クラスター分析とフレーム分析による語彙のジャンル別特徴 ： 「現代日本語書き言葉均衡コーパス」を用いて

内田, 諭
九州大学大学院言語文化研究院 ： 准教授

藤井, 聖子
東京大学大学院総合文化研究科 ： 教授

https://doi.org/10.15017/1500408

出版情報：言語文化論究. 34, pp.21-34, 2015-03-20. 九州大学大学院言語文化研究院
バージョン：
権利関係：

クラスター分析とフレーム分析による語彙のジャンル別特徴
――「現代日本語書き言葉均衡コーパス」を用いて――

内田　　諭・藤井　聖子[1]

要　旨

本研究は、『現代日本語書き言葉均衡コーパス』における語彙頻度情報に基づき、語彙使用の特徴をジャンル別に明らかにすることを目的とする。基本語彙の特徴を明らかにするため、量的なアプローチとして動詞および名詞の高頻度語彙を対象にコーパスのジャンル間の共起関係を調べ、統計的な類似度の尺度の１つであるコサイン係数によって類似性を測定し、クラスター分析を行った。その結果、名詞・動詞ともに書籍コーパスと教科書コーパスで大きな違いがあること、また文系・理系の分野間では類似性が見られることなどが明らかになった。次に、ジャンルごとの特徴を意味的に明らかにするため、各ジャンルにおける特徴的な動詞について、フレーム意味論およびFrameNetに基づき、フレームの観点から質的な分析を行った。その結果、各ジャンルには典型的に喚起されるフレームが存在することが明らかとなった。FrameNetにおけるフレーム間関係を用いて分析すると、例えば「文学」のジャンルでは「感情」「知覚」「身体動作」などに関するフレームが特徴的に見られることが分かり、これまで明らかにされてこなかったジャンルごとの意味的な特徴が浮かび上がる。

キーワード：ジャンル、クラスター分析、意味フレーム、フレームネット、BCCWJ

１．はじめに

本論文は、『現代日本語書き言葉均衡コーパス（以下BCCWJ）』における語彙頻度に基づき、ジャンル別の語彙使用の特徴を（1）クラスター分析によって統計的に類似度を測定し、（2）意味的な特徴をフレーム意味論の観点から明らかにすることを目的とする。これまでコーパスのジャンルごとの特徴抽出は、慣用表現（村田・山崎，2011）、接続表現（石黒ほか，2009）、文末表現（鄭ほか，2009）、出現語彙（石川，2001）、品詞分布・人称などの文法的特徴（Biber & Conrad, 2009: 177-208; Puschmann, 2010）など様々な観点から行われてきたが、フレームを手がかりにした分析はほとんど行われていない[2]。フレームとは、Fillmore（1982，1985など）が提唱するフレーム意味論の中心となる概念で、言語表現が喚起する背景知識である。現在、フレームの辞書としてFrameNet（http://framenet.icsi.berkeley.edu/）が構築されており、FrameNetの記述に照らし合わせれば、単語が喚起するフレームを特定することができ、「意味」を中心とした分析ができる。また、フレーム意味論に依拠することで、各ジャンルの意味的な特徴について、言語理論に裏打ちされた体系的な分析が可能となる。

本研究では、BCCWJの書籍コーパスおよび教科書コーパスから、日本十進分類法（NDC）・８教

科の区分[3]に従って抽出した動詞・普通名詞の頻度上位語を分析の主な対象として用いる。書籍コーパスと教科書コーパスを選択した理由は、一般的に流通している汎用性の高い書籍コーパスと、特定の領域に限定された教科書コーパスの違いを語彙項目の観点から明らかにできると考えたからである。また、動詞と名詞を分析の対象とする理由は、それぞれフレーム喚起語（cf. 4.2節）として認定できる場合が多く、フレーム意味論の観点から、最適な分析対象だからである。

2節でBCCWJの概要を示した後、3節ではジャンルごとの大まかな特徴と類似性を捉えるために量的なアプローチとして語の共起回数から算出したコサイン係数を用い、クラスター分析によって類似度測定を行う。4節では、動詞に焦点を当て、各ジャンルに特徴的に出現する語を抽出し、フレームの観点から質的な分析を加え、ジャンルの特性を論じる。5節はまとめである。

2.『現代日本語書き言葉均衡コーパス』

『現代日本語書き言葉均衡コーパス』（Balanced Corpus of Contemporary Written Japanese）は、国立国語研究所が中心となって構築された日本語の大規模コーパスである[4]。2006年に開発が始まり、2011年に1億語規模の汎用コーパスとして公開された。その内実は、従来からのコーパスの研究対象であった新聞・雑誌・書籍（出版（生産実態）サブコーパス）に加え、広く図書館などを介して流通している書籍（図書館（流通実態）サブコーパス）や教科書・白書・ブログなど（特定目的サブコーパス）を含み、ランダムサンプリングによって均衡化が図られている[5]。本論文では『BCCWJ領域内公開データ2008年度版』を用い、書籍コーパスと教科書コーパスを対象に分析を行う。このデータでは、同一の基準で書籍コーパスおよび教科書コーパスの語彙リストが提供されており、それぞれのコーパスの特徴と類似度を調査するという本稿の目的に適っている[6]。

書籍コーパスは、日本十進分類法（NDC）により、総記、哲学、歴史、社会科学、自然科学、技術・工学、産業、芸術・美術、言語、文学、に分類されたジャンル別コーパスがある。教科書コーパスは、国語、数学、理科、社会、外国語、技術家庭、芸術、保健体育の8科目に分類される。それぞれの語数は表1の通りである。

