The University of Osaka
Institutional Knowledge Archive

Title	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
Author(s)	田畑, 智司
Citation	言語文化共同研究プロジェクト. 2017, 2016, p. 5-21
Version Type	VoR
URL	https://doi.org/10.18910/62075
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

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

The University of Osaka

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

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，機械学習ツールキットの MALLET を用いて，LDA を実行することにより生成されたトピックモデルが，FLOB コーパスのサブコーパスとなっている 15 のレジスターおよび下位の標本テクストに内在する意味構造をどの程度適切に抽出しうるかを検討する。まず，生成された個々のトピックモデルの解釈を試み，それら各トピックが FLOB コーパスの各言語使用域の標本テクスト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 word clusters リスト，ネットワークグラフ，ワードクラウド，使用域毎のトピック密度グラフなどによる視覚化を行うことで，トピックモデリングがコーパスの局所的な特徴抽出，記述に有効であることを示す。マクロ的な視点からは，FLOB コーパスにおける言語使用域のクラスタリング，トピックのクラスタリングをヒートマップの形で要約し，構築されたトピックモデルが FLOB コーパスの標本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　トピックモデル，共起語彙項目，意味構造，言語使用域

The Semantic Structure of the FLOB Corpus:
Using Probabilistic Topic Models for Corpus Description

Tomoji Tabata

Graduate School of Language and Culture, University of Osaka

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

Abstract　Topic modelling is a machine learning method for uncovering hidden semantic structures in a corpus of texts. Based on a probabilistic inference algorithm called latent Dirichlet allocation, the technique makes it possible to identify sets of frequently co-occurring words, or topics, that characterize a text as well as classify texts into groups defined by inferred sets of strongly associated topics. One of the major advantages topic modelling has over traditional key-word detection techniques employed in many stylometric or corpus linguistic studies is that topic models do not simply help to classify texts, but enable us to visualize complex yet meaningful interrelationships between vocabulary items, topics, and more importantly, association between topics and texts in the form of network diagrams and heatmaps. The present study applies topic modelling to the FLOB corpus with a view to analyzing latent semantic structures underlying in the corpus and mapping its subcorpora (or, registers) in the network of words, topics, and texts. What is of special interest is that

by means of this approach it is now 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 open up a new avenue of inquiry into key semantic patterns in a large collection of texts, thereby demonstrating 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 interdisciplinary stylistics.

Keywords topic model, lexical co-ocurrence, semantic structure, register

1. はじめに

1.1. トピックモデルとは何か

所与のテーマに即した文章を書く時，書き手はボキャブラリーの中から，そのテーマに関連する語句を選び出して文を作り，さらにその文を他の文とつなぎ合わせて文章を作り上げる。例えば，ラグビーユニオン (Rugby Union)[1]についての文章を書く場合，書き手は，「パス」，「スクラム」，「タックル」，「トライ」，「コンバージョン」，「モール」，「オフサイド」，「ノックオン」，「反則」など，ゲームのルールに関する用語群から選択した語句を用いる一方，「ワラビーズ」，「ヘッドコーチ」，「スタジアム」，「観客」，「気温」，「風向き」など参与者や気象条件に言及したり，さらには「ワールドカップ」，「代表キャップ数」，「デビュー戦」など参与者の経歴などにも言及するであろう。これらの語句は使用頻度は高低様々であるにしろ「ラグビーユニオン」というノードを中心にした「トピック」を構成する語群 (word cluster) であると言える。また，文書は往々にして複数のトピックによって構成される。例えば，新聞の紙面には，事件報道，社会面，政治面，国際面，社説，文化面，科学面，経済面，スポーツ記事，連載小説，悩み相談，読者による投稿など実に多種多様なタイプの文章が掲載されており，各面にはそれぞれ異なる語群が現れる。また，刊行日，月，年が異なれば，各面に現れる語群，すなわちトピックも当然異なるものになるであろう。このように，新聞という文書は，刊行する新聞社の編集方針が明確に定まっているとしても，語彙的には決して ‘monolithic’ ではなく，多数のトピックの集合体として構成されているものであると言える。

ラグビーユニオンに関しての文章や新聞各面の例が端的に示すように，文書中で共起する語句は同じ話題，トピックを構成しているという前提に立脚し，確率論的アルゴリズムによって共起する語群，すなわちトピックを抽出するのがトピックモデルである。統計学的アルゴリズムを基に，文書を特徴付ける語群を抽出し，文書の内容推定や類型化を行う手法を提案した Deerwester et al. (1990) や Papadimitriou et al. (1998) による潜在的意味索引 (Latent Semantic Indexing; LSI) は，*tf-idf* (Term Frequency-Inverted Document Frequency) (Salton and McGill eds.,1983, Manning and Schütze, 1999; 田畑, 2010) によって重み付けをした語彙頻度プ

[1] 日本でいう，いわゆる，15 人制または 7 人制の「ラグビー」のこと。この他に，スクラムやタックル，キックなどに関して異なるルールが適用される 13 人制のチームで行うラグビーリーグ (Rugby League) という競技が存在するため，区別して表記する。

ロファイル (term-frequency matrix) を特異値分解 (singular value decomposition) することで，文書とそれに対応する特徴的語彙項目をマッピングする手法であり，トピックモデルの先駆に位置付けられる。LSI を改良した Hofmann (1999) の確率論的潜在意味解析 (Probabilistic Latent Semantic Analysis) をさらに汎化したのが，Blei et al. (2003) による潜在的ディリクレ配分法 (Latent Dirichlet Allocation, LDA)[2] である。

> A topic model takes a collection of texts as input. It discovers a set of "topics" — recurring themes that are discussed in the collection — and the degree to which each document exhibits those topics. (Blei, 2012a)

Blei et al. が提案した LDA に基づくトピックモデルは，文書が複数のトピックからなることを前提として，大規模な文書集 (collection of documents) に潜在するトピックを構成する語群を推定するとともに，文書におけるそれらの語群の密度 (topic density) を指標として，文書のクラスタリングを行うことを可能にする。最近では，トピックモデリングのアルゴリズムは，テクストデータだけでなく，遺伝子データ，画像，ソーシャルネットワークのパターン解析にも応用されている (Blei, 2012b)。

1.2. Digital Humanities におけるトピックモデルの応用

Meeks and Weingart (2012) は，トピックモデリングはまさに 'distant reading' であること，個々のテクストではなく，コーパス全体に焦点を当て，テクストをいわば単語が詰まったバケツ容器のごとく扱い，魅惑的だが茫漠な分析結果を解釈しやすいトピックの形で提供することを指摘し，トピックモデリングは提喩として Digital Humanities (DH) を象徴する手法であると述べている。

> Topic modeling could stand in as a synecdoche of digital humanities. It is distant reading in the most pure sense: focused on corpora and not individual texts, treating the works themselves as unceremonious "buckets of words," and providing seductive but obscure results in the forms of easily interpreted (and manipulated) "topics." (Meeks and Weingart, 2012)

オンラインの学術誌 *Journal of Digital Humanities* の第 2 号 (2012) 第 1 分冊にはトピックモデルの特集が組まれ，上掲の論文の他，Brett (2012), Crymble (2012), Goldstone and Underwood (2012), Graham and Milligan (2012), Rhody (2012) が掲載され，DH コミュニティーにトピックモデルが深く浸透する契機となった。翌 2013 年にはさらに，学術誌 *Poetics* が Volume 41, Issue 6 (December, 2013) を "Topic Models and the Cultural Sciences" と題した特集号とし，8

[2] Blei (2012a) はこの用語について次のような註釈を与えている。

> The name "latent Dirichlet allocation" comes from the specification of this generative process. In particular, the document weights come from a Dirichlet distribution — a distribution that produces other distributions — and those weights are responsible for allocating the words of the document to the topics of the collection. The document weights are hidden variables, also known as latent variables.

