BCCWJ における出典情報とトピックおよびレジスターとの関係

ホドシチェク・ボル　（東京工業大学大学院社会理工学研究科）†
仁科　喜久子　（東京工業大学留学生センター）

Comparison of Metadata with Topic and Register in the BCCWJ

Hodošček Bor　(Tokyo Institute of Technology)
Nishina Kikuko　(Tokyo Institute of Technology)

１．はじめに

『現代日本語書き言葉均衡コーパス』（BCCWJ）には様々なメタ情報が付与されており、その中にはメディア名（サブコーパス名）、出典情報（作家名、出版年、出版社名）、またはジャンルを記述する NDC（日本十進分類表）情報などがある。ある現象のジャンル別傾向を調査するときにメディアに頼ることはしばしばあるが、メディアの間ではそれぞれジャンルとして非常に似ている文書もあれば、著しく異なるものもある。そこで、本稿では BCCWJ の様々なメディアにおけるメタ情報とトピックおよびレジスターとの関係を分析する。

２．トピックとレジスター

まずトピックとレジスターについて簡単に述べる。言語の構造を単純化していえば、文は内容的な要素（内容語）とそれらの内容を構成する機能的な要素（機能表現）からなるといえる。例えば、普通名詞「野球」、動詞「投げる」は内容語、接続詞「なぜならば」、副詞「極めて」は機能表現である（Biber and Conrad, 2009；松吉ら, 2007）。

内容語は本稿で扱うトピック、つまり話題に関連する。言語学では「一文中の主題」として「topic」という用語を用いることがあるが、本論のトピックは文書単位における「話題」に近いものである。

一方、機能表現はレジスターに深く関わるものといえる。本稿では、レジスターを「言語の共時的、機能的な変異であり、状況によって語彙・文法の使用が変化するものおよびコミュニケーションの目的とコンテクストにおいて明瞭に定量的なパターンによって特徴づけられる」と定義する（Biber and Conrad, 2009）。日本語には機能表現が多種多様に存在し、豊富であることから、言語の機能的な変異であるレジスターを分析する上で機能表現を用いることが考えられる。また、文書の文体に深く関わる品詞比率もレジスターを分析する上で有効であると考えられる。

２．１．トピックモデル

トピックモデルは、確率的生成モデルであり、代表的なものとしては Latent Dirichlet Allocation（LDA）がある。LDA はトピック分布の多項分布でモデル化し、トピックの分布に対してディリクレ分布を仮定する（Blei et al., 2003）。本稿では、Yahoo! LDA を用いて 1000 トピックのモデルを BCCWJ で学習した（Smola and Narayanamurthy, 2010）。トッピクモデルの素性は、形態素解析辞書 UniDic の品詞名によって名詞（数詞を除く）、動詞（非自立可能なものを除く）、形容詞（非自立可能なものを除く）、形状詞（助動詞語幹のものを除く）、副詞の品詞から語（短単位）を選択した。表 1 は、LDA モデルにおけるそれぞれのメディアごとのトピックの関連語をそれぞれ示したものである。

例えば、書籍（PB、LB、OB）のグループでは、「顔、目、手、声」「言う、思う」などが共通して出現しており、小説などの創作における具体的な人間の行為、想念などに関連していることが推測できる。韻文（OV）では、「月、花、秋、夜」「赤、白、風、空」など詩歌の題材に見られるトッピクが抽出されている。国会会議録（OM）では「大臣、政府、委

† hodoscek.b.aa@m.titech.ac.jp

員」のような役職のグループ、「言う、訳、こと、風」など言語行為に関する語群が抽出されている。

このように LDA によるトピックの抽出は、それぞれのメディアの特色を提示していることが分かる。

表 1: 各メディアにおけるトピック（高頻度順）

	1 位	2 位	3 位
PB	顔，目，声，そう，手	言う，事，思う，そう，時	事，於く，つく，因る，物
LB	言う，事，思う，そう，時	顔，目，声，そう，手	言う，声，顔，事，そう
OB	言う，事，思う，そう，時	顔，目，声，そう，手	言う，出る，男，電話，入る
PM	人気，スタイル，感，デザイン，使う	バッグ，スカート，ニット，パンツ，スタイル	シャツ，ブランド，ジャケット，プリント，カラー
PN	優勝，回，大会，初，決勝	大統領，米国，関係，政府，外交	首相，自民，コイズミ，政治，総理
OC	方，教える，どう，分かる，出る	言う，今，人，時，知る	言う，事，思う，そう，時
OY	所，もう，後，前，気	言う，事，思う，そう，時	今日，明日，笑い，頑張る，まあ
OW	於く，為，つく，行う，図る	年，パーセント，増加，図，別	事業，整備，年度，施設，実施
OV	夜，日，花，秋，月	風，白い，中，空，赤い	姿，巨大，物，光，今
OT	実験，調べる，事，考える，分かる	計算，数字，数，答え，桁	運動，力，速度，時，物体
OP	月，日，市，申し込み，センター	課，■■，月，平成，市	時，日，午後，分，午前
OL	条，項，規定，当該，於く	条，項，規定，因る，業務	事業，事，指定，定める，大臣
OM	委員，つく，事，大臣，政府	言う，訳，そう，事，風	案，国会，提出，つく，法案

