

of this study present internal evidence of how the two authors were ‘unequally partnered’ with each other.

Dickens と Collins の共著作品への文体統計学的アプローチ

田 畑 智 司^{†1}

本稿では、語彙の生起パターンを統計解析することにより、英国ビクトリア朝の作家 Charles Dickens と Wilkie Collins の作品を識別する指標語彙を抽出し、それをもとに二人の共著作品の特徴や他の作品との関係や位置づけを考察する。語彙頻度という internal evidence と、信頼度の高い書誌学的情報、伝記的記述などの external evidence を比較考察し、文体統計学的手法の妥当性を吟味するとともに、二人の共著作品への貢献度の推定を試みる。本研究では、まず、Random Forests を用いて Dickens と Collins の文体識別に有効な語彙項目を抽出する。併せて、従来コーパス言語学で行われてきた対数尤度比やカイ二乗検定をもとにした ‘key’ words 分析を文学作品の特徴語抽出に適用する際に直面する問題点を指摘する。そして、Random Forests によって抽出した作者判別マーカーを変数として、多変量解析を実行し、共著作品と Dickens および Collins それぞれの作品との関係を視覚化する。文体統計学的分析の結果からは二人の作家の不均衡な関係が見て取れる。

Stylometry of co-authorship: Charles Dickens and Wilkie Collins

TOMOJI TABATA^{†1}

The Victorian author Charles Dickens collaborated with his younger contemporary Wilkie Collins on a number of stories, typically for the Christmas Numbers of his journals: *Household Words* and *All the Year Round*. While some of their collaborative pieces were written with the assistance of other ‘staff writers’ including Elizabeth Gaskell, Adelaide Anne Proctor, *et al.*, four works are known to have been co-authored by Dickens and Collins alone (Nayder, 2002). These collaborative writings vary in design and style from one another as well as in theme and setting. In some cases, even one chapter can be read differently from another due to the varying proportion of contribution by each of the duo.

The present study applies computational techniques in an effort to detect stylistic differences in the collaborations of Dickens and Collins. Random Forests was employed to spotlight lexical variables that strongly differentiate between the two authors. In my experiments running Random Forests on a set of 48 texts (24 Dickens texts *versus* 24 Collins texts), the Dickens texts and the Collins texts were classified into two distinct clusters with the accuracy of 96–100%. Some 100 marker words identified during the process of the Random Forests run were then used as lexical variables so as to examine relationships between the collaborative texts and the canonical works of Dickens and Collins. The results

1. はじめに

1.1 Dickens と他の作家のコラボレーション

英国ヴィクトリア朝の人気作家 Chales Dickens (1812–1870) は、主宰する雑誌 *Household Words* と *All the Year Round* の Christmas Numbers (クリスマス特集号) 向けに、Wilkie Collins (1824–89) 他の ‘staff writers’ の手を借りて数多くの短編を著している。それらの作品の多くは、Dickens が、自身の着想を基に、書き手にプロットやキャスト、描写の視点を細かく「指揮」(“conducted by Charles Dickens”) して執筆させた草稿に加筆修正を施して仕上げたものである。*Mugby Junction* と *No Thoroughfare* (1867) を除き、そうした短編作品は匿名で発表されているが、そのうち、*Household Words* の Christmas Numbers に掲載された作品については、会計帳簿の記録などからはほぼ著者の特定が可能である (Thomas, 1982: 140)。Thomas (1982: 140–152) は会計帳簿および Stone (1968) を参考に、Christmas Numbers への寄稿者一覧を提示している。

本稿では、Stone (1968), Thamas (1982), Nayder (2002) によって明らかにされている書誌学的情報をもとに、Dickens と Collins 二人の協力によって著されたことが判明している作品 4 点 (表 1) を取り上げる。

表 1 Dickens と Collins の共著作品 4 点
Table 1 Four collaborative works by Dickens and Collins

No.	発表年 (Date)	標題 (Title)	略号 (Abbrev.)	総語数 (Word-tokens)
1	1857, 1866, 1874	<i>The Frozen Deep</i> ^{*1}	(FrozenDeep)	28,434
2	1857	<i>The Lazy Tour of Two Idle Apprentices</i>	(LazyTour)	40,214
3	1857	<i>The Perils of Certain English Prisoners</i>	(PCEnglishPrisoners)	36,409
4	1867	<i>No Thoroughfare</i>	(NT)	47,914

