「論 文」

Dickens の *The Mystery of Edwin Drood* と T. P. James によるその続編の文体的評価 ―多変量解析を用いて―

後藤　克己

Abstract

Three years after Dickens' death, Thomas Power James (henceforth James) added a continuation to *The Mystery of Edwin Drood*, claiming that it was written by the 'spirit-pen of Charles Dickens, through a medium'. This study attempts to clarify whether the continuation can be considered a posthumous work of Dickens as James suggested, at least as far as the linguistic features are concerned. Word preferences in the continuation are analyzed for similarity with those in *The Mystery of Edwin Drood*. Methods used are multivariate analyses of the frequencies of frequent words in three corpora - *The Mystery of Edwin Drood*, the continuation, and *Our Mutual Friend,* with the third corpus added as a reference. The analyses display distinct clustering of sections included in the continuation, on one hand, and those in the two Dickens' works, on the other, highlighting differences in terms of word preferences. In addition to the findings that the continuation is stylistically distinct from the original so that James' claim is not supported, this paper also confirms that (i) the most distinct results are obtained by word-list of narrative – not speech or combinations of both, and that (ii) high frequency words, up to 100, sufficiently distinguish the authors.

1．はじめに

1.1　リサーチクエスチョン（RQ）

Charles Dickens は *The Mystery of Edwin Drood*（以下，*ED*）を Serial Fiction として 1870 年 4 月から逐次発表していたが，作品を完成させることなく同年 6 月に世を去った。その 3 年後，米国 Vermont 州 Brattleboro の印刷工 Thomas Power James（以下，James）は，*ED* に第 2 部として 23 章もの大部となる続編

を加え完全版として発表した（Dickens［James］, 1873）。James はこの続編を, “By the Spirit Pen of Charles Dickens, through a Medium.”, つまり「James が, 亡き Dickens の霊に導かれてテクスト化したもの」とアピールしている。本研究では, この続編をめぐって次の 3 点について考察する。

（RQ1）テクストに生起する語彙頻度を多変量解析の手法を用いて分析することで, *ED* と続編の文体の異同性を語彙嗜好の観点から評価し, 続編の文体が Dickens の霊に導かれたものと考えうるほどに, Dickens のそれと類似しているかどうかを判定する。

（RQ2）頻出語彙数の多変量解析による著者推定では, 多くの場合テクストから発話部を除き, 分析対象を地の文に生起する語彙に限定している（Hoover, 2001, 2002 他）が, 発話テクストの除去は, コーパスによっては 1 重引用符とアポストロフィの区別, 引用符を欠いた Free Direct/Indirect Discourse の抽出などが必要となり, 特に分析対象コーパスが大規模な場合大きな負担となる。そこで発話部を除外せずテクスト全体の語彙で分析した場合と, 地の文のみの語彙を用いた場合の結果を比較し, 分析対象を地の文に限ることの優位性について分析の精度の観点から評価する。

（RQ3）頻出語彙数の多変量解析による著者推定で用いる語彙の数（タイプ数）については, 推定方法は異なるものの, 上位 150 語（Burrows, 2002: pp. 276–67）, あるいは同 800 語（Hoover, 2004: p. 455）で最良の結果が得られたとの研究がある。一方で, 複数作品をそれぞれ複数セクションに分割してクラスター分析した場合, 上位 10 語でも各セクションがそれぞれの作者に正しくグループ化されるとの報告もある（Hoover, 2003b: p. 343）。そこで, 語彙頻度データが, その頻度ランクの高低によって分析結果にどう影響するかを評価し, 著者推定において十分と考えられる語彙の数を見究める。

1.2　研究の背景

1.2.1　James の続編への批評

James による続編（以下, 続編）については, James が「Dickens の霊による作[1]」（Dickens［James］, 1873, 中表紙）とアピールしたこともあり, 物語性, 人物造型の一貫性, 言語的側面等の観点から以下に概括するように多くの批評がなされた。

B., W. H.[2]（1874: p. 219, pp. 221–23）は, 続編を「Dickens 風の文体の癖をかろうじて真似ていて, そこここに注意深くしつらえた感傷を, しかるべく添えてはいるが, 言葉づかいはウィットに欠け, むだ口が目につく」と評している。

また，主要な登場人物（Mr. Grewgious, Miss Twinkleton, Rosa, Jasper など）が，一見 *ED* と似たキャラクターとして描かれているものの，例えば「Grewgious の発話からはユーモアが失われて陰気／単調なものとなり，Rosa のそれからは面白さが消えて単に愚かさが目立ち，Mr. Crisparkle の発話からは快活さが消え重々しい常套句に終始している」として，人物造型の一貫性の欠如を指摘している。さらに言語的側面について，文法面で自動詞と他動詞の誤用（例，lay），主語・動詞の数の不一致，不規則動詞の過去変化形の誤用などを挙げ，さらに，realise［realize］[3] を perceive の意味で用いていること，happen でなく transpire を，as soon as でなく directly を使用していることなど，多くの Yankeeisms（アメリカ英語風の語彙使用）を指摘している。

また，Gadd（1905: p. 272）は，続編について，「まるで原典［*ED*］のもつ非凡さをうかがわせるものがない」と批評し，また人物造型の不自然さを指摘するとともに，言語的側面では「気紛れとも言うべきひどい文法乱れ」を指摘しているが，その内容について具体的に言及していない。

次に，Walters（1913: p. 217）は続編を，「無価値であり，書き間違い・非慣用語法が目立ち，ひどくアメリカ的な型にはまった作」と Gadd（1905）による関連言説を引用して否定的に評している。ただし，「書き間違い・非慣用語法，ひどいアメリカ的型はまり」について具体的に例示していない。

そして「シャーロック・ホームズ」シリーズの著者であり，また，心霊現象の観点から Dickens, Oscar Wilde, および Jack London の遺作を研究した Doyle（1927: pp. 343–44, p. 349）は，批評は交霊云々にではなく物語自体についてなされるべきとして，「似てはいるが退屈な Dickens 作品になっている。わくわく感，ひらめき，伸び伸び感がない。しかし趣向や文章上の癖に Dickens らしさが残っている」と評している。また，原典と同様に語りが突然現在形になること，登場人物にグロテスクなニックネームをつけ，それを大文字で綴っていることを指摘するとともに，“traveller” のようにエル 2 つで綴られていることを例にあげて綴りがアメリカ英語風でない等々と述べている[4]。そして続編を，Dickens についての知識がなく，また新聞の小記事さえ書いたことがないと伝えられる米国人［James］が著したということは，深いトランス状態にあったことの証左だとして，「Dickens の霊による作」とアピールされた続編に一部好意的な批評[5] をしたうえで，全体として「［続編は］決定版とは言い難いものの，丁寧に考察するに価するものだ」と批評を締めくくっている。

