The University of Osaka
Institutional Knowledge Archive

Title	英国Classic Fictionコーパスの潜在的トピック : LDAによるテクストクラスタリング
Author(s)	田畑, 智司
Citation	言語文化共同研究プロジェクト. 2020, 2019, p. 47-58
Version Type	VoR
URL	https://doi.org/10.18910/76991
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

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

The University of Osaka

英国Classic Fictionコーパスの潜在的トピック：LDAによるテクストクラスタリング

田畑　智司

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

E-mail:　tabata@lang.osaka-u.ac.jp

あらまし　本稿では，Charles Dickensの文体を史的・通時的な観点から研究するために編纂した英国classic fictionコーパスであるOsaka Reference Corpus for HIstrorical/Diachronic Stylistics (ORCHIDS) に対して，Latent Dirichlet Allocation (LDA) によるトピックモデリングを実行し，比較対象となる英国18, 19世紀の代表的fictionに参照して，Dickens作品を特徴付ける意味構造を検討する。まず，生成された各トピックがORCHIDSの標本テクストやサブコーパスとどのように結びついているかを考察する。トピックを構成する単語リストと作家やテクストとの結びつきを探るとともに，主成分分析による視覚化を行うことで，LDAがDickens作品や18, 19世紀作品サブコーパスの文体を特徴付ける意味構造の抽出，記述に有効であることを示す。マクロ的視座からは，ORCHIDSにおけるサブコーパスのクラスタリング，トピックのクラスタリングを散布図の形で要約し，構築されたトピックモデルがORCHIDSを構成するテクストの特徴付け，潜在的意味構造の発見に寄与するということを明らかにする。

キーワード　ディケンズ，文体，トピックモデル，共起語彙項目，意味構造

Latent Topics in British Classic Fiction:
Using LDA to classify texts into meaningful groups

Tomoji Tabata

Graduate School of Language and Culture, University of Osaka

1-8 Machikaneyama-cho, Toyonaka, Osaka, 560-0043 Japan

Abstract　The present study applies Latent Dirichlet Allocation to a corpus of British classic fiction, the Osaka Reference Corpus for HIstrorical/Diachronic Stylistics (ORCHIDS) with a view to analyzing latent semantic structures underlying in the corpus and mapping its constituent texts in the network of words, topics, and texts. What is of special interest is that by means of this approach it is possible to shed new light on thematic structures composed by a large number of infrequent words, which would otherwise escape the net of key-word statistics due to infrequency of occurrence. Emerging results from this research are expected to lead to a new line of inquiry into key semantic patterns that characterize Dickens's semantic style as well as thematic features of 18th- and 19th-century classic fiction. The results also suggest a possibility of building a bridge between findings from machine learning text

mining and stylistics, distant reading and close reading, with an empirical interplay of insights that will benefit historical/diachronic stylistics.

Keywords Dickens, style, topic model, lexical co-occurrence, semantic structure

1. はじめに

確率論的トピックモデリングの代表的アルゴリズム Latent Dirichret Allocation (LDA) が大規模文書データの潜在的意味構造のマイニングに有効であることはすでに Blei *et al.* (2003), Blei (2012a/b), Fothergrill *et al.* (2016) などによって示されており，LDA を応用した文学研究 (Rhody, 2012; Weingart and Meeks, 2012; Tangherlini and Leonard, 2013; Goldstone and Underwood, 2013)，テクスト研究 (Schmidt, 2012; Jaworska and Nanda, 2016; Törnberg and Törnberg, 2016)，文体研究 (田畑, 2017b; Tabata, 2017) への応用の可能性を探る試みがなされてきている。近年進展著しい Digital Humanities におけるトピックモデルの応用，コーパス言語学における応用の主な潮流については，田畑 (2017b: 3–4) において概観を示した。

本稿では，英国において小説が文学ジャンルとして大きく興隆した 18–19 世紀を代表する作品を収録したコーパス ORCHIDS をもとに，LDA によって構築されたトピックモデルによっていかなる作品のクラスタリングが生じるか，また主要なトピックがコーパスのどのような側面を照らし出すのかを探索的に検討する。

2. コーパス

史的通時的文体論研究のためのコーパスとして編纂された ORCHIDS (Osaka Reference Corpus for HIstorical/Diachronic Stylistics）の構成については，Tabata (2015) にて詳細な解説を行なっているが，当該コーパスはその後，作品の入れ替え，追加，品詞情報アノテーションが施され，改訂を行なった。Appendix 1–3 にコーパス収録の作品一覧を示している。Dickens 作品を収めたサブコーパス (Appendix 1) は品詞情報アノテーションを施したことによるトークン数増（次の段落においてふれる）以外の変化はない。18 世紀サブコーパス (Appendix 2) は，Defoe の *Robinson Crusoe* (1719) から *The Farther Adventures of Robinson Crusoe* を切り離し，独立したテクストとして取り扱う一方，Eliza Haywood, *The Fortunate Foundlings* (1744) を追加し，計 26 作品，総トークン数は 4,622,929 語となった。19 世紀サブコーパス (Appendix 3) は，George Eliot の作品中著しくトークン数が少ない *Brother Jacob* (1864) を外し，代わって長編 *Ramola* に入れ替え，Thackeray の長編小説 3 点，*The History of Pendennis* (1848–1850), *The Newcomes* (1854–1855), *The Virginians* (1857–1859) を加えた結果，34 作品，総トークン数 6,655,409 語に増加した。

ORCHIDS の大きな改訂として，TreeTagger[1]による品詞情報アノテーションを施した。TreeTagger では，英語テクストに付与する品詞標識として，主に自然言語処理分野で広く