^{†1} 大阪大学・大学院言語文化研究科
GSLC, University of Osaka

^{*1} 本稿は 2012 年 1 月 27–28 日開催の第 93 回人文科学とコンピュータ研究発表会の資料として用意したものである。図表等のデータに誌面を大部分を充当する必要があるため、詳細な説明は研究会当日の口頭発表にて行う。

1.2 書誌学的情報に基づく authorship

表 1 に挙げた作品のうち、*The Frozen Deep* と *No Thoroughfare* はもともと演劇として書かれたものである。*The Frozen Deep* は、その当時大きな話題となった Sir John Franklin 率いる 1845 年の南極探検隊が遭難し全滅したことを示す証拠品を入手し、食糧枯渇のため人肉食が行われた可能性を示唆した Rae 博士の報告に対して、探検隊は先住民に襲われ命を落としたと主張する Dickens が *Household Words* 誌上で Rae 博士と議論を戦わせたことを発端として着想を得た劇である。この作品は、Dickens がプロットおよびキャストを考案し、Collins に行かせた下書きに大幅な改訂を施して 1857 年に上演された後、1866 年と 1874 年に (Dickens との間に距離のできた)Collins が改訂した版があるほかに、短編小説に書き改められたものがあるなど、複数の異本が存在する (Nayder, 2002: 10)。ト書きを除いて、ほぼ登場人物の dialogue だけで作品が構成される戯曲は、事物・情景・動作・心理などの描写を含む narrative と登場人物の dialogue を巧みに組み合わせた小説との言語的な差異が大きいため、そのままでは小説と比較を行うことは有意義ではない。そこで本研究では、*The Frozen Deep* については小説化されたものを用いた。

No Thoroughfare についても、*The Frozen Deep* 同様小説版を分析対象とした。この作品では Act II を Collins が、Act III を Dickens が、それぞれ単独で執筆しているが、Act I, IV, V は二人の共同執筆である (Thomas, 1982: 152)。

The Perils of Certain English Prisoners は、二人の分担章が明確である。Thomas (1982: 146) および Allingham (2011) によれば、Chapter I, III を Dickens が、Chapter II を Collins が単独で執筆している。

The Lazy Tour of Two Idle Apprentices では、Dickens が単独で執筆している Chapter IV を除き、他の章は Dickens と Collins の共同執筆による (Nayder, 2002: 106)。

これら書誌学的資料 (external evidence) をまとめたものが表 2 である。

表 2 書誌学的情報に基づく共著作品 4 点の分担
 Table 2 Attribution of chapters based on the bibliographic records (the external evidence)

No.	発表年 (Date)	標題 (Title)	Part	Authorship
1	1857, 1866, 1874	<i>The Frozen Deep</i>		Collins & Dickens ⇒ Collins
2	1857	<i>The Lazy Tour of Two Idle Apprentices</i>	Chapter I	Dickens & Collins
			Chapter II	Dickens & Collins
			Chapter III	Dickens & Collins
			Chapter IV	Dickens
			Chapter V	Dickens & Collins
3	1857	<i>The Perils of Certain English Prisoners</i>	Chapter I	Dickens
			Chapter II	Collins
			Chapter III	Dickens
4	1867	<i>No Thoroughfare</i>	Act I	Dickens & Collins
			Act II	Collins
			Act III	Dickens
			Act IV	Dickens & Collins
			The Curtain Falls	Dickens & Collins

2. 語彙頻度情報に基づく作者の判別

本節では、Dickens および Collins それぞれの canonical な作品 24 点ずつからなるコーパスを基盤として、二人の作者の作品を判別する上で貢献度の高い語彙項目を抽出する。

表 3, 4 は、それぞれ Dickens, Collins のサブコーパスを構成する作品一覧である。二つの表の第 3 列の省略記号は後掲する散布図、樹状図において作品ラベルとして使用している。第 4 列は作品カテゴリーを表す。Dickens, Collins 共に、(月刊分冊式の)小説 (Fiction) の他、Sketches や History なども含めている。