略称：LB、PB、OB：書籍；PM：雑誌；PN：新聞；OC：Yahoo! 知恵袋；OY：Yahoo! ブログ；OW：白書；OV：韻文；OT：教科書；OP：広報紙；OL：法律；OM：国会会議録

２．２．レジスター

本稿では、レジスターとして下記の３種類の異なるデータを均等に重み付けをし、計量する。

a) 松吉ら（2007）の「つつじ：日本語機能表現辞書」に含まれる機能表現

b) Srdanović ら（2008）で用いる推量副詞

c) 品詞比率からなる Modifier Verb Ratio (MVR) という指標

以下、それぞれについて述べる。

２．２．１．機能表現

「つつじ」では、機能表現が階層構造によって構成されている。本稿では、つつじのレベル L2 の区分から異なり表現 435 種類を用いた。表２は各メディアごとの上位５位までの機能表現を示している。

表 2: 各メディアにおける機能表現（高頻度順）

	1 位	2 位	3 位	4 位	5 位
PB	から	こと	よう	という	です
LB	から	こと	という	よう	です
OB	から	こと	です	よう	という
PM	から	です	という	こと	よう
PN	から	など	こと	では	という
OC	です	から	ので	って	こと
OY	です	から	ので	こと	よう
OW	こと	について	から	において	として
OV	から	よう	なり	とき	こと
OT	よう	から	など	こと	には
OP	など	から	です	こと	とき
OL	とき	において	により	による	なら
OM	という	こと	です	から	よう

２．２．２．推量副詞

表３では、Srdanović ら（2008）で使用したものと同じ推量副詞（合計 18 種類）を用い、各メディアごとの上位５位までの推量副詞を示す。

表 3: 各メディアにおける推量副詞（高頻度順）

	1 位	2 位	3 位	4 位	5 位
PB	あるいは	必ず	絶対（に）	恐らく	きっと
LB	あるいは	絶対（に）	必ず	恐らく	きっと
OB	あるいは	必ず	絶対（に）	きっと	恐らく
PM	絶対（に）	必ず	あるいは	きっと	多分
PN	絶対（に）	必ず	あるいは	きっと	必ずしも
OC	絶対（に）	必ず	多分	きっと	もしかして／たら／すると
OY	絶対（に）	多分	きっと	どうも	必ず
OW	あるいは	必ずしも	絶対（に）	必ず	大抵
OV	あるいは	きっと	恐らく	多分	どうやら
OT	あるいは	絶対（に）	必ず	必ずしも	恐らく
OP	必ず	絶対（に）	あるいは	きっと	必ずしも
OL	必ず	あるいは	よほど	ひょっとして／たら／すると	もしかして／たら／すると
OM	あるいは	どうも	恐らく	必ずしも	絶対（に）

この分布を見ると例えば Yahoo! 知恵袋と Yahoo! ブログ、新聞、雑誌などでは、「絶対に」という感情的な表現が高頻度に出現し、白書（OW）、国会（OM）では「あるいは、必ず（しも）、どうも、多分、おそらく」などの婉曲的な表現が出現している。このような特色もレジスターを区別する有用なデータとなると考えられる。

２．２．３．MVR

MVR は文章中における用の類（動詞）とそれらを修飾する相の類（副詞、連体詞、形容詞、形状詞）の比率（100× 相の類の比率／用の類の比率）であり、文章の文体を計る指標とされる（樺島忠夫、寿岳章子, 1965; 冨士池ら, 2011; Hodošček, 2011）。名詞比率が低い場合、MVR が高いほど「ありさま描写的」、低いほど「動き描写的」であることから、ジャンルによる特色を読み取ることができると考えられる。また、名詞比率が高いと「要約的」という。本稿では、サ変名詞などの影響を少なくするために品詞比率を計算する際、BCCWJ の長単位データを用いた。

表 4: 各メディアにおける名詞比率と MVR

		PB	LB	OB	PM	PN	OC	OY	OW	OV	OT	OP	OL	OM
N	平均値	31.06	28.66	25.81	35.00	40.12	25.01	28.35	44.07	34.67	33.23	49.56	42.75	30.77
	SD	7.01	6.13	4.35	7.16	5.86	6.79	10.30	4.76	6.68	6.15	4.30	3.38	4.90
MVR	平均値	74.98	75.75	76.17	86.23	51.45	94.92	116.08	59.74	54.34	61.55	57.98	30.99	69.71
	SD	24.63	21.54	20.64	32.79	19.43	86.04	151.25	25.49	42.83	23.49	9.15	12.76	15.19

