「日本語コーパス」における語彙のジャンル別特徴
―クラスター分析とフレームの観点から―

内田　諭[i ii]　　藤井　聖子[i]

[i] 東京大学大学院総合文化研究科言語情報科学専攻

[ii] 日本学術振興会特別研究員

1. はじめに

本発表は、『現代日本語書き言葉均衡コーパス（構築中：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明らかにすることを目的とする。具体的には、書籍コーパスおよび教科書コーパスから、各々日本十進分類法(NDC)・8 教科の区分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。分析手法として共起関係から算出した修正ファイ係数を用い、クラスター分析によって類似度測定を行う。また、各ジャンルにおける特徴語を抽出し、フレームの観点から分析を加える。以下、2 節で BCCWJ の概要を示し、3 節で各コーパスの頻度上位の動詞および名詞についてのクラスター分析を提示する。その結果を受けて、4 節で各ジャンルの特徴語を抜粋し、特に動詞に関してフレームに基づいた分析を行う。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、国立国語研究所が構築している日本語の大規模コーパスである[1]。2006 年に開発が始まり、2011 年までに 5000 万語（最終的には 1 億語）規模の書き言葉コーパスを構築することを目標としている。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・雑誌に加え、書籍全般や教科書、白書などを含み、ランダムサンプリングによって均衡化が図られている[2]。本発表では『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。

書籍コーパス（22,967,145）[3]は、日本十進分類法(NDC)により、総記（521,436）、哲学（1,403,199）、歴史（2,141,841）、社会科学（5,447,856）、自然科学（1,074,332）、技術・工学（1,115,821）、産業（700,269）、芸術・美術（1,107,179）、言語（398,497）、文学（8,775,301）、に分類されたサブコーパスがある。教科書コーパス（466,429）は、国語（104,545）、数学（50,591）、理科（66,984）、社会（106,468）、外国語（12,767）、技術家庭（64,627）、芸術（36,170）、保健体育（24,277）の 8 科目に分類される。

本稿では上記の 17 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[4]。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. 語彙のクラスター分析

3.1. 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、各ジャンル上位の 100 語を抜き出し、このリストに含まれるか否かによって質的データに変換した。上位 100 語には、動詞は平

[1] 詳しくは前川(2008)、前川・山崎(2008)などを参照。

[2] 詳しくは丸山 et al.(2006)参照。

[3] 括弧内はコーパスののべ語数を示す。

[4] ただし、「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、番号なしのものが 281,414 語含まれる。

均 78.2%、名詞は平均 32.4%の語彙トークンがカバーされたことになり、それぞれのジャンルの大まかな特徴を見るには十分であると考えられる。また、この操作により低頻度の特異な語彙を排除することができる。質的データに変換した理由は、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ためである。

この上位 100 語のリストに対して、一つのテーブルに集計し、それぞれのジャンル間における語彙の共起回数を計算した。次に共起回数の表を基に修正ファイ係数を用いて相関を計算し、クラスター分析を行った。修正ファイ係数とは、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]、b=[1,0]、c=[0,1]、d=[0,0]とすると次式によって求められる[5]。

$$\phi \text{ rev} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

この数値を用いることで、複数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（上田 2008a）を使用した。また、クラスター分析には R を用い、ユークリッド距離を基準に、分類度の高い最遠隣法（最長距離法）を用いて描画した（cf. 徳永 1999, 浅野・江島 1996）。

3.2. 結果

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。それぞれ図 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。

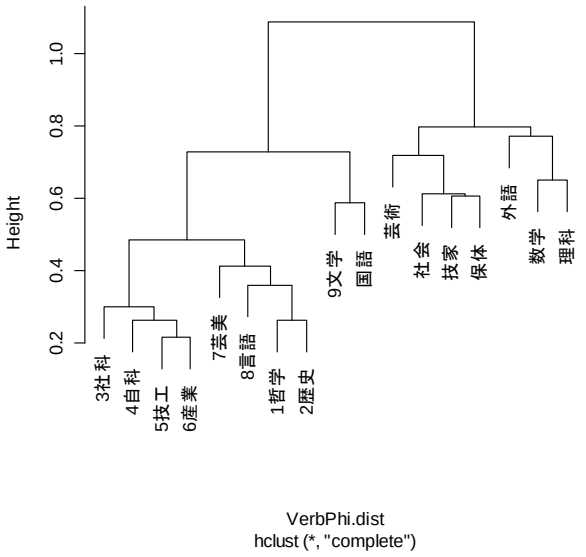

図 1　動詞のクラスター

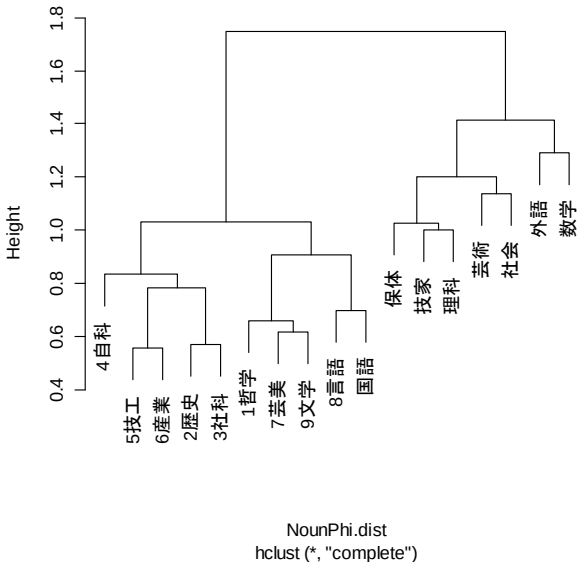

図 2　名詞のクラスター

3.3. ジャンル間の類似度

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。このことは教科書における学習言語の特異性を示しているといえる。ただし、