表1　BCCWJの各ジャンルの語数

No	ジャンル	延べ語数	異なり語数
0	総記	521,436	27,358
1	哲学	1,403,199	37,338
2	歴史	2,141,841	53,257
3	社会科学	5,447,856	55,934
4	自然科学	1,074,332	31,509
5	技術・工学	1,115,821	35,496
6	産業	700,269	29,274
7	芸術・美術	1,107,179	40,092
8	言語	398,497	24,337
9	文学	8,775,301	71,304
n	番号なし	281,414	19,882
	合計	22,967,145	425,781

教科	延べ語数	異なり語数
国語	104,545	12,379
数学	50,591	1,813
理科	66,984	4,348
社会	106,468	9,478
外国語	12,767	1,883
技術家庭	64,627	6,017
芸術	36,170	5,823
保健体育	24,277	3,564
合計	466,429	45,305

本稿では上記の16（総記を除く[7]）のジャンルについてコーパスの付属資料である語彙頻度データを用いて分析を行う。なお、この語彙表は、書籍コーパスはプレーンテキストのものを、教科書コーパスは中学教科書を対象に各ジャンル別の語彙・品詞別の頻度を一覧にしたものである。

3．高頻度語のクラスター分析

3.1　分析対象と手法

本稿では、2節で提示したジャンルの特徴を明らかにするために、動詞・名詞それぞれについて、各ジャンルの頻度上位100語を抜き出し、これを分析対象とした。上位100語の全語彙に対するカバー率は、動詞は平均78.2%、名詞は平均32.4%である。上位100語に限定することにより、各ジャンルの低頻度の特異な語彙を排除することができ、不必要に類似度が低くなることを避けることができる。また、高頻度語に焦点を当てることで、汎用的な語彙が抽出されるため、ジャンル間の規模の差をある程度吸収することができる。さらに、4節での分析の対象となる動詞については出現頻度としてはおよそ8割の語彙をカバーしており、一般的な傾向を調査するには十分な分量だといえる。

次に、各ジャンルについて、上位100語の頻度を質的データに変換した。質的データに変換した理由の1つは、「する」、「なる」、「こと」、「もの」などの高頻度語の影響で不必要に類似度が高くなることを抑えるためである。また、質的なデータに変換することでコーパス間の母数の不均衡や、偶然の順位差による影響をできるだけ少なくすることにもなる[8]。質的データに変換することで、適用する統計的手法が必然的にノンパラメトリックなものとなるが、Kilgarriff（2001）は、言語の分布はジップの法則に従うため、コーパス間の比較を行う際に、正規分布を前提とした統計手法ではなく、いかなる分布の形式も前提としないノンパラメトリックな手法が有効であることを指摘しており、本稿の手法もこれに従うものである。

統計処理をする基礎データを作成するため、それぞれのジャンルの上位100語のリストを結合し、出現する語の項目一覧を作成した（動詞は、「見る」「来る」「書く」など325項目、名詞は「人」「語」「時」など682項目）。この基礎データでは語彙項目を行に、コーパスのジャンルを列に取り、出現の有無を0か1で記録している。例えば、「会う」という動詞は［歴史］、［芸術・美術］、［文学］の上位100語に出現するため、それぞれのジャンルに1としてコードし、それ以外のジャンルでは0とした。この表を基に、それぞれのコーパスで上位100語がいくつ共通しているかという語彙項目の共起回数を計算した。例えば、［歴史］と［文学］は上位100語中51語が共通しているため、51という値を付し、行列として記録した。次に共起回数の行列を基にコサイン係数を用いて相関を計算した。コサイン係数とは、複数のデータを複数の項目で比較するとき、ある2つのデータ間において、特定の項目が存在しない場合である［0,0］の影響を極力抑えた数値で、a＝［1,1］（両方に存在する）、b＝［1,0］（一方に存在する）、c＝［0,1］（他方に存在する）、とすると次式によって求められる。

$$\text{コサイン係数} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

この数値を用いることで、複数項目間の属性相関を求める場合でも［0,0］データが多くなることによる不自然な係数の上昇を抑えることができ、各コーパス間の関係をより正確に算出することが可能になる。なお、計算はExcelのVBAプログラムであるNumeros（cf. 上田，2013）を使用した。表1および表2は、動詞・名詞の各コーパス間のコサイン係数を示したものである。

表２　頻度上位動詞のコーパス間のコサイン係数

[PHIX_V]	1哲学	2歴史	3社科	4自科	5技工	6産業	7芸美	8言語	9文学	外語	技家	芸術	国語	社会	数学	保体	理科
1 哲学	1.000																
2 歴史	0.830	1.000															
3 社科	0.820	0.800	1.000														
4 自科	0.790	0.780	0.820	1.000													
5 技工	0.750	0.740	0.810	0.830	1.000												
6 産業	0.770	0.770	0.850	0.850	0.870	1.000											
7 芸美	0.770	0.840	0.760	0.750	0.730	0.780	1.000										
8 言語	0.780	0.770	0.780	0.760	0.740	0.770	0.740	1.000									
9 文学	0.670	0.700	0.620	0.650	0.620	0.650	0.790	0.640	1.000								
外語	0.580	0.560	0.590	0.570	0.580	0.590	0.580	0.590	0.530	1.000							
技家	0.550	0.520	0.570	0.580	0.630	0.570	0.500	0.530	0.470	0.520	1.000						
芸術	0.540	0.560	0.530	0.560	0.550	0.540	0.590	0.550	0.510	0.500	0.550	1.000					
国語	0.720	0.710	0.640	0.690	0.670	0.660	0.730	0.670	0.670	0.640	0.590	0.620	1.000				
社会	0.630	0.620	0.640	0.620	0.600	0.620	0.570	0.560	0.460	0.540	0.590	0.540	0.640	1.000			
数学	0.510	0.480	0.520	0.570	0.530	0.550	0.510	0.470	0.470	0.520	0.520	0.460	0.540	0.490	1.000		
保体	0.570	0.530	0.590	0.610	0.600	0.580	0.520	0.510	0.460	0.460	0.580	0.510	0.550	0.590	0.460	1.000	
理科	0.500	0.500	0.550	0.580	0.560	0.560	0.500	0.510	0.430	0.480	0.580	0.510	0.610	0.570	0.560	0.560	1.000

