現代のミステリー小説との比較、特にトピックに関わる比較考察はなされていない。

そこで本稿では、中国のミステリー小説に着目し、1940 年代に活躍した程小青と、21 世紀の著名なミステリー作家である鬼馬星 [2]、雷米 [3]、蜘蛛 [4] の作品を対象として、LDA に基づくトピックモデルを用いて解析を行い、得られた分析結果をもとに比較考察を試みる。それにより、作者によるミステリー小説創作における潜在的なトピック選択の違いを明らかにしたい。

2 分析対象

本研究では程小青、鬼馬星、雷米と蜘蛛の中長編作品を研究対象として取り上げる。雷米と蜘蛛は全作品それぞれ 5 作と 6 作を取り上げ、程小青は電子版 8 作を全てダウンロードした。鬼馬星の作品は本研究で取り上げる他の作家と比較して著しく多いため、対象とする作者間のファイル数のバランスも考え、代表作を 7 作選定し [5]、コーパスを作成した。

表 1 は作家、作品名と語数などを示したものである。3 列目は筆者がつけたラベルである。程小青の作品は C、鬼馬星の作品は G、雷米の作品は L、蜘蛛の作品は Z から始まるラベルとした。可視化の図表を描くときに作品名の代わりにラベルを使用すると、より直感的にグループ間の特徴の有無を見ることができる。

---

[2] 鬼馬星（1972-2017 本名：马雨默）は中国の著名な女性ミステリー作家である。2000 年に「文学新人大賞」の一等賞を受賞した。代表作：莫兰シリーズなど

[3] 雷米（生年不詳）は中国刑事警察学院の教師であり、ミステリー小説の男性作家として有名である。代表作：『心理罪』シリーズ

[4] 蜘蛛（1978- 本名：王黎伟）はミステリー小説の男性作家として有名である。代表作：『十宗罪』シリーズ

[5] 鬼馬星の作品は多数あるが、全部取り上げるとトピックの確率計算において鬼馬星にトピックが偏る恐れがないとは言えない。そのため、他の作家とほぼ同じ数を決め、異なるシリーズより代表作を選定することにした。

表 1：作者と作品情報

作者	作品名	ラベル	出版年	総語数	異なり語数
程小青[6]	白衣怪	C_baiyiguai	1946	54,998	8,270
	催命符	C_cuimingfu	1946	43,283	7,024
	断指团	C_duanzhituan	1946	27,928	6,598
	矛盾圈	C_maodunquan	1946	43,983	7,515
	青春之火	C_qingchun	1946	26,793	5,485
	霜刃碧血	C_shuangren	1946	35,254	6,645
	舞后的归宿	C_wuhou	1946	62,473	8,342
	紫信笺	C_zixinjian	1946	28,006	6,062
鬼馬星	风的预谋	G_fengdeyumou	2006	224,193	16,761
	猫的复生	G_maodefusheng	2008	79,283	9,819
	迷宫蛛	G_migongzhu	2009	263,546	18,490
	暮眼蝶	G_muyandie	2009	129,080	12,325
	纽扣杀人案	G_niukou	2008	110,478	11,218
	淑女之家	G_shunv	2009	96,840	10,642
	葬礼之后的葬礼	G_zangli	2009	155,355	13,601
雷米	暗河	L_anhe	2011	82,501	12,412
	城市之光	L_chengshi	2012	106,503	15,261
	第七个读者	L_diqigeduzhe	2010	31,485	6,303
	画像	L_huaxiang	2012	86,961	12,088
	教化厂	L_jiaohuachang	2012	37,503	7,544
蜘蛛	十宗罪 1	Z_shizongzui1	2010	91,010	13,689
	十宗罪 2	Z_shizongzui2	2011	98,084	15,535
	十宗罪 3	Z_shizongzui3	2012	95,773	15,164
	十宗罪 4	Z_shizongzui4	2013	83,664	14,094
	十宗罪 5	Z_shizongzui5	2015	82,676	14,399
	十宗罪 6	Z_shizongzui6	2018	56,446	11,603

3 研究手法

本研究では、機械学習による自然言語処理ツールキット MALLET（Machine Learning for LanguagE Toolkit）を実装し LDA を実行する。トピック数を設定し、全てのファイルを整えて MALLET を使えば、プログラムが自動的にファイルごとのトピックの割合、それぞれのトピックにおける単語の割合を計算し、テキスト形式として出力する。

[6] 霍桑シリーズに属する程小青の作品は全て 1946 年の世界書局により出版された『霍桑探案全集袖珍丛刊』に収録されている。

MALLETを実行する前に、いくつかの前処理が必要である。最初に必要な処理は分かち書きである。LDAはコーパスに収録されているテキストをサンプリングしながらトピック構成を推定するモデルなので、中国語のテキストに応用するためには句を単語ごとに分かち書きしなければならない。Howl-Anderson[7]はSecond International Chinese Word Segmentation Bakeoff[8]によって提供されたデータをもとに13個の中国語分かち書きツールの精度を評価した。その評価方法を使い、本研究で用いるコーパスに適用したところ、jieba[9]の精度[10]が0.739、再現率[11]が0.896、F値[12]が0.81で最良だという結果を得た。そのため、本研究では最も精度の高いjiebaを選定し、分かち書きを試みた。

