



Title	中国のミステリー小説をめぐる計量的分析：程小青と鬼馬星の作品を中心に
Author(s)	黄, 晨雯
Citation	言語文化共同研究プロジェクト. 2020, 2019, p. 31-45
Version Type	VoR
URL	https://doi.org/10.18910/76990
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

<https://ir.library.osaka-u.ac.jp/>

The University of Osaka

中国のミステリー小説をめぐる計量的分析

—程小青と鬼馬星の作品を中心に—

黄 晨雯

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

Email: u469911c@lang.osaka-u.ac.jp

あらまし 本稿では、中国 20 世紀以降の著名なミステリー作家である程小青と鬼馬星の作品計 34 点に対して量的分析を行うことを目的とする。主に代名詞と名詞の情報に焦点を当て、言語差異を見つけ出す試みをする。まずは主成分分析を実行し、テキスト間の相互関係、また単語間の相互関係を示す散布図を観察することによって、両作家の作品における明確な差異を解明した。特に女性を指す三人称代名詞の使用について大差が見られる。時代の影響により言葉遣いも変化するものである。さらに、LDA モデルを実行し生成された 50 個の名詞トピックに対して考察を行う。中国 20 世紀以降のミステリー小説において、身体言語と室内に関わるトピックは非常に大きな割合を占めている。程小青を特徴づける身体言語のトピックから、作品における人物の特徴まで探ることができた。寝室の痕跡と解釈したトピックも程小青を代表しているが、複数のトピック内容が 1 つのトピックに混在していることからトピック数の調整による細分化も必要なのではないかという示唆も受けた。

キーワード 言語特徴, 主成分分析, トピックモデル

中国近代悬疑小説の計量分析研究

—以程小青和鬼馬星的作品为中心—

黄 晨雯

大阪大学大学院言語文化研究科

〒 560-0043 日本大阪豊中市待兼山町 1-8

摘要 本文旨在通过计量手法对中国近代著名的推理小说家程小青和鬼马星的作品共 34 本小说进行定量分析。本文主要聚焦在作品中的代名词和名词上, 通过主成分分析手法, 可以看出在代名词和名词的使用率上, 两位作家的作品之间有着明显的区别。尤其是在对于女性第三人称的用法的不同上可以总结为时代变化对于用词的影响。本文还应用了主题模型分析, 设定了 50 个名词主题。从结果上看, 在悬疑小说中, 人物肢体语言以及有关房间构造的主题占了非常大的比例, 可以说是悬疑小说的代表性主题。在有关肢体语言的多个主题中, 通过主题的分布可以看出程小青的相关主题带有非常强烈的作品特色, 可以通过主题中的词语分布归纳出场景和人物

特点。有关寢室痕迹的主题也是程小青作品的特色，但是该主题中混杂着多个主题的关键词，有必要增加主题数来达到细分主题的目的。

关键词 语言特征，主成分分析，主题模型

1. 研究目的

程小青（1893-1976）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として名をあげており、さらにシャーロック・ホームズの登場人物設定に倣って著名な霍桑シリーズを創作し、20世紀における中国のミステリー小説の発展に貢献している。鬼馬星(1972-2017 本名：马雨默)は21世紀初頭における中国の代表的な女性ミステリー作家であり、莫兰や简东平を主人公とする有名なミステリーシリーズを創作した。

黄（2020）では中国のミステリー作家4人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、時代やモチーフの設定によるトピックの相違を明らかにしたが、品詞別にトピックを検討する試みはしていない。本研究は程小青と鬼馬星のミステリー作品を扱い、まずは主成分分析を通して、頻度差による作家間の言語差異を明らかにしたい。さらに、LDAを実行し名詞トピックを解析することで新たな情報を見つけ出すことを目的とする。

2. データ

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、程小青の霍桑シリーズの8作品と鬼馬星の中長編推理小説26作品の計34点である。程小青の作品は努努书坊¹、鬼馬星の作品は52书库²よりダウンロードした。

