計量分析による村上春樹文学の語彙構成と歴史的変遷

Lexical Formation and the Historical Shifts within the Works of Haruki Murakami Identified using Quantitative Analysis

工藤彰^{1*}, 村井源¹, 往住彰文¹

Akira KUDO, Hajime MURAI, Akifumi TOKOSUMI

*1 東京工業大学大学院社会理工学研究科

Graduate School of Decision Science and Technology, Tokyo Institute of Technology

〒152-8552 東京都目黒区大岡山2-21-1

E-mail: {akudo,h_murai,akt}@valdes.titech.ac.jp

近年、文学や音楽、美術、映画、漫画等のさまざまな芸術を研究するにあたって、微細なデータを扱い、一貫性を有した科学的分析手法に基づいたフィクションの研究に注目が集まっている。本研究では、文学テキストの表層に見られる語彙を作家の文体的特性が宿る最小の構成要素とみなし、作風の変遷を計量的に明らかにすることを目的としている。対象としたデータは、現代の日本文学を代表する小説家村上春樹の長篇とし、そのテキスト中から抽出した語彙を品詞と意味カテゴリーに分類したのち、それぞれクラスター分析を用いて近似的な作風を持つ作品群を明示化した。その結果、品詞分類からは「初期三部作」をはじめとした通時的な区分によってのクラスターが、また意味分類からは「初期三部作」の続篇とも呼ばれる一作品を含んだ「鼠四部作」のクラスターが形成された。

Similar to the research trends within various artistic realms, such as art, literature, music, movies, and animation, in recent years, fiction studies are increasingly employing more sophisticated techniques of computational analysis. The aims of the present study are to quantitatively analyze the surface vocabulary of a literary text into the minimum components that reflect the writer's characteristic style and to identify style transitions. The target data consists of the full-length works of Haruki Murakami, as a representative of modern Japanese literature. The works are grouped in terms of writing styles through a clustering analysis of the textual vocabulary which was initially classified according to both word class and semantic categories. While the analysis results based on word-class categories yielded cluster that reflect a diachronic division, such as an “early trilogy”, the results based on semantic categories yielded a cluster for the “Nezumi Tetrabiblos” that adds another work which is referred to as a sequel to the “early trilogy”.

キーワード: フィクション, 文学, 村上春樹, 計量書誌学, クラスター分析

Fiction, literature, Haruki Murakami, bibliometrics, cluster analysis

1 背景

これまで文学や音楽、美術、映画、漫画をはじめとした諸芸術の意味や解釈の周囲で、さまざまな研究や批評が展開されてきた。近年、中でもとりわけフィクションをめぐる言説は加速の一途をたどっており、文学研究にとどまらず記号論や表象文化論、社会学、美学、言語哲学、分析哲学あるいは精神分析等の多様な学問領域から議論がなされている。認知科学の分野において、フィクションとは、人間である作家の思考概念や創造力の表出と捉えることができ、数理的ないし統計的手法の進展にともない、微細なデータを網羅的に処理し一貫性を有した分析手法で、対象の潜在的な特性や傾向を客観的に顕在化させる記述を可能にしはじめている。そこで信頼性の高い客観的手法が確立できれば、今後、諸芸術におけるフィクションの読解技術への汎用性や新たな方法論の開発、そして最終的には人工知能への搭載をも視野に入れることができる。その実現を目指すべく、まずは作家の直接的な創造力の産物であるテキストをデータとして扱うことができる文学に焦点を当てたい。

さらに、従来の文芸批評や文学研究が言及するのが困難であった文学の周縁的な領域から、テキストの新たな意味や解釈を計算し創出するための応用可能な環境を設計するためにも、具体的に、一人の作家が発表した複数のテキストを、比較検討できるサンプルとして用意する必要がある。また実際に、コンピュータによる文学の解析として、物語構造から意味のあるまとまりを分割し抽出する研究が進んでいる[1]。

以上を踏まえた上で、本研究の目的は、文学テキスト上に見られる語彙を、認知科学的視点に立脚した作家の文体的特性が宿る、データとしての最小の構成要素とみなし、統計的处理によって算出された意味概念の総和やネットワークをもとに文学を捉え、フィクションの計算処理技術を向上させる長期的展望を視野に入れた試みの一環にあるものとする。