また、対象作品計26作をそれぞれ先頭から1000語ごとに切り分け、総じて2042個のファイルを作成した。実際の分析はこの2042個のファイルに対して行う。LDAはテキストに複数のトピックが存在することを前提とし、トピックモデルを構築するアルゴリズムであり、ファイル全体をコンテクストとして扱うため、テキストファイルを細かく切り分ける必要がある。

最後の前処理は分析対象から除外すべき単語のリストを作成することである。テクストにおいて代名詞・助詞・助動詞・連体詞・副詞などの出現率は高いが、他方これらの語群は語彙的意味を担わない機能語であり、トピックの質的解釈にあたりノイズになることが多いため、ストップワードとして分析対象から除外することにする。小説の場合、人名・地名などの固有名詞が非常に多く存在するため、固有名詞が主要構成要素となる共起語群がトピックと推定されてしまうケースが生じる。特に主人公1人の名前だけでトピックが構成されるようなケースが発見された。そのようなケースは量的分析を必要としない自明の事実であり、探索的な分析の対象ではない。全ての人名を取り除くことは困難であるが、主人公の名前を分析対象外とすることにした。

前処理が整った後、トピック数を設定すればモデルが実行できる。トピック数の設定は分析の粒度を設定することでもある。少なく設定した場合、一つのトピックに多数の話題が混在し、内容的一貫性が損なわれるため、解釈が困難になる。他方、推定するトピック数がデータの分散を大きく超える数に設定されると、局所性の高いトピックが生成され、全体像の把握にはふさわしくない出力となってしまう。そのため、トピック数

---

[7] http://github.com/howl-anderson/ Chinese_tokenizer_benchmark 2019.9.20最終アクセス
[8] 2005年の夏に行われた第二回国際中国語分かち書き大会のことである。提供されたデータには「Academia Sinica」、「City University of Hongkong」、「Peking University」と「Microsoft Research」の4つのコーパスが含まれている。
[9] jiebaは中国語を対象とする分かち書きツールの一種である。
[10] 分かち書きツールにより正しく分割できた単語の数をcとする。誤って分割された単語をeとする。正解数をNとする。精度とは、c/Nの値であり、高いほど良いとされている。
[11] 再現率は、c/（c+e）の値であり、高いほど良いとされている。
[12] F値は（2*精度*再現率）/（精度+再現率）の値であり、高いほど良いとされている。

を変化させて実験を繰り返し最適なトピック数を検討する必要がある。テストを繰り返した結果、全体的に解釈が最もしやすいと思われるトピック数200個を設定し、その結果をもとに考察を試みる。

4 結果と考察

分析プログラムMALLETを実行すると、出力としてそれぞれのトピックを構成する主要単語とその重みのデータ、文書ファイルとトピックの関連度を表す数値データが生成された。表2は各トピックを構成する中心単語とアルファ値[13]を表すデータの一部の抜粋である（上位5つにアルファ値が大きいトピックと、下位5つにアルファ値が小さいトピックを提示した)。中心単語をもとに各トピックのラベルをつけることが可能である。例えばトピック140はキーワード「警察側」「進める」「調査」「発見」「見つけた」などから警察の捜査に関するトピックだと判断できる。また、アルファ値を通して、トピックの大きさがイメージづけられる。例を挙げると、トピック5（α =0.33056）はアルファ値が非常に大きく、複数の作品に渡って出ていると考えられる。一方、トピック150（α =0.00274）は非常に小さいため、局所的コンテクストにのみ出現している可能性が高い。

表2：トピックごとの中心単語（抜粋）

トピック	アルファ値	キーワード
5	0.33056	知道 问 告诉 看见 那里 回来 等 问道 找 打 看 拿 发现 肯定 两个 死 用 那天 见
77	0.28696	知道 喜欢 问 男人 女人 看 发现 认识 名字 关系 告诉 特别 朋友
130	0.20888	笑 道 看 问 地说 问道 今天 事 一眼 说话 看见 马上 吃 起来 喜欢 一句
67	0.19431	声音 起来 脸 眼睛 感到 手 身体 一声 双手 头发 仿佛 身上 挣扎 慢慢 看到 死 发出 头
140	0.18861	警方 进行 调查 发现 经过 工作 一名 找到 以及 一起 犯罪 线索 两名 平时
16	0.00613	酒店 镜子 碎尸案 蓝可儿 刁爱青 魔术师 浴室 蓝 尸体 蓝京 警方 可儿 大丽花
180	0.00453	披风 队伍 游行 赌场 老大 小偷 陈帅 杜 看守所 赌 赌徒 有人 出现 爱国 日本 赌博 楼顶 越狱 用