¹ <https://www.kanunu8.com/files/writer/9824.html>

² <https://www.52shuku.me/zuozhe/guimaxing/>

表 1: 作品情報一覧

作者	作品名	ラベル	出版年	総語数	異なり語数
程小青	白衣怪	C_baiyi	1946	54,998	8,270
	催命符	C_cuiming	1946	43,283	7,024
	断指团	C_duanzhi	1946	27,928	6,598
	矛盾圈	C_maodun	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_molan_fengde	2006	224,193	16,761
	被害人姓名: 倪震	G_beihairen	2008	79836	9362
	纽扣杀人案	G_jian_niukou	2008	110,478	11,218
	猫的复生	G_molan_maode	2008	79,283	9,819
	宴无好宴	G_molan_yanwu	2008	126958	12504
	暮眼蝶	G_jian_muyan	2009	129,080	12,325
	淑女之家	G_jian_shunv	2009	96,840	10,642
	黑背鱼之谜	G_kufayi_heibeiyu	2009	58836	7993
	木锡镇	G_kufayi_muxi	2009	53251	6979
	幽灵船	G_kufayi_youling	2009	107075	11099
	迷宫蛛	G_lujing_migong	2009	263,546	18,490
	八鲜汤	G_molan_baxian	2009	20754	4366
	葬礼之后的葬礼	G_molan_zangli	2009	155,355	13,601
	蔷薇犯罪事件	G_qiangwei	2009	78039	9165
	魔幻图书馆	G_shaonian_mohuan	2010	45441	6736
	百合花房秘语	G_shaonv_baihe	2010	52312	7023
	演出	G_shaonv_yanchu	2010	72181	8601
	犀牛旅社	G_kufayi_xiniu	2011	71196	8474
	少年理发师	G_shaonian_lifa	2011	45771	6298
	蓝戒之谜	G_shaonv_lanjie	2011	63740	7932
	笼鸟	G_lujing_longniao	2012	112820	11460
	雾锁天途	G_shaonv_wusuo	2012	88227	10191
	虫屋	G_kufayi_chongwu	2013	60805	7654
	隔墙玫瑰	G_molan_geqiang	2013	118553	11905
	被偷走的秘密	G_beitouzou	2014	88134	9670
	喜悦岛	G_xiyuedao	2015	110857	11766

中国語テキストを処理する際には、分かち書きが非常に重要な一環である。アメリカのスタンフォード大学により開発された Stanford CoreNLP という分かち書きツールを使い、注釈タグや作品紹介などの情報を取り除いたプレインテキストに対して、プログラミング言語 python

で分かち書きと品詞タグを付与する加工を行う。品詞のタグセットは Chinese Penn Treebank part-of-speech tagset³に基づいたものとされている。

以下に示す文は分かち書きと品詞付与を施したテキストの例である。分析時に使用するデータはこのように分かち書きと品詞付与が行われたものである。

十月_NT 二十三日_NT , _PU 傍晚_NT 五点钟_NT 光景_NN , _PU 我_PN 忽_AD 接到_VV 我_PN 的_DEG 老友_NN 霍桑_NN 打来_VV 的_DEC 一_CD 个_M 看似_VV 轻松_VA 滑稽_VA 的_DEC 电话_NN 。_PU “_PU 包朗_VV , _PU 今夜_NT 你_PN 如果_CS 没有_VE 旁_LC 的_DEG 紧要_JJ 事_NN , _PU 请_VV 向_P 尊_NN 夫人_NN 请_VV 两_CD 小时_NN 假_NN , _PU 到_VV 我_PN 这里_PN 来_VV 走_VV 一_CD 趟_M 。_PU 我_PN 有_VE 一_CD 种_M 奇怪_JJ 的_DEG 东西_NN 给_VV 你瞧_NN 。_PU ”_PU

(程小青の『催命符』の冒頭部分)

3. 主成分分析

3.1 代名詞

本研究は、量的手法を用いて程小青と鬼馬星の作品における言語特徴を捉えることを試み、特に品詞別の語彙分析に重点を置く。

表2は34点の作品それぞれにおける代名詞（_PN タグが付く単語）の生起度を表す語彙頻度（テキスト 1000 語あたりの相対頻度）プロファイルの一部である。105 タイプもあり目視で分析することは容易ではないため、主成分分析を実行し、視覚化したデータをもとに分析を行う。

主成分分析とは、多変数を持つデータの背後に隠れた特徴を明らかにするための教師なし学習の手法である。具体的には、情報をできるだけ失うことなく変数の数を減らし、次元の削減を実現する。元のデータを最も多く保存する主成分を第1主成分といい、次に高い主成分を第2主成分と定義している。主成分分析の結果は、第1主成分と第2主成分の値を使った2次元散布図などの形式で可視化される。データ解析の方法として、文体研究における主成分分析の有効性は田畑（2012）や土山（2019）で認められた。そのため、本節では主成分分析を実行した結果をもとに考察を試みる。