更に続編発表の 1 世紀後，Wolkomir（1973）は当時の新聞記事，前述した Doyle の著作，さらに Vermont の史家である Hill（1961: pp. 178–82）の著作から，「Dickens の霊による作」にかかわる記述を引用して肯定的に評している。なお，Cox（1998: p. 569）はこの評について，掲載された *Psychic* ならではのもの，とコメントしている。

1.2.2　頻出語彙の生起数による著者推定に用いられる異なり語の数

Burrows（2002: pp. 275–77）は独自の手法 “Delta”[6] を用いて，英国復古期（17C 後半）の 25 人の作者による韻文を参照コーパスとし，32 作の長詩について，頻度上位 40 ～ 150 語で著者推定を行った。その結果，上位 150 語を用いた場合に最もよい結果が得られたとしている。

Hoover（2004: p. 456）は，その “Delta” を散文に適用して，20 人の作者による米国小説（19C 末～ 20C 初頭）を参照コーパスとし，39 作について上位 20 ～ 800 語で著者推定した。その結果，推定の確度は語彙数の増加につれて高くなり，700 語で最もよい結果が得られたとしている。また Hoover（2001, 2002, 2003a）は，「14 人の作者による 29 作」のような多くの作者・作品で構成されたコーパスに頻出する語，語シーケンス，および語コロケーションの頻度をクラスター分析して行う著者推定では，上位 600 ～ 800 語で最良の結果が得られたとしている。

一方，Hoover（2002: p. 157）は Burrows（1987）を引用して，全トークン数のほぼ 20％を占める *the, and, of, a* および *to* の語は，直観的には意味的・文体的に重要でないように思えるが，作者，作品さらには 1 つの作品内で登場人物さえ互いに区別する，と述べている。さらに Hoover（2003b: p. 343）は，そのメインテーマである同一作品内における文体差異を，上位 50 ～ 800 語で分析する前段として，クラスター分析による著者推定の有効性を評価する中で，作者の異なる 6 作品の地の文を，それぞれ 25,000 語のセクションに分割して，各セクションに頻出する語彙の生起数をクラスター分析した場合，上位 10 語でも各セクションをそれぞれの作者に正しくグループ化していることに言及している。

2. 分析方法

本研究では，分析対象作品の語彙嗜好に類似性があるかどうかを明らかにするため，それらの作品に生起する語彙の頻度を多変量解析の手法を用いて分析する。

2.1　分析ツール

続編が James の手によるものであることは疑いない。本研究の RQ1 は，続編の文体が James の主張するように Dickens の霊に導かれてテクスト化されたもの，と考えうるほどに類似しているか否かである。一般に著者推定は，候補となる著者の作品を特徴づける言語データを，問題作品の特徴と比較して行われる。Dickens の作品は多く遺されていて，その文体の特徴を表すデータ（教師データ）は容易に得られるが，James の手による作品は，ごく短い断片を除き他に見つかっていない[7]。従って作品をカテゴライズする教師あり学習（supervised learning）の手法はとれない。しかし，RQ1 は続編の文体が Dickens のそれと似ているかどうかの問題であることから，似た者同士でグループ化する手法，具体的には多次元尺度法 (multidimensional scaling (MDS）) と階層的クラスター分析（以下，クラスター分析）を用いた[8]。これらは先行研究でも用いられている（田畑，2016; Hoover, 2001, 2002 他）。

2.2　分析データ（コーパス）

分析作品は *ED* と続編の 2 作に，同じく Dickens による作品で *ED* より数年前の作品である *Our Mutual Friend*（1864–65）（以下，*OMF*）を参照用に加えた計 3 作とした[9]。*ED* と続編については，Dickens［& James］(1873)[10] を自ら OCR スキャンして得た電子テクストを元の書籍と目視照合し，誤変換された文字を修正してコーパス化した。また *OMF* のテクストについては，Project Gutenberg からダウンロードした電子テクスト[11] を使用した。

これらを地の文と発話部に分離[12] し，分析結果のプロットや樹形図で，語彙嗜好が作品間で類似しているかどうかを把握しやすくするため，それぞれその規模に応じて概ね 1 セクションが 15,000 トークンとなるように章単位で分割した（*ED* と続編はそれぞれ 4 セクション，*OMF* は 12 セクションの計 20 セクション）。また，各セクションの語彙嗜好を際立たせるため，語彙の生起数は表記形でなくレマ[13] でカウントした。これら 3 作の地の文・発話部別のトークン数とタイプ数（レマ）を表 1 に示す。

表 1　3 作品の規模データ

	地の文			発話部		
作品区分	*ED*	続編	*OMF*	*ED*	続編	*OMF*
トークン数	54,115	71,826	174,311	49,511	57,403	159,471
同、比率	52%	56%	52%	48%	44%	48%
タイプ数	5,440	4,041	8,744	3,547	3,416	6,017

なお，語彙の内，人物・場所などの固有名詞や，作品テーマに関連する語彙は，作者の文体とは関係なく，むしろその異同性の分析結果をゆがめることになりかねないので除外する必要があるが，手作業でこのような語彙を選別するのは容易ではなく，また選別に際して恣意的な判断が混入する懸念が残る（田畑，2016: p. 60）。そこで本研究では，Hoover（2004）を参考に，「全体の語数の 70％以上が 1 作品に偏在する語彙を除外する」手法を用い，人物名についてはさらに手作業で補完的に除外した。なお，続編にはその性格から *ED* と同じ固有名詞が含まれるため，この「70% の偏在基準」を，(i) *ED*+ 続編，*OMF* の 2 作品，(ii) *ED*，続編，*OMF* の 3 作品，の 2 ステップで適用した。また地の文に生起する人称代名詞の頻度は，語彙嗜好とは無関係な語りの人称の違いによる影響を受ける（Hoover, 2001: pp. 427–28）ため除外した[14]。これらの語彙選別を経たテクスト全体・地の文・発話の 3 種類の語彙頻度データを，語彙（行）と 20 のセクション（列）で表にし，各語彙の作品ごとの相対頻度の計で降順にソートし，上位 50, 100, 200, 300, 500, 750, および 1,000 の 7 つの語頻度ランクで切り出して分析に用いる語彙データとした（表 2, 地の文の例）。