表3 Dickens のテキスト 24 点
Table 3 The set of 24 Dickens works

No.	Texts	Abbr.	Category	Date	Word-tokens
1	<i>Sketches by Boz</i>	(D33_SB)	Sketches	1833-6	187,474
2	<i>The Pickwick Papers</i>	(D36_PP)	Serial Fiction	1836-7	298,887
3	Other Early Papers	(D37a_OEP)	Sketches	1837-40	66,939
4	<i>Oliver Twist</i>	(D37b_OT)	Serial Fiction	1837-9	156,869
5	<i>Nicholas Nickleby</i>	(D38_NN)	Serial Fiction	1838-9	321,094
6	<i>Master Humphrey's Clock</i>	(D40a_MHC)	Miscellany	1840-1	45,831
7	<i>The Old Curiosity Shop</i>	(D40b_OCS)	Serial Fiction	1840-1	217,375
8	<i>Barnaby Rudge</i>	(D41_BR)	Serial Fiction	1841	253,979
9	<i>American Notes</i>	(D42_AN)	Sketches	1842	101,623
10	<i>Martin Chuzzlewit</i>	(D43_MC)	Serial Fiction	1843-4	335,462
11	<i>Christmas Books</i>	(D43b_CB)	Fiction	1843-8	154,410
12	<i>Pictures from Italy</i>	(D46a_PFI)	Sketches	1846	72,497
13	<i>Dombey and Son</i>	(D46b_DS)	Serial Fiction	1846-8	341,947
14	<i>David Copperfield</i>	(D49_DC)	Serial Fiction	1849-50	355,714
15	<i>A Child's History of England</i>	(D51_CHE)	History	1851-3	162,883
16	<i>Bleak House</i>	(D52_BH)	Serial Fiction	1852-3	354,061
17	<i>Hard Times</i>	(D54_HT)	Serial Fiction	1854	103,263
18	<i>Little Dorrit</i>	(D55_LD)	Serial Fiction	1855-7	338,076
19	<i>Reprinted Pieces</i>	(D56_RP)	Sketches	1850-6	91,468
20	<i>A Tale of Two Cities</i>	(D59_TTC)	Serial Fiction	1859	136,031
21	<i>The Uncommercial Traveller</i>	(D60a_UT)	Sketches	1860-9	142,773
22	<i>The Great Expectations</i>	(D60b_GE)	Serial Fiction	1860-1	184,776
23	<i>Our Mutual Friend</i>	(D64_OMF)	Serial Fiction	1864-5	324,891
24	<i>The Mystery of Edwin Drood</i>	(D70_ED)	Serial Fiction	1870	94,014
Sum of word-tokens in the set of Dickens texts:					4,842,337

2.1 特徴語 ('key' words) 抽出にまつわる諸問題

テキストやジャンル、言語使用域の特徴を記述する際に用いられる手順として特徴語 ('key' words) の抽出は重要なステップである。Henry & Roseberry (2001: 110) はテキストの特徴語 ('key' words) 次のように定義している。

'Key words' are defined as words that 'appear in a text or a part of a text with a frequency greater than chance occurrence alone would suggest'.

ターゲットとするテキストやテキスト群を参照コーパス(またはテキスト)と比較し、(典型的には)対数尤度比やカイ二乗値をもとにテキストに生起する語彙を篩いに掛け、統計学的に有意な頻度差のある語彙項目を洗い出す手法が数多くの先行研究で行われている (Dunning,

1993; Rayson & Garside, 2000; Henry & Roseberry, 2001; 高見, 2003; Scott & Tribble, 2006; etc.)

