文法形式に基づいた日本語文体の多次元的抽出

中俣尚己（大阪大学）
n.nakamata.ciee@osaka-u.ac.jp

1. 研究の背景

1.1 文体研究と「語の文体」

文体研究とはどのような研究であろうか。CiNii で「文体論」を検索すると 1,500 件以上ヒットする一方、「計量文体論」では 20 件にとどまる。このうち、最もまとまったものは陳(2012)であり、雑誌や新聞を対象に出版社ごとの違いを分析したものである。馬場(2022b)は計量文体論について、「単語や文の長さ、品詞の使用率、単語の使用率、文字種・句読点などの様々な特徴に基づいて、著者推定、執筆時期推定、文学作品・作者の類型、同一作者の文体変化、文章ジャンルの特徴などに関する分析」を行うものとしている。基本的には「文章」を対象にした研究であり、2024 年時点では、英語を対象にしたものが多い。

一方で、日本語の文体研究では、馬場(2022b)が指摘するように文章よりも「語の文体」(宮島 1977)に注目したものが多い(宮島 1977, 井上 2009, 佐野 2016, 馬場 2018, 中俣 2020)。これは、日本語では論説文で使ってもよい語・いけない語の区別がはっきりしている（井上 2009）というような特性を持つためであると思われる。「語の文体」の研究としては柏野(2015)のデータに基づいて、馬場(2022a)が 13 万語の文体値データを公開しており、これを一つの到達点と呼ぶこともできるだろう。

さらには浅原(2015)、馬場(2022b)のように「語の文体」を定めてから、それぞれの語の多寡でジャンルの文体を評価するような研究も見られる。

1.2 「語の文体」研究の限界

しかし、「語の文体」の研究に全く問題がないわけではない。馬場(2022a)を例に説明すると、これは柏野(2015)が BCCWJ の図書館サブコーパスの文章に作業者の主観で付した「専門度」「客観度」「硬度」「くだけ度」「語りかけ性度」という５つの評価軸に依拠している。まず、この５つの評価軸が文体の違いを捉えるのに適切かという妥当性の問題がある。また、馬場(2022b)によれば、「専門度」「客観度」「硬度」「くだけ度」はそれぞれ中程度から強い相関があることがわかっており、結局のところ一次元的な尺度となっていることが問題である。他の研究も同様で、井上(2009)は白書で使われるか否かという軸、佐野(2016)は語彙密度という軸を使っており、いずれも一次元的である。このような一次元的な尺度は、ある語を論説文で使ってよいか、というような実用的な課題を解決するのには向く。しかしながら、多様な書き言葉・話し言葉の全容を記述するには足りない。

なお、石黒(2015)は理論的に「情報の伝達」と「感情の伝達」に問題を切り分け、「かたさ／やわらかさ」「あらたまり／くだけ」という二次元で文体を捉えることを提案している。

1.3 Biber の多次元的文体研究

英語の文体研究に目を向けると Biber(1988)は多次元的文体研究を提唱し、その後の研究のスタンダードとなった。これはラジオ放送、電話会話から SF 小説、論文まで多様なジャンルのテキストを集め、そこから単語の頻度や各種の指標などの客観情報のみを抽出する。その後、因子分析により次元（評価軸）を取り出し、それによってジャンルの特徴を分析していく。Biber(1988)は 1.Involved vs Information Production, 2.Narrative vs Non-narrative, 3.Explicit vs Situation Depended, 4.Overt Expression of Persuasion, 5. Abstract vs Non-Abstract, 6.On-Line Information Elaboration という 6 つの次元(Biber et. al. 1999 では 5 次元)を提唱しており、この概念はその後も広く使われている。この評価軸は主観ではなく客観情報のみから抽出されており、因子間の相関も抑えられている。

だが、英語では 40 年近く前から存在するこの手法を日本語の書き言葉・話し言葉に対して実施した研究は管見では見当たらない。日本語でもより客観性の高い文体の評価軸を抽出するため、本研究を実施した。