編の原著論文 (original research articles) を掲載して刊行した。それらのうち，McFarland, et al. (2013) はトピックモデルがテクスト研究，さらに言語使用の場面間の特徴の識別 (a new way to study text and differentiate language domains) に貢献することを示した論考である。また，Jockers and Mimno (2013) は 19 世紀の小説 3,200 点余りからなるコーパスから数百のトピックを抽出した研究である。彼らは，書き手の gender を識別するのに有効なトピックとして，'female fashion', 'Tears and sorrow', 'Drink as in liquor and beer and tobacco', 'Pistols and other guns' を含むテーマ 25 点を特定している。この *Poetics* 特集号では，ニュース記事，科学雑誌，論文アブストラクト，ブログ投稿などのテクストから，詩，小説に至る様々な素材が対象とされており，DH におけるトピックモデルの応用例が俯瞰できる。

2016 年に Poland, Kraków で行われた Digital Humanities 2016 では，Long papers, Short papers, Poster 発表，Panel 発表を合わせるとタイトルに明示的に topic model(s) を含むものが 6 件，発表内容に何らかの形でトピックモデルを使ったものは，Pre-conference workshop の発表も含めて約 40 件[3]に達しており，DH の中心を占めるテクスト分析の手法として定着しつつあると言ってよいであろう。

1.3. コーパス言語学研究における応用

コーパス言語学におけるトピックモデルの応用は DH に比べるとやや遅れ気味であるが，コーパスの類似性 (corpus similarity) を測定するツールとしてトピックモデルの評価を試みた研究 (Fothergill et al., 2016) や，大規模インターネットフォーラムで取り交わされる言葉のやり取りに見られる Islamophobia や anti-feminism を表象する言説の批判的談話分析 (Critical Discourse Analysis, CDA) にトピックモデルを応用する取り組み (Törnberg and Törnberg, 2016)，学術誌 *Global Environmental Change* (*GEC*) 掲載の研究論文からなる特殊コーパスを分析し，論文の導入部と結論部におけるトピックの生起パターンの違いや，*GEC* の通時的トピック変異の問題などを詳細に分析した議論 (Murakami et al., 2017 forthcoming) などが行われるようになってきた。これらの先行研究が示すように，特殊目的コーパスの意味構造の解析におけるトピックモデルの有効性から判断して，LDA に基づくトピックモデルは，多岐にわたる言語使用域のテクストによって構成される汎用コーパスの潜在的意味構造解析にも有用である可能性が考えられる。

2. FLOB コーパスの潜在的意味構造の解析と言語使用域の特徴付け

2.1. FLOB コーパス

本研究では，トピックモデルを適用した汎用コーパスの分析・記述のパイロットスタディとして，FLOB コーパス (The Freiburg-LOB Corpus of British English) の解析を試みる。FLOB コーパスは，Christian Mair がプロジェクトリーダーとなって，ドイツ，Albert-Ludwigs-Universität Freiburg にて開発された総語数 100 万語の汎用コーパスである。1961 年発行のアメリカ英語の書き言葉テクストを収録した世界最初の電子コーパス The Brown University Standard

[3] *Digital Humanities 2016 The Book of Abstracts*. Available at http://dh2016.adho.org/abstracts/

表 1: FLOB コーパスの構成

Label	Register	No. of texts
A	Press: Reportage	44
B	Press: Editorial	27
C	Press: Review	17
D	Religion	17
E	Skills, Trades and Hobbies	38
F	Popular Lore	44
G	Belles Lettres, Biographies, Essays	77
H	Miscellaneous	30
J	Science	80
K	General Fiction	29
L	Mystery and Detective Fiction	24
M	Science Fiction	6
N	Adventure and Western	29
P	Romance and Love Story	29
R	Humour	9
	Total number of texts in the corpus:	500

Corpus of Present-Day American English（通称 Brown コーパス），同じく 1961 年発行のイギリス英語の書き言葉テクストを収録した Lancaster-Oslo-Bergen Corpus (通称 LOB コーパス) のコーパスデザイン（表 1 参照）を踏襲し，1991 年発行のイギリス英語の書き言葉テクストを収録したコーパスである。当コーパスは各 2,000 語の標本テクスト 500 点を収録しており，各標本テクストは，表 1 の Label A から R の言語使用域 15 タイプのいずれかに属している。収録されている使用域はさらに 9 つの情報散文 (informative prose) カテゴリーと 6 つの創作散文 (imaginative prose) に二分することができる。Jockers and Mimno (2013) は，LDA に基づくトピックモデルが，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，長編の文書を対象にすると，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うこと，結果として広範囲に及び過ぎて茫漠なトピックが生成されてしまうことを指摘している。

> Topic models employ what is called a “bag of words” approach to text analysis. The words in a text are treated as if they had been tossed together into an imaginary bag. Because the model looks at which words tend to co-occur, the bigger the bag the more likely it is to have a set of words appear as collocates. For this reason, a modeling of full novels tends to result in topics of a broad and general nature. For topic modeling purposes, the unit of analysis should be a segment of text that is large enough to measure word co-occurrences but small enough that it can reasonably be assumed to contain a small number of themes. Full texts were too large to discover patterns of word use: topics derived from full novels were vague and “washed out.” (Jockers and Mimno, 2013: 753–754)

この点を踏まえると，各標本テクストがほぼ2,000語に均等化されているFLOBコーパスは，トピックモデリングの対象として理想的な構成に近いデータセットであると考えることができるであろう。

2.2. MALLETによるトピックモデリングとデータの前処理 (Pre-processing)

本稿は，University of Massachusetts, Amherstで開発された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)[4] Version 2.0.7を用いてLDAを実行した結果を報告する。MALLETはJavaで書かれたコマンドラインベースのプログラムであり，Sequence tagging, 文書の分類, トピックモデリングを実行することができる。MALLETには様々な実行オプションと，最適化オプションが用意されており，機能語などを分析から除外するためのStop words listがあらかじめ同梱されている。Stop words listはテクストファイル形式になっているので，ユーザが自分で項目を追加あるいは削除したり，カスタマイズが可能な仕様になっている。