表4 Collins のテキスト 24 点
Table 4 The set of 24 Collins works

No.	Texts	Abbr.	Category	Date	Word-tokens
1	<i>Antonina, or the Fall of Rome</i>	(C50_Ant(onina))	Historical	1850	166,627
2	<i>Rambles Beyond Railways</i>	(C51_RBR)	Sketches	1851	61,290
3	<i>Basil</i>	(C52_Basil)	Fiction	1852	115,235
4	<i>Hide and Seek</i>	(C54_HS)	Fiction	1854	159,048
5	<i>After the Dark</i>	(C56_AD)	Short stories	1856	136,356
6	<i>A Rogue's Life</i>	(C57_ARL)	Serial Fiction	1856-7	47,639
7	<i>The Queen of Hearts</i>	(C59_QOH)	Fiction	1869	145,350
8	<i>The Woman in White</i>	(C60_WIW)	Serial Fiction	1860	246,916
9	<i>No Name</i>	(C62_NN)	Serial Fiction	1862	264,858
10	<i>Armada</i>	(C66_Armadale)	Serial Fiction	1866	298,135
11	<i>The Moonstone</i>	(C68_MS)	Serial Fiction	1868	196,493
12	<i>Man and Wife</i>	(C70_MW)	Fiction	1870	229,376
13	<i>Poor Miss Finch</i>	(C72_PMF)	Serial Fiction	1872	162,989
14	<i>The New Magdalen</i>	(C73_TNM)	Serial Fiction	1873	101,967
15	<i>The Law and the Lady</i>	(C75_LL)	Serial Fiction	1875	140,788
16	<i>The Two Destinies</i>	(C76_TD)	Serial Fiction	1876	89,420
17	<i>The Haunted Hotel</i>	(C78_HH)	Serial Fiction	1878	62,662
18	<i>The Fallen Leaves</i>	(C79_FL)	Serial Fiction	1879	133,047
19	<i>Jezebel's Daughter</i>	(C80_JD)	Fiction	1880	101,815
20	<i>The Black Robe</i>	(C81_BR)	Fiction	1881	107,748
21	<i>I Say No</i>	(C84_ISN)	Fiction	1884	119,626
22	<i>The Evil Genius</i>	(C86_EG)	Fiction	1886	110,618
23	<i>Little Novels</i>	(C87_LN)	Fiction	1887	148,585
24	<i>The Legacy of Cain</i>	(C89_LOC)	Fiction	1888	119,568
Sum of word-tokens in the set of Collins texts:					3,466,156

しかし、こうした特徴語抽出の手法は長編小説を取録した作家コーパスの比較を行う際に問題に直面する。表5は、Dickens の特徴語のうち 'keyness' (ここでは対数尤度比) 上位 40 項目を挙げている。表5は Collins の作品に比べて Dickens の作品において overuse (過剰使用) されている項目である。紙面の都合上割愛するが、Collins を中心にして同様のリストを作成している。表5の第4列、DF (Document Frequency) は当該の語が生起する文書数である。表の 15, 18, 24, 34, 35, 40 位にランクインしている固有名詞 (Dombey, Pecksniff, Boffin, Nickleby, Clennam, Squeers) は極めて高い keyness を示しているにもかかわらず、一

作品にしか生起しない*1.

一作品にしか生起しない語は Dickens の特徴語リスト上位 100 項目中 19 項目, Collins の特徴語リストでは上位 100 項目中 35 項目を占める. このように特定 (あるいはごく少数) の作品にしか生起しないものの, 著しく頻度の高い項目 (特に固有名詞) が「特徴語」とされてしまう問題を回避するためには, Document Frequency の閾値を設定したり, 固有名詞を除外するというオプションも考えられるが, データ選択の恣意性の問題や, 作業コストを考慮すると, より合理的な手段で特徴語を抽出する方法が望ましい. かつ, サブコーパス間で相対頻度が一貫して異なる語に焦点を当てる手法が望ましい. そこで, 本研究では Random Forests (Breiman, 2001) を特徴語抽出のツールとして使用する.

2.2 Random Forests によるテキスト分類と作者判別マーカーの抽出

Random Forests は, ensemble learning による回帰・分類ツールであり, 金・村上 (2007), 小林・田中・富浦 (2011) などテキスト分類に用いられ, 高い分類精度を誇る手法であるが, 分類に加えてデータ分類に貢献度の高い変数を出力することができる. 本研究では, ランダムフォレストにかける語彙変数の数を 1000 項目から 100 項目まで変化させて実験を行った. 分類精度は 96-100%であった. 変数 300 項目の時に最も高い平均精度が得られた. 図 1 および表 6 は実行結果の要約である. 多次元尺度構成法の散布図では Dickens と Collins のテキストが明確に識別できる. Collinsn 作品のうち outliers となっているのは, 処女作の Antonina (古代ローマを舞台にした歴史物) と二作目の *Rambles Beyond Railways* (スケッチ) であり, いずれも後の作品とは, 大きく主題・作風が異なるテキストである. それ以外の作品は密集したクラスターを形成している.