[1] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ なお C5 (BNC) タグセットを使用するためには，https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/english-bnc.par.gz よりダウンロードした圧縮ファイルを展開し，パラミターファイルをデフォルトのものと入れ替える必要がある。

用いられている Penn Treebank タグセットと，BNC (British National Corpus) で採用されいる CLAWS (Constituent Likelihood Automatic Word-tagging System) の C5 タグセット[2]の 2 種類が利用可能である。ORCHIDS の改訂にあたって，大規模共時均衡コーパスである BNC, BNC2014, COCA (Corpus of Contemporary American English)，通時コーパス COHA (Corpus of Historical American English) など，fiction のデータを大量に収録した参照コーパスとの比較可能性を担保するために C5 タグセットを採用した。C5 タグセットは CLAWS の初期設定である C7 タグセットの 137 種に及ぶ品詞・ワードクラス分類のうち，特に名詞，代名詞，副詞類のカテゴリーをスリム化し少数のタグにまとめるとともに，VBDR (*were*), VBDZ (*was*) のようにタグと語形とがほぼ 1 対 1 対応するような細分化したタグを廃し，全てのカテゴリーを 3 文字（アルファベット大文字および数字（012））で表示する 62 種類のタグにまとめたものである。品詞情報アノテーションを施した結果，例えば *I've* などの contraction 語形は，*I*_PNP *'ve*_VHB のように 2 語に分割されるため，Tabata (2015) に記載したコーパスの語数から増加が認められる。

3. パラミター設定と MALLET による LDA の実行

トピックモデリングの結果は，モデルを構成するトピック数やモデル構築のための iteration 回数，ターゲットとする語彙項目，分析対象とするコーパスのサイズ，あるいはコーパスに収録されるテクスト（セグメント）サイズ，ハイパーパラミター最適化の有無，さらには最適化の間隔などによって大きく左右される。設定項目の組み合わせは無数におよび，一意に最適解となりうる単純明快な設定は存在しえない。ほとんどの場合，どのようなパラミターを設定するかは，どのようなリサーチクエスチョンを設定するかによる。これまでに，構築するモデルのトピック数や iteration の回数の最適値，さらにはトピックの質を，データの内的指標，例えば perplexity（平均分岐量）や "held-out likelihood" を総語数で除した値，LL/token 比[3](Bhatia and Kaluza, 2018: 244–6) や，抽出したトピックを構成する語彙の一貫性 (coherence) で評価する方法 (Mimno *et al.*, 2011) が示されている。LL/token 比はトピック数の増加に準じ，向上するとされているが，トピック数の変化による LL/token 比の関係を調査した Motz *et al.* (2018) では，概ねトピック数 50 前後で LL/token 比は飽和に近づくという結果を報告している。また，一方で，LL/token 比と人間によるモデルの評価の間には相関が認められないという指摘もある (Kapadia, 2019) もある。[4]

本研究では，英国 classic fiction の潜在的意味的特徴を探る目的で，以下のような条件を設定した。

---

2 http://ucrel.lancs.ac.uk/claws5tags.html

3 本研究で使用した Mallet ではモデルの尤度を表す指標として LL/token（対数尤度を総語数で除した数値）が示される。この値が 0 に近いほど，モデルの予測性能が高いと評価される。"LL/token indicats the model's likelihood, divided by the total number of tokens, indicating how likely the data is given the model. Increasing values mean the model is improving." (Bhatia and Kaluza, 2018)

4 "[R]recent studies have shown that predictive likelihood (or equivalently, perplexity) and human judgment are often not correlated, and even sometimes slightly anti-correlated." (Kapadia, 2019)

1. トピック数: 50
2. iteration 回数: 1,000
3. テクスト分割サイズ: 1,000 語
4. hyperparameter 事前設定値の影響期間 (burn-in): 50
5. hyperparameter 最適化間隔: 10
6. ターゲットとする語彙項目は名詞，形容詞，一般動詞に限定（品詞タグの情報を活用）

LDA に基づくトピックモデルは，いわゆる “bag of words” アプローチを採り，文書中で共起する語句を学習してモデルを生成することから，長編のテクスト全体を対象にすると，あまりにも多くの項目が共起項目として計算されてしまい，結果として茫漠なトピックが生成されてしまう。それを未然に防ぐために，LDA を実行するにあたって，コーパスに収録されているテクストの分割を行なった。テクストの分割サイズは，Jockers and Mimno (2013: 754) に基づき 1,000 語とし，1,000 語未満の最終セグメントは棄却した。コーパス全体で 16,179 個のセグメントが得られた。描写される場面や出来事，人物が推移していく小説というレジスターを考慮すると，1,000 語というサイズは紙媒体の図書で概ね 3 ページ前後に相当し，一定のまとまりをもつトピックをカバーする分量であると判断してよいであろう。もっとも，機械的に 1,000 語で分割するため，各セグメントとセンテンス，あるいは段落とは対応しないことを予め念頭においておく必要はあるだろう。

4. 実行結果

機械学習言語処理ツールキット MALLET を用いて LDA を実行した結果得られたトピックモデルを構成する 50 点のトピックを Table 1 に示した。なお，LL/token 比は最初の値 (iteration 10 回後) の −11.26967 から徐々に減少し，iteration 1,000 回後は −9.5205 であった。第 1 カラムはトピック番号 (ID) を，第 2 カラムにはハイパーパラミター α の値を記している。α の値の大小は，トピックの普遍性 ⇔ 局所性を表す。なお，α 値の表記は，0. を省略し，小数点以下 4 桁目以下を四捨五入している。第 4 カラムではトピックの上位 20 点の key words を weight が大きいものから順に並べている。各単語に付した品詞情報標識は紙面の都合上，品詞を示す最初の 1 文字で示す[5]が，動詞 (v) の場合のみ，時制等を識別するためにタグの 1 文字目と 3 文字目の情報を付している。Table 1 の第 3 カラムには当該トピックと最も結びつきの強い作家名あるいは作品名などを挙げている。