[13] アルファ値はトピックの大きさを表す指標である。アルファ値が大きいほどトピックが多くの作品に渡って現れている。一方、小さいほど局所的なコンテクストにしか出ていないことを示している。

117	0.00407	绑匪 小希 赎金 绑架 家属 爸爸 绑架案 孩子 要求 木屋 树林 人质
24	0.00372	寺庙 和尚 僧人 方丈 施主 武僧 假钞 大师 茶 主持 老和尚 云寺 庙里 静 接线员 功德 香客 棍法 庙
150	0.00274	集市 村民 尸块 冷冻 冰柜 所长 戴 老板 包子铺 老板娘 超市 碎尸案 韩梅梅 二懒 肉 冻死

より細かく考察するためには、数値データの可視化が必要となってくる。トピックを構成する単語の重みを示すデータと、ファイルごとのトピックの比率を表す数値データをもとに、統計解析ソフトRを用いてワードクラウドやボックスプロット、ヒートマップなど可視化グラフを作成する。可視化グラフによって、より直観的に単語のウェイトを見ることができ、またトピックとファイルの関係性、トピックによる作品の分類結果などを考察することができる。

まずは作品とトピックの全体像を把握するために、ヒートマップとクラスタリングを結合した図1を提示する。

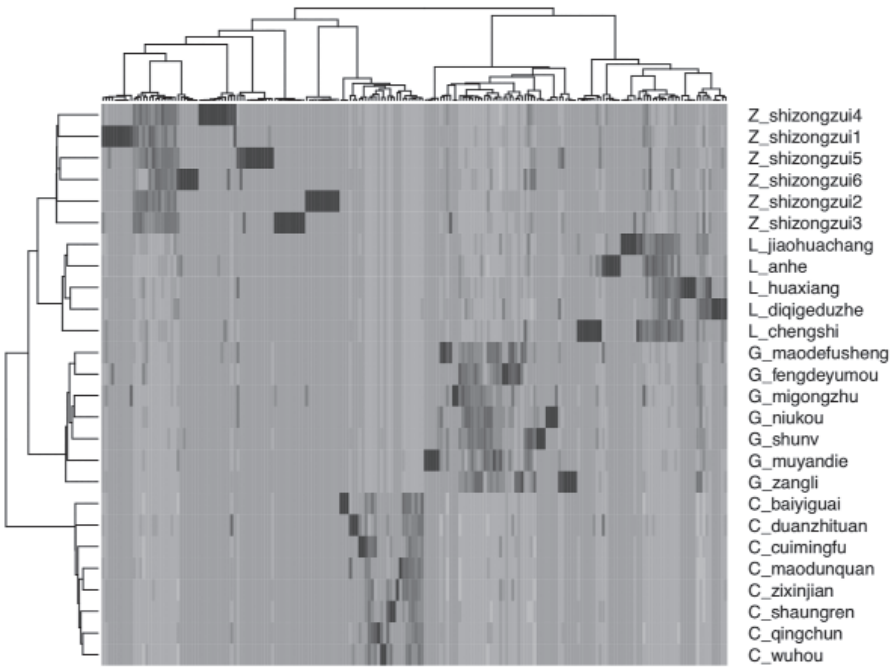

図1：ヒートマップ

図1では、横軸にトピック[14]、縦軸に作品を配置し、トピック間、また作品間の類似

[14] トピックが200個もあり数が多いため、トピックの番号を横軸に表示することができない。図1では全体像を提示し、トピックに対するミクロな考察は次の節で行う。

性をクラスタリングの結果によって可視化している。横軸ではトピック間の関係性を表し、同一のクラスターに属するトピックは生起するコンテクストやテクストが類似している傾向があることを示している。縦軸では作品の関係性を表し、類似した作品は近く、類似していない作品は遠く配置されている。枝分かれから明らかなように、まず一番下のクラスターを構成する作品（C から始まる作品）と他の作家の作品とは大きく 2 つに分かれている。程小青の作品は他の 3 人と端的に違うことが読み取れる。さらに、鬼馬星、雷米と蜘蛛のうち、雷米と蜘蛛の作品、つまり Z と L から始まる作品がペアとして分岐し、比較的に類似していることも図からわかる。コーパスを構成する作家の作品は全て独立したクラスターとして提示されている。このことは、作家個々のスタイルが一貫して他の作家と異なることを示していると解釈できる。上述したような分類結果になっている原因については、トピックのミクロな考察を通して分析する。

ヒートマップはセルの濃淡の違いでトピック密度の差を表している。作品においてトピックの占めた割合が高いほど、セルの色が濃くなる。一つの作品にだけ色が濃く見えるトピックは特定の作品の主題や人物、叙述・描写される場面、出来事を色濃く反映するトピックである。一方、複数の作品に渡って割合が高いトピック、特に違う作家間において密度の差があるトピックについては、作家もしくは作家が所属する流派などの特徴を反映している可能性が高いため、重点的に検討すべきだと考えられる。