表 7 には Random Forests の出力結果から MeanDecreaseGini を基準に, 上位 100 項目の特徴語 (作者判別マーカー) を列挙したものである.

*1 それ以外にも, Pickwick (9), Nickolas (12), Sam (27), Weller (30), Dorrit (31), Florence (39) などごく少数のテキストにしか生起していない. そもそも, こうした固有名詞は (少なくとも文体論的観点からは) 特徴語と呼ぶべき語とはいえない.

表 5 Dickens の key words 上位 40 項目: Collins と比較した場合, 対数尤度比 (LLR) 降順
Table 5 40 most significant key words of Dickens compared with Wilkie Collins (sorted according to log-likelihood ratio)

Rank	Word	Frequency	DF	Proportion	LLR
1	upon	12,990	24	0.27%	8871.058
2	and	176,688	24	3.65%	8215.058
3	mr	31,312	24	0.65%	5151.845
4	very	14,312	24	0.30%	3639.361
5	so	20,986	24	0.43%	3479.701
6	a	109,288	24	2.26%	2666.071
7	but	26,202	24	0.54%	2580.812
8	said	30,698	24	0.63%	2457.399
9	pickwick	2,198	2	0.05%	2373.409
10	great	6,975	24	0.14%	2015.407
11	much	7,268	24	0.15%	1912.053
12	nicholas	1,740	4	0.04%	1878.858
13	they	17,630	24	0.36%	1863.395
14	tom	1,846	21	0.04%	1712.521
15	dombey	1,420	1	0.03%	1533.321
16	replied	3,875	24	0.08%	1447.446
17	john	2,113	24	0.04%	1432.579
18	pecksniff	1,250	1	0.03%	1349.755
19	gentleman	4,601	24	0.10%	1344.429
20	or	16,102	24	0.33%	1278.461
21	king	1,687	21	0.03%	1275.183
22	joe	1,172	17	0.02%	1251.145
23	martin	1,121	7	0.02%	1210.460
24	boffin	1,105	1	0.02%	1193.183
25	being	6,904	24	0.14%	1181.222
26	though	3,841	24	0.08%	1180.777
27	sam	1,210	4	0.02%	1143.636
28	many	4,225	24	0.09%	1141.551
29	down	8,408	24	0.17%	1088.151
30	weller	992	2	0.02%	1071.165
31	dorrit	971	2	0.02%	1048.489
32	`em	1,331	21	0.03%	1047.537
33	old	9,624	24	0.20%	1041.342
34	nickleby	939	1	0.02%	1013.936
35	clennam	938	1	0.02%	1012.856
36	were	17,745	24	0.37%	989.771
37	indeed	2,503	24	0.05%	966.672
38	such	7,351	24	0.15%	961.150
39	florence	1,083	4	0.02%	940.432
40	squeers	861	1	0.02%	929.711

表 6 Random Forests 実行結果
Table 6 A result of running Random Forests

Call:			
randomForest(formula = dat\$AuthGroup ~.,			
data=dat[,2:301], proximity=T, importance=T, mtry=20)			
Type of random forest: classification			
Number of trees:	500		
No. of variables tried at each split:	20		
OOB estimate of error rate:	0%		
Confusion matrix:			
	Collins	Dickens	class.error
Collins	24	0	0
Dickens	0	24	0