本稿の基礎となった研究では，固有名詞がKey wordsとなってトピックを占有することを避けるため，史的・通時的文体研究に特化したコーパス **O**saka **R**eference **C**orpus for **Hi**storical/**D**iachronic **S**tylistics (ORCHIDS)から取得した固有名詞とその属格形，約3,000項目をStop words listに加えた。とはいえ，あらかじめ用意したStop words listだけでFLOBコーパスに現れる全ての固有名詞をカバーすることは不可能である。そこで，固有名詞が顕在化するトピックが生成された場合，その解釈に固有名詞がどのように影響するか，データ解析の過程で併せて検討することとした。

```
L01 1 <#FLOB:L01>He had a similar small apartment in Nice under yet
L01 2 another name.<p/>
L01 3 <p_>As soon as he was unpacked he rang Dickie Ashton's number.  A
L01 4 maid answered and he put down the receiver straight away.  He took
L01 5 the replacement Steyr from behind a refrigerator, left the
...
L01 37 <p_>Jacko struck Gorley quite hard, grabbed him, and in a stronger
L01 38 cockney twang than usual, said, <quote_>"I'm sorry, mate.  My fault;
L01 39 in too much of a bloody hurry."<quote/> He held on to Gorley's arm
L01 40 and added, <quote_>"You all right?"<quote/><p/>
```

図 1: FLOB コーパス標本テクストのフォーマット

オリジナルのFLOBコーパス (POSタグなし，Horizontal format)は図1のようなフォーマットになっている。各使用域カテゴリー毎に標本テクストは一つのファイルに連結されており，全部で15ファイルに収められている。各行最初の桁は言語使用域カテゴリーを示すラベル (A–R)，続く2–3桁目がテクスト番号，スペースを挟んで5–7桁目が各テクストの行番号である。行番号は右寄せ配置になっているため，テクスト番号と行番号（最大3桁）の間隔はスペース1つから3つである。例に挙げた標本テクストL01の最終行の次には直ぐに

[4] http://mallet.cs.umass.edu/index.php

「L02 1 <#FLOB:L02><hl>3 ...」という具合に，標本テクスト L02 が続いている。MALLET 実行前の事前加工処理として，各使用域カテゴリーのファイルは全て個々の標本テクストに分割した。本稿執筆時点の MALLET には，行頭のマークアップや<*>タグをスキップする機能は実装されていないため，正規表現を使用してあらかじめ削除しておく必要がある。この実験では，アポストロフィを語の一部と定義して単語の計量を行うようオプション指定を行った。それに付随して，二重引用符が単語に付着してピックアップされてしまうのを抑止するために，事前に二重引用符も削除した。なお，これは macOS 環境固有の問題であると思われるが，分析対象のコーパステクストが所収されているフォルダ（ディレクトリ）に .DS_Store という不可視ファイルが存在していると，モデルの生成時に狂いが生じることがある。そのため，MALLET 実行前には必ず Terminal.app で不可視ファイル.DS_Store を削除しておく必要がある。これら一連の前処理を以下にまとめた。

1. 固有名詞とその属格形約 3,000 項目を stop words list に追加
2. ファイルを個々の標本テクストに分割し，全 500 標本をフォルダに所収
3. テクストマークアップ（タグを含む）の除去
4. 二重引用符削除
5. コーパステクストを所収するフォルダから不可視ファイル .DS_Store を削除（macOS 固有の問題である可能性）

2.3. MALLET の実行

トピックモデリングでは抽出するトピックをどれくらいの数に設定するかによって生成されるトピックモデルの粒度は大きく変動する。コーパスに潜在するトピックに比して抽出するトピック数が著しく少ない場合，生成されるモデルの粒度が粗く，広範囲に及び過ぎてしまい，コーパスに内在する意味構造の差異を検出できなくなってしまうことがある。他方，コーパスに潜在するトピックに比して抽出トピック数が著しく多い場合，生成されるモデルが細分化され過ぎてしまい，トピックに対する重みが大きい語が少なくなり結果の一般化が困難になる場合がある。最適な生成トピック数は，分析対象とするコーパスの規模，収録サンプル数，潜在的なトピックの文書内変異，ジャンル内変異など，複合的な要因に左右されるため，一意に最適値となりうる解は存在しない。本研究では，試行実験を繰り返しながら，生成するトピック数の最適化を試みた。生成するトピックの数を 50 に設定したところ，コーパスがカバーする情報散文と創作散文の対比，個々の言語使用域を特徴付けるトピック (register variaation)，言語使用域の内的変異 (intra-register variation)，さらにはトピックを構成する語彙項目間，トピック間，さらには標本テクストとトピックの関係性が反映される解析結果が得られた。

3. 解析結果とその解釈

3.1. 生成されたトピックモデル

MALLET Version 2.0.7 を使用して FLOB コーパスのトピックモデル生成を行なった結果得られた 50 項目のトピックとそれらを構成する word clusters のうち，そのトピックへの