表３　頻度上位名詞のコーパス間のコサイン係数

[PHIX_N]	1哲学	2歴史	3社科	4自科	5技工	6産業	7芸美	8言語	9文学	外語	技家	芸術	国語	社会	数学	保体	理科
1 哲学	1.000																
2 歴史	0.630	1.000															
3 社科	0.530	0.640	1.000														
4 自科	0.490	0.490	0.480	1.000													
5 技工	0.500	0.520	0.550	0.590	1.000												
6 産業	0.460	0.500	0.610	0.470	0.640	1.000											
7 芸美	0.560	0.570	0.440	0.450	0.480	0.430	1.000										
8 言語	0.520	0.520	0.540	0.530	0.530	0.460	0.500	1.000									
9 文学	0.590	0.510	0.380	0.400	0.400	0.390	0.600	0.420	1.000								
外語	0.230	0.220	0.220	0.200	0.210	0.240	0.290	0.370	0.250	1.000							
技家	0.160	0.170	0.220	0.250	0.330	0.280	0.180	0.250	0.150	0.170	1.000						
芸術	0.240	0.250	0.240	0.270	0.260	0.210	0.370	0.300	0.220	0.170	0.190	1.000					
国語	0.430	0.440	0.360	0.390	0.420	0.350	0.500	0.560	0.450	0.380	0.240	0.380	1.000				
社会	0.280	0.470	0.470	0.310	0.360	0.440	0.260	0.300	0.200	0.150	0.220	0.260	0.270	1.000			
数学	0.120	0.120	0.120	0.180	0.200	0.170	0.140	0.180	0.110	0.130	0.170	0.160	0.200	0.120	1.000		
保体	0.260	0.220	0.220	0.380	0.270	0.280	0.190	0.240	0.210	0.110	0.290	0.210	0.260	0.250	0.160	1.000	
理科	0.190	0.170	0.190	0.350	0.290	0.230	0.210	0.230	0.190	0.150	0.300	0.250	0.230	0.230	0.210	0.280	1.000

３．２　クラスター分析の結果

表１および表２を基に、統計ソフトウェアR（3.0.2）を用いて、クラスター分析を行った。その際、ユークリッド距離を基準に、ウォード法を用いて描画した（cf. 徳永，1999; 浅野・江島，1996）。各ジャンルのコーパスの上位100語によるクラスター分析の結果を図１と図２に示す。図１は動詞に関するデンドログラムで、図２は名詞に関するものである。

図１と図２から明らかになったことは、動詞・名詞ともに教科書が書籍とは異なった傾向を示すということである。図１・図２ともに［数学］［理科］［社会］［技術家庭］［保健体育］［外国語］［芸術］とそれ以外がそれぞれクラスターを作っており、その他のクラスターと結合する高さが高く、大きく異なるということを示しており、教科書における学習言語（教育を目的とした記述）の特異性が示唆される。ただし、教科書の中では［国語］が書籍コーパスと近い振舞いを示しており、動詞の場合は［文学］と、名詞の場合は［言語］とクラスターを形成していることが見て取れる。これは［国語］の教科書に書籍等から抜粋した随筆や小説や文法事項の説明などが掲載されているこ

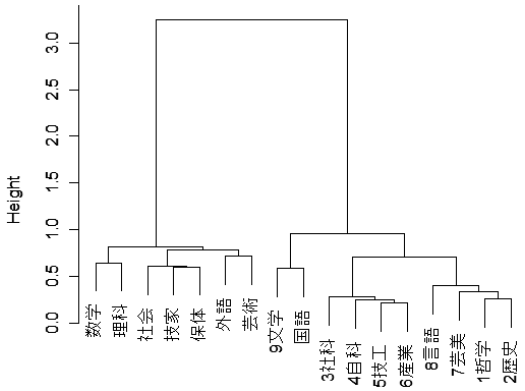

図１　動詞による各ジャンルのクラスター

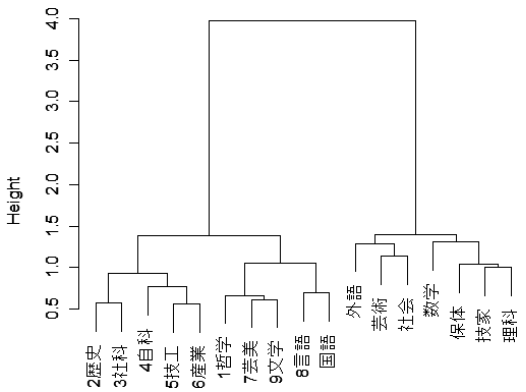

図２　名詞による各ジャンルのクラスター

とが理由として考えられる。さらに、注目すべき点は、動詞のクラスターが［数学］［理科］、［５技工］［６産業］という比較的理系のものがまとまる一方で、［１哲学］［２歴史］など文系の内容がクラスターを形成しているということである。これは、専門用語が多く系統の似た分野が類似の分布を示すと考えられる名詞の場合だけではなく、動詞においてもジャンルによって独特の分布を示すということを示唆している。