表 2　語彙頻度データの概略構造

		ED 4 セクション				続編 4 セクション				*OMF* 12 セクション			
		O1N	O2N	…	O4N	P1N	P2N	…	P4N	M1N	M2N	…	MCN
語頻度（レマ）	the	786	826	…	759	847	1221	…	933	870	1032	…	713
	and	456	514	…	507	432	721	…	493	491	713	…	505
	be	347	371	…	358	420	640	…	504	378	517	…	396
	of	436	464	…	382	301	451	…	363	433	606	…	375
	to	271	336	…	351	410	645	…	479	324	433	…	326
	a	335	329	…	330	281	445	…	302	421	472	…	323
	in	260	274	…	247	201	319	…	265	309	370	…	278
	have	115	204	…	207	240	326	…	280	165	258	…	204
	that	111	149	…	169	237	358	…	309	139	220	…	202
	with	161	183	…	197	122	191	…	150	184	280	…	211
	mr	134	155	…	151	146	203	…	148	137	202	…	138
	as	95	153	…	140	62	83	…	37	133	188	…	225
	say	121	197	…	237	89	172	…	100	142	283	…	166
	at	73	110	…	110	106	158	…	114	125	172	…	126
	on	109	89	…	115	98	116	…	66	112	134	…	97
	for	75	90	…	83	109	186	…	119	65	102	…	76
	s	54	62	…	99	76	139	…	95	64	124	…	95
	by	73	85	…	59	56	95	…	53	87	102	…	70
	but	37	72	…	66	66	96	…	68	60	101	…	63
	not	28	50	…	60	75	132	…	80	46	65	…	47
	…	…	…	…	…	…	…	…	…	…	…	…	…
	…	…	…	…	…	…	…	…	…	…	…	…	…

3. 分析結果

3.1　3つの作品における語彙嗜好の異同性

3 つの作品について，地の文・テクスト全体・発話の 3 種類のデータの上位 100 語を MDS で分析した結果を図 1 に示す。ここで，プロットのセクションラベルは以下のとおりである。（以下，同）

(i)　先頭は作品の区別で，‘O’ は *ED*（Original part），‘P’ は続編（Posthumous part），‘M’ は *OMF*（Mutual）

(ii)　2 番目の 1,2,･･･9,A,B,C は作品内のセクションの区別

(iii) 3 番目は語彙データの種別で，‘N’ は地の文，‘T’ はテクスト全体，‘S’ は発話

また図下部のハイフンで連ねた 2 つの比率値は，それぞれ 1 軸－ 2 軸の寄与率である。

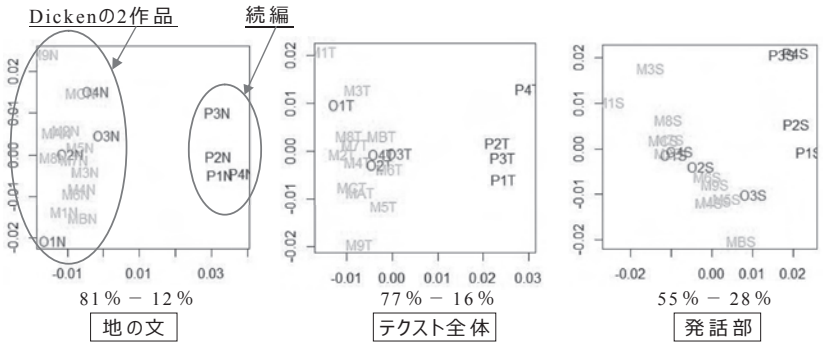


図 1　上位100語による MDS プロット

つぎに同様に上位 100 語についてクラスター分析して得た樹形図を図 2 に示す。ここで図下部のハイフンで連ねた 2 つの数値は最終段階のクラスター対の距離（clustering height[15]；以下，クラスター高），およびそれとその一つ前のクラスター高の比（最後のクラスター高／（最後－ 1）のクラスター高；以下，ハイト比）である。

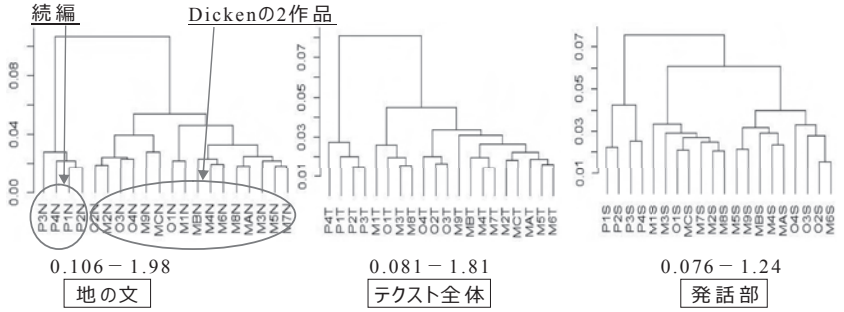


図２　上位100語による樹形図

図１・図２の各左端の地の文語彙によるプロット／樹形図には，Dickens の２作品のセクションは１つのクラスターとなり，続編のセクションはそれとは別クラスターを形成することが明瞭に表れている。また，最後のクラスター高（0.106），およびハイト比（1.98）が，この３種類の分析の中で最も大きく，地の文語彙による分析で２つのクラスターの区別性が最も高くなることを示している。

また，同２図中央の，テクスト全体語彙によるプロットでは，地の文の場合より，やや各セクションが拡散し，また，樹形図では最後のクラスター高が24% 程度狭まっている（0.081）。しかし，ハイト比は 1.81 で 9% の減少に留まっていて，依然として２つのクラスターの区別性が高いことがうかがえる。