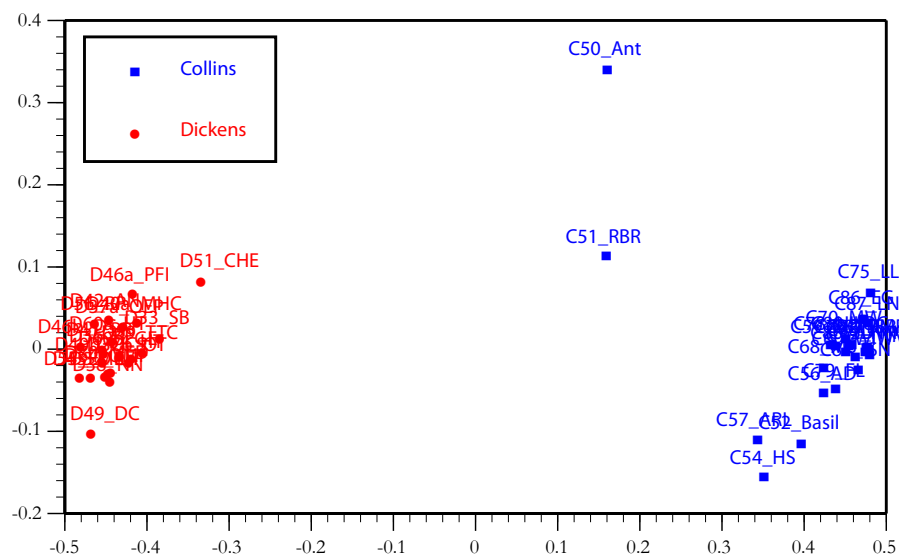


図 1 多次元尺度構成法による Dickens 作品, Collins 作品の空間布置 (Random Forest の親近度行列出力に基づく)
Fig. 1 A multi-dimensional scaling plot based on the proximity matrix generated by RF: Dickens versus Collins

表 7 Random Forests の結果を基に取り出した作者判別マーカー上位 100 項目: MeanDecreaseGini 降順ソート
Table 7 100 Important variables: Dickens markers and Collins markers (in the descending order of MeanDecreaseGini)

Dickens markers:
very, many, upon, being, much, and, so, with, a, such, indeed, air, off, but, would, down, great, there, up, or, were, head, they, into, better, quite, brought, said, returned, rather, good, who, came, having, never, always, ever, replied, boy, where, this, sir, well, gone, looking, dear, himself, through, should
Collins markers:
first, words, only, end, left, moment, room, last, letter, to, enough, back, answer, leave, still, place, since, heard, answered, time, looked, person, mind, on, woman, at, told, she, own, under, just, ask, once, speak, found, passed, her, which, had, me, felt, from, asked, after, can, side, present, turned, life, next, word

3. 共著作品と canonical works の比較

Random Forests の結果抽出した 100 項目の語彙変数を使って階層クラスター分析, 主成分分析を実行した結果が図 2, 3 である. 二つの図は共著作品 4 点と Dickens, Collins の単著作品との類似度・近接度を視覚化している.

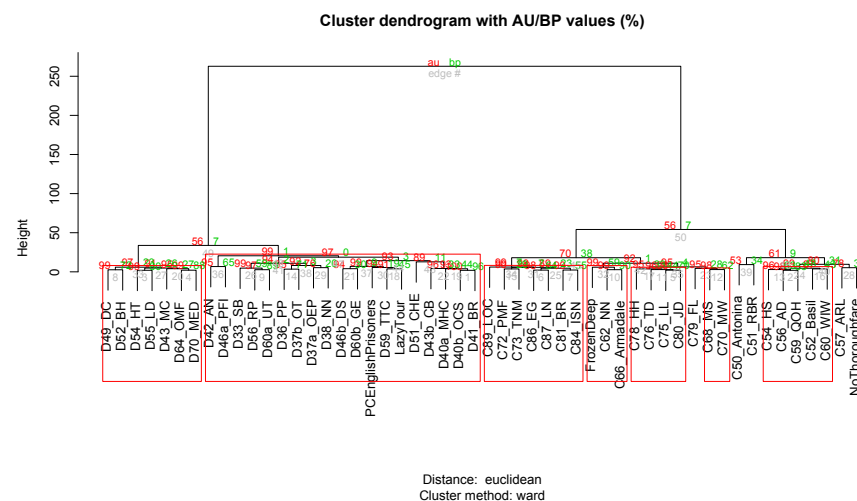


図 2 Dickens 作品, Collins 作品と共著作品 4 点: R パッケージ pvclust による階層クラスタリング
Fig. 2 24 Dickens texts, 24 Collins texts, and the four collaborations: Hierarchical clustering with AU/BP values