³ <https://www.sketchengine.eu/chinese-penn-treebank-part-of-speech-tagset/>

表 2: 代名詞の語彙頻度プロファイル

作品名	総語数	他_pn	我_pn	她_pn	你_pn	我们_pn	自己_pn	他们_pn	这_pn	什么_pn	
総頻度	3557768	82597	75351	71099	47934	8525	8407	8252	7686	7120	
C_催命符	57096	21.420064	24.765308		0	14.116576	4.2735043	0.9983186	0.8757181	2.2243239	0.9632899
C_断指团	37271	17.81546	30.720936	0.0268305	12.342035	5.2051193	0.6171018	3.4879665	2.1732714	0.7512543	
C_白衣怪	71498	20.657921	25.553162	0.0279728	11.818512	3.5665333	0.7412795	0.7552659	2.13992	0.7552659	
C_矛盾圈	57151	17.025074	28.031006		0	12.335742	4.2168991	0.9273679	2.3271684	2.3971584	1.1198404
C_紫信箋	36302	21.321139	20.549832	0.0275467	7.9334472	5.5644317	0.7437607	1.2396011	2.4516556	0.5784805	
C_舞后的归宿	82484	19.434072	25.835314	0.0121236	11.020319	2.6550604	0.9941322	0.9335144	2.5701954	1.2850977	
C_霜刃碧血	46923	15.578714	25.701681		0	11.231166	2.9196769	0.7245914	1.001641	2.8344309	1.12951
C_青春之火	35574	17.934447	22.741328		0	10.06353	3.9073481	1.0963063	0.8152021	2.8672626	0.8995334
G_喜悦岛	146369	26.330712	22.197323	19.232214	10.733147	3.0265972	1.9676298	3.5048405	2.5278577	2.1862553	
G_少女莫兰_演出	94534	20.331309	40.747244	23.991368	14.746017	3.1946178	2.0310153	2.4012525	2.0627499	2.2531576	
G_少女莫兰_百合花房秘语	69316	2.6256564	15.912632	37.06215	15.234578	2.9286168	3.0007502	1.9331756	2.3226961	2.6977898	
G_少女莫兰_蓝戒之谜	82878	19.088298	16.530322	22.877	12.47617	2.3649219	2.3166582	2.6786361	1.9667463	2.4252516	
G_少女莫兰_雾锁天途	115689	23.044542	19.310393	17.15807	15.489805	1.970801	2.4980767	3.1377227	1.8584308	2.4980767	
G_少年_少年理发师	62287	15.492799	14.288696	19.939955	8.4447798	5.3622746	2.1673865	4.2384446	2.3118789	2.4403166	
G_少年_魔幻图书馆	61767	25.580002	16.513672	7.2530639	11.219583	3.1570256	2.4446711	7.0102158	2.201823	2.6551395	
G_简东平_暮眼蝶	165869	25.707034	18.773852	16.57332	10.954428	2.6587247	2.9300231	2.9179654	2.5260899	2.1221566	
G_简东平_淑女之家	131499	19.483038	21.589518	26.045825	13.376528	1.4752964	2.8441281	1.262367	1.7262489	2.0456429	
G_简东平_纽扣杀人案	146116	25.726135	21.907252	23.022804	14.947028	1.4372143	2.9086479	1.7451888	2.1968847	2.0531632	
G_莫兰_人骨八鲜汤	27879	9.074931	15.280319	29.520428	12.554252	3.4793214	2.6184583	0.5021701	1.5782489	2.1521575	
G_莫兰_宴无好宴	168118	30.002736	20.533197	19.569588	14.983523	2.2662654	2.3495402	2.3316956	1.9093732	1.8915286	
G_莫兰_猫的复生	106345	19.643613	12.873196	33.410127	9.9205416	1.4951338	2.8022004	2.5106963	2.0687385	2.2191923	
G_莫兰_葬礼之后的葬礼	215145	22.88689	22.617305	26.665737	16.398243	1.7662507	3.4395408	1.3711683	2.0776685	1.9661159	
G_莫兰_隔墙玫瑰	156663	17.783395	20.2026	29.228344	14.636513	1.6851458	2.5149525	2.7128294	2.0681335	2.4000562	
G_莫兰_风的预谋	84554	31.210824	17.846583	20.259243	13.790004	0.8397001	3.2641862	1.3009438	2.2943917	1.8213213	
G_蔷薇犯罪事件	102282	23.542754	27.541503	24.06093	13.130365	1.7793942	2.7277527	1.6327409	1.798948	1.9358245	
G_被偷走的秘密	118332	23.129838	17.493155	28.808775	15.169185	1.8169219	1.9859379	2.7296082	2.6789034	1.8507251	
G_被害人姓名倪震	107038	21.497038	16.844485	22.122984	12.967357	2.1113997	2.4103589	2.0179749	1.6909882	1.8311254	
G_酷法医_幽灵船	143799	20.702508	15.765061	19.311678	13.004263	2.0167039	2.5174028	2.3087782	1.7594003	1.7733086	
G_酷法医_木锡镇	71025	16.55755	45.124956	15.881732	11.277719	2.717353	2.0556142	2.3794439	2.1400915	2.9426258	
G_酷法医_犀牛旅社	95926	23.413882	17.805392	24.643996	11.998832	2.908492	1.8660217	2.5019286	2.2830098	2.1579134	
G_酷法医_虫屋	81577	17.002341	25.092857	39.055126	15.200363	2.231021	2.2065043	2.1206958	2.5619966	2.574255	
G_酷法医_黑背鱼之谜	78161	31.319968	16.043807	16.018219	12.896457	1.7272041	2.2773506	2.2773506	2.1877919	1.7783805	
G_陆幼_笼鸟	152343	36.667257	18.038243	17.933216	15.727667	2.5993974	2.2843189	3.3017598	2.1070873	2.1333438	
G_陆幼_迷官蛛	347958	29.247783	19.809862	20.005288	15.251841	2.0059892	2.7532058	2.3623541	2.1381891	2.1123239	

統計解析ソフト R によって、主成分分析の実行および散布図の作画を行う。最も寄与率⁴の高い第 1 主成分 (55.87%) と第 2 主成分 (14.41%) の得点を二次元散布図に配置したものが図 1 と図 2 である。