ちなみに，同２図右の，発話部語彙によるプロット／樹形図では Dickens ２作品と続編は別クラスターを形成しているものの，プロットでは各セクションがテクスト全体の場合より更に拡散している。樹形図で最後のクラスター高は28% 程度の減少に留まっている（0.076）ものの，ハイト比は 1.24 へと大きく減少し，２つのクラスターの区別性が低下していることがうかがえる。ここで，ハイト比が１に近いことは，クラスタリングの最終段階におけるクラスター対の隔たり（ここでは，Dickens の２作品と続編の間）が，その１つ前のクラスタリングにおけるクラスター間の隔たり（ここでは Dickens の２作品のセクションによるクラスター間）と大差ないこと，つまり Dickens の２作品と続編の語彙嗜好の差異が，Dickens の２作品の中での語彙嗜好の差異と比較して抜きん出たものではないことを示している。

同様に，上位 500 語について分析した結果を図３と図４に示す。

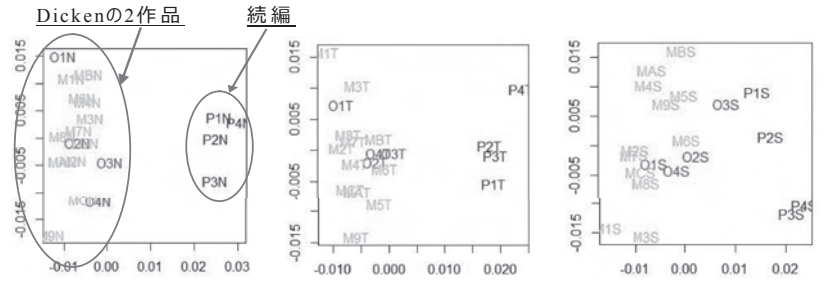


図 3　上位500語による MDS プロット

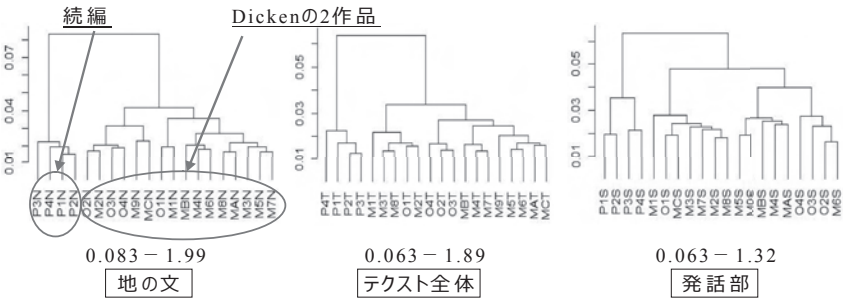


図 4　上位500語による樹形図

上位 500 語による分析においても，上位 100 語の場合と同様な結果と傾向が 3 種類のデータに見られる。語彙数が増加したことで Dickens の 2 作品と続編のクラスター間の距離は 100 語の場合の概ね 80％に減少[16] しているものの，クラスターの区別性の尺度になると考えられるハイト比は殆ど同じである。これら 2 つの語彙ランク以外による分析でも同様であり，上位 50 語～ 1,000 語の 7 つの語彙ランクのデータで一貫して，Dickens の 2 作品のセクションは一つのクラスターを形成し，続編のセクションはそれとは別クラスターを形成していた。

さらに地の文・テクスト全体・発話の 3 つの語彙データを一括した，上位 500 語のデータ[17] による MDS のプロットにおいても，Dickens の 2 作品と続編のセクションはそれぞれ別クラスターを形成し，それらの距離は地の文，テクスト全体，発話の順に狭まっていた（図 5）。

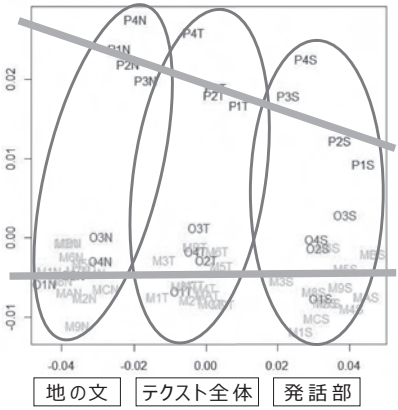


図 5　上位500語の 3 種類の語彙データを一括したデータによる MDS プロット

3.2　頻度ランクによる語彙嗜好の違いと分析結果への影響度

前 3.1 節の上位 50 語～ 1,000 語の 7 つの語彙データによる分析結果には，一貫して Dickens の 2 作品と続編の間の語彙嗜好の違いが表われている。7 つの頻度ランクについて, 地の文の語彙データをクラスター分析して得た樹形図の, クラスター高とハイト比を比較した（表 3，図 6，図 7）。

表 3　頻度ランクとクラスター高・ハイト比

語彙頻度ランク	クラスター高		ハイト比
	last	last-1	
1-50	0.121	0.063	1.94
1-100	0.106	0.054	1.98
1-200	0.094	0.048	1.97
1-300	0.089	0.044	1.99
1-500	0.083	0.042	1.99
1-750	0.080	0.040	1.99
1-1000	0.078	0.039	1.99

凡例
last: 最終段階のクラスタリングにおけるクラスター高
last-1: 最終段階の一つ前のクラスタリングにおけるクラスター高
ハイト比 : last と last-1 のクラスター高の比（last/（last-1））
（以下，同）

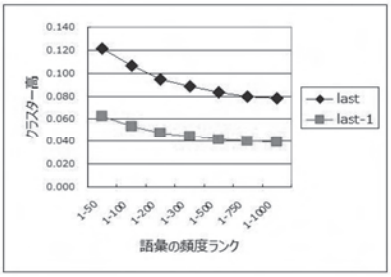


図 6　頻度ランクとクラスター高

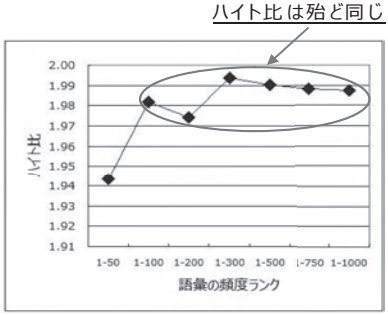
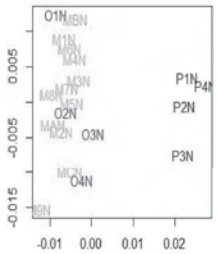