Table 1 において，大きな α 値をとるトピックは普遍的に生起する傾向があるため，特定の作家やテクストとは必ずしも結びつかないが，α 値が小さいトピックほど，特定のテクストやセグメントに局所限定的に生起する傾向があることが表されている。

[5] a (adjective): AJ0 (unmarked), AJC (comparative), AJS (superlative); n (noun): NN0 (neutral for number), NN1 (singular), NN2 (plural); v (verb): VVB (base form of lexical verb except infinitive), VVD (past tense form of lexical verb), VVG (*–ing* form of lexical verb), VVI (infinitive of lexical verb), VVN (past participle form of lexical verb), VVZ (*–s* form of lexical verb) なお，Table 1 では全て小文字で表示されている。

Table 1: ORCHIDS より構築したトピックモデルを構成する 50 トピック

No.	α	Author/Text	Key words, or words with the greatest weight
0	.132		door_n room_n house_n bed_n night_n went_vd window_n morning_n came_vd fire_n stairs_n open_a candle_n little_a took_vd key_n time_n opened_vd way_n going_vg
1	.044	Defoe	came_vd great_a men_n found_vd made_vd way_n went_vd told_vd time_n other_a little_a place_n good_a began_vd country_n saw_vd things_n island_n shore_n part_n
2	.114		young_a ladies_n lady_n little_a people_n great_a music_n evening_n other_a day_n girls_n time_n company_n gentlemen_n party_n old_a girl_n dinner_n happy_a came_vd
3	.028	BronteC	seemed_vd eye_n thought_vd pupils_n little_a english_a evening_n master_n certain_a monsieur_n felt_vd school_n saw_vd french_a day_n asked_vd continued_vd liked_vd answered_vd pleasure_n
4	.128		money_n pounds_n man_n business_n sum_n good_a time_n fortune_n year_n poor_a old_a paid_vn house_n years_n life_n hands_n property_n pocket_n small_a little_a
5	.184	Haywood	time_n great_a person_n made_vd found_vd other_a gave_vd lady_n fortune_n manner_n took_vd made_vn passion_n day_n told_vd thought_vd young_a good_a means_n whole_a
6	.053	Dickens_BH	says_vz man_n sir_n comes_vz old_a hand_n looks_vz little_a goes_vz lady_n takes_vz good_a friend_n head_n young_a woman_n looking_vg way_n knows_vz eyes_n
7	.012	Swift_Stella	letter_n dined_vd md_n secretary_n morning_n day_n duke_n went_vd came_vd queen_n good_a night_n great_a told_vd treasurer_n town_n court_n days_n business_n sent_vd
8	.093		wine_n table_n dinner_n glass_n tea_n good_a bottle_n bread_n time_n supper_n water_n little_a fire_n company_n took_vd breakfast_n hot_a drank_vd cold_a small_a
9	.074	Thackeray	old_a young_a little_a colonel_n gentleman_n poor_a good_a friend_n lady_n great_a boy_n man_n says_vz family_n came_vd house_n mother_n gentlemen_n fellow_n world_n
10	.065	Trollope	said_vd man_n course_n fawn_n knew_vd house_n made_vn cousin_n matter_n said_vn told_vn moment_n friend_n word_n world_n thought_vd wife_n duke_n men_n rate_n
11	.049	Richardson_P.	said_vd good_a lady_n dear_a master_n poor_a ladyship_n sir_n madam_n thing_n sure_a happy_a honour_n dear_n girl_n great_a heart_n came_vd goodness_n thought_vd
12	.279	Dickens_et_al	mind_n time_n other_a course_n case_n life_n part_n made_vn point_n circumstances_n state_n way_n subject_n man_n question_n means_n great_a fact_n certain_a character_n
13	.062	Dickens	old_a boy_n head_n boys_n dog_n time_n master_n jew_n small_a cried_vd gentleman_n replied_vd brass_n eyes_n great_a little_a got_vn kit_n eye_n dwarf_n
14	.058	GEliot	said_vd felt_vd seemed_vd life_n new_a mind_n sense_n made_vd way_n things_n other_a feeling_n sort_n men_n tone_n thought_n husband_n come_vn strong_a consciousness_n
15	.044	Fielding	answered_vd cries_vz man_n great_a lady_n gentleman_n says_vz sir_n honour_n madam_n said_vd good_a squire_n other_a poor_a reader_n fellow_n sure_a woman_n colonel_n
16	.166	Dickens	night_n old_a time_n day_n man_n came_vd light_n long_a went_vd dark_a seemed_vd dead_a bed_n child_n place_n fire_n life_n looked_vd saw_vd mind_n
17	.028	Collins_Moon.	time_n house_n count_n sergeant_n way_n diamond_n place_n man_n moment_n lady_n words_n room_n cuff_n sir_n asked_vd answered_vd looked_vd side_n mind_n person_n
18	.094	18thC	great_a book_n world_n men_n work_n books_n man_n other_a history_n good_a life_n true_a learning_n age_n reader_n knowledge_n author_n time_n art_n things_n
19	.025	GEliotGaskell	said_vd mother_n bit_n poyser_n got_vn work_n good_a folks_n things_n lad_n time_n old_a ud_n poor_a father_n gone_vn job_n way_n went_vd words_n
20	.046	Smollett	gentleman_n company_n hero_n person_n young_a adventurer_n whole_a great_a began_vd order_n friend_n knight_n time_n consequence_n other_a appearance_n house_n place_n view_n pipes_n
21	.141	Richardson	man_n good_a character_n other_a person_n young_a woman_n great_a men_n world_n nature_n love_n gentleman_n mind_n lady_n life_n opinion_n pride_n others_n fortune_n
22	.079		horse_n horses_n coach_n road_n way_n carriage_n journey_n day_n inn_n miles_n going_vg came_vd coachman_n rode_vd town_n house_n morning_n man_n landlord_n time_n