4.1 時代の影響による差異

前述したように、程小青と他の 3 人とは格段に違っている。程小青のトピックを考察すると、小説の内容によるトピックの違いはもちろんあるが、時代の変化から生み出したミステリー要素の相違と言葉遣いの相違が発見された。

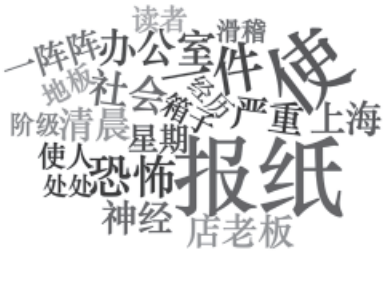

図 2：topic186（新聞記事）のワードクラウド

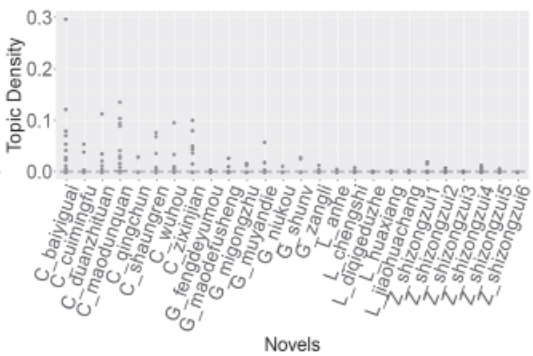

図 3：topic186 の箱ひげ図

まずtopic186を見てみよう。図2はtopic186の中心単語から構成されたワードクラウド[15]であり、図3は箱ひげ図[16]である。「新聞」「社会」「上海」「オフィス」「恐怖」「読者」「深刻」などのキーワードより、「新聞記事」の話題だと解釈できる。このトピックはほぼ程小青の作品にしか出現していないことは図3からわかる。その大きな要因として考えられるのは、媒体の変化であろう。中国報告網の2017年の報告[17]によると、新媒体の急速な発展により、ニュースを入手する手段として、紙媒体よりウェブ媒体がはるかに好まれるようになっている。程小青が活躍した当時では、ニュースの伝達手段として新聞のような紙媒体が主流であったため、出現率にこうした明確な違いが現れたのであろう。

図4：topic172（自動車）のワードクラウド

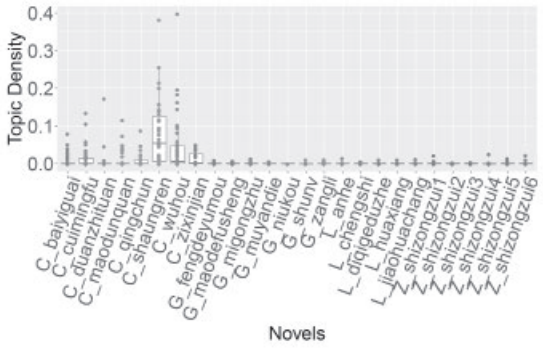

図5：topic172の箱ひげ図

また、topic172とtopic73（p.10に掲載）の違いが非常に興味深い。この2つのトピックはどちらも車に関わるトピックである。しかし、箱ひげ図の分布から明らかなように、これら2つの話題のコーパスにおける出現パターンは対照的である。Topic172は程小青の作品にしか生起していないのに対して、topic73は程小青以外の作家の作品に平均的に多く出現している。これらはいずれも車に関するトピックであるが、それぞれのトピックを構成する語彙には明確な相違が見られる。まず、Topic172において最も大きなウェイトを占める語「汽车」は今日では車のことを指し、使用されてはいるがやや前時代的な表現である。直接「车」、もしくは車のブランドで呼ぶことがより好まれていると思われる。程（2015）によると、当時の上海において「汽车」は重要な交通手段と

[15] ワードクラウドは単語のウェイトを可視化するツールである。単語のウェイトが大きければ、字のフォントも大きい。それほど重要でない単語は、小さく見える。
[16] 箱ひげ図はファイルとトピックの関係を示す図である。点がファイルのことを表し、横軸に作品、縦軸にファイルが占めた割合が配置されている。箱ひげ図により、トピックにおける作品ごとの違いが見られる。
[17] http://market.chinabaogao.com/chuanmei/1223309AH017.html 2019.11.17最終アクセス

して利用されている。ただし、当時の「汽车」はほぼレンタカー事業と関わっている。「汽车」を借りて自分が運転することもあり、運転手を雇用して運転させることもあるという。つまり、topic172 の「汽车」は「レンタカー」の意味に近く、今日で言う「汽车」や「车」とは違う概念であろう。自動車産業の相違、要するに時代の表象による違いが topic172 と topic73 との対比から読み取れるのではないか。