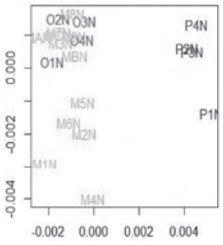
図 7　頻度ランクとハイト比

分析語数が増えるにつれクラスター高が減少している[18]（図 6）が，ハイト比は 100 語と 1,000 語の間でごく僅かな差しかなく，約 2 倍の比を保っている（図 7）。つまり語彙数が 100 語と 1,000 語の間で大きく変わっても，分析結果に大きな差は表れていない。

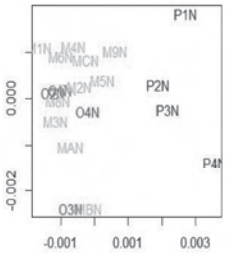
そこで，3.1 節の結果から Dickens の 2 作品と続編の語彙嗜好は異なっているとした上で，この 1,000 語のそれぞれの頻度ランクの語彙データが，どのようにこの結果に反映されたかを分析した。すなわち，前節の分析結果から区分性が最も高いと考えられる地の文について，語彙の属するランクレベル（つまり，生起数の多い高ランク語彙，少ない低ランク語彙など）による語彙嗜好の違い，および分析結果への影響度を確認することとし，上位 1,000 語までの語彙を 100 語ごとにセグメント化したデータを用いて前節と同様に分析した[19]（図 8, 図 9）。なお, 以下の図で, プロット／樹形図の下に付記したデータの内容は, 前節と同じである。



83%—11%



61%—12%



41%—14%

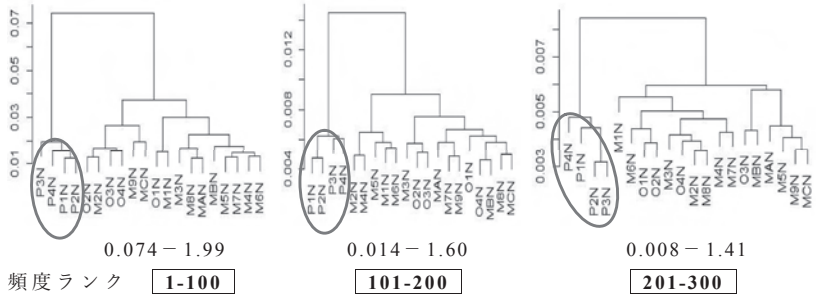


図 8　地の文語彙の上位300語を100語ごとにセグメント化したデータによる分析結果

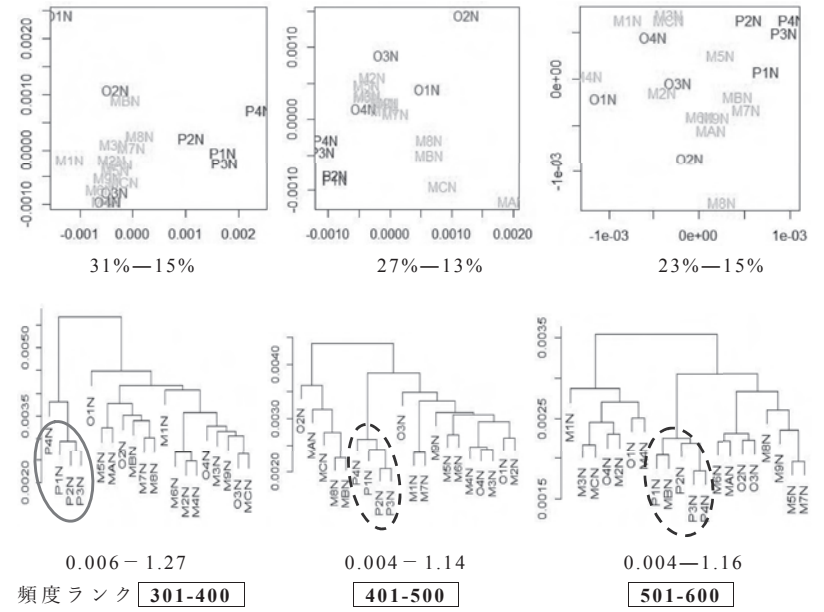


図 9　地の文語彙の上位301～600語を100語ごとにセグメント化したデータによる分析結果

上位 400 語までの 4 つの 100 語セグメントによる分析では，前節の上位 50 語～ 1,000 語による分析で一貫して得られたのと同様に，Dickens の 2 作品と続編のセクションが最終段階のクラスタリングで別クラスターとなり，続編のセクションが単独のクラスターとしてグループ化されている。一方，401 語以降の 100 語セグメントでは，続編のセクションは最早最終段階の単独クラス

ターになっていない。（601 語以降の 100 語セグメントでも同様であった。）最後のクラスター高は，ランク 1–100 では最大の 0.074 であったが，次のランク 101–200 では 0.014 と約 1/5 に大きく減少している。また，ハイト比は，ランク 1–100 の最大の 1.99 から，ランク 101–200 以降の 1.60，1.41，･･･ と徐々に減少している。

一方，続編のセクションが最後の単独クラスターとしてグループ化されなかったランク 401 以降の 100 語セグメントでは，最後のクラスター高はごく小さく，またハイト比は 1 に近い。従ってランク 401 以降の語のクラスタリングへの影響度はごく低いものと考えられる。

本研究ではセクション間の距離をユークリッド法で測っている。セクション間の語彙頻度のバラつきは一般に高頻度語ほど大きく，またセクション間の距離は，各語彙の頻度差の２乗値の計の平方根であるため，一般に高頻度語ほど距離への影響度が高い。さらに，作品には少数の頻度上位語が際立って多く生起する傾向があるため，クラスター高は高頻度語のデータをより反映したものとなる。本節の分析結果もこれと符合しており，分析結果に大きく影響するのは，ランク 1–100 のセグメントの語頻度データであって，ランク 101 以降のセグメントに属する語データによる影響は低いと考えられる（表 4，図 10，図 11）。

表 4　100語セグメントのクラスター高とハイト比

語彙頻度ランク	クラスター高		ハイト比	続編の区分性
	last	last-1		
1-100	0.074	0.037	1.99	最終クラスタリングでDickensの２作品と続編を区分
101-200	0.014	0.0091	1.60	
201-300	0.0084	0.0060	1.41	
301-400	0.0057	0.0045	1.27	
401-500	0.0044	0.0038	1.14	続編セクションは最終の単独クラスターにならない
501-600	0.0035	0.0031	1.16	
601-700	0.0027	0.0025	1.06	
701-800	0.0030	0.0024	1.22	
801-900	0.0023	0.0022	1.07	
901-1000	0.0021	0.0020	1.03	