2. 多次元的分析の方法

2.1 指標の選定

Biber(1988)の多次元的文体研究の特徴は 1. Past tense, 2.Pefect Aspect など、67 種類の Linguistic Feature によって分析を行った点にある。これらの指標は英語の文体研究ではその後も引き続き使われることが多いが、それをそのまま日本語に応用することは不可能である。「た」を例に挙げても、これを過去テンスと完了アスペクトに分類することはできない。また、指標の中には 42. Adverbs のようなものもあるが、中俣(2020)が示すように副詞の中にも大きな文体差があり、これをまとめてしまえば見落としが生じる可能性もある。さらに、アスペクト形式などは非縮約形「ている」と縮約形「てる」にも文体差があると考えるべきであろう。

そこで、方針として、表層形に注目し、「構造格」から３形式、「意味格」から３形式、のようにある程度細分化した文法カテゴリーから最大で３形式ずつ機能語を選び、100 語あたりの調整頻度（テキスト内の出現頻度÷テキスト内の出現語数×100）を指標として利用することにした。３形式は BCCWJ 語彙表において頻度が多い順に選定した。Biber(1988)には平均語長（単語が何文字か）や、語彙多様性(Type/Token Ratio)といった個別の語に依存しない指標もわずかに存在するが、今回は文法形式と文体の関係を見るために、文法形式の頻度に絞る。選んだ指標は次ページの表 1 の通りであり、合計 89 形式である。

2.2 サンプルテキストの選定

使用したコーパスは『現代日本語書き言葉コーパス（BCCWJ)』『日本語話し言葉コーパス(CSJ)』『日本語日常会話コーパス(CEJC)』の３種類である。これらのコーパスから次ページの表 2 の通り書き言葉 5 ジャンル、話し言葉 5 ジャンルの 10 ジャンルを選定した。

表1 利用した文法的指標(89 形式)

カテゴリ	形式	カテゴリ	形式	カテゴリ	形式
構造格	が、を、に	意味格	Nと、Nで、Nから	疑問詞	いつ、何、誰
連体助詞	NのN	提題助詞	Nは、Nも、Nって	副助詞	だけ、など、ほど
テンス	た	否定	Vない、Vん、ません	動詞丁寧形	ます
人称代名詞	あなた、わたし、僕	アスペクト	ている、てしまう、ていく	アスペクト(縮約)	てる、ちゃう、てく
指示連体詞	そのN、このN、あのN	指示代名詞	それ、あれ、これ	条件節	と、ば、たら
ヴォイス	れるられる、せるさせる、可能動詞	複合格助詞	Nとして、について、にとって	C類節	けど、けれど、し、から
時間副詞	すぐ、まだ、もう	程度副詞	ちょっと、まったく、少し	連体節	っていうN、というN
コピュラ	だ、である、です(終止形)	接続詞	しかし、そして、また	接続助詞	て(補助動詞除く)
恩恵	てあげる、てくれる、てもらう	恩恵(敬語)	ていただく、てくださる	同時節	たまま、つつ、ながら
準体表現	のだ、のである、のです(終止形)	準体表現(縮約)	んだ、んです(終止形)	終助詞	よ、ね、か(文末)
証拠性判断	そう様態、よう様態、みたい	評価性判断	なければならない、ばいい、べき	推量性判断	かも、だろう、と思う

表2 ジャンルとサンプル数(N=880)

ジャンル	サンプル[1]	平均語数*	ジャンル	サンプル	平均語数
白書(BCCWJ, コア)	62	664	学会講演(CSJ, コア)	70	3,094
雑誌(BCCWJ, コア)	86	600	模擬講演(CSJ, コア)	107	2,104
自然科学(BCCWJ 出版・書籍[2]より抽出)	100	620	用談(CEJC)	96	4,689
文学(BCCWJ 図書館[2]より抽出)	100	638	会議・会合(CEJC)	59	4,397
新聞(BCCWJ, コア)	100	632	雑談(CEJC)	100	4,365

そのうえで、各ジャンルから最大で100サンプルのテキストをランダムに選定した。そして、「中納言」を用い、選ばれた各サンプルにおける指標の出現頻度を調査した。各ジャンルのサンプル数と語数は表2の通りで合計880の文章サンプルを利用した。