図 6：topic73（車）のワードクラウド

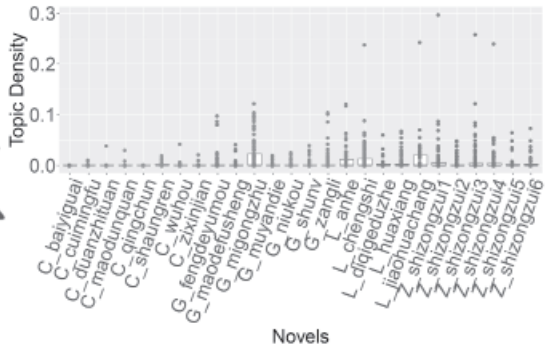

図 7：topic73 の箱ひげ図

また topic172 に現れた「车夫[18]」（運転手）とは対照的に、topic73 に「司机」（運転手）が出現している。時代の変遷につれての言葉遣いの変化もトピックの分布により明らかになっている。その上、topic172 における重要度として「车夫」はやや低い。それに対して、「司机」が topic73 において高いウェイトを占め、車との共起もより緊密であることが言える。これも上述した自動車産業の変化と関係しているのではないであろうか。

Topic172 と topic73 はどちらも車、いわゆる交通手段に関わるトピックである。これら以外に交通手段に関するトピックはないのかという疑問が生じるかもしれない。そこで、200 個のトピックを全て調べたところ、「地下鉄」を中心とする topic55 が発見された。ただし、topic55 はほぼ蜘蛛の『十宗罪 1』にしか出現しておらず、特定した事件としか関わっていないことが図 9 から明らかである。

程（2015）によると、程小青の小説では容疑者の後をつけるシーンが多く、交通手段として「汽车」を借りることが多い。つまり、程小青の作品において車は肝要なミステリー要素として使用されていると言っても過言ではない。一方、交通手段の選択が大いに増加した今日でも交通手段に関するトピックとして車の一般性が現れたことから、中国のミステリー小説では、地下鉄やバスなどの公共交通機関ではなく、一般車両（によ

[18] 車両を運転する人の旧称である。出典は『儒林外史』である。

る追従など)がサスペンス要素として利用される傾向が強いことを表しているであろう。

図 8：topic55(地下鉄)のワードクラウド

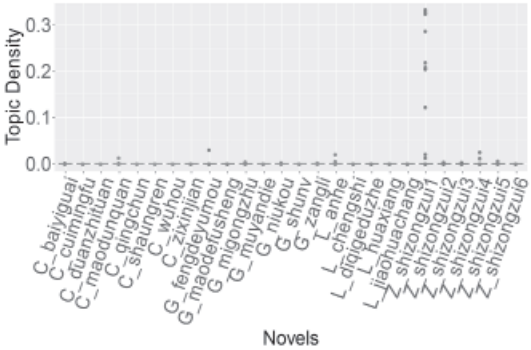

図 9：topic55 の箱ひげ図

4.2 モチーフ設定の差異

ミステリー小説といっても、モチーフの設定の違いは作家間に当然存在している。特にどのような人物を主人公として描くかは、作品のモチーフと密接に関わる重要な要素である。程小青の作品では、主人公として描かれる人物は探偵である。鬼馬星の作品においては、対照的に一般人、たとえば作家や記者を主人公にする傾向があり、恋愛のもつれも多く描かれている。一方、雷米の作品では「方木」という犯罪学の大学生を主人公としているが、シリーズの中期より刑事になった。全体としては警察とともに行動する設定である。蜘蛛の作品シリーズの主人公は全て警察関係者である。つまり、サブジャンルとして、程小青の作品は探偵小説、鬼馬星は恋愛推理小説、雷米と蜘蛛は刑事小説であるとジャンル分けすることもできる。では、そのようなサブジャンルに注目して考察を行う。

まず事件の偵察方法について、トピックでは明らかな相違が見られた。例えば、topic28 を見てみよう。キーワード「動画」「監視」「画面」「録画」「パソコン」「アングル」などより、topic28 は「監視カメラ」を中心とするトピックであると解釈できる。右側の箱ひげ図からわかるように、程小青の作品においてこのトピックはほぼ皆無である。鬼馬星の作品に少しは言及されており、平均的に多く割合を占めているのは雷米と蜘蛛の作品である。このトピックを構成したキーワードは科学技術の進歩による産物を代表したものであり、時代との関わりが大きいと思われる。21 世紀ころから徐々に電子製品による偵察手法が使われ始めたため、2000 年以降、特に 2010 年からの作品（雷米と蜘蛛）に多く出現していることは容易に理解できる。