図 1 はテキスト間の関係を可視化した図である。図 1 では、程小青の作品ラベル (C_から始まるもの) を青色、鬼馬星の作品ラベル (G_から始まるもの) を赤色でマークしている。変数の頻度パターンの近いテキストが近く布置され、頻度パターンの異なるテキスト同士は遠く布置される。第 1 主成分、つまり横軸の特徴を見ればわかるように、程小青の作品が左側に偏っているのに対して、鬼馬星の作品は右側に集中している。作家間の個人差が非常に鮮明だと考えられる。縦軸においては、作品間、特に鬼馬星の作品においてずれがみられるが、特徴といえるほどの結果は出ていない。

⁴ 全情報量のうち、該当する主成分が占める情報量の割合のことである。

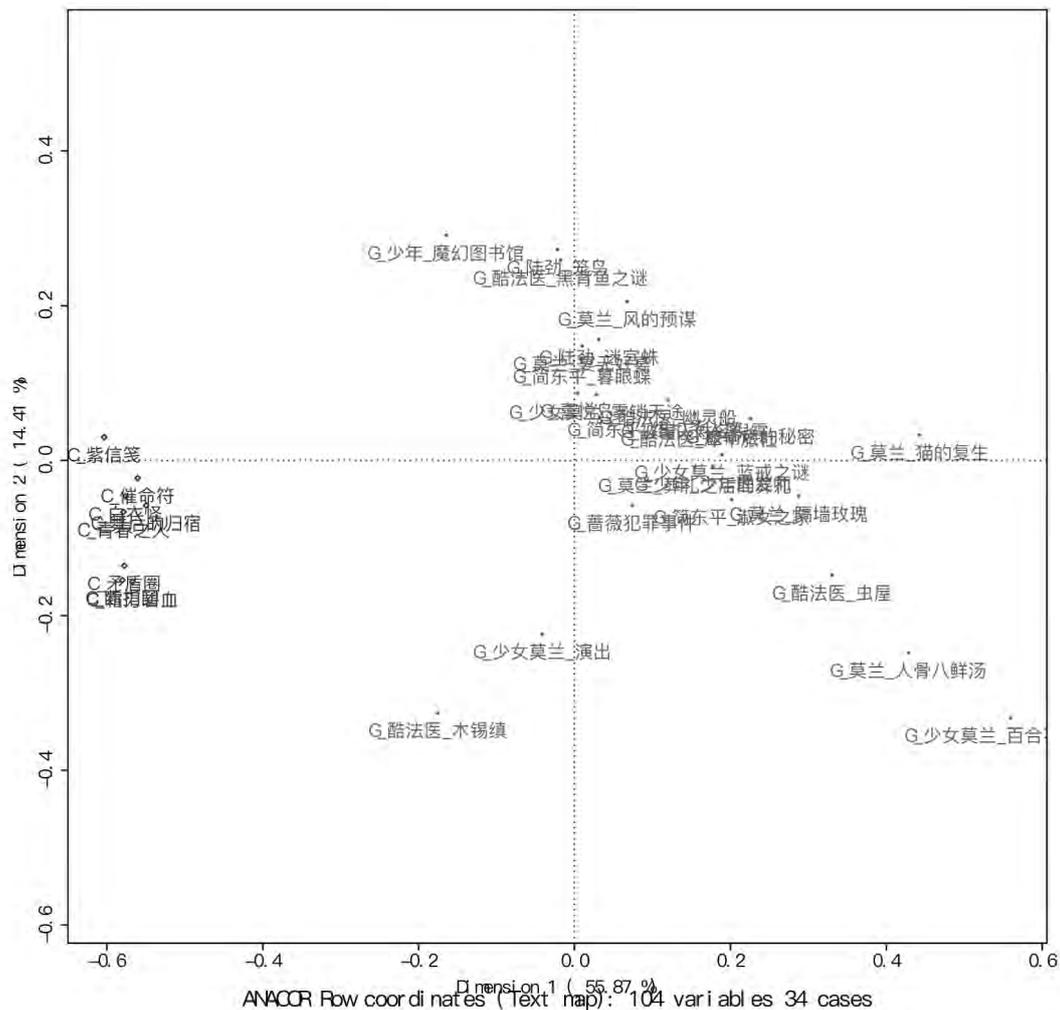


図 1: テキスト間の相互関係

図 2 は単語間の関係を可視化した図である。赤いサークルは単語それぞれのウェイトを具現化したものである。単語の重みが大きいかほどサークルが大きく表示されている。図 1 と照らし合わせて見ればわかるように、横軸では大概-0.4あたりが両作家の境界線であり、図 2 において左半分に出現した単語は程小青を代表し、右半分の単語は鬼馬星を特徴づけると言えよう。図 2 の横軸の単語分布を観察すると、「我」（私）、「我们」（私達）が真ん中の大きなサークルの中心に位置しており、どちらにも平均的に多用されていることが分かる。「彼女（ら）」を意味する「她」「她们」が図 2 の右側に寄せており、鬼馬星の作品に特徴的に出現する傾向がある。特に「她」（彼女）が右側のサークルの中心にあり、頻度的には鬼馬星にきわめて愛用されていると解釈してよいだろう。