表 2: FLOB コーパスより抽出した 50 個のトピックと各トピックを構成する word clusters

No.	Topic label	Key words, or words with the greatest weight
0	national politics	labour government major british election secretary leader political general public thatcher week cent yesterday home britain poll power conservative president mp parliament office support ...
1	university	school royal university students report palace geoffrey held time wrote letter student term course evidence ron senior attended end general concerned terms year boys case member due ...
2	theatrical	theatre play stage shakespeare peter plays performance drama production theatrical actor actress romantic sign actors dramatic june characters vere comedy comic enter successful acting ...
3	body language	head feet face water front body air hand white cold sun fell blue heavy rest top ground hold lay legs beneath foot standing light dropped edge mine caught weather floor sea shot clear ...
4	fishing	water fishing island boat fish sea line east ship cameron crew coast tissee boats west carried rock peter april shore found black harbour low miles ships tide fishermen arrived box ferry ...
5	health & medicine	patients hospital medical patient heart baby thalidomide treatment research clinical mental testing body health disease babies pregnancy risk evidence parents johnson found mass chronic ...
6	feminine narrative	mother life love day years night home long felt father woman found bed happy months loved daughter girl told face street returned remember clothes lived hair husband child dark grew ...
7	television	television film christian bbc field north films chrissie cinema east late video south quality appeal desert oster ago movie camera dundee tv (and bird image southern pollen america ...
8	trade & industry	industry uk production technology europe plant international products countries commercial trade economic food bullet market high world business ec economy energy environmental ...
9	scholarly research	point question theory part line position general (and principle definition form particular terms object discussion necessarily system rules parts structures kind argue course behaviour basis ...
10	system design	design system control aircraft equipment project work systems cost time requirements performance weight large maintenance building power experience features required several built ...
11	local public policies	local services people health government community members national public care authority independent authorities charter staff information standards work area programme year ...
12	theory & criticism	history language reading text work nature writing understanding speech science read historical different view narrative study texts like criticism words present fiction audience events ...
13	narrated action	looked door like knew head hand face turned thought voice hands right sat stood suddenly moment walked open mouth hair office smiled found heard dark stopped pulled smile dead moved ...
14	publishing	book books writing written points write letters picture published art page article copy list style makes series writers point original magazine volume papers photographs subject articles ...
15	scientific approach	development important time way place part particular major process view possible difficult based period different work interest including early means problem problems range individual ...
16	experiments & analysis	case data number different results analysis shown groups study model effects effect studies table present (see higher standard found low large terms internal performance factors differences ...
17	materials chemistry	boundary energy formula sites pyrite grain reaction large particle equation processes high material formation nucleus system stress particles atoms materials reactions systems incident ...
18	global issues	soviet union workers security people wage plan chinese united world europe gorbachev international american china states trade economic foreign square aid communist leaders damage ...
19	narrative suspense	like time right way find look long course wrong feel making doesn't business words moment mind question kind talking bit word remember run show car easy play start chance answer ...
20	land development	land century local change mortality people human infant regions rates areas resources south natural population poor rural poverty degradation timber trees cause regional area inequality ...
21	driving infringements	drivers pressure unit change tunnel drinking driving limit route present area trains alcohol table high felt units previous sand comfort sands flow recorded mm railway section train stock ...
22	pop fiction setting	like ricky black pair firmly night inside abby guess hell minton talley shirt cake dancer daisy laugh machine special robot rule hundred break clothes dyson robby shadow smoke flash ...
23	disease	population knowledge household census aids cases individuals disease native change estimates households people infection figures infected long play active sexual cent game middle ...
24	belief	true god world people way sense truth moral physical human belief knowledge means christian thought subject mind argument called certain hand doctrine idea theory account claim beliefs ...
25	religion	religious religion tradition words muslims stress poetry muslim love particular settlers douglas einstein marx light latin poet long find followers syllables word bohr classical held syllable ...
26	education	education group college training higher students teachers courses schools years academic colleges work science level course teaching qualifications national reading growth working groups ...
27	family	family years house father time life son born died wrote brother married enjoyed wife father's lived friends met called career hall large brothers death place castle gave home living ...
28	sport matches	top team harry cup year game match win world test final week star won winning minutes county race ben season big players tony start played ago racing round despite fourth grand ...
29	music & musicals	music radio audience opera musical show american british played dance act band pop festival broadcasting sound bbc singing charts scene director film concert programmes voice week ...
30	police	police told road found home crime yesterday attack court car station several hours murder admitted driver gun pilot pub accident claimed west escape plane judge death driving officer ...
31	wemen's lives	children women men age social experience relationship life sex male behaviour female children's feelings people find lives feel ways family child cent differences divorce group parents ...
32	skin care	skin care colour size pattern piece cut pure simple white top japanese fit eye look body dolls front thread making added fabric leaves pieces range garment products basic guide bust ...
33	city	city years year part high end centre town people day park fire brought site built today north found village building christmas spring main special square travel winter area find places ...
34	narrative past	time like told thought people asked work way house gave called day evening morning round home big call friends wanted leave live started coming fact lot front money end decided woman ...
35	art	church century art churches work painting early late works artist nineteenth period practice collection common eighteenth gallery painted artists portrait prayer medieval tradition figures ...
36	attention	task noise attention tradition effects effect target position led tasks subjects engaged reading state normal lay teinds word directed preferred saccades eye increased intent lords perform ...
37	political institution	political power social society state movement class politics interests societies groups economic capitalist elite cultural system generation radical policy single bloc community rights ...
38	cooking	cook mushrooms minutes water heat birds add potatoes ml dry corn sheep cut oil large earth golden lambs oz sauce half bring butter growing place cheese fresh mixture stone taste crop ...
39	biology	species temperature mass reaction extinction transfer rate gas preservation (fig external pressure concentration heat resistance sea soft temperatures deep coal common water early range ...
40	biochemistry	al studio fviii shown min kd plasma band described breastfeeding pins binding thrombin degreec treatment activity concentrate samples breasts peptides breast chain pain showed atpase ...
41	public administration	act information office authority report department relevant required management right costs issues home procedures planning scheme ensure public study application civil force proposed ...
42	issues & concerns	world life time modern like private end free sense way age form brought simply real kind attempt death part public instance began believed style turn natural popular concern culture lost ...
43	business & finance	pounds company year market million years cent sales price business sold british companies group cash offer prices bought sell share buy months uk loss profits interest paid deal largest ...
44	war	war british military german men britain ireland irish gulf famous official gordon soldiers number special civil douglas officer force several victory campaign combined foreign africa rudolf ...
45	UNINTERPRETABLE	word gate airport (p parity air works collection logic figures museum horse figure heathrow road bit bus output bits odd windsor tables boer buses black double craven gates sets songs truth ...
46	UNINTERPRETABLE	child animals signal contact food echo agriculture animal future thimbles act live circuit parent responsibility making time end farming path connection receive residence favour section care ...
47	social security	pounds million tax fund number cent income financial benefit rate year cost department increase security debt expenditure rates national billion paid social insurance money building ...
48	lawsuit	case court law property cases request judges article appeal legal justice agreement parties material assistance earlier involved paid payment proceedings requested respect evidence judge ...
49	gardening	garden rose dog wine flowers air white trust dogs place houses gardens rooms fruit style wall chardonnay roses wilkie produce year flower village look several arrangement plants preston ...

word-weight が高い単語から降順に並べたものが表 2 である。各トピックは word-weight の高い単語から低い単語まで多数の項目で構成されている。生成したモデルが，コーパスの規模や潜在的意味の変異を概ねカバーできている場合，当該トピックに属する全ての単語を調査しなくとも，word-weight が高い項目とそれらが生起している使用域カテゴリー，標本テクストを参考にしてトピックの解釈を行うことができることが多い。ただし，構成する単語に一貫性を認めることが困難なトピックや，ほとんどの標本テクストにおいて関連語の密度 (topic density (d): $0 \leq d \leq 1$ の値で表示) が低いトピックは解釈，一般化が困難であることも少なくない。実際，本研究で得られたトピックのうち，No. 45, 46 (*cf.* 表 2) は解釈が不可能であった。

3.2. ネットワークグラフによる関係性の可視化

図 2: FLOB の潜在トピックと関連する Key words との関係（Word weight: $w \geq 50$ の項目）

先述したように，往往にして一つの文書に複数のトピックが存在すること，そして，テクストを構成している個々の語彙項目も，高頻度の項目になればなるほど，あるいは多義性のある単語の場合，一つの語が複数のトピックに属しているケースが数多く存在する。それゆ