表４では、Yahoo! 知恵袋と Yahoo! ブログにおける MVR の標準偏差が大きいのに対し、広報誌、新聞、国会会議録、法律書における MVR の標準偏差が小さいことが明らかになった。つまり、Yahoo! 知恵袋とブログは、多様なテキストが混在している一方で、広報誌や新聞などでは、事のありさまを描写する文書から成りたっていることが推測できる。名詞比率からは、広報誌、白書、法律および新聞が要約的なメディアであることが分かった。

３．考察

BCCWJ 中の様々なメディア間の差異を計量するために、前述のトピックモデルとレジスターの観点から BCCWJ に含まれる全サンプルをメディアごとにまとめて観察した。表５はスピアマンの順位相関係数でメディア間のトピックとレジスターのそれぞれの相関を示したものである。

表 5: メディアにおけるトピックおよびレジスターの相関

トピック												
	PB	LB	OB	PM	PN	OC	OY	OW	OV	OT	OP	OL
LB	0.85											
OB	0.50	0.66										
PM	0.58	0.56	0.43									
PN	0.56	0.55	0.29	0.59								
OC	0.25	0.18	0.28	0.54	0.33							
OY	0.18	0.22	0.34	0.60	0.37	0.77						
OW	0.33	0.17	−0.09	0.10	0.49	0.06^{ns}	−0.11					
OV	0.28	0.42	0.42	0.27	0.16	0.16	0.33	−0.20				
OT	0.52	0.54	0.35	0.34	0.38	0.19	0.20	0.24	0.29			
OP	0.14	0.09	-0.02^{ns}	0.17	0.37	0.10	0.13	0.29	−0.01	0.13		
OL	0.24	0.04^{ns}	−0.15	−0.08	0.29	0.04^{ns}	−0.17	0.61	−0.22	0.12	0.23	
OM	0.18	0.10	-0.04^{ns}	-0.01^{ns}	0.36	-0.03^{ns}	−0.15	0.53	−0.18	0.08	0.21	0.56

レジスター												
	PB	LB	OB	PM	PN	OC	OY	OW	OV	OT	OP	OL
LB	0.93											
OB	0.89	0.90										
PM	0.88	0.88	0.87									
PN	0.78	0.77	0.78	0.81								
OC	0.81	0.80	0.79	0.83	0.74							
OY	0.86	0.85	0.83	0.85	0.76	0.87						
OW	0.66	0.65	0.65	0.67	0.76	0.60	0.62					
OV	0.64	0.64	0.65	0.66	0.69	0.59	0.62	0.63				
OT	0.74	0.73	0.74	0.76	0.82	0.70	0.71	0.77	0.71			
OP	0.72	0.71	0.71	0.74	0.76	0.72	0.72	0.71	0.61	0.78		
OL	0.41	0.40	0.40	0.42	0.49	0.38	0.38	0.60	0.44	0.53	0.51	
OM	0.72	0.72	0.72	0.72	0.72	0.70	0.70	0.68	0.57	0.69	0.71	0.44

* 注意：$n.s.$ 以外の値はすべて $p < .05$

4．まとめ

以上の分析から、あるメディアがほかのメディアと大凡どの程度トピックおよびレジスターが異なるかが分かった。今後の課題としては、メディアごとのサンプルに分析を拡大することが必要である。

文　献

樺島忠夫、寿岳章子（1965）『文体の科学』綜芸舎

冨士池優美、小西光、小椋秀樹、小木曽智信、小磯花絵（2011）「長単位に基づく媒体・カテゴリ間の品詞比率に関する分析」特定領域研究「日本語コーパス」平成 22 年度公開ワークショップ（研究成果報告会）予稿集，pp. 273–280.

松吉俊、佐藤理史、宇津呂武仁（2007）「日本語機能表現辞書の編纂」自然言語処理，Vol. 14，No. 5，pp. 123–146.

Biber, Douglas, and Susan Conrad (2009) Register, Genre, and Style. Cambridge: Cambridge Textbooks in Linguistics.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003) “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol. 3, pp. 993–1022.

Hodošček, Bor (2011) “Word Class Ratios and Genres in Written Japanese, Revisiting the Modifier-Verb Ratio”, Acta Linguistica Asiatica, Vol. 1, No. 2, pp. 53–62.

Smola, A., and S. Narayanamurthy (2010) “An Architecture for Parallel Topic Models”, In The Proceedings of the VLDB Endowment (PVLDB), Vol. 3, No. 1, pp. 703–710.

Srdanović, I., B. Hodošček, A. Bekeš, and K. Nishina (2009)「ウェブコーパスと検索システムを利用した推量副詞とモダリティ形式の遠隔共起抽出と日本語教育への応用」自然言語処理，Vol. 16，No. 4，pp. 29–46.

関連 URL

つつじ：日本語機能表現辞書	http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/
Yahoo! LDA	https://github.com/shravanmn/Yahoo_LDA