回目の彼女は「她」で表現されている。汉语大辞典⁵によると、「伊」は三人称、特に女性を指す古めな言い方である。つまり、程小青は異なる三人称の言い方を使用し、「她」をめぐる差異は言葉遣いの相違に起因しているのではないかと考えられる。Stanford CoreNLP によって「伊」は NR、つまり固有名詞というタグで付与されており、「伊」の使用に差がみられるか否かを気に留めながら、名詞の特徴を探っていく。

3.2 名詞

図3と図4は名詞を主成分分析で視覚化したものである。名詞のタイプ数は7万以上あり、全てインポートすると単語ラベルが混雑しすぎて分析することが困難になるため、分析対象を上位200タイプまでに限定している。

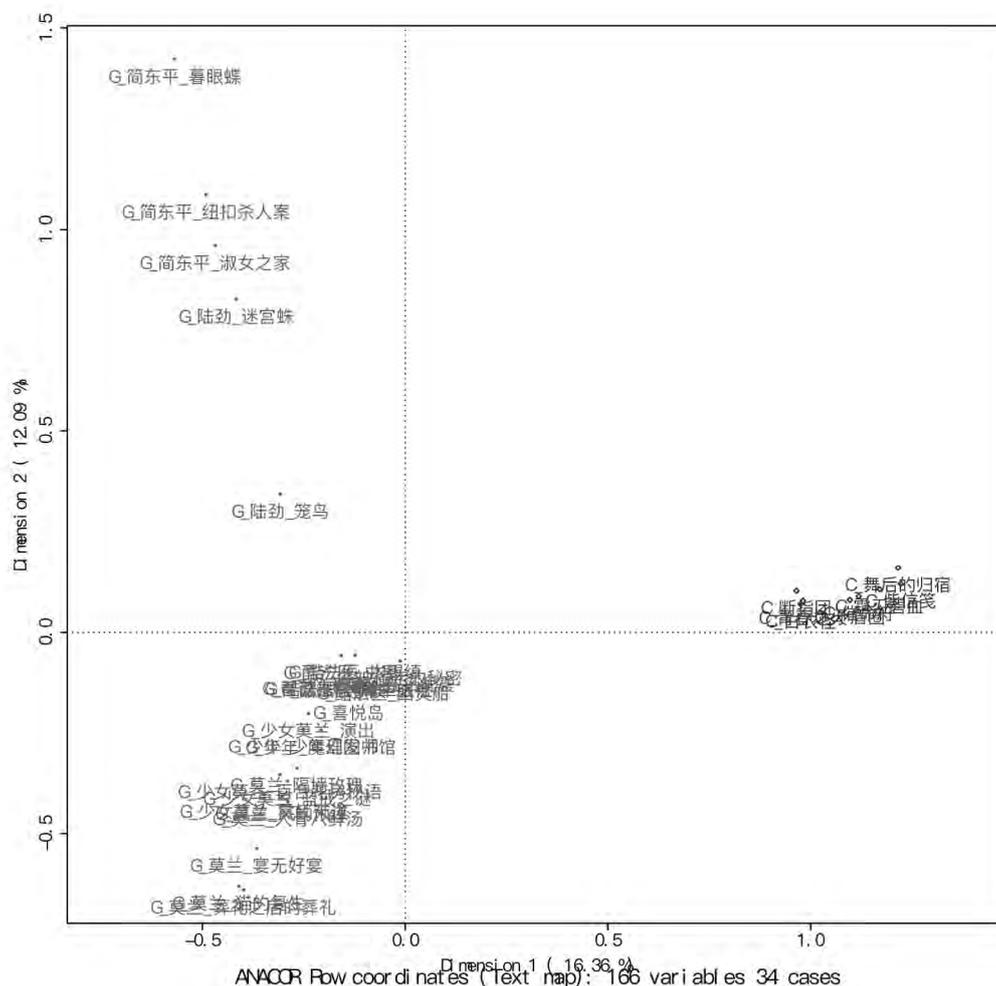


図3：テキスト間の相互関係

図3からわかるように、第1次元では、鬼馬星の作品と程小青の作品は左右明確に分かれている。代名詞の分布と同様に、名詞の言葉遣いも作家によって相当異なっている。第2次元

⁵ <http://www.hydcdd.com/zidian/hz/2798.htm>

において、右半分の程小青の作品は中央あたりに配置され分散が非常に小さい。それに対して、左半分の鬼馬星の作品では差異を見せている。上部に簡东平と陆劲を主人公とした作品、下部に莫兰などを主人公とした作品が布置されている。鬼馬星の作品は縦軸の上下にシリーズで分かれていることが読み取れる。