23	.041	Trollope	parliament_n man_n great_a men_n house_n country_n people_n member_n public_a government_n political_a members_n office_n bill_n party_n honourable_a speech_n friend_n side_n election_n
24	.086	Richardson_C.	man_n dear_n letter_n lady_n heart_n brother_n creature_n thing_n mother_n family_n favour_n time_n sister_n friend_n given_vn uncle_n dear_a honour_n reason_n friends_n
25	.021	Dickens	little_a yard_n head_n headstone_n work_n carrier_n returned_vd dolls_n sloppy_a boy_n hand_n marshalsea_n dressmaker_n got_vd eye_n got_vn fire_n turnkey_n madame_n eyes_n
26	.146	Dickens19thC	man_n cried_vd hand_n head_n men_n hands_n blood_n dead_a arms_n face_n moment_n body_n fell_vd ran_vd poor_a saw_vd came_vd fire_n ground_n death_n
27	.204		heart_n life_n love_n poor_a tears_n words_n happy_a eyes_n mind_n world_n soul_n heaven_n happiness_n nature_n sorrow_n hand_n grief_n death_n dear_a day_n
28	.077		little_a trees_n garden_n house_n side_n sun_n green_a ground_n old_a stone_n water_n village_n road_n air_n way_n grass_n country_n walk_n summer_n great_a
29	.025	Dickens	ai_n vbz_n said_vd man_n says_vz took_vd got_vn wery_a boy_n way_n wos_n good_a got_vd mum_n name_n war_n bit_n night_n pipe_n knowed_vd
30	.024	Smollett_TFI	great_a city_n other_a place_n country_n people_n town_n water_n english_a part_n mountains_n number_n small_a built_vn french_a sea_n roman_a agreeable_a seen_vn different_a
31	.109	Austen	sister_n time_n thing_n good_a great_a feelings_n friend_n seemed_vd sure_a felt_vd pleasure_n happy_a day_n moment_n other_a mother_n family_n brother_n subject_n happiness_n
32	.031	Swift	great_a other_a country_n majesty_n nature_n whole_a master_n people_n reason_n feet_n court_n certain_a use_n body_n number_n emperor_n animal_n manner_n persons_n author_n
33	.060	Dickens_CHE	king_n great_a army_n men_n people_n prince_n french_a english_a duke_n war_n made_vd earl_n time_n came_vd enemy_n queen_n country_n soldiers_n made_vn took_vd
34	.044	Dickens_late	young_a lady_n gentleman_n family_n head_n dinner_n manner_n company_n state_n little_a honour_n gentlemen_n dear_n general_a going_vg society_n occasion_n barnacle_n ma'am_n great_a
35	.043	Dickens_early	said_vd replied_vd gentleman_n lady_n young_a inquired_vd gentlemen_n man_n great_a little_a old_a boy_n ladies_n exclaimed_vd sir_n countenance_n other_a friend_n ma'am_n morning_n
36	.120	Dickens_mid	said_vd little_a dear_n child_n aunt_n old_a poor_a good_a mother_n dear_a thought_vd came_vd face_n went_vd time_n head_n going_vg papa_n returned_vd looking_vg
37	.133	Dickens	said_vd sir_n man_n old_a returned_vd gentleman_n good_a cried_vd friend_n replied_vd young_a head_n great_a looking_vg hand_n name_n time_n way_n fellow_n manner_n
38	.013	Stern_Tristram	uncle_n father_n said_vd quoth_n corporal_n world_n honour_n thing_n man_n replied_vd cried_vd brother_n hand_n slop_n continued_vd way_n nature_n chapter_n head_n whole_a
39	.041		captain_n ship_n boat_n sea_n board_n water_n shore_n wind_n ships_n great_a men_n voyage_n vessel_n tide_n river_n deck_n gills_n sail_n going_vg cabin_n
40	.163		letter_n day_n morning_n time_n letters_n came_vd went_vd night_n house_n wrote_vd written_vn paper_n days_n gone_vn took_vd note_n name_n town_n news_n evening_n
41	.292	Dickens19thC	said_vd hand_n face_n eyes_n looked_vd looking_vg door_n moment_n voice_n room_n head_n turned_vd saw_vd words_n sat_vd man_n went_vd hands_n stood_vd took_vd
42	.011		stage_n theatre_n play_n manager_n audience_n tragedy_n baron_n piece_n part_n theatrical_a pit_n scene_n actor_n night_n honner_n performance_n actors_n curtain_n scenes_n comic_a
43	.104	19thC	hair_n white_a black_a eyes_n little_a face_n large_a dress_n head_n wore_vd looked_vd blue_a red_a long_a gold_n figure_n coat_n fine_a dressed_vn small_a
44	.298	19thC	said_vd man_n good_a way_n thing_n thought_vd told_vd woman_n sure_a going_vg people_n things_n went_vd poor_a came_vd knew_vd wife_n time_n little_a word_n
45	.031	Trollope	church_n bishop_n archdeacon_n clergyman_n man_n old_a hospital_n parish_n warden_n men_n sermon_n new_a service_n dean_n cathedral_n curate_n good_a rector_n congregation_n signora_n
46	.043		doctor_n patient_n surgeon_n man_n physician_n medical_a fever_n bed_n ill_a sick_a health_n nurse_n day_n case_n poor_a medicine_n illness_n apothecary_n death_n good_a
47	.050	Defoe_Journal	people_n prison_n time_n prisoner_n other_a justice_n court_n city_n house_n law_n taken_vn poor_a case_n great_a plague_n magistrate_n evidence_n death_n prisoners_n houses_n
48	.161		father_n mother_n son_n old_a brother_n child_n family_n daughter_n children_n young_a years_n sister_n wife_n poor_a boy_n house_n life_n good_a husband_n man_n
49	.087	Dickens_j	old_a people_n great_a street_n streets_n place_n little_a houses_n windows_n men_n man_n other_a town_n way_n looking_vg city_n walls_n day_n house_n children_n