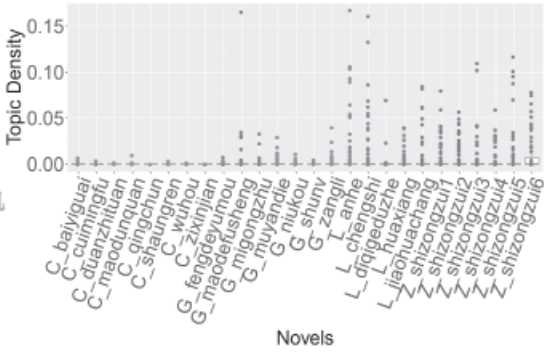

図10：topic28（監視カメラ）のワードクラウド　　**図 11：topic28 の箱ひげ図**

ただし、このような相違はモチーフの設定とも関連しているのではないか。昨今、警察が防犯カメラ・監視カメラで撮影された映像を利用するのはいわゆる一般的な捜査方法とされているが、主人公が警察官でない場合、監視カメラなどに触れる機会も限られるため、登場のチャンスも少ないと考えられる。鬼馬星の作品において topic28 の割合が雷米と蜘蛛に及んでいないことは鬼馬星が他の作家とは異なる背景を持つ主人公を中心に物語を展開させているからであろうと解釈できる。面白いことに、G_maodefusheng（『猫的复生』）という鬼馬星の作品には、同一作品の他のセグメントとはトピック構成が大きく異なる数値データを示すセグメントが 1 点ある。そのファイルを実際に読んでみると、先端技術を利用した警察の調査場面が描かれている。電子機器を利用した偵察手法が刑事小説の肝心な要素だということは確かであろう。

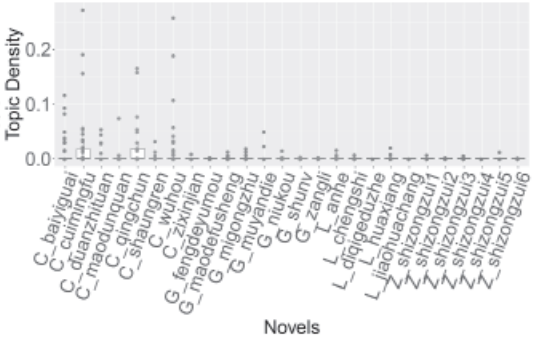

図 12：topic91（事件現場の家具）のワードクラウド　**図 13：topic91 の箱ひげ図**

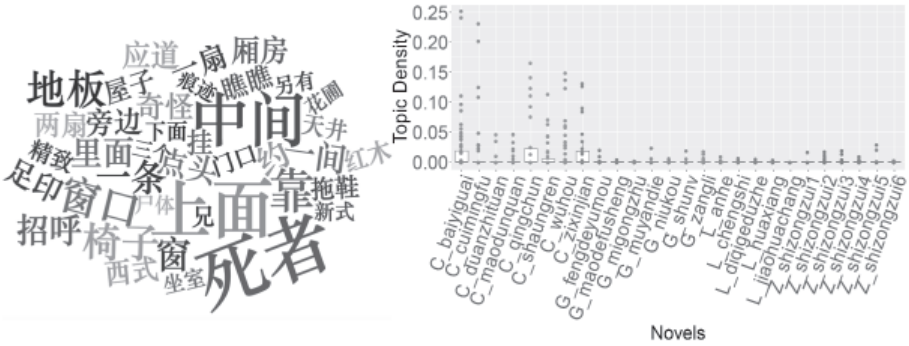

図 14：topic123（被害者の寝室）のワードクラウド　図 15：topic123 の箱ひげ図

Topic91 は事件現場の家具に関するトピックで、topic123 は被害者の寝室をめぐる語群によって構成されている。この 2 つのトピックはどちらも程小青の作品において多く出現していることは図 13 と図 15 から明らかである。本雅明（2006）は居室を探偵小説と関連させ、「家具の位置は犯罪の一環であり、部屋の仕組みも被害者の逃走ルートと関わってくる」[19] と説明し、室内空間の描写に重点を置く探偵小説の特徴を述べている。周（2015）も程小青の小説における都市文化の要素を探り、居室の痕跡を重点的に描写している程小青の特徴を叙述している。トピックモデリングの結果として Topic91（事件現場の家具）と Topic123（被害者の寝室）が程小青に高い比重を占めていることが先行研究を客観的に裏付けしているのではないか。

もちろん、単に探偵小説の特徴というより、西洋探偵小説の特徴が作家間の相違を生み出したと言った方がより適切だと思われる。黒田（2017）ではトピックモデルを通してコナンドイルの推理小説と歴史小説を比較していることは前にも述べた。その結果では家・家具に関するトピックだと判断したトピックがホームズシリーズの方に多く出現していることが発見されたという。先述したように、程小青はホームズシリーズの翻訳者として有名であり、彼の作品もコナンドイルの設定を模倣し、探偵を主人公とし、助手的なパートナーを入れている。Topic91 と topic123 の分布傾向を吟味すると、程小青が西洋の探偵小説の要素を十分に取り入れている点が、他の作家の作品との大きな違いを生んでいると考えられる。

上述した Topic91 と topic123 には両方とも「死者」という重要なサスペンス要素が出現している。実際、「死者」という言葉が他のトピックにもキーワードとして出ていることが観察できた。トピックの構成の違いにより、解釈と分布も完全に異なっている。