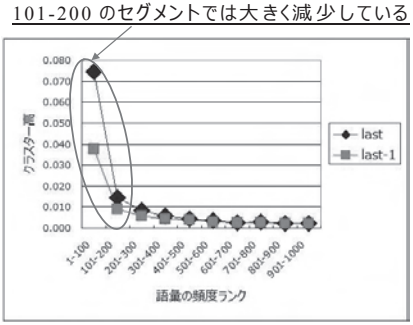


図10　100語セグメントのクラスター高

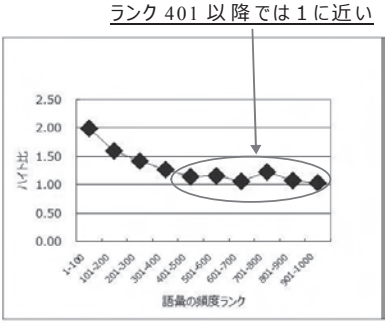


図11　100語セグメントのハイト比

次に，頻度ランク上位 25 語について分析した（図 12 ～図 14）。

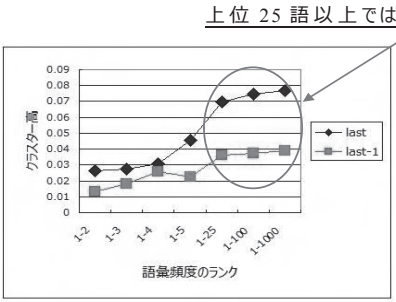


図12　頻度ランクとクラスター高
（上位25語）

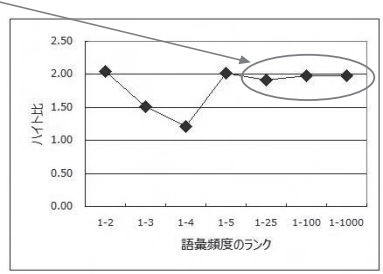


図13　頻度ランクとハイト比
（上位25語）

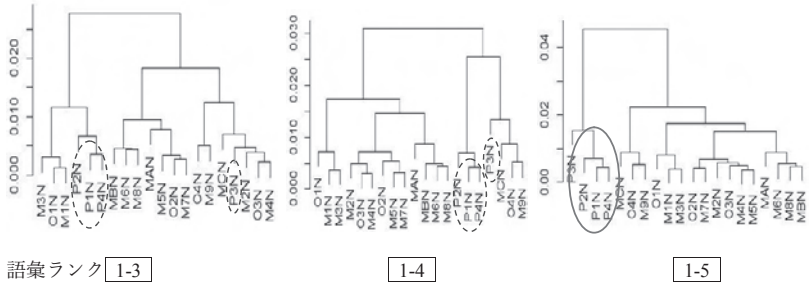


図14　上位 5 語までの樹形図

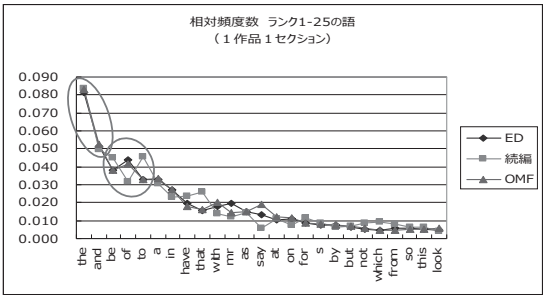


図15　上位25語の相対頻度数

一般にクラスタリングに大きく影響するのは，高頻度かつセクション間で頻度差が大きい語である。本研究で扱った 3 作品の場合，図 15 のように上位の 2 語（*the, and*）は際立って高頻度だが，ランク 3 以降の *be, of, to* とは異なり，作品間の頻度差が小さい。そのためランク 4 までの語彙による樹形図では，続編セクションが最終の単独クラスターとなっていない（図 14）。一方，ランク 1–25 以降では 2 つのクラスター高およびハイト比は安定しており（図 12，図 13），また Dickens の 2 作品と続編の区分が MDS プロットおよび樹形図に明瞭に表れていることから，上位 25 語程度が絞り込みの限界と考えられる。なお，同サイズのテクストで，*the* と *and* の生起数が 2 倍異なることもある（Hoover, 2002: pp. 157–58）ことから，一般に分析語数を上位 5 語程度の少数にまで絞り込むことは適当でないと考えられる。

前 3.1 節および本節では，データの距離・結合法に，それぞれテキストマイニングで多く用いられるユークリッド距離・ウォード法を用いている。比較のため，使用した 3 作品のコーパスの上位 100 語と 500 語について，dist 関数の 3 つの距離（euclidean, manhattan, canberra）と hclust 関数の 4 つの結合法（ward.D2, single, complete, average）を組み合わせで分析した（表 5，表 6）。

表 5　上位100語による比較

結合法 (hclust)	距離 (dist)	クラスター高 last	クラスター高 (last-1)	ハイト比
ward.D2 ウォード法	euc	0.106	0.054	1.98
	man	0.654	0.278	2.35
	can	45.0	18.8	2.39
single 最近隣法	euc	0.039	0.028	1.38
	man	0.238	0.162	1.47
	can	17.7	13.7	1.29
complete 最遠隣法	euc	0.065	0.050	1.28
	man	0.368	0.265	1.39
	can	26.2	19.5	1.34
average 群平均法	euc	0.048	0.037	1.30
	man	0.298	0.202	1.47
	can	21.8	15.9	1.37

表 6　上位500語による比較

結合法 (hclust)	距離 (dist)	クラスター高 last	クラスター高 (last-1)	ハイト比
ward.D2 ウォード法	euc	0.083	0.042	1.99
	man	0.798	0.365	2.19
	can	292	182	1.61
single 最近隣法	euc	0.031	0.023	1.37
	man	0.340	0.261	1.30
	can	162	159	1.02
complete 最遠隣法	euc	0.051	0.038	1.33
	man	0.468	0.345	1.35
	can	205	176	1.17
average 群平均法	euc	0.039	0.030	1.30
	man	0.393	0.297	1.33
	can	182	169	1.08