5. 主成分分析によるトピックとテクストのクラスタリング

Table 1 はトピックと特定の作家やテクストとの強い結びつきが認められることを示唆するものである。そこで，主成分分析 (PCA) を実行して生成されたトピック 50 点とテクスト（セグメント）との関係性の視覚化を試みた。まず，1,000 語に分割したセグメント 16,179 個のデータをもとに PCA を実行した結果得られた上位 3 主成分の得点を 3 次元散布図に布置したのが Fig. 1 である。データ数が著しく多いため，上位 3 主成分の寄与度は 5.04%, 3.85%, 3.46%と低い値にとどまっている。図中の青のデータポイントは 18 世紀サブコーパス収録作品のセグメント，黒は Dickens 作品，赤は Dickens を除く 19 世紀サブコーパス収録作品のセグメントである。3 つのサブコーパスのうち，特に Dickens と他の 19 世紀作家の作品セグメント間にはかなりの程度オーバーラップが認められるものの，概ね 18 世紀と 19 世紀の 2 大クラスターに識別可能である。この散布図は，それぞれのサブコーパスを特徴付けるトピックの生起傾向があることを示唆していると解釈できる。

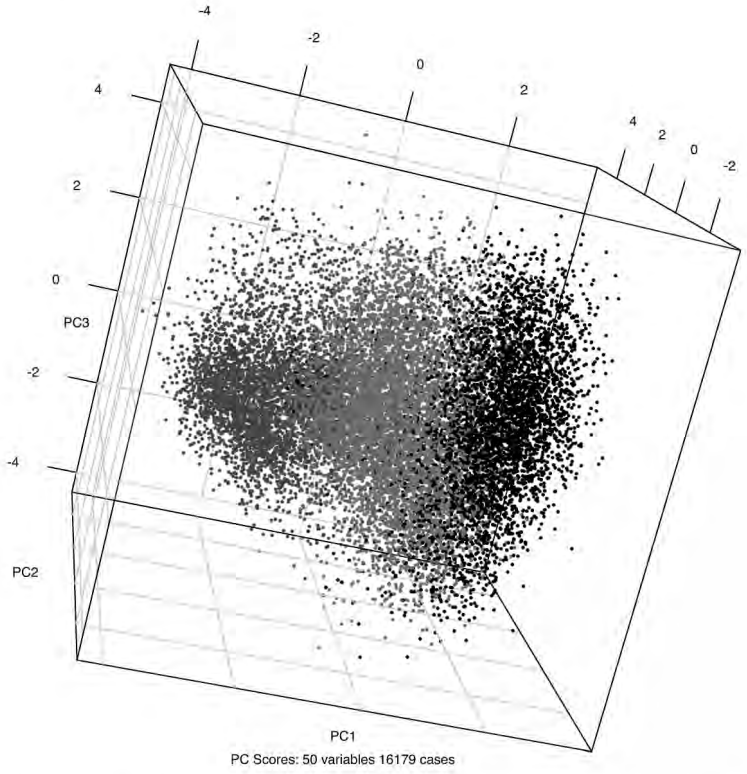

Fig. 1: テクストを 1,000 語毎に分割したデータの主成分スコア散布図

次に，各テクストの平均トピック含有率 (topic proportion) をもとめ，その値でテクスト全 84 点の傾向を探った。PCA の結果得られた上位 2 主成分をプロットしたものが，Figs. 2–3 である。Fig. 2 では，3 つのサブコーパスがほぼ独立したクラスターを構成していることが観察できる。Dickens による児童向けの歴史物語である *A Child's History of England* は，Table 1 より，トピック 33 (*king*_n, *great*_a, *army*_n, *men*_n, *people*_n, *prince*_n, *french*_a, *english*_a, *duke*_n, *war*_n, *made*_vd, *earl*_n, *time*_n, *came*_vd, *enemy*_n, *queen*_n, *country*_n, *soldiers*_n,

*made*_vn, *took*_vd) がテクストを特徴付ける語彙項目であることが示されているが，Fielding に代表される 18 世紀の retrospective narratives と類似した特徴が認められることは高頻度語彙あるいは Language Action Types (LATs) の生起率をデータとした分析 (田畑, 2017a) でも示されており，トピックモデリングでも一貫した傾向が認められる。その他，Sterne, *A Sentimental Journey through France and Italy* が 19 世紀作品クラスターの上に位置しているが，Sterne の作品と 19 世紀作品群とは第 2 主成分においては一定の差が認められるものの，第 1 主成分においては 19 世紀作品群のほぼ真ん中あたりの値を取っている。