４．フレームによる特徴語の分析

前節ではジャンル間の類似性を、出現語彙の共通性から統計的手法を使って分析した。特に、動詞のクラスター分析において、近接すると考えられる分野がクラスターを形成しており、ジャンルに特定の動詞表現があることが示唆された。本節では、動詞とジャンルの連関を意味的に分析するため、フレームを用いた分析を提示する。まず、ジャンルごとを特徴づける語彙を選定した後、ウェブ上で構築されている FrameNet（http://framenet.icsi.berkeley.edu/）の記述をもとにフレームの分析を進める。

４．１　ジャンルを特徴づける語彙

近藤（2011）では、BCCWJ の「教科書コーパス」（2010年12月９日版（非公開））を対象に、各教科に特徴的な語彙のリストを提供している。この研究では、特徴度の算出には対数尤度比を補正した以下の数値（近藤, 2011: 5より引用）が用いられている（ln は自然対数を表す）。

２(alna+blnb+clnc+dlnd−(a+b)ln(a+b)−(a+c)ln(a+c)−(b+d)ln(b+d)−(c+d)ln(c+d)+(a+b+c+d)ln(a+b+c+d))

a：当該テキストでの当該語の度数
b：参照テキストでの当該語の度数
c：当該テキストの延べ語数−a
d：参照テキストの延べ語数−b
※ただし、ad-bc＜０の場合、−１を乗じる補正を行う。

この指標では、「特徴度が 0 であれば、当該テキストと参照テキストで当該語の出現の程度は等し」く、「特徴度が正の値で、かつ値が高ければ高いほど、当該テキストにおいて高頻度という意味で特徴的な語と見なされる」（近藤, 2011: 5）[9]。提供されている特徴語のリストには、特徴度が10.83 より大きい語（p<.001）で、数詞・人名などではないものが含まれており、ジャンルごとの特徴を表す語を概観することができる。しかしながら、この手法では、低頻度の語が抽出され、ジャンルによってはリストが膨大になるという問題点がある。例えば、高校の外国語の教科書では、以下のような語が特徴語としてリストされている[10]。

> 意見 (5)、インタビュー (5)、英語 (5)、選ぶ (7)、学校 (9)、記号 (6)、機能表現 (1)、空所 (6)、具体例 (2)、時間的順序 (2)、主題文 (3)、順 (3)、将来 (5)、説明 (8)、相違点 (3)、其々 (6)、対照 (3)、つ (20)、次 (15)、つく (19)、展開 (5)、展開法 (7)、発信型ライティング (1)、パラグラフ (7)、ホームページ (7)、物語文 (1)、ライティング (1)、類似点 (3)、例証 (3)

このリストには、「機能表現」、「発信型ライティング」、「物語文」、「ライティング」など頻度 1 の単語も含まれている。これらの語は、他の教科では出現しないため、外国語の教科書の特徴語として算出されたと考えられるが、その語彙自体の頻度が少なく、教科を特徴づける重要語であるとはいえない。

本論文の分析では、各ジャンルの上位100語を対象とするため、低頻度の単語を抽出するという問題はなく、また汎用的な高頻度語に限られるため、特徴語のリストは不必要に大きなものとはならずポイントが明確になる。本稿では、「単独のジャンルにのみ出現する語」を単独出現語（cf. 石川，2001）と定義し、それを特徴語として扱うこととする。単独出現語は、そのジャンルの出現頻度上位100語に入っていながら、他のジャンルでは上位に入っていないものであり、そのジャンルを特徴づける高頻度語であるといえる。例えば、「冷える」という動詞は、［理科］にのみ出現する（表 4 ）。

表 4 「冷える」の出現状況

単　語	1 哲学	2 歴史	3 社科	4 自科	5 技工	～	理科
冷える	0	0	0	0	0	～	1

以下は、各ジャンルにおける動詞の単独出現語のリストである。

書籍コーパス
［1 哲学］ ［なし］
［2 歴史］ 記す
［3 社科］ 生ずる、通ずる
［4 自科］ 無くなる
［5 技工］ 混ぜる、煮る、編む
［6 産業］ 育てる、咲く、伸びる、占める、売る
［7 芸美］ 撮る
［8 言語］ 扱う、居る、挙げる、限る、指す、数える
［9 文学］ 驚く、見詰める、済む、思い出す、振る、聞こえる、落ちる、連れる、頷く