え，トピックモデリングによって生成されたトピック，トピックを構成する word clusters，そしてトピックとコーパスの標本テクストとの複雑な関係性を把握するには，ネットワークグラフやワードクラウド，ヒートマップなどデータの視覚化を支援するツールの助けを借りる必要がある。図 2 は FLOB コーパスのトピックを構成している語のうち，生成した 50 個のトピックの一つ以上に対する word-weight が 50 以上のもの，計 1,119 項目について，単語 (source)，属するトピック (target)，そのトピックへの word-weight (weight) の 3 変数からなるデータを視覚化アプリケーションの Gephi にインポートし，word clusters とトピックとの関係性をネットワークグラフの形にレイアウトしたものである。トピックはそれぞれ番号で示されており，グラフ中でノード (nodes) を成している。数字のフォントサイズは当該トピックの相対的規模[5]を表している。トピックを構成する語もまたノードとして表されており，トピックと単語とを結ぶエッジ (edges, 辺) の幅はそのトピックと単語との相対的な結びつきの強さを示す。大規模なトピックは多くの語と太いエッジで結び付けられている傾向がある。たとえば，トピック 15 (“scientific approach”) では，*development, important, time, way* などが主要な key words (図 3 参照) であり，これらを含め，142 項目の単語がノードであるトピック 15 とつながって表示されている。逆に，小規模なトピックがごく少数の語と太いエッジで連結されているケースもある。たとえば，トピック 44 (“war”) は，*war* とひときわ太いエッジで結びついているが，その他は *british, military, german* の順でエッジは細くなり，

development
important
time
way
approach
develop
considerable
recently
established
conditions
strong
practice
sources
move
questions
recent
aspects
result
largely
activity
ways
short
highly
context
attention
successful
responsible
experience
widely
serious
increasingly
introduction
quality
group
current
increase
potential
practical
relationship
similarly
provide
appears
increasing
crucial
developing
division
recognised
closely
producing
considering
emphasis
appropriate
significant
research
needed
adopted
activities
continued
failure
represent
present
level
clear
source
influence
effective
prevent
fully
end
stage
working
change
extent
concerned
field
term
impact
giving
understand
partly
reached
aware
majority
contained
knowledge
element
essential
section
promote
awareness
minimum
expressed
growing
valuable
benefits
understanding
confidence
wider
enterprise
tendency
demand
achieve
encouraged
leading
resources
deal
provides
allowed
immediate
fundamental
issues
common
original
position
ability
central
past
direct
variety
extended
stronger
degree
feature
support
open
several
obvious
remains
principal
stage
efforts
focus
introduced
achieved
encouraging
scope
necessity
effectively
series
easily
represented
separate
rapidly
emerging
relevance
affecting
developments
possibility
remained
exists
background
tended
replaced
grounds
requires
included
incorporated
communication
contribute
consists
exception
participation
advanced
readily
reducing
related
categories
widespread
extend
extensive
involves
techniques
changing
experienced
initially
subsequently
stages
scale
situation
drawn
form
status
similar
context
highly
short
attention
responsible
successful
widely
serious
experience
quality
group
introduction
increasingly
main
implications
content

それ以外はわずか 3 項目がグラフに含まれているに過ぎない。ただし，図中のトピックの規模には，「より周辺的な」単語[6]の word-weights の総和は反映されていないため，結果の解釈を行う際には，図 2 が必ずしもトピックの全体像を表しているわけではないということを念頭に置いておくべきである。

全 50 個のうち，トピック 40 (“biochemistry”) は 1 項目 (al) のみ，トピック 45 (“UNINTERPRETABLE”) は word-weight が 50 以上の key word は一つもなかったため図 2 には表示されていない。

各トピックの解釈・ラベルづけを行うに当たっては，表 2 に示す word clusters，図 2 のようなネットワークグラフ，単語の重みを視覚的に示すワードクラウドを照らし合わせることで語群が共有しているテーマを抽象化することができる。

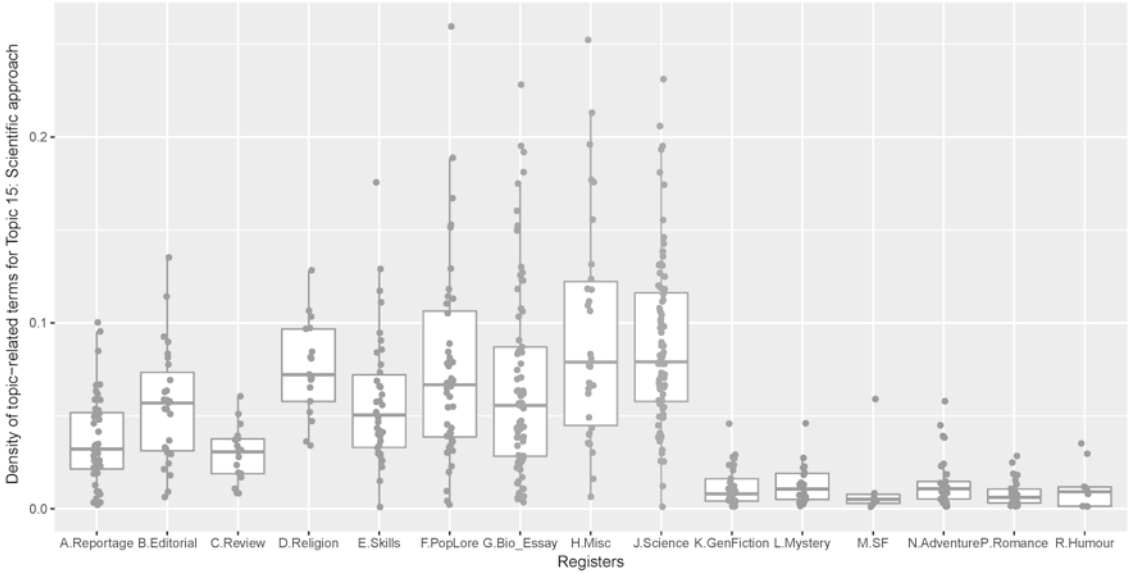

図 4: Density of No. 15 Topic labelled “scientific approach”

図 4 はトピック 15 に関連する語の密度 (topic density) が言語使用域ごとにどのような分布になっているかを示している。トピック 15 を構成している関連語群 (word cluster) は，特に F (Popular Lore), G (Belles Lettres, Biographies, Essays), H (Miscellaneous), J (Science) の使用域でそれぞれ大きく分散してはいるが，情報散文のカテゴリー (A–J) と創作散文のカテゴリー (K–R) とで一貫した密度の相異が認められる。

他方，再び図 2 に視点を戻すと，この図において規模の大きいもう一つのノードは右下に位置しているトピック 34 である。さらに注視すると，トピック 34 は，13, 19, 6，そしてやや小さなノードの 3, 22 などと，いくつかの語を介してネットワークを形成していることがわかる。トピック 34 を構成している word cluster をワードクラウドとして表示したのが図 5 である。このトピックを特徴付ける語として目立つのは，*thought, told, asked, happened, turned, replied, discovered* など過去形の動詞，そして *time, like, people, way, house* などであ

[6] ここでは，いずれのトピックに対しても word-weight が 50 未満しかない単語を指している。

る。このトピックは “narrative past” と解釈できる。このトピックの key words の一つである *like* はトピック 13 (“narrated action”), 19 (“narrative suspense”), 22 (“pop fiction setting”) でも word-weight の大きな語であり，これらのトピックを結びつけるノードとなっているのは着目に値する（図 6）。