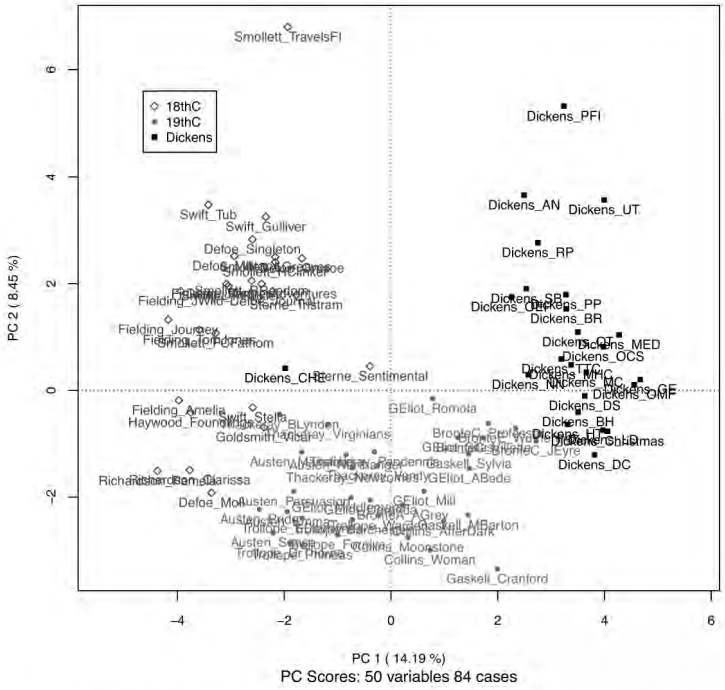

Fig. 2: 50 個のトピックの平均含有率に基づくテクストのクラスタリング

別の説明変数データに基づいたこれまでの研究ではこうした類似性は確認できていないため，これについては別途詳細に考察を要する。Thackeray, *Barry Lyndon* が 18 世紀作品に酷似したトピックを持っていることが示されている。同様の傾向は異なる説明変数に基づく田畑 (2017a) の分析においても指摘されている。また，19 世紀作品では，Emily Brontë, *Wuthering Heights* が Dickens のクラスターに最も近い位置に布置されていることは興味深い。トピック数を増減させた場合に，この作品の位置取りがどのように変化するかについては，稿を改めて論じることにする。トピック含有率のテクスト平均値データをもとに，機械学習分類器 Random Forests によるデータ分類を試みたところ，主成分分析の結果とほぼ一致するアウトプットが得られた (Table 2)。誤分類されたテクストは上述の 4 点である。分類正確率は 95%を上回っており，トピック含有率データをもとに高い精度でテクストのクラスタリングが可能であることを裏付ける結果であると解釈できる。

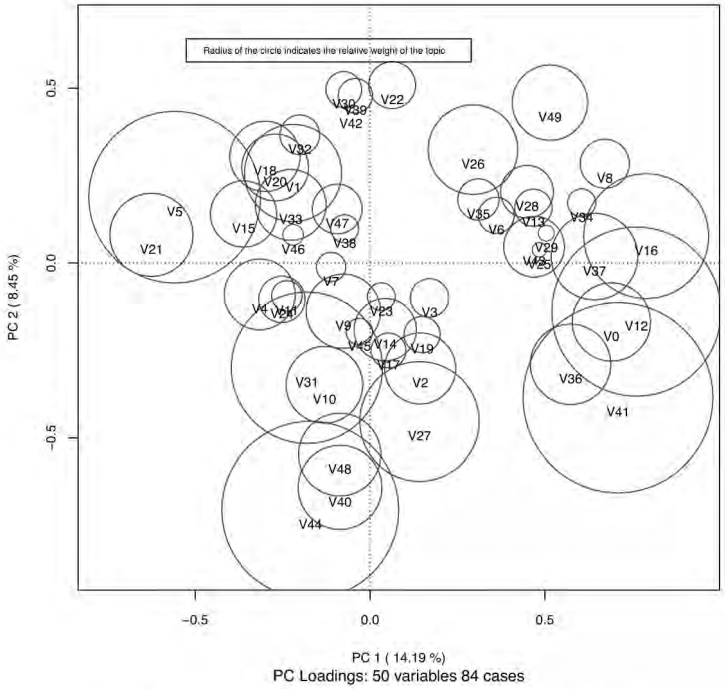

Fig. 3: 主成分負荷量：50 個のトピック間の関係

Table 2: Random Forests 実行結果

Call: randomForest(formula = ORCHIDS$text.group ∼.,
data=ORCHIDS, proximity=T, importance=T, ntree=1e+05)
Type of random forest: classification
Number of trees: 1e+05
No. of variables tried at each split: 9
OOB estimate of error rate: 4.76%
Confusion matrix:

	18th-C.	19th-C.	Dickens	class.error
	25	1	0	0.03846154
	1	32	1	0.05882353
	1	0	24	0.04166667

6. 結び

本稿では，LDA によるトピックモデリングを実行し，英国 18, 19 世紀の代表的 fiction ならびに，Dickens 作品間にどのようなクラスターが生じるか分析を行った。生成されたトピックのうち，α が低いトピックは特定のテクストやサブコーパスと結びつく傾向があることが確認できた。もっとも，それぞれのトピックに適切なラベルを付与するためには，トピックを構成する key words 間の意味的関係性を，結びつきの強い作家やテクストを考慮に入れて解釈する必要があり，これについては別稿にて論じることとする。テクストを 1,000 語毎に機械的に分割したセグメントのトピック含有率，および各テクストの平均トピック含有率のいずれも，同様のクラスタリング傾向が生じることが明らかとなった。主成分分析による視

覚化を通して，少なくとも，LDA が Dickens 作品や 18，19 世紀作品サブコーパスの文体を特徴付ける意味構造の抽出，記述に有効であることを示す結果が得られた。マクロ的視座からは，ORCHIDS におけるサブコーパスのクラスタリング，トピックのクラスタリングを散布図の形で要約し，構築されたトピックモデルが ORCHIDS を構成するテクストの特徴付け，潜在的意味構造の発見に寄与するということが明らかとなった。

References

[1] Bhatia, A. S. and B. Kaluza (2018) *Machine Learning in Java: Helpful techniques to design, build, and deploy powerful machine learning applications in Java*, 2nd Edition, Packt Publishing.