教科書コーパス

［外語］ 演ずる、使う、申し出る、吹き込む、断る、綴る、届く、聞き取る、聞き返す、褒める、誘う、頼む、話し掛ける

［技家］ 育つ、盛る、洗う、整える、罫書く、削る、振り返る、適する、遊ぶ、汚れる、深める、縫う、着る

［芸術］ 溢れる、感じ取る、慣れる、輝く、構える、込める、塞ぐ、伸ばす、親しむ、吹く、生み出す、弾く、舞う、優れる

［国語］ 泣く、降る、取り上げる

［社会］ 読み取る、果たす、築く、巡る、栄える、広まる、訪れる、目指す

［数学］ 消す、切り取る、追い付く、離れる、得る、埋める、直す、諮る、混じる、転がる、投げる、導く、折る、当て嵌まる、重なる、交わる、割る、成り立つ

［保体］ 楽しむ、及ぼす、繋がる、取り除く、取り組む、悩む、止まる、満たす、踊る、感ずる、履く、解す、高める、掴む

［理科］ 加わる、覆う、冷える、押す、温める、分かれる、下がる、結び付く、運ぶ、混ぜ合わせる、冷やす、燃える、流す、繋ぐ、取り出す、溶ける、熱する

これらを仔細に観察すると、ジャンル内の語同士に類似性があることが見て取れる。例えば、「理科」における「冷える」「温める」「熱する」などは、すべて温度変化に関わる動詞である[11]。以下、この類似性についてフレームの観点から分析し、ジャンルごとの特徴語の共通性はこれらの動詞が同一または類似のフレームを喚起するということから説明できるということを示す。

4.2 フレームと FrameNet

フレームとは、言葉を理解するための背景知識であると定義される（Fillmore, 1982）。フレーム意味論ではフレームは主に語句によって喚起されると規定されており、英語の語彙が喚起するフレームは FrameNet と呼ばれる現在構築中であるオンライン上のフレーム辞書に記述されている。例えば、次の例を考えてみよう。

(1) ［<*cook*>Matilde］ fried_Tgt[12]［<*food*>the catfish］［<*heating_instrument*>in a heavy iron skillet］
（Ruppenhofer et al., 2006: 5）

この例の場合、動詞 fry は Apply_heat フレームを喚起していると分析される。このようにフレームを喚起する語をフレーム喚起語（frame evoker）と呼ぶ。単語が喚起するフレームを特定することは、意味を「具体化」するということである。FrameNet の記述に依拠することで、語が持つ意味内容をフレームという形で具体的に示すことが可能となる。

フレームは、その喚起された状況に参与する要素を、フレーム要素（frame element）として持っている。このフレーム要素には文法的・意味的に重要な役割を果たすコアなもの（core frame element）と、周辺的なものがある（peripheral frame element）。Apply_heat フレームの場合、コアフレーム要素として *cook*, *food*, *heating_insrument* を持つ。これを文にアノーテーションを施したのが上記の（1）である[13]。さらに、フレームはフレーム間関係で結ばれており、上位フレームや下位フレームなどの関連するフレームが規定されている。これらの関係はコアフレーム要素を上位からすべて引き継ぐ Inheritance、一部を使用する Use、シナリオの一部である Subframe、自動詞と他動詞などの関係

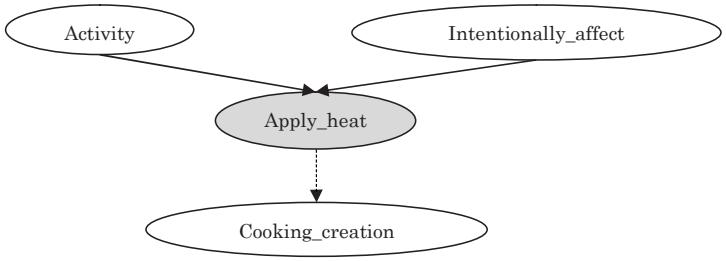

図３　Apply_heat フレームのフレーム間関係

性を示したIs Causative ofなどの関係がある（詳しくはFillmore et al., 2003, Ruppenhofer et al., 2010）。図3はApply_heatフレームのフレーム間関係の一部を抜き出したものである（実線はInheritance、点線はUseの関係を表す）。

フレーム間関係により、フレームを階層化・グループ化することができる。これにより、「意味が似ている」という直感的な判断は、同一フレームに属する、あるいは共通する上位フレームを有するなどの観点から、具体化する事が可能である。

本稿の分析は、日本語が対象であるが、英語版のFrameNetを用いる。その理由は、構築の初期段階である日本語のフレームネット（http://jfn.st.hc.keio.ac.jp/）に比べて、英語版のほうが安定しており、項目数が多いこと、汎用的な高頻度語が喚起するフレームについては日本語・英語で大きな差はないと考えられ、本研究の分析上不都合はないことなどによる[14]。

４.３　単独出現語が喚起するフレーム

それでは、各ジャンルの特徴語はどのようなフレームを喚起するのであろうか。まず、［産業］の特徴語である「伸びる」から考えてみることとする。コーパスより例を一つ引く。

(2) 調査結果によると、約九割のチェーンが医薬品の導入で売り上げが大幅、もしくはある程度伸びると予測した。

この「伸びる」に対応する英語はgrowであると考えられる。FrameNetによると、この語が喚起するフレームは、Expansionである。一方、もう一つの単独出現語の「売る（sell）」が喚起するフレームはCommerce_sellである。これらのフレームは他のジャンルには現れないものであり、［産業］コーパスの特徴を端的に表しているといえる。以下、ケーススタディとして［文学］と［理科］を取り上げ、フレーム間関係も含めて詳しく考察する。

４.４　フレームから見た「文学」ジャンルの特徴

文学のジャンルにおける単独出現語は以下の通りである（再掲）。

［９文学］ 驚く、見詰める、済む、思い出す、振る、聞こえる、落ちる、連れる、頷く

これらに対応するフレームを英語のFrameNetから抜き出すと表5のようになる。表5にはフレーム間関係を参考にそれぞれのフレームの上位フレームも提示した。これにより、これらのフレーム