備考 : euc: euclidean, man: manhattan, can: canberra

その結果，ウォード法の数値が最も大きくなっているが，続編と Dickens の 2 作品は，全ての距離・結合法の組み合わせで一貫して最終のクラスター対を形成しており，他の組み合わせを選んでも分析結果に影響を与えない。

4. 結　論

RQ1 については 3.1 節の分析で，最終のクラスタリングにおいて一貫して Dickens の 2 作品のセクションは 1 つのクラスターを形成し，続編のセクションはそれとは別クラスターを形成した。このことは Dickens の 2 作品と続編の語彙嗜好に明らかな差異があることを示しており，続編の文体が Dickens のそれとは異なっていることが裏付けられた。

RQ2 については 3.1 節の分析で，地の文語彙を用いた場合にテクスト全体語彙を用いた場合より明確な嗜好の差異が表れており，著者推定における地の文語彙の優位性をあらためて確認した。なお，テクスト全体語彙で得られる区分性は，地の文語彙の場合より低いものの差は小さく，分析対象となる著者／作品が限定的なケース，あるいは予備的／概略の区分を行う場合であれば，全体語彙を分析に充てることも可能と考えられる。

RQ3 については，3.1 節でみられた上位 50 語～ 1,000 語による分析で一貫して得られたクラスタリングパターンが，3.2 節の分析で，概ね上位 100 語あるいはそれ以下の語で形成されており，頻度ランクが下がるにつれパターン形成への影響度が大きく減少することが確認された。これは，MDS やクラスター分析を用いて行う著者推定は，頻度上位 100 語程度の分析で十分であることを

示唆している。なお前節の分析では上位 25 語でも上位 100 語と同様な結果が得られているが，頻度ランク 4 までの語による分析では結果が異なっている。研究のケースによって分析の対象となる作者・作品は異なり，上位 100 語に含まれる語も異なりうること，また一般に語彙数を減らすことが分析の容易さにつながるわけでもないことから，100 語以下に絞り込むメリットは少ないと考えられる。なお，表 7 は上位 100 語までの語であるが，いずれも平易な日常語と考えられる。

表 7　頻度ランク 1 ～100の語彙一覧[20]

ランク 1–25 の語
the, and, be, of, to, a, in, have, that, with, mr, as, say, at, on, for, s, by, but, not, which, from, so, this, look,

ランク 26–100 の語
an, hand, when, upon, do, would, all, out, if, then, take, or, come, one, into, who, little, there, face, man, go, miss, no, time, very, again, eye, some, make, up, more, old, return, see, after, other, down, could, head, reply, now, door, before, turn, what, though, way, leave, think, about, than, house, ask, such, young, two, back, room, know, day, find, seem, over, might, like, gentleman, much, answer, while, place, any, still, well, great, lady

上位 100 語に含まれる語のうち，頻度ランク 21 の *which* は，その殆どが関係詞であり，続編に *ED* の 2.2 倍の頻度（相対頻度, 以下同）で生起する。また，主に伝達動詞として用いられる *say, return, reply* および *ask* のうち，頻度ランク 13 の *say* は 7:3 で *ED* に多く生起し，同 65 の *reply* は 3:7 で逆に続編に多く生起する。（同 58 の *return* と同 78 の *ask* は，それぞれ *ED* と続編にほぼ同じ頻度で生起）。このような関係詞や伝達動詞の選択傾向が，*ED* と続編の文体的差異の例として考えられる[21]。

本研究で用いた，語の生起頻度をクラスター分析（ユークリッド距離・ウォード法を適用）する手法は，習慣的に用いられ，同じ作者の作品間で生起頻度が一貫すると想定される高頻度語（Hoover, 2003b: p. 341）が，分析結果に大きく寄与することから，著者推定に有効な手法と考えられる。一方，作者の文体に特徴的でも，生起頻度が低い語の寄与度は小さいため，同じ作者の作品間や，作品内における文体変異の分析のような，微妙な差異に注目する分析への適用には限界があると考えられる。そのような分析に有効な手法の見極めについては，今後の課題としたい。

謝　辞

本稿の内容は，英語コーパス学会第 43 回大会（2017 年 9 月，関西学院大学），同 44 回大会（2018 年 10 月，東京理科大学），および THE 4th ASIA PACIFIC CORPUS LINGUISTICS CONFERENCE（APCLC2018）（2018 年 9 月，高松）における研究発表に基づいている。ご指導いただいた大門正幸教授に感謝致します。また，匿名の査読者の方々から貴重なご指摘・ご助言をいただいた。ここに記して感謝します。

注

1. 以下，日本語訳は全て筆者による。
2. 著者のフルネームは不詳。当批評は W. H. B. と署名されている。
3. W. H. B. の指摘の中では realise と表記されているが，Dickens［James］（1873）では realize の綴りになっている。
4. Doyle は続編全体ではなく，続編の一部を基に批評している。そしてその点について，「Walters が指摘している futility について批評するには続編全体を精査する必要があるものの，illiteracy と Americanisms については一部による精査で足りる」と述べている（Doyle, 1927: p. 344）。
5. 続編について Doyle は交霊による作との見方をとるものの，その霊を Dickens とすることには懐疑的な立場をとっている。
6. 著者が不明な作品と既知な作品（候補となる参照コーパス）について，高頻度で生起する語の生起特性（z-score）の差を求め，その差の語全体での計 “Delta” が最も小さくなる著者を，不明作品の著者と推定する手法である。この手法は，著者候補の手がかりとなる外的証拠が乏しい中で，著者を多くの候補の中から絞り込む研究において有効なツールになりうる（Burrows, 2002: p. 267; Hoover, 2004: p. 454）とされている。本研究では，文体の異同性評価の候補著者は Dickens に限定されるため用いなかった。
7. 続編の他に，Dickens［James］（1873）の序文で言及している *The Life and Adventures of Bockley Wickleheap* の断片テクストが見つかったとされているが，続編の規模の 0.5％にも満たない 600 語程度のごく小規模の断片であり，James による作品の文体特徴を参照するコーパスとして扱えないと考えた。なお，James は当作品も Dickens の霊に導かれたものと主張している。
8. すべての分析は，R（version 3.4.4）の環境にて dist, cmdscale および hclust の関数を用いた。データ数は相対度数化し，標本間の距離はユークリッド距離（dist の euclidean）を，またクラスター間は Ward’s linkage（hclust の ward.D2）を指定した。
9. 匿名の査読者から，続編の文体が Dickens と似ていないと判断するのに，Dickens の 2 作品のみとの比較では，Dickens の作品間の変異の可能性が不明なため不十分，との指摘をいただいた。文体の類似性について確定的結論を導くには Dickens の多くの作品との比較が必要だが，本論では時間的制約から *ED* と執筆時期の近い *OMF* を比較対象とした。重要な指摘であり今後の課題としたい。