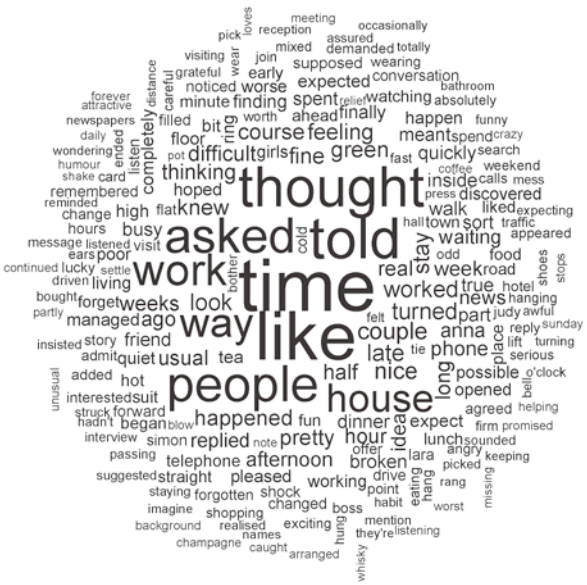

図 5: Wordcloud of No. 34 Topic labelled “narrative past”

Topic 13 “Narrated action” **Topic 19 “Narrative suspense”** **Topic 22 “Pop fiction setting”**

図 6: Topic 34 と関連するトピック (Fiction word clusters)

FLOB コーパスの個々の標本テクスト，各使用域におけるトピック 34 の密度を図 7 に示した。トピック 15 とは対照的に，創作散文のカテゴリー (K–R) と情報散文のカテゴリー (A-J) の間には一貫した差を認める。フィクションの中でも Romance & Love story (P) と Mystery & Detective fiction (L) が特にこのトピックの密度が高い一方，各小説カテゴリー内の internal variation が大きい。同様なカテゴリー内の変異は上に挙げたトピック 13, 19, 22 についても

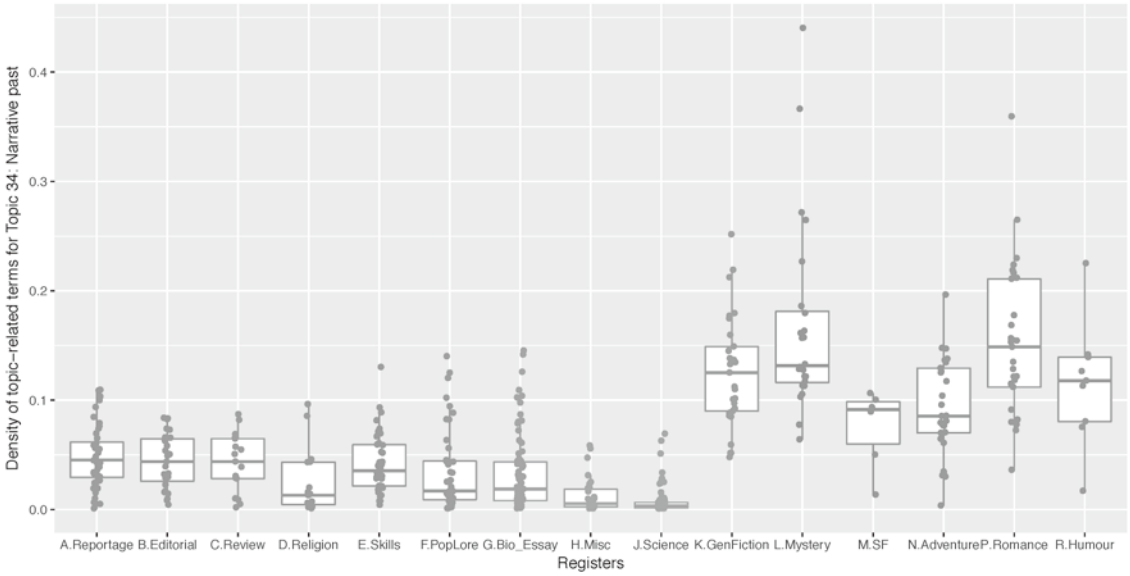

図 7: Density of No. 34 Topic labelled "narrative past"

見られる。フィクション作品は書き手の視点，語りの時制，登場人物や出来事，場面設定など変動要因が大きいことが理由の一つであると考えられる。

次に，生成したトピックモデルと個々の標本テクストとの関係を見てみたい。図 8 は，50 個のトピックと標本テクストとのネットワークグラフである。語彙項目とトピックのネットワークでも大きなノードとして図に表れていたトピック 15 がテクストとのネットワークにおいても同様に最も大きなノードとなっており，トピック 16 ("experiments & analysis") とつながり合っている。トピック 16 につながっているテクストは J (Science) に属する標本である。その下にはやや小さなノードであるトピック 41 ("public administration") が，11 ("local public policies"), 47 ("social security") とリンクしている。これらのノードには主に政府や役所の白書，報告書などが収められたカテゴリー H (Miscellaneous) に属する標本テクストが連なっている。これらは，さらに，8 ("trade & industry") や 10 ("system design") とも緩やかにつながっている。

グラフの中央やや上には，トピック 42 ("issues & concerns") が位置している。このトピックには，C (Press: Review), D (Religion), G (Belles Lettres, Biographies, Essays) などの使用域に属する標本がリンクしている。42 のすぐ隣にはトピック 29 ("music & musicals") が接している。このトピックは特に C (Press: Review) の標本において密度が高い。さらにその右側にトピック 27 ("family") があり，29 といくつか共通する標本テクストで結ばれていて，いずれも G (Belles Lettres, Biographies, Essays) の標本テクストとの繋がりも強い。標本テクストにおけるこれらのトピックの密度を図 7 のようなグラフで俯瞰すると，これらのトピックは特定の言語使用域を単独で特徴付けるのではなく，情報散文の複数の使用域にまたがって生起しており，使用域内での分散が大きいことが確認できる。

図 8 の右側には，大きい順に 34 ("narrative past"), 19 ("narrative suspense"), 13 ("narrated action"), 6 ("feminine narrative"), 3 ("body language"), 22 ("pop fiction setting") などフィクショ