[19] 瓦尔特・本雅明（2006）单行道 . 江苏：江苏人民出版社 . p.78

図 16：topic195（死体発見と現場）のワードクラウド

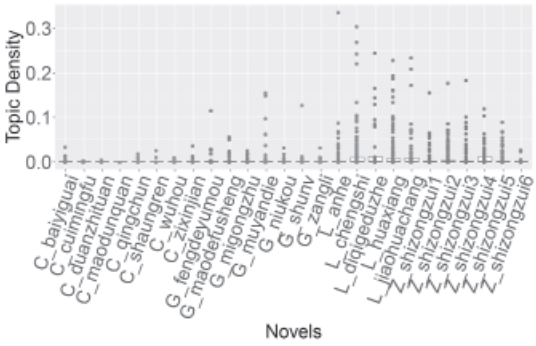

図 17：topic195 の箱ひげ図

Topic195 は「死者」「現場」「発見」「死体」「検死医」「痕跡」などより、死体発見と現場に関わるトピックだと解釈して良いであろう。このトピックは、まず程小青の作品にはほとんど生起していない。他方、鬼馬星の作品には多少生起するが、鬼馬星を代表すると言えるほど一貫して出現しているわけではない。それに対して、topic195 は雷米と蜘蛛の作品には比較的に多く現れている。キーワードから見れば殺人事件と深く関係していることは明らかであるが、Z_shizongzui6（『十宗罪 6』）を除き、ほか全ての作品は殺人事件を中心に展開しているため、殺人事件というモチーフが相違を起こしたわけではない。作家ごとのキーワードの出現頻度を調べたところ、「现场」（現場）という言葉が程小青の作品には 1 回も出現していないことは興味深い。「现场」が使われないのは程小青の習慣なのか、それとも時代が関係しているのかは今後同期の作家と比較して解明する必要があるように思われるが、現場捜査の描写において、他の作家は「现场」を多用していることは分かる。また、「检死医」は程小青の全作品に 9 回しか出ておらず、コンテキストを見れば死亡現場における検死医の登場場面は描かれていない。程小青と他の作家とは偵察過程の描写において大きく異なることはこのトピックからでも観測できる。また鬼馬星と比べ、より雷米と蜘蛛に偏っていることは、事件現場での捜査の描写が、警察を中心に物語を進める雷米と蜘蛛の作品のプロット展開により多く利用されていることを反映しているのではないか。

類似したトピックとして、topic42 が挙げられる。Topic42 は「死体」「死者」「検死医」「死亡」「発見」「時刻」「ネイル」「指」などより「死体と検死」のトピックだと解釈できる。Topic195 と比べ、死体をめぐる言葉が多数集まり、より死体の描写にウェイトを置いたトピックであろう。トピックの分布傾向からすれば、意外なことに雷米と蜘蛛に共通して特徴的であるとは言いがたい。むしろ蜘蛛の個人的な特徴であろう。最新作の Z_shizongzui6（『十宗罪 6』）はシリーズ内の他の作品と違う傾向を示しているが、

前述したように『十宗罪 6』は他の『十宗罪』シリーズ作品と違い、殺人事件を中心に展開していないため、特徴としても一概に見ることができないであろう。このように、類似したトピックにおける微妙な相違もトピック解析により明らかになる。

図 18：topic42（死体と検死）のワードクラウド

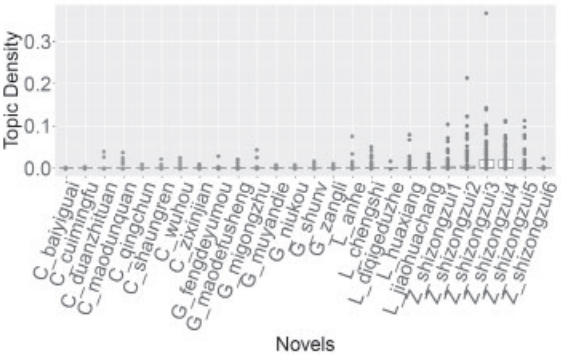

図 19：topic42 の箱ひげ図

さらに、鬼馬星に関しても非常に特徴的なトピックが存在している。例えば、topic130 と topic142 には「質問する」「答える」「と言う」など会話動詞を表す言葉が多く含まれている。つまり「発話の伝達」に関わるトピックと言ってよいであろう。一般的にフィクション作品において、直接話法で引用される登場人物の発話、あるいは間接話法や、ナレーターによるスピーチアクトの要約など、登場人物の発話を直接に、あるいは間接的に伝達する表現が特徴的に用いられている。そのためこうしたトピックを詳細に分析することはフィクションの言語表現を捉える上で肝要だと思われる。