2 村上春樹

本研究では、日本の現代文学を代表する小説家村上春樹の長篇を扱う。村上春樹という小説家は物語と文体に対して意欲的に取り組んできたことで知られる作家である。社会や時代の流行や世相を直接的にあるいは間接的に反映させてきたそのテキスト群は、日本の80年代以降の純文学シーンを代表する小説となっている。国内外での受賞歴も数多く、谷崎潤一郎賞や読売文学賞といった文学的に極めて評価の高い国内の文学賞だけでなく、フランツ・カフカ賞やフランク・オコナー短篇賞、エルサレム賞の受賞等の海外での評価や知名度も獲得している。現存する日本の作家では例外的に世界中で流通する文学性を持ちえた作家といえる。また昨今では、国内外問わず彼の小説に影響を受けた若手の作家も目立っている。さらには新作発表ごとに文芸批評のみならず、一般読者の膨大な量の感想がネット上に生まれており、村上の社会的認知や話題性は極めて高い。これらの理由から、村上春樹という作家の小説を、現在の日本文学を牽引する重要なテキストとみなし研究対象として採用することにした。

また作家の思考概念や創造力は当然作品

を産み出す行為の最中で流動的に変容し、それにふさわしいと決定した言葉、あるいは確率論的に導かれた言葉がテキストとして書き連ねられることになる。本研究では、作家がいかに関時的推移の中で言語体系を変化させているのか、その進行中の過程に視線を注ぐべく、通時的に比較可能とする手段として、処女作から8作目までの長篇を一作単位で扱うことにした。短篇やエッセイと比較した際に、長篇は、村上春樹自身、主戦場と位置づけており、本人の特性が最も色濃くあらわれるテキストだといえる。

本研究で8作目までの長篇を取り上げる理由としては、村上春樹という小説家の大きな作風の変化は、オウム事件や阪神大震災による90年代の転向にあるといわれるが[2]、それ以前の時期の作品群と限定し詳細に検討することによって、比較的作風の変化が緩やかに思われる前期作品群の中であっても、村上春樹のフィクションを記述する語彙が、どのようにテキスト上で変貌を遂げていったのか明白にするためである。

一般的に文芸批評で検討されることが多い主題や物語等の大きな枠組みのパラダイムの背後で、いかに微細な言葉がテキストの表象で相対的な遷移を見せるのか、それを科学的に確固たる数値として示し歴然たる所与として吟味することには大きな意義があるだろう。

3 データ

まずは、以下に対象とする8作品の基本的な統計的数値の詳細をあげる。形態素解析エンジンMeCabを使用し、テキストを語彙に分割後、さまざまな観点から表記上の集合として捉え処理し、新たに算出し直した。

以下、表と図から見てとれるように、小説

の一文の長さを語数と字数それぞれで調べた結果、平均して15語、25字前後で書かれていることがわかった。

また、日本語の使われ方に関していえば、ひらがなが最も多く増加傾向にあり、漢字、カタカナ、記号は減少の傾向を示した。これは平易でシンプルな文体を使って書きたいという村上春樹自身の発言[3]を裏付けるものである。

TTRは、テキスト自体の語数を踏まえたとしても、総じて年々減少しており、同じ語彙を繰り返し使用する文体に移行している。

また台詞率は、2、5作目を除きほぼ40%前後を維持している。2作目が際立って少ないことについては登場人物の内面まで描いた文体[4]が影響し、また5作目は入れ替わり立ち替わり現れる登場人物の多さがそのまま台詞の多さにつながったと考えられる。

表1 文数や単語数のデータ

	文数	単語数	文字数	平均文長(語)	平均文長(字)
1 風の歌を聴け	2348	34508	54761	14.7	23.3
2 1973年のピンボール	3052	50783	81569	16.6	26.7
3 羊をめぐる冒険	8181	136484	215115	16.7	26.3
4 世界の終り	15105	258681	414663	17.1	27.5
5 ノルウェイの森	16231	185462	297464	16.5	26.5
6 ダンス・ダンス・ダンス	18093	254603	407933	14.1	22.5
7 国境の南、太陽の西	6462	106407	168221	16.5	26
8 ねじまき鳥クロニクル	23442	430092	694773	18.3	29.6