なお、BCCWJ の平均語数が CSJ や CEJC と比べて少ないのは、調整頻度を正確に計算するため、固定長(1,000 文字)サンプルのみを利用したためである。実際には検索結果をダウンロード可能な 10 万語以内に押さえる必要もあった。また、山崎(編)(2014)によれば出版・書籍コーパスは専門書が多く含まれ、図書館コーパスは相対的にその割合が小さい。「自然科学」と「文学」の違いを明確にするため、前者はより専門的な出版・書籍コーパスからサンプリングし、後者は図書館コーパスからサンプリングしている。

2.3 因子分析の方法

880 サンプル×89 形式のデータセットに対して、R を用い、4 因子モデルで因子分析を行った。因子の抽出は最尤法、回転はプロマックス回転を用いた。重要なのは分析においてサンプルのジャンルのデータは用いていないという点である。因子分析では各テキストの頻度データのみを基に因子＝文体を特徴づける軸を決定する。その後、因子と形式の関係を因子負荷量から、因子とジャンルの関係を因子得点から分析するのである。

3. 因子分析の結果

抽出された 4 因子の寄与を表 3 に、因子間の相関を表 4 に示す。プロマックス回転の結果、因子間の相関はほとんど見られなかった。D1 と D2 には弱い負の相関がみられ、散布図を作成すると共に正の値で高いというサンプルは見られないことがわかった。

表 3 抽出された4因子

	D1(ML1)	D2(ML3)	D3(ML2)	D4(ML4)
因子寄与	14.577	4.812	4.230	2.581
説明力	0.556	0.184	0.161	0.099
累積説明力	0.556	0.740	0.901	1.000

表4 因子間の相関

	D1	D2	D3	D4
D1	1.000			
D2	-0.345	1.000		
D3	-0.125	-0.281	1.000	
D4	-0.180	-0.153	0.038	1.000

4. 考察

4.1 評価軸の分析

ここからは抽出された評価軸を分析する。まず、それぞれの評価軸に影響を与えている文法形式と、それぞれの評価軸によって特徴づけられるジャンルを表で示す。その後、評価軸によって特徴づけられるサンプルの例文を示し、評価軸を命名、分析する。

4.1.1 D1:共感的対話 VS 客観的伝達

+の形式	−の形式	+のジャンル	−のジャンル
終助詞、だ、から（接続）、んだ、縮約形、指示代名詞、推量性判断、疑問詞、たら、けど、ない、ちょっと、もう、みたい、ばいい、まだ、し、あのN、Nって、これ、だけ、そう（様態）	格助詞、NのN、受身、ている、など、て、について、という、よう(様態)	雑談(1.47) 用談(1.39) 会議(0.96)	白書(-1.22) 新聞(-0.78) 科学(-0.78) 学会(-0.73)

(1)　うち仲いんだよな早く出てかないとな。そんなうらやましい。まじでいいことよ。うん。そうかもね。だってあたしこないだあたし唐突にあたし社会人なったらうち出るからってばさってちょっとゆってみたの。　（雑談、T009_005b_030、D1=2.32）

(2)　また、これにより得られたデータの分析・評価により、森林の整備の状況等に関する評価手法等の検討を行う。（白書、OW6X_00071_8810、D1=-1.51）

D1 は説明力の強い対立軸である。要素が多いため命名は難しいが終助詞や縮約形を使って相手の共感を呼び起こしながら話す文体と、格助詞を使って論理関係を明確にし、受身形などで主観を廃して伝える文体の対立である。「共感的対話 VS 客観的伝達」と命名する。単純な音声対文字の対立でないことは、学会発表が客観的伝達の性質が強いことからわかる。Biber(1988)の D1:involvement vs information processing に相当する。

4.1.2 D2:語り VS 非語り

+の形式	−の形式	+のジャンル	−のジャンル
Nは、た、のだ、Vない、ている、構造格、て、Nも、しかし、ば、使役、てしまう、ほど、ながら、だろう	です、Nって、んです、ね、けど、っていうN	文学(1.59) 雑誌(0.56) 自然(0.41)	会議(-0.77)用談(-0.68)雑談(-0.68)白書(-0.67)