をグルーピングし、喚起されたフレームの傾向を知ることが可能になる。

日本語をそれぞれ英語に置き換え、喚起されるフレームをリストすると、Experiencer_subj, Perception_active, Activity_finish, Memory, Body_movement, Perception_experience, Motion_directional, Bringing である。これらのフレームのフレーム間関係の全体像を視覚化したのが図４である。

最上位に位置するフレームをみると、「文学」の単独出現語が喚起するフレームは、Emotions, Perception, Event, Motion にまとめることができる。これらの意味領域はこのジャンルに特徴的に見られるものであると結論付けることができるだろう。この結果は、直感的な感覚に沿うもので、文学作品では登場人物の「感情」（Emotion フレーム）、「知覚」（Perception フレーム）、「身体動作」（Motion フレームに関するもので特に Body_movement が該当）が描画されることが多いことが、これらのフレームの出現の理由であると解釈することができる。重要な点は、このような直感がフレーム意味論によって、理論的な理由付けを得ることができるということである。同時に、ジャンル分析の観点からは、先行研究が部分的な現象や形式的な特徴に基づいたものが多かったことに対して、「フレーム」という形で、意味内容を具体化できるという点が重要である。これにより、「どの意味（フレーム）がどのジャンルに特徴的か」ということを明らかにすることができる。

表５　「文学」の特徴語が喚起するフレーム

日本語	英　語	フレーム	上位フレーム
驚　く	be surprised	Experiencer_subj	Uses: Emotions
見詰める	stare	Perception_active	Inherits from: Intentionally_act, Perception
済　む	finish	Activity_finish	Inherits from: Intentionally_act, Process_end; Subframe of Activity
思い出す	recall	Memory	Uses: Eventive_affecting［Inherits from: Event］
振　る	shake	Body_movement	Uses: Motion, Observable_bodyparts
聞こえる	hear	Perception_experience	Inherits from: Perception
落ちる	fall	Motion_directional	Inherits from: Motion
連れる	take	Bringing	Uses: Cause_motion, Motion
頷　く	nod	Body_movement	Uses: Motion, Observable_bodyparts

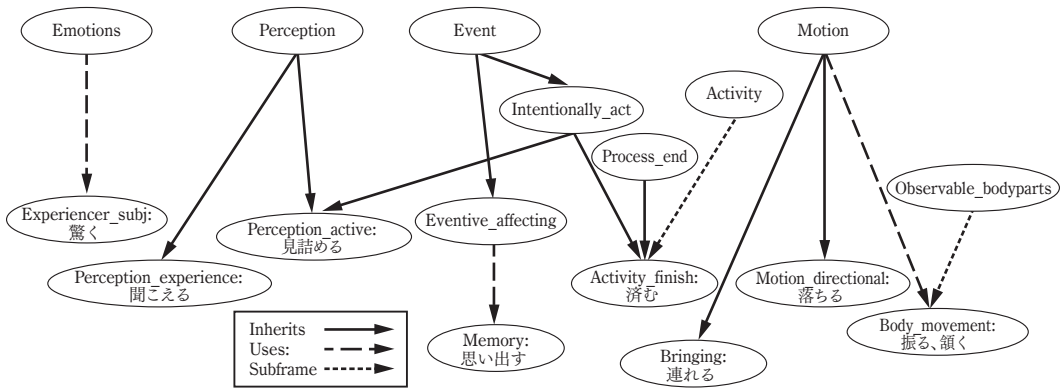

図４　「文学」の特徴語のフレームとフレーム間関係

４.５　フレームからみた「理科」ジャンルの特徴

次に教科書コーパスから［理科］のジャンルについて、フレームの観点から考察する。以下は［理科］の特徴語として抽出されたものである（再掲）。

［理科］ 加わる、覆う、冷える、押す、温める、分かれる、下がる、結び付く、運ぶ、混ぜ合わせる、冷やす、燃える、流す、繋ぐ、取り出す、溶ける、熱する

この中で、特に「冷える」、「温める」などの温度関係の語彙が特徴的に現れていることが見て取れる。これらの温度関係の語について、喚起するフレームと、その上位フレームを示したものが表６である。また、図５にはこれらのフレームのフレーム間関係を図示した。

その結果、教科書コーパスの［理科］のジャンルに関して、「変化」を表すフレーム（Transitive_action, Change_position_on_a_scale, Change_of_phase_scenario フレームなど）が特徴的であることがわかる。これは理科の教科書に温度や状態についての変化・変遷が特徴的に含まれているからだと考えられるが、直感に沿う結果であるといえる。

表６　「理科」の特徴語が喚起するフレーム

日本語	英　語	フレーム	上位フレーム
冷える	cool (vi)	Change_of_temperature	Inherits from: Change_position_on_a_scale
温める	warm (vt)	Cause_temperature_change	Inherits from: Transitive_action
			Is Causative of: Change_of_temperature
下がる	fall	Change_position_on_a_scale	Inherits from: Event
冷やす	cool (vt)	Cause_temperature_change	Inherits from: Transitive_action
			Is Causative of: Change_of_temperature
溶ける	melt	Change_of_phase	Subframe of: Change_of_phase_scenario
熱する	heat	Cause_temperature_change	Inherits from: Transitive_action
			Is Causative of: Change_of_temperature