図1 平均文長（語）と平均文長（字）の推移

表2 日本語の表記比率

	漢字(%)	ひらがな(%)	カタカナ(%)	記号(%)
1 風の歌を聴け	23.91	55.63	9.38	9.48
2 1973年のピンボール	24.96	56.03	10.62	7.85
3 羊をめぐる冒険	24.61	61.48	6.3	7.19
4 世界の終り	22.18	64.74	6.13	6.5
5 ノルウェイの森	21.39	65.9	5.3	6.94
6 ダンス・ダンス・ダンス	22.42	62.22	7.37	7.49
7 国境の南、太陽の西	22.55	66.84	3.96	6.46
8 ねじまき鳥クロニクル	22.77	65.89	4.89	6.15

図2 日本語の表記比率推移グラフ

表3 TTRと台詞率

	語彙数	語彙数/語数(TTR)	台詞率(%)
1 風の歌を聴け	3917	0.114	39.58
2 1973年のピンボール	5297	0.104	21.52
3 羊をめぐる冒険	8265	0.061	37.58
4 世界の終り	10608	0.041	41.62
5 ノルウェイの森	8263	0.045	54.07
6 ダンス・ダンス・ダンス	9967	0.039	39.77
7 国境の南、太陽の西	5282	0.05	35.91
8 ねじまき鳥クロニクル	13502	0.031	37.47

図3 TTRの推移グラフ

台詞率(%)

図4 台詞率の推移グラフ

4 クラスタ分析

4.1 品詞分類

クラスタ分析に用いるため、名詞、動詞、形容詞、副詞の4品詞を集計する。その目的は語彙が持っている最も基本的な品詞という属性をもとに通時的に比較することで、語彙の大きな変動を俯瞰的に捉えることにある。

表4 品詞の比率

	名詞	動詞	形容詞	副詞
1 風の歌を聴け	0.61	0.29	0.04	0.06
2 1973年のピンボール	0.63	0.28	0.04	0.05
3 羊をめぐる冒険	0.61	0.29	0.04	0.06
4 世界の終り	0.56	0.34	0.04	0.06
5 ノルウェイの森	0.58	0.31	0.04	0.07
6 ダンス・ダンス・ダンス	0.58	0.31	0.04	0.07
7 国境の南、太陽の西	0.59	0.30	0.04	0.07
8 ねじまき鳥クロニクル	0.60	0.30	0.04	0.07

4.2 意味分類

品詞分類同様クラスタ分析のデータとして、語彙を同義語や類義語といった、抽象度の高い意味の枠組みまで一步踏み込んでグルーピングしていくために、本研究は、テ

クスト中から最も作品の特徴やテーマが得られやすい名詞を選択し、『分類語彙表』を用いた。『分類語彙表』からは「人間活動の主体」「人間活動—精神および行為」部門に基づいて、16のカテゴリーに分類した。

表5 意味カテゴリーの比率

	人間	家族	仲間	人物	成員	公私	社会	機関
1 風の歌を聴け	0.32	0.02	0.05	0.01	0.19	0.03	0.03	0.01
2 1973年のピンボール	0.30	0.02	0.03	0.01	0.19	0.03	0.03	0.01
3 羊をめぐる冒険	0.28	0.02	0.04	0.02	0.19	0.04	0.03	0.01
4 世界の終り	0.34	0.02	0.03	0.01	0.09	0.08	0.03	0.01
5 ノルウェイの森	0.37	0.02	0.03	0.01	0.17	0.05	0.03	0.01
6 ダンス・ダンス・ダンス	0.31	0.02	0.04	0.01	0.23	0.02	0.03	0.01
7 国境の南、太陽の西	0.34	0.03	0.06	0.00	0.23	0.02	0.02	0.00
8 ねじまき鳥クロニクル	0.30	0.03	0.02	0.01	0.16	0.06	0.02	0.01

	心	言語	芸術	行為	交わり	待遇	経済	事業
1 風の歌を聴け	0.12	0.09	0.02	0.03	0.01	0.02	0.02	0.01
2 1973年のピンボール	0.14	0.10	0.02	0.04	0.01	0.02	0.02	0.03
3 羊をめぐる冒険	0.13	0.10	0.02	0.03	0.01	0.03	0.03	0.03
4 世界の終り	0.15	0.07	0.01	0.03	0.01	0.09	0.02	0.02
5 ノルウェイの森	0.11	0.08	0.02	0.02	0.01	0.05	0.01	0.01
6 ダンス・ダンス・ダンス	0.13	0.08	0.02	0.03	0.01	0.02	0.02	0.02
7 国境の南、太陽の西	0.11	0.07	0.02	0.03	0.01	0.02	0.02	0.01
8 ねじまき鳥クロニクル	0.15	0.09	0.01	0.03	0.01	0.05	0.02	0.02