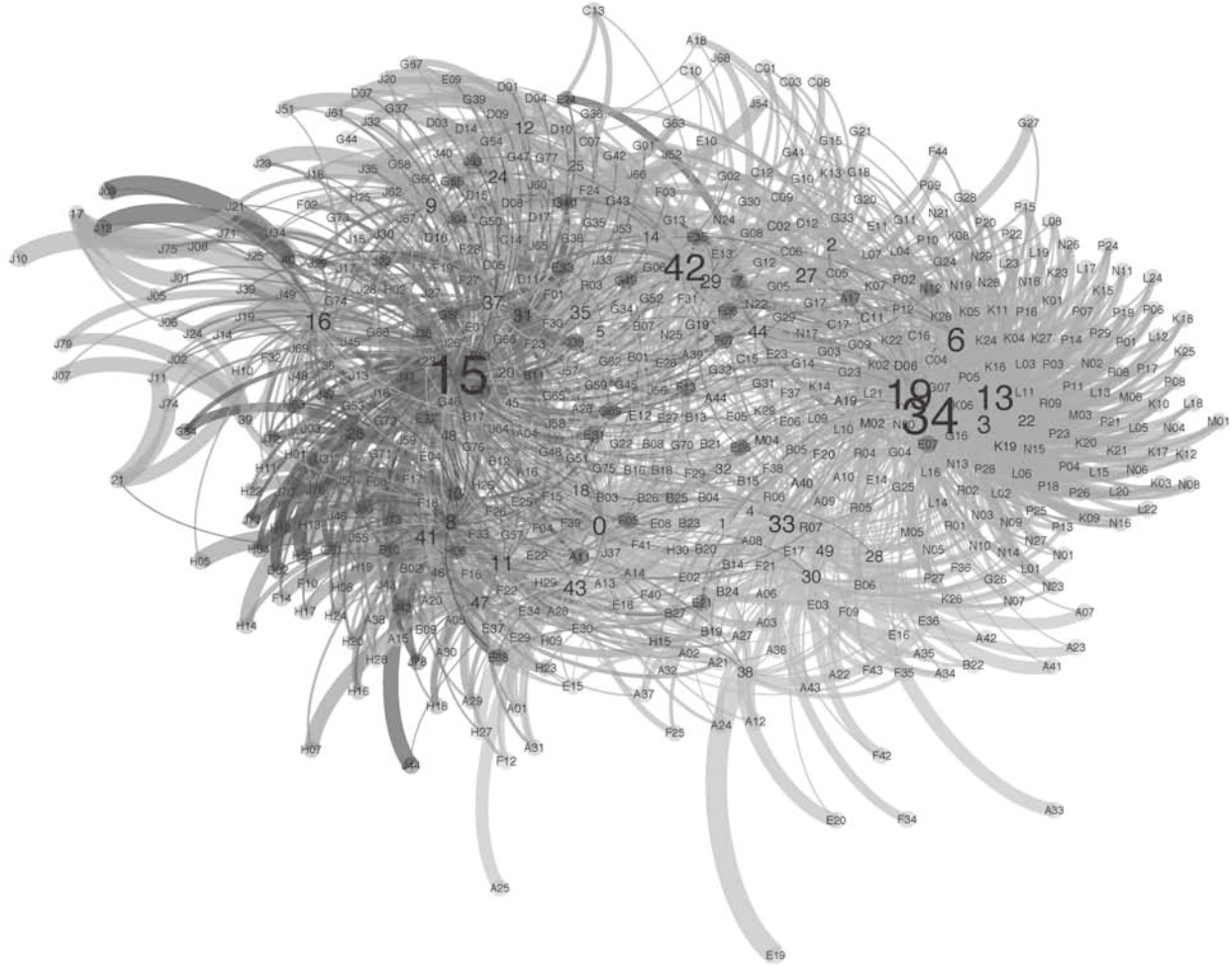

図 8: FLOB コーパスのトピックと標本テクストとの関係

ンの各ジャンルを特徴付けるトピックが相互にリンクする形で位置している。これらのトピックからやや左下に位置するトピック 33 (“city”), 30 (“police”), 28 (“sport matches”) にも細いエッジで接点があり，フィクションと他の言語使用域とのトピックの接点があることを示しており興味深い。

4. トピックモデルの評価

前節で提示したトピックモデルによる言語使用域の記述・特徴づけがどの程度妥当性を持つものかを評価するために，本節では個々の標本テクストにおける各トピックの密度を集約し，使用域の代表値によって FLOB コーパス全体をマクロ的な視点から捉えてみる。各言語使用域を代表する指標として，トピック密度の平均値をもとめた。50 個のトピックの各使用域における密度を 50 行 ×15 行のクロス集計表にした。集計表の数字に潜在するパターンを視覚化したのが図 9 のヒートマップである。全データを集計したところ，トピックによって密度に大きな差があることが判明した。少数の高密度のトピックによって低密度のトピック間の生起パターンの微妙な差異が見えづらくなってしまうことを抑止するために，クロス集計表を統計解析言語 R の scale() 関数によって基準化し，このヒートマップを描画している。

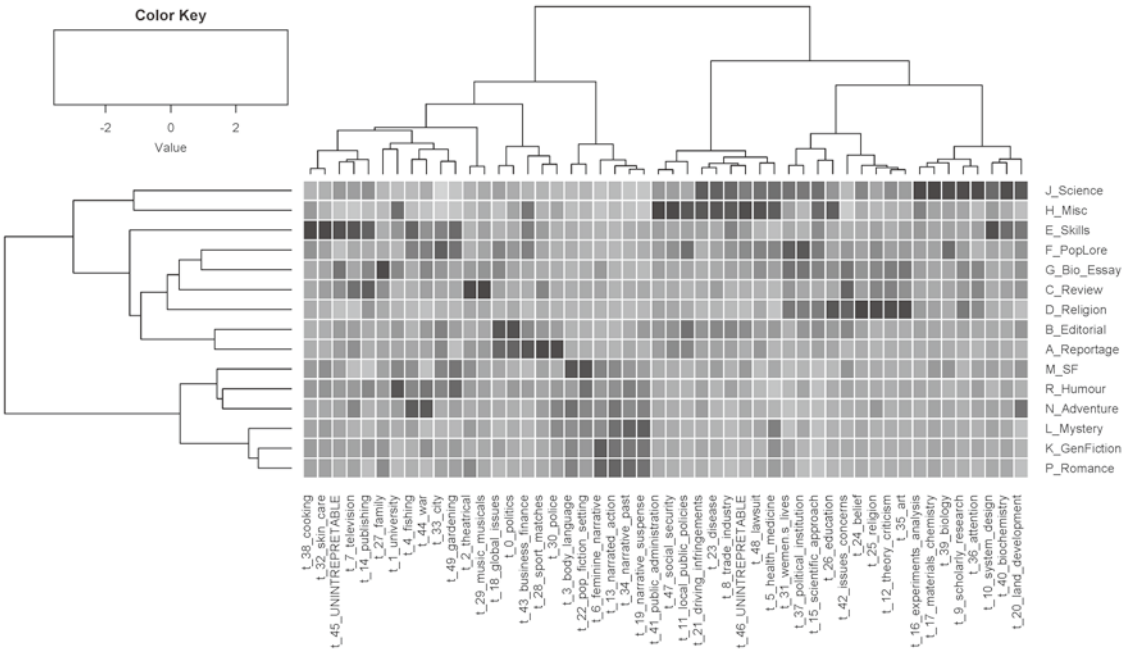

図 9: Heatmap of 50 topics across 15 registers (with mean weights scaled)