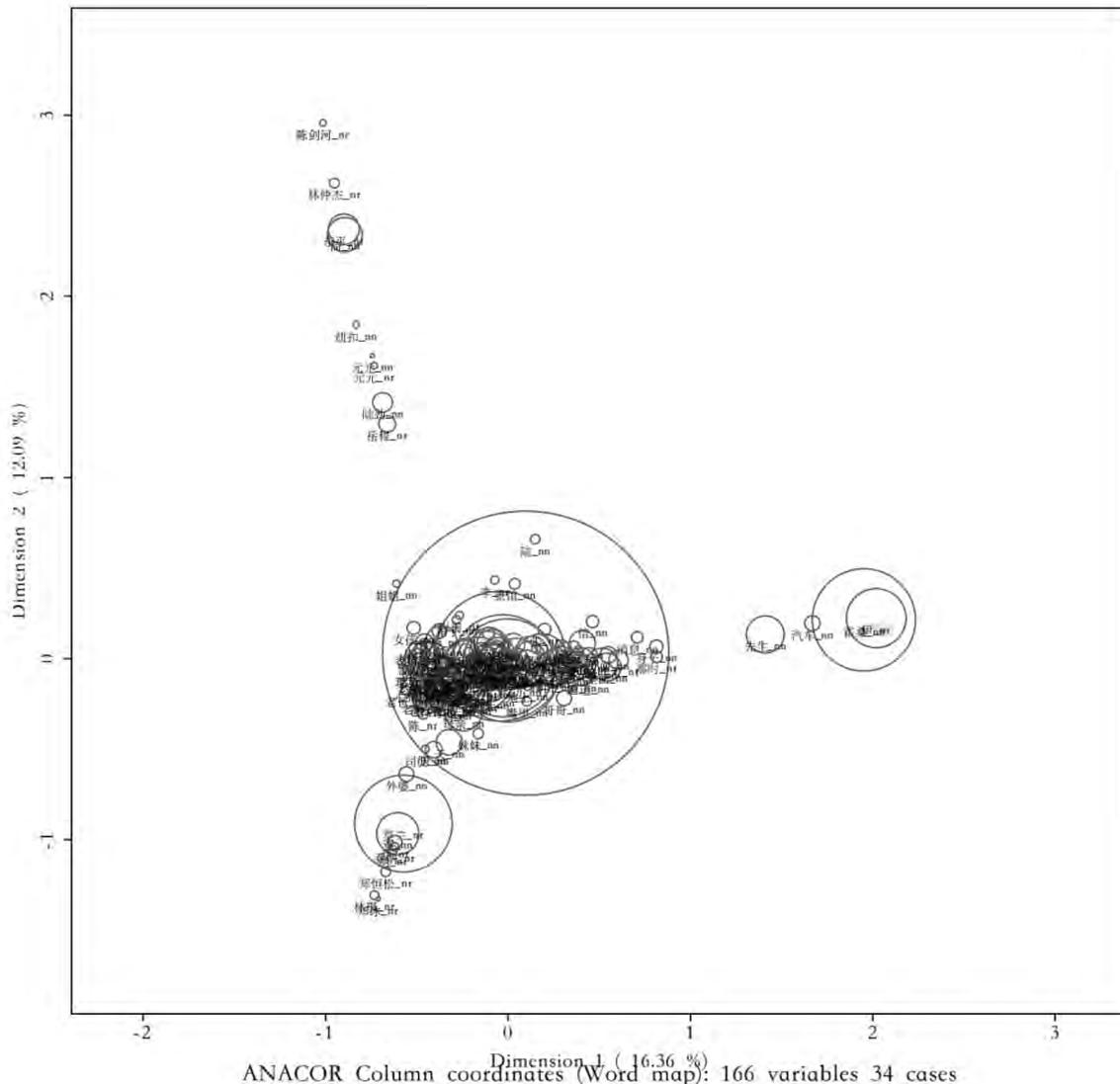


図4：名詞の相互関係

図4の横軸を観察すると、「伊」が最も右側に位置しており、程小青の作品の特徴としてあげられた。4.1の結果と合わせて分析すると、「伊」と「她」の使いが両極化していることがわかる。「彼女」を示す三人称の言い方は相当時代の影響を受けて変化しているのではないかと考えられる。また、「霍桑」という主人公の名前を除き、「先生」（男性の名前の後につける敬称）、「汽车」（くるま）も程小青側の特徴として挙げられている。「汽车」を愛用することは黄（2020）のトピック解析でも明らかにしたことであり、その裏付けにもなると思われる。

程小青の作品を特徴づける「先生」という語を図5のコンコーダンスで吟味すると、ほぼ会話文にしか出現していないことはまず明らかである。ニュアンス的には両作家に相違が感じられる。程小青の作品において、調査などを行う主人公たちに対して、話し手がやり取りの中で「先生」をよく用いることがコンテキストからわかる。これは主人公の身分に対して尊敬の意を持った呼び方だと考えられる。その一方で、鬼馬星の作品に出てくる「先生」は、コンテキストを観察すると、ただ男性に対する正式な呼び方だと思われ、特に身分も知らずに呼ぶ例が多い。