4.3 品詞分類クラスター

品詞分類後のデータを用いクラスター分析を行った。可視的な分析手法であるクラスター分析を用いることによって、諸パラメータの一覧的なデータからは比較しにくい総体としての対象間の近似性を明らかにすることができる。クラスター化する方法としてはWard法を採用した。

結果、1, 2, 3の通称「初期三部作」、7, 8, また4, 5, 6と完全に年代ごとで集団傾向を示すクラスターが形成された。このことからまさに通時的な区分とともに村上春樹のテキスト文体には特徴があるということが

確認された。

図5 品詞分類クラスター

4.4 意味分類クラスター

次に意味カテゴリーにも同様の分析をしたところ、1, 6, 2, 3, また4, 8, そして5, 7のクラスターが見られた。1, 2, 3, 6の4作品は「鼠」というキャラクターが繰り返し登場することから通称「鼠四部作」とも呼ばれており、この実験で扱った意味カテゴリーの語彙で見ても、高い類似性を持つことが明らかとなった。また4作目『世界の終りとハードボイルド・ワンダーランド』と8作目『ねじまき鳥クロニクル』は章によって複数の語りをテキストに組み込んだり[5]、「地下」や「井戸」といった深くて暗い場所の想像力をメタファーとして題材に扱っているといった点が共通している。5作目『ノルウェイの森』と7作目『国境の南、太陽の西』は過去の女性との決して成就されることはない恋愛をめぐる物語である。これらが示すように、一連の続篇小説であったり似通った作風の小説が、抽象的な批評の言説のみならず、文体上の語彙レベルにおいても、幾つかの意味カテゴリーに分布されるという非常に興味深い結果が出た。

図6 意味分類クラスター

5 結論と課題

本研究は、フィクションの科学的解明を目指すべく、文学を対象に、村上春樹の小説テキスト8作品を実データとして、計量的に品詞と意味カテゴリーの二つの分類から、作風の近似性を明示化するためにクラスター分析をおこなった。結果、品詞分類からは「初期三部作」をはじめとした時代ごとの傾向を示すクラスターが、また意味分類からは「初期三部作」の続篇とも呼ばれる6作目の『ダンス・ダンス・ダンス』を含んだ「鼠四部作」のクラスターが形成された。これらの結果が示すのは、意識的であれ無意識的にであれ、作家の文体上の軌跡が、時代の変遷や続篇の作品群というクラスターを示しながら、確かに痕跡となって検出されたということである。

今後の展望であるが、引き続き文学を研究データとして村上春樹の長篇を扱っていく予定である。中でも本研究で触れていない後期の4作品およびそれらを含んだ全12作品の変遷を明らかにしたい。また、本研究の通時

的な観点からは漏れ落ちてしまう一作内での道筋に連動した語彙構成を詳細に考慮する必要もあり、その意味で、小説というフォーマットに適したチャプターに焦点を絞った分析は有効である。さらには、登場人物に着目しその係り受け関係とネットワークを作成することで、テキスト間同士での相関的な近似性や物語上の意味的な位置づけの変化の精査も検討している。これらを含む複数の科学的手法を接ぎ木としながら、文学のテキストにおける表層的計算が可能とするフィクションの潜在的機構の浮上を目指したい。

謝辞

本研究は科研費「知識共有のための価値指向型オントロジーの多分野多言語化」(20300074)の助成を受けた。

参考文献

- [1] 赤石美奈：「文書群に対する物語構造の動的文解・再構成フレームワーク」，人工知能学会論文誌，Vol. 21, No. 5, pp. 428-438, 2006.
- [2] 黒古一夫：『村上春樹「喪失」の物語から「転換」の物語へ』，勉誠出版，294p. 2007.
- [3] 村上春樹，柴田元幸：『翻訳夜話』，文春新書，245p. 1999.
- [4] 加藤典洋：『村上春樹 イエローページ①』，幻冬舎，249p. 2006.
- [5] Jay, Rubin (畔柳和代訳)：『ハルキ・ムラカミと言葉の音楽』，新潮社，463p. 2006.