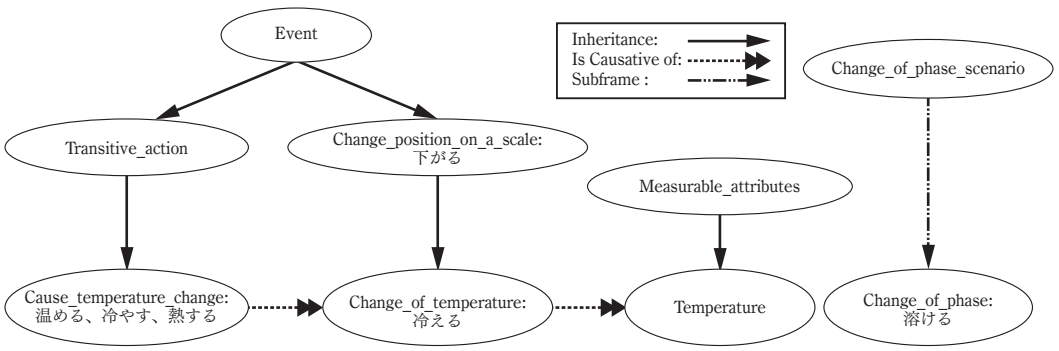

図５　「理科」の特徴語のフレームとフレーム間関係

4．6　その他のジャンルを特徴づけるフレーム

FrameNet は進行中のプロジェクトであるため、本稿で取り上げた単独出現語について全てのフレームを特定することはできない。従って部分的な分析となるが、以下、それぞれのジャンルを顕著に特徴づけると思われるフレームを、単独出現語のリストから抽出して提示する。

［6 産業］
成長（Expansion）、売買（Commerce_scenario）
［外国語］
コミュニケーション（Communication）
［社会］
繁栄（Thriving < State）

4．3 節で示した［6 産業］に関わるフレームである Expansion と Commerce_scenario は、企業の成長や衰退、商品の売買など、産業の文脈で頻出するフレームである。また、［外国語］に関しては、「聞き取る」、「聞き返す」、「話し掛ける」などの単語から、Communication フレームが喚起されるという結果もこのジャンルの特徴をよく表している。さらに［社会］では「栄える」から Thriving フレームが特徴的に現れていることがわかる。

5．まとめと今後の課題

本稿では BCCWJ の語彙頻度表からジャンル別に頻度上位100語を抜粋し、コサイン係数による類似度の分析を行った。クラスター分析の結果、(1) 国語の語彙分布は書籍コーパスに比較的近いが、概して教科書コーパスにおける各教科の語彙は書籍の各ジャンルとは異なった振舞いを示す（学習言語の特異性）ということが明らかになった。さらに、フレームの観点からの分析で、(2) ジャンルを特徴づけるフレームが存在する、ということが明らかになった。この結果は、フレームによるコーパス評価の有効性を示すものであり、コーパスのジャンル別の特徴を意味の観点からの具体的な分析を可能にするものである。

本稿では、ジャンルを特徴づける語彙として単独出現語のみに限ったが、全てのジャンルに登場する語、あるいは特定の複数の分野に共通して出現する語などの分析については今後の課題である。また、本研究では、名詞だけではなく、動詞にもジャンル別の特徴が見られることをフレームの観点から明らかにしたが、形容詞や副詞などのフレーム喚起語に関しても同様の分析を行うことは意義深いと考えられる。

注

1　東京大学大学院総合文化研究科言語情報科学専攻教授
2　Paltridge（1997）はフレームに基づいたジャンルの分析を提示しているが、質的な分析の手法の一部に会話の参与者やイベントに付随する意味役割などを分析の枠組みとして取り入れるに留まっており、具体的に喚起されたフレームへの言及はない。
3　本論文ではこの区分をジャンルと呼ぶ。
4　詳しくは前川（2008）、前川・山崎（2009）などを参照。

5 詳しくは丸山ほか（2006）参照。
6 http://www.ninjal.ac.jp/corpus_center/bccwj/freq-list.html では最終版の教科書コーパスの語彙リストが掲載されているが、書籍コーパスのリストは提供されていない。統一的な基準での分析を担保するため、本稿ではそれぞれのジャンルの語彙リストが提供されている『BCCWJ 領域内公開データ2008年度版』を用いることとした。データは BCCWJ の構築期のものであるが、サンプリングの大枠は終了した段階であり、また本研究の対象はデータの追加・削除があってもほとんど影響のないと考えられる高頻度語であるため、結論の一般性は確保できると考えられる。
7 「総記」の語彙頻度表は（何らかの理由で）哲学と同一の内容であったため、本稿の対象からは除外する。また、書籍コーパスで番号情報のないもの（n で表示されるもの）も対象から外す。
8 分析対象となるデータが構築期のものであることも考えると、最終版では頻度や順位に差がでる可能性があるが、上位100語の項目には大きな変化はないと考えられ、この点においても質的データに変換するメリットがある。
9 特徴語抽出におけるダイス係数・対数尤度比・相互情報量などの比較については、中條・内山（2004）に詳しい。
10 教科特徴語 _ 高 _ 外国語 .xls より抜粋（カッコ内は度数）。
11 多義語の場合、単語リストのみから意味を特定することは難しいが、本稿ではそれぞれの語彙を各分野の文脈にあてはめてもっとも適切と考えられる意味で代表させることとした。より精緻な分析については今後の課題である。
12 Tgt（target）はその単語がフレームを喚起しているということを表す。
13 フレーム要素には core, peripheral のほかに extra-thematic などのタイプがある。詳しくは、Ruppenhofer et al.（2006）および Fillmore et al.（2003）などを参照。
14 藤井（2005）は伝達を表すフレームについて日本語と英語の「フレームの構造」の相違点について指摘しているが、「喚起されるフレーム」については、汎用的な高頻度語に関しては日英語で共通性が高く、日本語の分析の際に英語版の FrameNet を用いても妥当な結果が得られると考えられる。この点に関しての詳細な考察は稿を改めることとする。