図 21 が如実に示すように、topic142 は鬼馬星の全作品において特徴的に多く出現している。程小青は時代が違うため、文体上の問題もあり比較はしづらいが、雷米と蜘蛛の作品と比べ鬼馬星が会話文、特に質問と回答がセットとなって構成されている会話をより頻用していることはトピックモデリングの結果に反映していると解釈できる。作家の個人スタイルが関係していると考えられるが､それだけに留まっているとは思えない。この節のはじめにも述べたように、鬼馬星は一般人を中心に物語を進め、恋愛に関する部分も作品には少なくない。Topic77 は恋愛をめぐるトピックであり、鬼馬星を代表するトピックである。「好き」「女」「男」「知り合う」「関係」などの主要単語のほか、会話トピックにおいても大きくウェイトを占めた「知る」「質問する」のような語が出現している。また、topic142 の中心単語にも「好き」「彼女」というキーワードが現れている。会話トピックに恋愛という主題も関係していることが言えるのではないか。つまり、刑事小説と比べ、こうした恋愛推理小説では、会話がより多くなされがちだと言っ

てよいであろう。もちろん、今後は恋愛小説との比較を行い、その正確さを検証する必要がある。

図 20：topic142（質問と回答）のワードクラウド

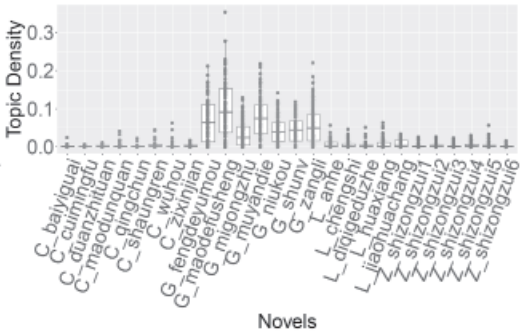

図 21：topic142 の箱ひげ図

5 まとめ

本稿では、トピックモデル手法を使用し、ミステリー小説創作における潜在的なトピックの選択に関して、作家の間では明らかな相違があることを証明した。

LDA モデルを用いることにより、トピック、つまり語群の特徴を考察し、単語の頻度計算などの計量手法だけでは発見が困難な意味的関係性を発見することもできる。同じ単語が複数のトピックに渡って出現することは少なくないが、高頻度の単語のみから実際考察できることに限りはある。しかし、トピックを通して解析する場合、その単語と共起する単語の違いから適切にトピックのラベルづけができ、ファイルとの関係性を検討することでより深い考察ができる。4.2 で言及したように、「死者」が複数のトピックに渡って出現しているが、「死者」と共起する言葉が違えば、トピックの解釈も全く異なり、使われる場面の違いも推定できる。トピックにおける共起というのは、必ずしもすぐ近くに位置しているわけではなく、同じ段落もしくは同じファイルに共起しやすいとも解釈できる。この点において、単に作家別の共起単語リストを観察するより、トピックモデルを使用した方がより柔軟に考察できるということがトピック解析のメリットだと思われる。

今後の課題としては、まずはコーパスの拡大である。本研究では 4 人の作家しか取り上げていないため、トピックの違いがあるとはいえ、そこに見られる特徴は個人差によるところがまだ大きいということは否定できない。対象とする作者数を増せば、よりグループ化された特徴、例えば細かな年代やジェンダーによる相違を発見できる可能性もあると思われる。

また、今回は1000語ごとにファイル分けを行ったが、今後さらに異なる条件下での分析を試みてより高質なトピックが得られることを期待する。さらに、作家間ではなく1つの作品におけるトピックの変化を考察することも興味深いであろう。

固有名詞のノイズワーズ処理について、主人公の名前だけを除外しているが、脇役で多出する人物のトピックがまだ残っていることについてはやはり問題である。今後の研究では、より機械的な手法で固有名詞を除外する方法について検討していきたい。

参考文献

[1] 黒田絢香「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」『テクストマイニングとデジタルヒューマニティーズ』大阪大学大学院言語文化研究科、2017、pp.23-41。

[2] 田畑智司「FLOBコーパスの意味構造：確率論的トピックモデルによる言語使用域の特徴づけ」『テクストマイニングとデジタルヒューマニティーズ』大阪大学大学院言語文化研究科、2017、pp.5-21。

[3] 程海燕 .（2015）. 论程小青侦探小说的本土化（Master's thesis）. Retrieved from http://cdmd.cnki.com.cn/Article/CDMD-10357-1015373632.htm

[4] 瓦尔特・本雅明 .（2006）. 单行道 . 江苏 : 江苏人民出版社 .

[5] 周楠 .（2015）. 近代侦探小说中的都市元素研究（Master's thesis）. Retrieved from http://cdmd.cnki.com.cn/Article/CDMD-10270-1015367498.htm

[6] Blei, D. M., Ng, A., & Jordan, M.（2003）. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

[7] Blei, D. M.（2012）Probabilistic Topic Models. Communications of the ACM, 55（4）, 77-84.

[8] Jockers, M. & Mimno, D.（2013）. Significant themes in 19th-century literature. Poetics, 41（6）, 750-769.