[2] Blei, D. M., Ng, A. and M. Jordan (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3: 993–1022.

[3] Blei, D. M. (2012a) Topic modeling and digital humanities, *Journal of Digital Humanities*, 2 (1), 2012. Available at http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/

[4] Blei, D. M. (2012b) Probabilistic Topic Models. *Communications of the ACM*, 55 (4): 77–84. doi:10.1145/2133806.2133826

[5] Brett, M. R. (2012) Topic Modeling: A Basic Introduction, *Journal of Digital Humanities*, 2 (1). Available at http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/

[6] Fothergill, R., Cook, P., and T. Baldwin (2016) Evaluating a Topic Modelling Approach to Measuring Corpus Similarity, *Procedings of LREC 2016, Tenth International Conference on Language Resources and Evaluation*, 23th-28th May, 2016, Portorož, Slovenia. Available at http://www.lrec-conf.org/proceedings/lrec2016/pdf/154_Paper.pdf

[7] Goldstone, A. and T. Underwood (2012) What Can Topic Models of PMLA Teach Us About the History of Literary Scholarship? *Journal of Digital Humanities*, 2 (1). Available at http://journalofdigitalhumanities.org/2-1/what-can-topic-models-of-pmla-teach-us-by-ted-underwood-and-andrew-goldstone/

[8] Jaworska, S. and A. Nanda (2016) Doing Well by Talking Good: A Topic Modelling-Assisted Discourse Study of Corporate Social Responsibility, *Applied Linguistics*, 2016: 1–28. doi: 10.1093/applin/amw014

[9] Jockers, M. L. and D. Mimno (2013) Significant themes in 19th-century literature, *Poetics*, 41 (6): 750–769.

[10] Kapadia, S. (2019) Evaluate Topic Models: Latent Dirichlet Allocation (LDA). (Online resource) Available at https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

[11] McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., and D. Jurafsky (2013) Differentiating language usage through topic models, *Poetics*, 41 (6): 607–625.

[12] Miller, I. M. (2013) Rebellion, crime and violence in Qing China, 1722–1911: A topic modeling approach, *Poetics*, 41 (6): 626–649.

[13] Mimno, D. M. , H. M. Wallach, E. M. Talley, M. Leenders, and A. K. McCallum (2011) Optimizing semantic coherence in topic models, *EMNLP '11: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–272.

[14] Motz, B., T. Busey, M. Rickert, and D. Landy (2018) Finding Topics in Enrollment Data, Paper presented at the 11th International Conference on Educational Data Mining (EDM), Raleigh, NC, Jul 16–20, 2018. Available at http://educationaldatamining.org/files/conferences/EDM2018/papers/EDM2018_paper_159.pdf

[15] Rhody, L. M. (2012) Topic Modeling and Figurative Language, *Journal of Digital Humanities*, 2 (1). Available at http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/

[16] Schmidt, B. M. (2012) Words Alone: Dismantling Topic Models in the Humanities, *Journal of Digital Humanities*, 2 (1). Available at http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/

[17] Tabata, T. (2015) Stylometry of Dickens's Language: An Experiment with Random Forests, in Arthur, P. L. and K. Bode (eds.) *Advancing Digital Humanities: Research, Methods, Theories*, Palgrave Macmillan, 28–53.

[18] Tabata, T. (2017) Mapping Dickens's Novels in a Network of Words, Topics, and Texts: Topic Modelling a Corpus of Classic Fiction, *Proceedings of the 7th International Conference of the Japanese Association for Digital Humanities* (JADH2017), 73–78.

[19] 田畑 智司 (2017a)「修辞項目のアノテーションを活用したテクスト分析—Language Action Types を通して見る文体—」『英語コーパス研究』第 24 号（英語コーパス学会）101–122.

[20] 田畑 智司 (2017b)「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付け—」田畑智司編 言語文化共同研究プロジェクト 2016『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) 5–21.

[21] Tangherlini, T. R. and P. Leonard (2013) Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research, *Poetics*, 41 (6): 725–749.

[22] Törnberg, A. and P. Törnberg (2016) Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum, *Discourse & Society 2016*, Vol. 27 (4): 401–422.

[23] Weingart, S. B. and E. Meeks (2012) The Digital Humanities Contribution to Topic Modeling, *Journal of Digital Humanities*, 2 (1). Available at http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/