(3)　私が知っている女性はかなりの減量に成功した。彼女はこう話している。「前より体調がよくなり、元気も出てきました。（自然、PB14_00100_15010、D2=3.22）

(4)　コリンは、大きな深呼吸をひとつしてから、ドアのベルを鳴らした。もったいをつけてでもいるように、ファーガソン医師はなかなか出てこない。コリンは緊張をほぐそうとした。ここが正念場だ。ぶち破るべき壁なんだ。（白書、OW6X_00071_8810、D2=3.20）

(5) 今回はあの私の住んでいる町っていうタイトルで、別名マイタウン青葉。続けていきます。でえーとね、位置としては神奈川県の横浜市なんですけど、電車のアクセスはまず東急の田園都市線と後はあざみ野っていう駅から市営地下鉄が出ていてあのニュータウンの方に行けるんですね。（模擬、S03F0062_170、D2=-1.72）

D2 は正方向には提題助詞と「た」が含まれ、さらに出来事の連鎖を表す「て」、同時の「ながら」、出来事に対する捉え方を表す「のだ」、転換を表す「しかし」、評価的意味を含む「てしまう」「ほど」が含まれる。これは Biber(1988)の D2:Narrative vs Non-Narrative に対応する「語り VS 非語り」の評価軸と命名できる。南(2009)は「語り」の 6 つの特徴をまとめているが、そこには「出来事（なにが起きたか)」に加え、「設定（いつ、どこで、何を)「評価」が含まれるのである。なお、(3)と(4)は翻訳文であった。

4.1.3 D3:新情報の解説

+の形式	−の形式	+のジャンル	−のジャンル
んです、可能、けれど、っていうN、そのN、Nって、と思う、て、てしまう	N.A.	模擬講演(1.88) 学会(0.41)	白書(-1.06) 新聞(-0.78)

(6)　真夏ことなんですがえーと夜遅く庭先であの洗濯物干してましたらあの寝ていた娘きゃーって悲鳴を上げたんですねそして何事かと思いましたらあーの太ってる猫がぽんと娘の寝てる上に乗っちゃったらしいんですそれがあのーその猫が来るう来てるんだってきっかけみたいだったんですね（模擬講演、S01F1522、D3=3.35）

D3は形式を見ると、Nという情報について「んです」で解説を行っているというように読み取れる。実際にはこれは模擬講演を特徴づける文体で、上位70サンプルは全て模擬講演である。模擬講演には(6)のように「〜んです」で話を展開する一群のサンプルが存在する。過去の話が多く、一種の語りと位置づけられるが、格助詞は時折省略され、引用なども少ない。「ちゃう」より「てしまう」、「けど」より「けれど」のように完全にくだけているわけではないのは、模擬講演というジャンルの特徴であろう。「新情報の解説」と命名した。

4.1.4 D4:聞き手への配慮

+の形式	−の形式	+のジャンル	−のジャンル
ます、ません、というN、です、ていただく、この、のです、よう（様態）、これ	N.A.	学会(1.45)会議(0.39) 模擬講演(0.43)	白書(-0.87) 新聞(-0.71)

(7)　あたし何度女子力学んだかみたいなって思う。はい学ばせていただきましたって思って。ほんとにほんとなんか人によりますけど（会議、K001_006_87760、D4=4.48）

(8)　ちょっと問題があります。問題といいますか、何が問題かと言いますと、日本語のあってのは皆さん御存じのようにあの前後の対立がありません。日本語は前寄りのあであっても後ろ寄りのあであっても一応あなんですね。（学会、A01M0074、D4=2.56）

D4は「です」「ます」という丁寧語の使用によって特徴づけられる文体であり、「ていただく」や「よう」も多い。また、聞き手を意識するからこそ、近称詞も出現するのだろう。ジャンルでは学会の得点が高いがD3と異なり、様々なジャンルに出現する。「聞き手への配慮」と命名した。