[5] ファイ係数 $\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$ において、d→∞ とし、分子・分母をそれぞれ d で割ることでこの式になる。詳しくは、上田(2008b)。

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。さらに、注目すべき点は、動詞のクラスターが「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、専門用語が多く近い分野が類似した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

本節では、本稿の比較対象である 17 のジャンルの上位各 100 語のリストからそれぞれのジャンルに独自に出現する語（以下、特徴語とする）を抽出し、それらをフレームの観点から分析を行う。

4.1. ジャンルごとの特徴語

表 1 は書籍コーパスの産業・文学および教科書コーパスの社会・理科の各ジャンルの動詞の特徴語を一部抜粋したものである。

表 1　ジャンルごとの特徴語

6 産業	9 文学	社会	理科
伸びる 売る	驚く 思い出す 聞こえる	築く 栄える	冷える 温める 熱する 燃える

これらを仔細に観察すると、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[6]。前節で動詞がジャンルごとに特徴的な分布を示すということをみたが、本稿の主張は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である。

4.2. フレームと FrameNet

フレームとは、端的に言うと、言葉を理解するための背景知識である(Fillmore 1982)。このフレームは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 という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[7]。例えば、次の例を考えてみよう。

(1)[<*cook*>Matilde] friedTgt [8] [<*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便宜的にフレーム喚起語(frame evoker)と呼ぶ。

フレームはその喚起された状況に付随する要素をフレーム要素(frame element)として持っている。Apply_heat フレームの場合、*cook, food, heating_insrument* というフレーム要素を含んでおり、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[9]。

日本語のフレームネットは現在目下構築中であるが、本稿では、日本語から対応する英語のフレームを探しそれをもとに分析することとする。

4.3. 特徴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

[6] 多義性の問題があるが、本稿は概要の提示を目的としているため、そのジャンルで典型的であると思われる文脈に当てはめて考えることとする。

[7] http://framenet.icsi.berkeley.edu/

[8] “Tgt”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。

[9] フレーム要素には core, non-core, 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。

徴語である「伸びる」から考えてみよう。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、もう一つの特徴語「売る(sell)」が喚起するフレームは Commerce_sell である。

このように考えると、各ジャンルの特徴語からそのジャンルを特徴付けるフレームを同定することができる。表 1 の他のジャンルについて考えてみると、まず、「文学」のジャンルでは「驚く(surprise)」が感情を表す語を多く含む Experiencer_obj フレームを、「聞こえる(hear)」は、知覚を表す語を多く含む Perception_experience フレームを喚起すると分析できる。また、「社会」では、「栄える(prosper)」が Thriving フレームを喚起する。さらに､「理科」では「冷える(cool) 」､「温める(warm)」、「熱する(heat)」が Cause_temperature_change フレームを喚起する。

以上のことから､「産業」では成長や商取引を表すフレームが､「文学」では感情や知覚を表すフレームが､「社会」では興隆を示すフレームが、そして「理科」では温度変化を表すフレームが特徴的に観察されるということがわかる。なお、名詞もフレーム喚起語であるが、本稿では詳しい分析は割愛する。

5. まとめ

本稿では「日本語コーパス」の語彙頻度表からジャンル別に上位 100 語を抜粋し、修正ファイ係数による分析を行った。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喚起語が存在する、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分析を可能にするものである。

今後の課題として、名詞や形容詞などのフレーム喚起語に関しても同様の分析を行う必要があると考えている。

＜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.

＜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.

Fillmore, C. J. (1982) ‘Frame semantics.’ In Yang, I. (eds.), *Linguistics in the Morning Calm: Selected Papers from SICOL-1981*. Seoul: Hanshin. 111-137.

Fillmore, C. J., C. R. Johnson, and M. R. L. Petruck (2003) “Background to Framenet” *International Journal of Lexicography* 16. 235-250.

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(2008)「『現代日本語書き言葉均衡コーパス』」『国文学解釈と鑑賞』932(74 巻 1 号). 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』

Ruppenhofer, J., M. Ellsworth, M. R. L. Petruck, C. R. Johnson and J. Scheffczyk (2006) ‘FrameNet Ⅱ: Extended Theory and Practice.’ http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=126

徳永健伸(1999)『情報検索と言語処理』 (言語と計算 5)東京大学出版会.

上田博人(2008a) 「言語現象の相関関係を観察する」『言語情報分析（応用）』 http://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo/2ouyou/b10soukan.doc

上田博人(2008b) 「Excel VBA による言語統計分析: VBA プログラム集」『言語情報分析（応用）』 http://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo/5numeros/numeros.xls