追記：本稿は「言語処理学会第15回年次大会」（鳥取大学）で発表した内容を基に加筆・修正したものである。また、「文部科学省科学研究費特定領域研究『日本語コーパス』プロジェクト」（研究代表者：前川喜久雄）の成果の一部である。

参 考 文 献

浅野長一郎・江島伸興（1996）.『基本多変量解析』日本規格協会.
Biber, D. & Conrad, S.（2009）. *Register, genre, and style*. Cambridge: Cambridge University Press.
Fillmore, C. J.（1982）. Frame semantics. In Yang, I.（ed.）, *Linguistics in the morning calm: Selected papers from SICOL-1981*. 111-137. Seoul: Hanshin.
Fillmore, C. J.（1985）. Frames and the semantics of understanding. *Quaderni di Semantica*, **6**（2）, 222-254.
Fillmore, C. J., Johnson, C. R. & Petruck, M. R. L.（2003）. Background to Framenet. *International Journal of Lexicography*, **16**（3）, 235-250.
藤井聖子（2005）.「日本語フレームネットにおける『伝達』領域での分析」『日本認知言語学会論文

集』**5**, 625-628.
石黒圭・阿保きみ枝・佐川祥予・中村紗弥子・劉洋 (2009).「接続表現のジャンル別出現頻度について」『一橋大学留学生センター紀要』, **12**, 73-85.
石川慎一郎 (2001).「テキスト・ジャンルと構成語彙」*Kobe English Language Teaching*, **16**, 3-17.
Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, **6** (1), 97-133.
近藤明日子 (2011).「『教科書特徴語リスト』について」<http://www.ninjal.ac.jp/corpus_center/bccwj/data- files/frequency-list/Domain-Specific-Vocabulary.zip>（アクセス日：2014年 5 月 7 日).
前川喜久雄 (2008).「KOTONOHA『現代日本語書き言葉均衡コーパス』の開発」『日本語の研究』, **4** (1), 82-95.
前川喜久雄・山崎誠 (2009).「現代日本語書き言葉均衡コーパス」『国文学：解釈と鑑賞』, **74** (1), 15-25.
丸山岳彦・柏野和佳子・山崎誠・前川喜久雄・稲益佐知子・秋元祐哉 (2006).「代表性を有する書き言葉コーパスのサンプリング手法について」『言語処理学会第12回発表論文集』, 150-153.
村田年・山崎誠 (2011).「『手』の慣用句を指標とした文章ジャンルの判別 ── 現代日本語書き言葉均衡コーパスを用いて ──」『日本語と日本語教育』, **39**, 75-88.
Paltridge, B. (1997). *Genre, frames and writing in research settings.* Amsterdam/ Philadelphia: John Benjamins.
Puschmann, C. (2010). *Thank you for thinking we could*: Use and function of interpersonal pronouns in corporate web logs. In Dorgeloh, H. & Wanner, A. (eds.) *Syntactic variation and genre.* 167-191. Berlin/New York: Walter de Gruyter.
Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R. & Scheffczyk, J. (2010). *FrameNet II: Extended theory and practice*. <https://framenet2.icsi. berkeley.edu/docs/r1.5/book.pdf>（アクセス日：2014年 5 月 7 日).
鄭惠先・小池真理・舩橋瑞貴 (2009).「『現代日本語書き言葉均衡コーパス』に見られる『〜てならない』『〜てたまらない』『〜てしかたない』『〜てしようがない』の使い分け：日本語学習者に対する指導への応用」『北海道大学留学生センター紀要』, **13**, 4-21.
徳永健伸 (1999).『情報検索と言語処理』東京大学出版会.
中條清美・内山将夫 (2004).「統計的指標を利用した特徴語抽出に関する研究」『関東甲信越英語教育学会研究紀要』, **18**, 99-108.
上田博人 (2013).『エクセルによる言語数量データ分析：基礎・応用・開発』<http://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo/4-numeros/numeros.pdf>（アクセス日：2014年 5 月 7 日).

言語資料出典

『BCCWJ 領域内公開データ2008年度版』国立国語研究所.

Genre Specific Characteristics of Lexicon from the Perspectives of Cluster Analysis and Frame Analysis:

A Case Study of BCCWJ

Satoru UCHIDA and Seiko FUJII

This study aims to clarify the genre specific characteristics of lexicon based on the frequency list of *Balanced Corpus of Contemporary Written Japanese* (BCCWJ). A quantitative approach was adopted to reveal the tendency of frequently-used words by examining co-occurrences of verbs and nouns among each genre of the corpus, which was converted into cosine coefficient to conduct a cluster analysis. The results indicate a clear difference between book subcorpora and textbook subcorpora, and at the same time, similarities within scientific genres and arts genres respectively, both for nouns and verbs. Then, a qualitative approach was employed with the genre specific verbs based on frame semantics and FrameNet to illustrate their semantic characteristics, which implied the existence of genre-specific frames. Using the frame-to-frame relations in FrameNet, it is clear that the genre of literature, for example, is closely associated with the frames of Emotions, Perception, and Body-movement indicating semantic characteristics of a particular genre, a point which previous studies have not successfully explained.

Key words: genre, cluster analysis, semantic frames, FrameNet, BCCWJ