C_舞后的归宿	伊问我说：“	先生	，什么事呀？”
C_催命符	“那也决不会的。	先生	，他上楼去干什么事？我老实说，他是怕我哥哥的。
C_舞后的归宿	“当然不知。我问他可要叫汽车，他也只摇摇头，不说一句话。	先生	，他到底干的什么事？我早就疑心他。”
C_舞后的归宿	“昨天半夜以后，大概是一点钟光景，所以就算今天也可以。	先生	，他是什么样人？我们也觉他很奇怪。他干了什么事？”
C_霏刃碧血	阿三感激地说：“	先生	，以后我一定不敢。
C_舞后的归宿	“	先生	，你——你——“他分明要问我在他房里做什么，却因着有所顾忌而不敢直截地说出来。
C_紫信箋	“他断续地说：‘我——我怕他不——不像醉啊！我——我怕得很！	先生	，你——你自己——’
C_舞后的归宿	老毛想了一想，答道：“那还是去年年底——	先生	，你不必疑心，这不会假。
C_霏刃碧血	伊侧过脸瞧瞧我，半信半疑地答道：“	先生	，你不要骗我，我——我——”
C_矛盾圈	“伊就在我们隔壁第六号王家里啊。不过伊已经走了，	先生	，你为什么要找伊？”
C_舞后的归宿	。回来时雨大透了，这双鞋子便完全浸湿。	先生	，你为什么问到我的鞋子？”
C_催命符	甘东坪答道：“不吸的。我们家里只有订菸吸纸烟。	先生	，你为什么问到这个？”
C_舞后的归宿	“我一个人啊——	先生	，你为什么问得这样仔细？莫非当真疑心我——”
G_黑背鱼之谜	“	先生	，你是警察吗？我是这个商场保安部的负责人，我姓赵，请出示一下证件好吗？”
G_少年理发师	保险经纪似乎也被五叔的气势吓住了，“先生，	先生	，先别急，先别急……”
G_少年理发师	“	先生	，我不知道你是谁，“电话一通，就传来保险经纪烦躁的声音，“但我要跟你弄清楚，
G_幽灵船	“	先生	，有什么需要？”
G_披偷走的秘密	“	先生	，请问你怎么称呼？”她跟着他下楼。
G_暮眼蝶	“你好，	先生	。请问你是要住宿吗？我们这里既可以整夜住宿，也有钟点房，请问你需要哪一种。”

図5：「先生」のコンコーダンス

また、図4の縦軸を観察すれば、上と下に極端に分散している特徴語はほぼ人名でしかなく、大概の名詞は中心部に寄せられていることがわかる。固有名詞以外の名詞ではそれほどの差が見られない。

主成分分析の分布図で、個々の上位単語の特徴を見つけることはできるが、頻度が上位でないが特徴をもつ単語、またはグループ化した特徴を見極めることは難しいと心得ている。そのため、トピックモデリング手法を使って名詞のみから構成されたトピックを解析することによって、より深まった考察ができることを望む。

4. トピック解析

Jockers and Mimno(2013)、田畑(2017)、黒田(2017)、黄(2020)などは文学研究におけるLDAトピックモデルの応用の有効性と意義を証明した。この節では、LDAに基づくトピックモデルを実行し、名詞のみに焦点を当て程小青と鬼馬星の作品をトピックという視点から解析してみる。

前節の主成分分析では全ての名詞を対象としたが、結果からも明らかなように、人名のような固有名詞が特徴語として挙げられている。トピックにおいても、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してしまうケースが生じやすい。それをできる限り防ぐため、固有名詞などを表すタグを取り除き、普通名詞を代表するNNのタグが付いた単語のみを対象

とする。また、対象作品計 34 点をそれぞれ 500 語ごとに切り分け、総じて 975 個のファイルを作成した。黄 (2020) ではテキストファイルを 1000 語ごとに切り分けたが、本研究では名詞のみを対象として扱うためテキストファイルをより細かく切り分ける必要があると考えられる。実際の分析はこの 975 個のファイルに対して行う。トピック数は 50 と設定し、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) を実装して LDA を実行する。

また、統計解析ソフト R を用いてワードクラウドやジッター図、ヒートマップなどの可視化グラフを作成し、トピックと単語、またはトピックとファイルの関係性を考察することができる。