Appendix

Appendix 1: Dickens component of ORCHIDS

No.	Author	Texts	Abbr.	Category	Date	Word-tokens
1	Dickens	*Sketches by Boz*	(SB)	Sketches	1833–6	189,535
2	Dickens	*The Pickwick Papers*	(PP)	Serial Fiction	1836–7	305,135
3	Dickens	Other Early Papers	(OEP)	Sketches	1837–40	67,622
4	Dickens	*Oliver Twist*	(OT)	Serial Fiction	1837–9	160,056
5	Dickens	*Nicholas Nickleby*	(NN)	Serial Fiction	1838–9	326,802
6	Dickens	*Master Humphrey's Clock*	(MHC)	Miscellany	1840–1	47,677
7	Dickens	*The Old Curiosity Shop*	(OCS)	Serial Fiction	1840–1	220,435
8	Dickens	*Barnaby Rudge*	(BR)	Serial Fiction	1841	257,126
9	Dickens	*American Notes*	(AN)	Sketches	1842	102,358
10	Dickens	*Martin Chuzzlewit*	(MC)	Serial Fiction	1843–4	341,605
11	Dickens	*Christmas Books*	(CB)	Fiction	1843–8	156,976
12	Dickens	*Pictures from Italy*	(PFI)	Sketches	1846	72,828
13	Dickens	*Dombey and Son*	(DS)	Serial Fiction	1846–8	346,798
14	Dickens	*David Copperfield*	(DC)	Serial Fiction	1849–50	360,386
15	Dickens	*A Child's History of England*	(CHE)	History	1851–3	163,966
16	Dickens	*Bleak House*	(BH)	Serial Fiction	1852–3	358,860
17	Dickens	*Hard Times*	(HT)	Serial Fiction	1854	104,751
18	Dickens	*Little Dorrit*	(LD)	Serial Fiction	1855–7	342,086
19	Dickens	*Reprinted Pieces*	(RPR)	Sketches	1850–6	92,457
20	Dickens	*A Tale of Two Cities*	(TTC)	Serial Fiction	1859	137,227
21	Dickens	*The Uncommercial Traveller*	(UT)	Sketches	1860–9	143,949
22	Dickens	*The Great Expectations*	(GE)	Serial Fiction	1860–1	187,141
23	Dickens	*Our Mutual Friend*	(OMF)	Serial Fiction	1864–5	331,694
24	Dickens	*The Mystery of Edwin Drood*	(ED)	Serial Fiction	1870	95,181
					Sum of word-tokens in the set of Dickens texts:	**4,912,651**

Appendix 2: 18th Century component of ORCHIDS

No.	Author	Texts	Date	Word-tokens
1	Defoe	*Robinson Crusoe*	1719	121,839
2	Defoe	*The Farther Adventures of Robinson Crusoe*	1719	100,655
3	Defoe	*Captain Singleton*	1720	111,074
4	Defoe	*A Journal of the Plague Year*	1722	93,937
5	Defoe	*Moll Flanders*	1722	138,563
6	Defoe	*The Military Memoirs of Capt. George Carleton*	1728	80,770
7	Fielding	*Joseph Andrews*	1742	125,650
8	Fielding	*Jonathan Wild*	1743	69,685
9	Fielding	*A Journey from this World to the Next*	1749	45,062
10	Fielding	*Tom Jones*	1749	347,080
11	Fielding	*Amelia*	1751	212,589
12	Goldsmith	*The Vicar of Wakefield*	1766	63,331
13	Haywood	*The Fortunate Foundlings*	1744	101,742
14	Richardson	*Pamela*	1740–1741	442,580
15	Richardson	*Clarissa Harlowe*	1748	941,637
16	Smollett	*Roderick Random*	1748	192,418
17	Smollett	*Peregrine Pickle*	1751	343,489
18	Smollett	*Ferdinand Count Fathom*	1753	158,292
19	Smollett	*Sir Launcelot Greaves*	1760	90,016
20	Smollett	*Travels through France and Italy*	1766	121,307
21	Smollett	*Humphry Clinker*	1771	151,253
22	Sterne	*Tristram Shandy*	1759–1767	188,366
23	Sterne	*A Sentimental Journey through France and Italy*	1768	41,061
24	Swift	*A Tale of a Tub and the History of Martin*	1704	44,233
25	Swift	*Gulliver's Travels*	1726	104,359
26	Swift	*A Journal to Stella*	1766	191,941
			Sum of word-tokens in the set of 18th Century texts:	**4,622,929**

Appendix 3: 19th Century component of ORCHIDS

No.	Author	Texts	Date	Word-tokens
1	Austen	*Sense and Sensibility*	1811	120,325
2	Austen	*Pride and Prejudice*	1813	123,045
3	Austen	*Mansfield Park*	1814	160,919
4	Austen	*Emma*	1815	161,400
5	Austen	*Northanger Abbey*	1817	77,806
6	Austen	*Persuasion*	1817	83,841
7	BrontëA	*Agnes Grey*	1847	68,941
8	BrontëC	*Jane Eyre*	1847	187,329
9	BrontëC	*Villette*	1853	195,440
10	BrontëC	*The Professor*	1857	88,604
10	BrontëE	*Wuthering Heights*	1847	118,739
12	Collins	*After Dark*	1856	137,456
13	Collins	*The Moonstone*	1868	198,444
14	Collins	*The Woman in White*	1859–1860	249,036
15	Gaskell	*Mary Barton*	1848	164,363
16	Gaskell	*Cranford*	1853	71,689
17	Gaskell	*Sylvia's Lovers*	1863	194,908
18	GEliot	*Adam Bede*	1859	220,445
19	GEliot	*The Mill on the Floss*	1860	212,080
20	GEliot	*Silas Marner*	1861	73,371
21	GEliot	*Daniel Deronda*	1876	313,885
22	GEliot	*Romola*	1862–1863	229,423
23	GEliot	*Middlemarch*	1871–1872	321,535
24	Thackeray	*Barry Lyndon*	1844	127,186
25	Thackeray	*Vanity Fair*	1847–1848	308,451
26	Thackeray	*The History of Pendennis*	1848–1850	361,837
27	Thackeray	*The Newcomes*	1854–1855	369,576
28	Thackeray	*The Virginians*	1857–1859	360,982
29	Trollope	*The Warden*	1855	72,838
30	Trollope	*Barchester Towers*	1857	200,444
31	Trollope	*Doctor Thorne*	1858	220,694
32	Trollope	*Can You Forgive Her?*	1864–1865	321,029
33	Trollope	*Phineas Finn*	1867–1868	265,795
34	Trollope	*The Eustace Diamonds*	1871–1873	273,553
		Sum of word-tokens in the set of 19th Century texts:		**6,655,409**