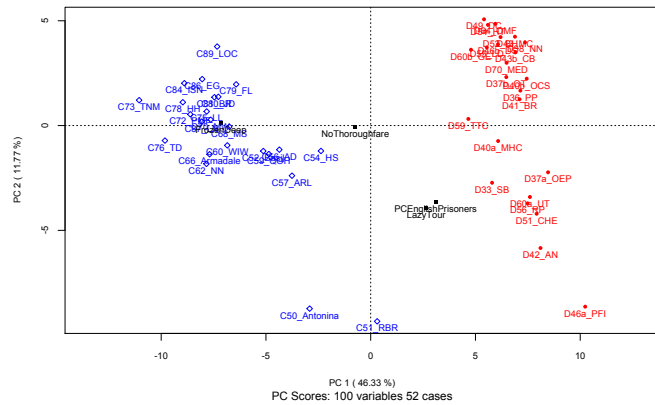


図 3 主成分分析による Dickens 作品, Collins 作品と共著作品 4 点の空間布置: 主成分得点
Fig. 3 Principal Component Analysis on 24 Dickens texts, 24 Collins texts, and the four collaborations: PC scores

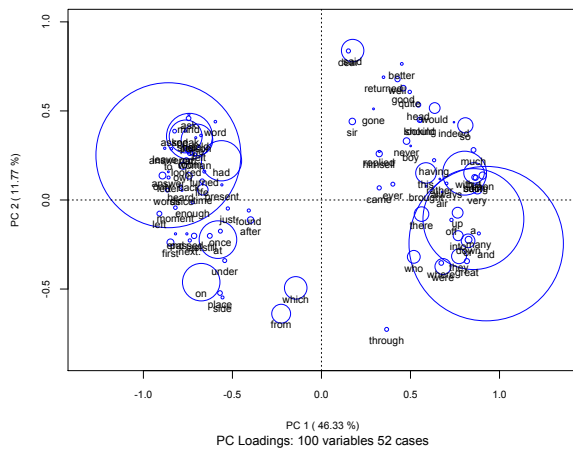


図 4 主成分分析による Dickens 作品, Collins 作品と共著作品 4 点の空間布置: 主成分負荷量
Fig. 4 Principal Component Analysis on 24 Dickens texts, 24 Collins texts, and the four collaborations: PC loadings

3.1 章分割したデータ

The Frozen Deep を除いて章毎に分割したテキストを用いて再び階層的クラスター分析, 主成分分析を行った結果を図 5, 6 に示している. 書誌学的記録をもとにした表 2 のと比較すると興味深いことが見えてくる. 単著の章は各作家単著との類似性が高い一方, 二人の作家が共同執筆した章は Dickens 側に偏っている. Dickens が Collins の描いた部分にも大きく手を入れていることが読み取れる結果である.

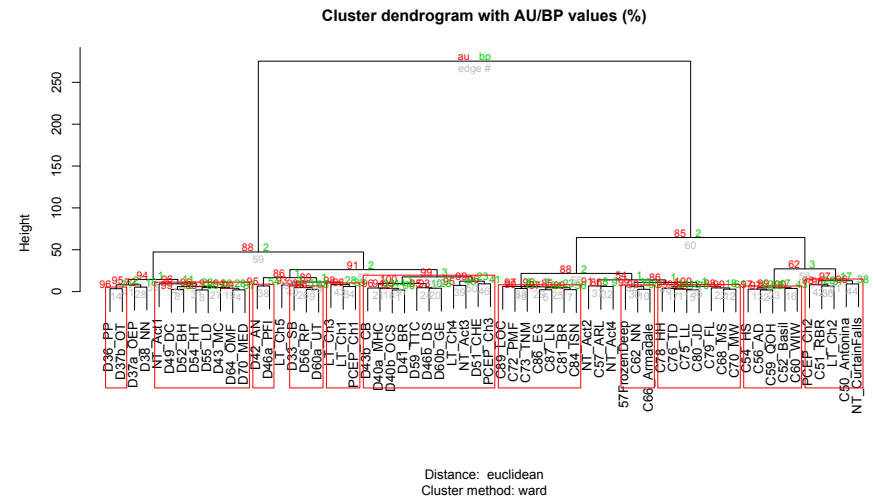


図 5 Dickens 作品, Collins 作品と章分割した共著作: R パッケージ pvclust による階層クラスタリング
Fig. 5 24 Dickens texts, 24 Collins texts, and 14 collaborative parts: Hierarchical clustering with AU/BP values