10. 実際にスキャンしたのは，1874 年の版である。
11. EBook#883, Release 2006/4/27, Last Updated 2016/9/25（元となった版の出版社名，刊行年は電子テクストに記載されていない）
12. 引用符による括りで発話部を切り分けており，引用符を伴わない Free Direct/Indirect Discourse の箇所は地の文に，また地の文に含まれる，強調のために引用符で括られた語句は発話部に切り分けている。なお発話部を機械的に「"*"」で検索したのでは，閉じの引用符を欠くなど括りの始端と終端が整っていない箇所があると，それ以降のテクストが正しく切り分けられない。そこで検索結果を目視で確認して，欠けた箇所や余剰な箇所に「仮の引用符；後日識別できるよう，例えば +"」を挿入するなどして，括りを整備した後に一括して切り分けた。
13. AntBNC Lemma List（antbnc_lemmas_ver_001 ）Laurence Anthony's Website, URL: http://www.laurenceanthony.net/software/antconc/
14. 本研究の対象となる 3 作品は，いずれも 3 人称語りの作品であるが，通例に従い除外した。なお，発話部についても除外した。
15. clustering height は hclust で出力された height データおよび共表形行列から得た。（ページ数の制限があるため本論には結果のみを記す）
16. hclust 分析の都度，処理対象語について相対頻度化しているため，セクション間距離への影響度の高い頻度上位語の相対頻度値は，分析語数が増えることで小さくなり，その結果クラスター高が減少する。
17. 分析データは，テクスト全体語彙の上位 500 語について，3 つの語彙データの生起頻度を 1 つに集約したもの。（地の文・発話について，それぞれ少なくとも上位 300 語が含まれている。）
18. 注 16 と同じ
19. 本節では上位 1,200 語までの語彙数で相対度数化したうえで，1,000 語までの各 100 語セグメントで分析しているので，前節とは，同じ上位 100 語の分析結果でも，寄与率，クラスター高などの値に多少差がある。
20. ランク 1–25 に含まれる "s" の頻度には，*is, has* などの縮約によるものと s 属格のサフィックスを含んでいる。そのため元来別の語彙の頻度が合計されており，今後の研究では注意して扱う必要がある。なお，この "s" の頻度データを除外して分析した場合，クラスター高などの数値はごくわずかに変化するが，クラスタリングに差異はない。
21. 本研究では語彙嗜好を際立たせるため，語彙頻度をレマでカウントしており，分析した語彙データに表記形別の頻度は含まれていない。この伝達動詞 4 語の頻度を同じコーパスで表記形別に再カウントしたところ，*say, return*, および *ask* は，*ED* で過去形が 73% を占め，続編で逆に現在形が 69％を占めている。一方，*reply* は *ED*・続編ともに現在形が多いものの，続編では 90％を占め，他の 3 語と同様に現在形が優勢となっている。ここから，*ED* と続編の語りの時制には現在形と過去形が混在しているが，それらの比が *ED* と続編で逆転している可能性がうかがえる。

参考文献

B., W. H. (1874). “Review. Part Second of *The Mystery of Edwin Drood*.” *The Southern Magazine*, 14 (February 1874): pp. 219–23.

Burrows, John (1987) *Computation into Criticism – A Study of Jane Austen's Novels and an Experiment in Method.* Oxford: Clarendon Press.

Burrows, John (2002) “‘Delta’: a Measure of Stylistic Difference and a Guide to Likely Authorship.” *Literary and Linguistic Computing*, 17(3): pp. 267–87.

Cox, Don Richard (1998). *Charles Dickens's The Mystery of Edwin Drood*：*An Annotated Bibliograph.* New York: AMS Press, Inc.

Dickens, Charles [& T. P. James] (1873). *Part Second of the Mystery of Edwin Drood. By the Spirit-Pen of Charles Dickens, Through a Medium. Embracing, Also That Part of the Work Which Was Published prior to the Termination of the Author's Earth-Life*. Brattleboro, VT.:T. P. James.

Doyle, Arthur Conan (1927). “The Alleged Posthumous Writings of Great Authors.” *The Bookman 66,* New York. [Online].

URL: https://www.unz.org/Pub/Bookman-1927dec-00342

Gadd, George F. (1905). “The History of a Mystery. A Review of the Solutions to ‘Edwin Drood’ (Chs. 3 & 4).” *The Dickensian*, 1: pp. 270–73.

Hill, Ralph Nading (1961). *Contrary Country: A Chronicle of Vermont.* Brattleboro, Vermont: The Stephen Greene Press.

Hoover, David L. (2001) “Statistical Stylistics and Authorship Attribution: an Empirical Investigation.” *Literary and Linguistic Computing*, 16(4): pp. 421–44.

Hoover, David L. (2002) “Frequent Word Sequences and Statistical Stylistics.” *Literary and Linguistic Computing*, 17(2): pp. 157–80.

Hoover, David L. (2003a) “Frequent Collocations and Authorial Style.” *Literary and Linguistic Computing*, 18(3): pp. 261–86.

Hoover, David L. (2003b). “Multivariate Analysis and the Study of Style Variation.” *Literary and Linguistic Computing*, 18(4): pp. 341–60.

Hoover, David L. (2004). “Testing Burrows's Delta.” *Literary and Linguistic Computing*, 19(4): pp. 453–75.

田畑智司（2016）「共著作品における Dickens の文体」堀正広・赤尾一郎（監修），堀正広（編）『英語コーパス研究シリーズ　第 5 巻 コーパスと英語文体』: pp. 53–71. 東京 : ひつじ書房.

Walters, John Cuming (1913) *The Complete Mystery of Edwin Drood: History, Continuations, and Solutions.* Boston: Dana ESTES & Company.

Wolkomir, Richard (1973). “Charles Dickens' Great Mystery.” *Psychic*, 4: pp. 16–17.

（後藤　克己　中部大学大学院　国際人間学研究科　Email: gotok@isc.chubu.ac.jp)