図 9 では横軸にトピック，縦軸に言語使用域を配置し，それぞれの関係性をクラスター分析によって視覚化している。使用域，トピックいずれもデンドログラムの項目布置に合わせて再配置されているので，概ね隣り合うものどうしは似た特徴を備えたものだという解釈ができる。[7]ヒートマップは濃淡でトピック密度の差を表しており，濃いセルほど当該トピックの密度が高いことを示している。

言語使用域のデンドログラムは情報散文と創作散文とを明確に識別しており，生成されたモデルが，この二つのカテゴリー間に存在する潜在的トピック構造の相異を検出するのに寄与していることは明らかであろう。フィクションのクラスター内では P (Romance & Love story) と K (General fiction)，N (Adventure & Western) と R (Humour) がそれぞれペアとして分岐し，二つの枝から，それぞれ L (Mystery & Detective fiction) と M (Science fiction) が分岐している。他方，情報散文のクラスターでは，A (Press: Reportage) と B (Press: Editorial) が ‘nearest neighbours’ となっており，F (Popular lore)，G (Belles Lettres, Biographies, Essays)，C (Press: Review) は別のクラスターを形成している。これらから離れて D (Religion)，さらに距離を置いて E (Skills)，最も離れてゆるくクラスタリングしているのが J (Science) と H (Miscellaneous) である。

トピックのクラスタリングも言語使用域の場合と同様に，まず大きく二つに枝分かれしている。これらは単純に情報散文と創作散文の言語的特徴の違いを反映しているわけではないが，情報散文内のクラスタリングや分岐がどのトピックの密度と密接に関わっているか判断

[7] ただし，大きく枝分かれしたクラスターごとにまとまっているので，たとえば，M_SF と A_Reportage, t_41_public_administration と t_10_narrative_suspense のように全く性質が異なる項目が，「見かけ上」隣り合わせに配置されている。解釈を行う際にはデンドログラムを基準に位置関係を把握する必要がある。

する材料を提供してくれる。書誌情報という外的な基準で分類されたレジスター間の変異が，LDA によって自動的に生成されたトピックの密度という内的基準でも確認できるという事実は，構築されたトピックモデルがこのコーパスに潜在している意味構造を分類記述するのに有効であることを如実に示すものであると評価できるのではないだろうか。

5. 結び

本稿では，MALLET を用いて LDA を実行することにより生成されたトピックモデルが，FLOB コーパスのサブコーパスとなっている 15 のレジスターおよび下位の標本テクストに内在する意味構造をどの程度適切に抽出しうるかを検討した。まず始めに，生成された個々のトピックモデルの解釈を試み，それら各トピックが FLOB コーパスの各言語使用域の標本テクストとどのように結びついているかを考察した。トピックを構成する word clusters リスト (表 2)，ネットワークグラフ (図 2, 8)，ワードクラウド (図 3, 5, 6)，使用域毎のトピック密度グラフ (図 4, 7) などによる視覚化を行うことで，トピックモデリングがコーパスの局所的な特徴抽出，記述に有効であることを示した。マクロ的な視点からは，FLOB コーパスにおける言語使用域のクラスタリング，トピックのクラスタリングをヒートマップ（図 9）の形で要約し，構築されたトピックモデルが FLOB コーパスの標本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示された。

コーパス研究でよく使用される対数尤度比検定 (LLR) や χ 二乗検定，U 検定など，典型的な key-word detection の手法とこのモデルの違いは，生起が infrequent な項目，つまり，単語単体では頻度が低く統計の網に掛かりづらい語彙項目を，特定のトピックのもとに集約して扱えること，そしてその集約が分析者の恣意的な判断に依拠するのではなく，機械学習によって語彙項目の共起パターン（換言すると，語彙項目のクラスタリング）をコーパスから自動的に取得できるという点である。このようにトピックモデリングは，コーパスの潜在的意味分析やコーパス内の言語変異をトピック濃度で記述することを支援してくれる有益なツールであるが，その一方で抽出する意味の解釈は多くの場合，分析者の洞察を必要とする。抽出，生成したトピックは最終的に分析する人間の解釈の問題に帰すという点で，トピックモデリングはむしろ人文学の営みと親和する。Meeks and Weingart (2012) が述べたのとは異なる意味でトピックモデリングは Digital Humanities の核心にあると言えるだろう。

文　献

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

[2] 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/

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

[4] Bonilla, T. and J. Grimmer (2013) Elevated threat levels and decreased expectations: How democracy handles terrorist threats, *Poetics*, 41 (6): 650–669.

[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] Crymble, A. (2012) Review of Paper Machines, produced by Chris Johnson-Roberson and Jo Guldi, *Journal of Digital Humanities*, 2 (1). Available at http://journalofdigitalhumanities.org/2-1/review-papermachines-by-adam-crymble/

[7] Deerwester, S., Dumais, S., Landauer, T., Furnas, G., and R. Harshman (1990) Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41 (6): 391–407.

[8] DiMaggio, P., Nag, M., and D. Blei (2013) Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts, *Poetics*, 41 (6): 570–606.

[9] 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

[10] 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/

[11] Graham, S. and I. Milligan (2012) Review of MALLET, produced by Andrew Kachites McCallum, *Journal of Digital Humanities*, 2 (1). Available at http://journalofdigitalhumanities.org/2-1/review-mallet-by-ian-milligan-and-shawn-graham/

[12] Hofmann, T. (1999) Probabilistic Latent Semantic Analysis, *UAI '99 Proceedings of the Fifteenth Conference on Uncertainty in Artificial Tntelligence*, 289–296.

[13] 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

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

[15] 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.

[16] Manning, C. D. and H. Schütze (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, MA.: The MIT Press.

[17] Marshall. E. A. (2013) Defining population problems: Using topic models for cross-national comparison of disciplinary development, *Poetics*, 41 (6): 701–724.

[18] Meeks, E. and S. B. Weingart (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/

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

[20] Mohr, J. W., Wagner-Pacifici, R., Breiger, R. L., and P. Bogdanov (2013) Graphing the grammar of motives in National Security Strategies: Cultural interpretation, automated text analysis and the drama of global politics, *Poetics*, 41 (6): 670–700.

[21] Murakami, A., Thompson, P., Hunston, S., and D. Vajn (2017 forthcoming) What is this corpus about?, *Corpora*, Issue 12.2. Pre-publication copy.
Available at http://pure-oai.bham.ac.uk/ws/files/25529512/TopicModelPaper_ACCEPTED_VERSION.pdf

[22] Papadimitriou, C., Raghavan, P., Tamaki, H., and S. Vempala (1998) Latent Semantic Indexing: A probabilistic analysis, *PODS '98 Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, 159–168. doi:10.1145/275487.275505

[23] 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/

[24] Salton, G. and M. McGill (eds.) (1983) *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

[25] 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/

[26] 田畑 智司 (2010)「TF-IDF 値を通してみるテクストの特徴－文体論研究への応用可能性を探る (その 1) －」田畑智司（編）『多変量アプローチによるテクストの計量研究』平成 21 年度統計数理研究所共同利用研究（研究課題番号 21 －共研－ 4307），47–62.

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

[28] 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.