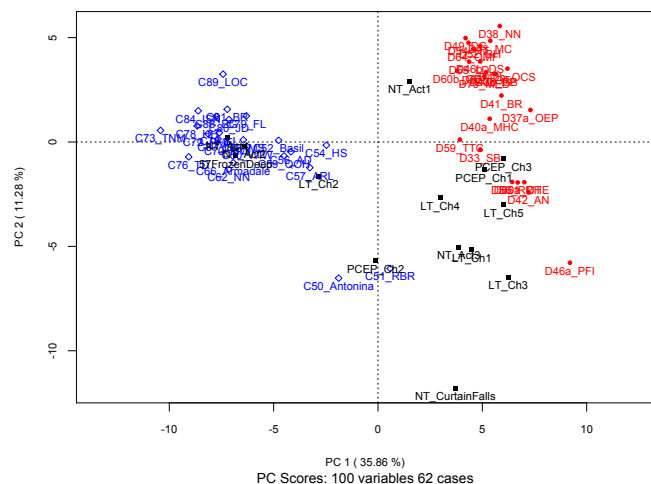


図 6 主成分分析による Dickens 作品, Collins 作品と章分割した共著作の空間布置: 主成分得点

Fig. 6 Principal Component Analysis on 24 Dickens texts, 24 Collins texts, and 14 collaborative parts: PC scores

参考文献

- 1) Allingham, P.V.(2011) A Comprehensive List of Dickens's Short Fiction, 1833-1868. *The Victorian Web*. Online resource. (Last accessed 20 December 2011.) Available online at <http://www.victorianweb.org/authors/dickens/pva/5.html>
- 2) Breiman, L. (2001) Random forests, *Machine Learning*, 45: 5–23.
- 3) Dunning, T. (1993) Accurate methods for the statistics of surprise and coincidence, *Computational Linguistics*, 19/1: 61–74.
- 4) Henry, A. and Roseberry, R.L. (2001) Using a small corpus to obtain data for teaching genre in M.Ghadessy, A.Henry and R.L.Roseberry (eds.) *Small Corpus and ELT*. Amsterdam/Philadelphia, Pa.: John Benjamins. 93–133.
- 5) 金明哲・村上征勝 (2007) 「ランダムフォレスト法による文章の書き手の同定」『統計数理』第 55 卷 2 号, 255–268.
- 6) 小林雄一郎・田中省作・冨浦洋一 (2011) 「ランダムフォレストを用いた英語科学論文の分類と評価」『情報処理学会研究報告 IPJS SIG Technical Report』第 90 号 (2011-CH-90): 53–68.

- 7) Lane, M. (1956) Introduction. In *The Oxford Illustrated Dickens, Christmas Stories*. Oxford: OUP.
- 8) Nayder, L. (2002) *Unequal Partners: Charles Dickens, Wilkie Collins, and Victorian Authorship*. Ithaca/London: Cornell UP.
- 9) Scott, M. and Tribble, C. (2006) *Textual Patterns: Key words and corpus analysis in language education*. Amsterdam/Philadelphia: John Benjamins.
- 10) Stone, H. (ed.) (1968) *Charles Dickens' Uncollected Writings from "Household Words" 1850–1859*. 2 vols. Bloomington: Indiana UP.
- 11) 高見敏子. (2003) 「『高級紙語』と『大衆紙語』の corpus-driven な特定法」『北海道大学大学院国際メディア研究科・言語文化部紀要』44: 73–105.
- 12) Thomas, D. A. (1982). *Dickens and the Short Story*. Philadelphia: University of Pennsylvania Press.
- 13) Rayson, P. and Garside, R. (2000). Comparing Corpora Using Frequency Profiling, *Proceedings of the Workshop on Comparing Corpora, Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, 1–8 October 2000, Hong Kong. 1–6. Available online at <http://www.comp.lancs.ac.uk/computing/users/paul/phd/phd2003.pdf>.