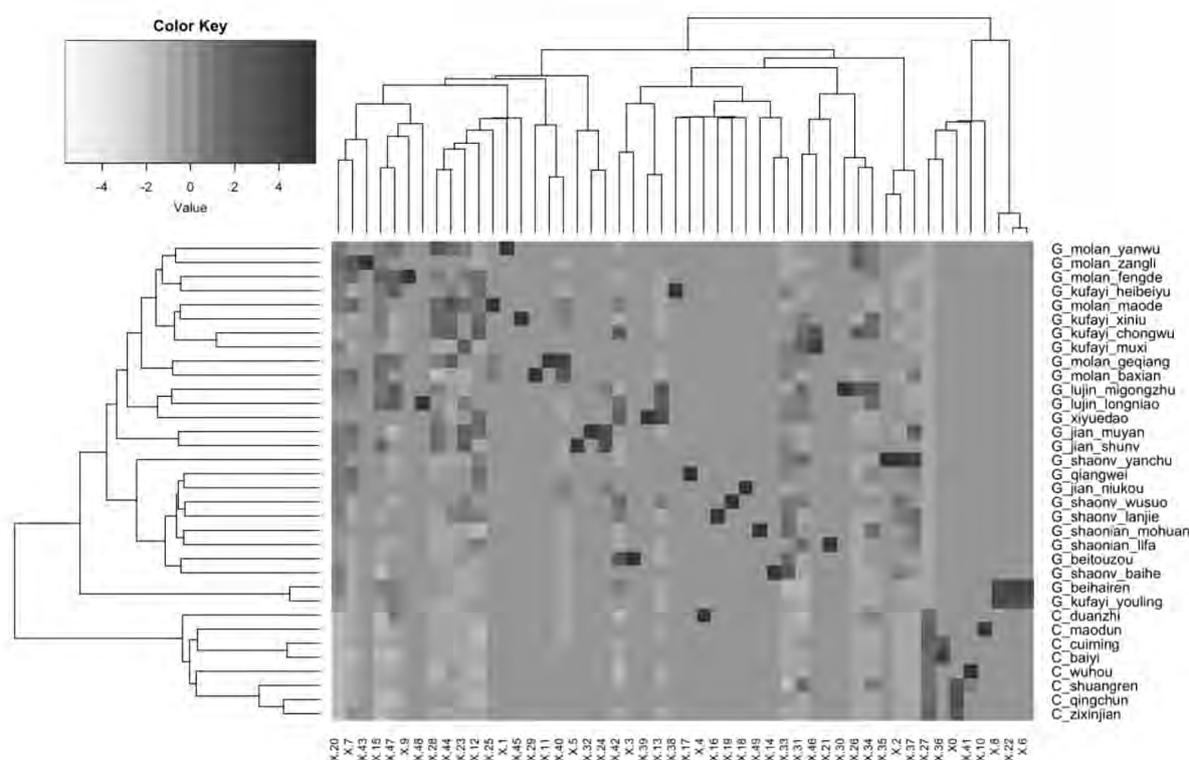


図 6：ヒートマップ⁶

図 6 はトピックと作品の関係を表すヒートマップである。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。作品におけるトピックの割合が高いほど、当該セルの色が濃くなる。横軸にトピック、縦軸に作品を配置し、トピック間、さらに作品間の関係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キストやテキ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、似た特徴を備えた作品は近接し、類似しない作品は遠く配置される。G_から始まる作品群（鬼馬星）と下側の C_から始まる作品群（程小

⁶ 作画時に中国語の文字化けを防ぐため、縦軸で扱う作品のラベルは作品名ではなく、表 1 に示したラベルを使用している。横軸ではトピックの数字を表している。

青) がトピック構造の相違によって明確に分岐していることは図6から判断できる。

ヒートマップはトピックと作品の関係性を見極めることができるが、トピックと単語の関係を俯瞰するためには、ネットワーク図が必要となる。

図7は、視覚化アプリケーションの Gephi によって作成されたネットワーク図であり、50個のトピックとそれらを代表する単語を上位1000語まで配置している。数字はトピックを表しており、トピックを構成する単語と数字の間はエッジで結びついている。エッジの太さはトピックにおける単語の重みと関連しており、重みが大きいほどエッジが太くなる。数字の大きさもトピックの大きさと対応している。例えば、トピック20は最も大きいトピックだと容易に読み取れる。また、共通する項目の多いトピックは近くに配置されている。



図7：ネットワーク図

語構成から見れば、図7の左上側の20、27と47はボディーランゲージに関わるトピックであり、共通項目も多く非常に類似していると言える。ただし、ファイルにおけるトピックの分布を観察すると、非常に興味深い結果が見られた。

文献

- [1] 程海燕 (2015) 「论程小青侦探小说的本土化」 修士論文 安徽大学
- [2] 黄晨雯 (2020) 「中国のミステリー小説におけるトピック解析の試み」 『言語文化学』 1-17
- [3] 周楠 (2015) 「近代侦探小说中的都市元素研究」 修士論文 上海師範大学
- [4] クジラ飛行机 (2016) 『Python によるスクレイピング&機械学習[開発テクニック]』 ソシム株式会社
- [5] 黒田絢香 (2017) 「小説テキストの計量的分析 アーサー・コナン・ドイルの作品から」 『言語文化共同研究プロジェクト 2016』 23-41
- [6] 小林雄一郎 (2017) 『R による優しいテキストマイニング：機械学習編』 オーム社
- [7] 田畑智司 (2012) 「テキストマイニングからテキスト分析へ：Collins との共著作品における Dickens の文体」 『電子化言語資料分析研究 2011-2012』 3-17
- [8] 田畑智司 (2017) 「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」 『言語文化共同研究プロジェクト 2016』 5-21
- [9] 土山, 玄 (2019) 「森鷗外の文体的特徴の変化に関する計量的な考察」 『人文・自然研究』 13:107-105
- [10] Blei, D.M., Ng, A. and Jordan, M. (2003) Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 3:993-1022.
- [11] Blei, D.M. (2012) Probabilistic Topic Models. *Communications of the ACM*, 55(4): 77-84.
- [12] Graham, S. and Milligan, I. (2012) Review of MALLETT, produced by Andrew Kachites McCallum, *Journal of Digital Humanities*, 2(1).
- [13] Hofmann, T. (1999) Probabilistic Latent Semantic Analysis, *UAI'99 Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 289-296
- [14] Jockers, M. and Mimno, D. (2013) Significant themes in 19th-century literature, *Poetics* 41: 750-769