4.2 他の研究結果との比較

今回の分析結果は他の文体研究の結果とも整合するものである。中俣(2020)は164の副詞を使って、BCCWJ内のレジスターに対し主成分分析を行ったものであり、その結果第1軸として「双方向的VS一方向的」という軸が抽出されているが、名称は異なるものの、これはD1「共感的対話VS客観的伝達」と同じと見なしてよい。中俣(2020)ではブログやYahoo!知恵袋を共感的対話と見なしており、これは共感的な書き言葉と言える。学会講演が客観的な話し言葉であることと対照的である。格助詞の省略などは書き言葉か話し言葉かという媒体の違いではなく、D1の文体差を反映したものである。

また、中俣(2020)の第２軸は「自己表現的 VS 公共的」と命名されているが、ジャンルで言えば、文学 VS 広報誌・国会会議録・白書の対立である。中俣(2020)は副詞のみに絞っていたため、「語り」という特徴を見いだせなかったが、PC2 が高い副詞とは「何ひとつ」「どんなに」「むろん」「じつに」といった書き手が主観的に評価する副詞であり、これは D2「語り」の特徴の一つとみなせる。

また、Biber(1988)と比較すると、D1 と D2 が形式は異なるものの対応を見せている。一方、Biber の D3 は名詞句や関係節が多用される"Explicit"と副詞が多用される"Situation-Dependent Reference"の対立である。本研究の D3 は模擬講演に強く特徴づけられた新情報の解説であった。中俣(2020)の第３軸は雑誌に特徴づけられ、評価的な副詞が多く現われていた。日本語と英語では丁寧語の有無など文体の表現方法に大きな違いがあり、完全に一致するはずがない。にもかかわらず、言語や手法を問わずに抽出された D1「共感的対話 VS 客観的伝達」と D2「語り VS 非語り」はかなり頑健な文体の評価軸として認めてよいのではないかと考えられる。いわばこの 2 軸が、日本語の（あるいは言語の？）文体の大分類として存在し、その下にジャンルごとの中分類が、さらには個人ごとの小分類が存在するという構造が提案できる（大江ほか 2020)。

4.3 日本語文体の全体像

D1 と D2 を用いて今回分析した 10 ジャンルをプロットしたものが図 1 である。

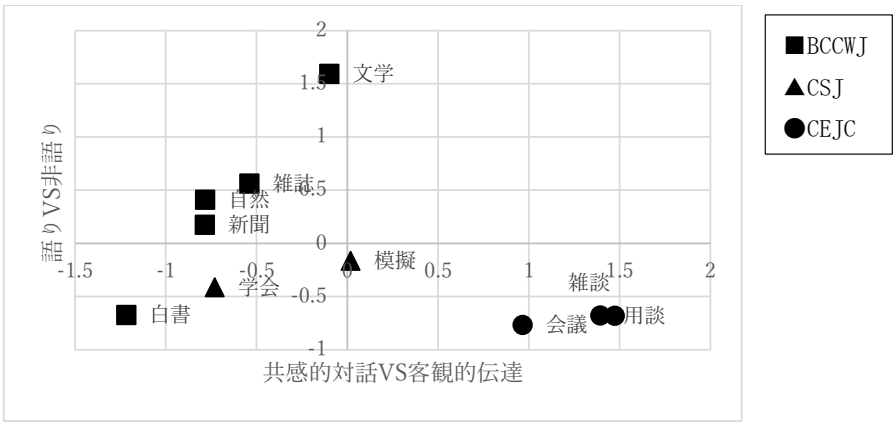

図1 D1 と D2 によるジャンルの分類

まず、図 1 の右上が空白であり、共感的な語りという文体は存在していない。これが D1 と D2 に-0.345 という弱い負の相関が見られた所以である。他研究も同様の結果である。

概ね右側に CSJ,左側に BCCWJ が配置され、その中間に CSJ が配置されているが、実際には模擬講演、学会講演はそれぞれ D3,D4 によって別次元で特徴づけられていることを忘れてはならない。この点をふまえると、話し言葉は書き言葉よりも文体的に多様ということができる。あるいは独話と対話は文体的には非常に異なると言えるかもしれない。本研究の D2 は基本的には小説の地の文に対応し、話し言葉における語りが D3 に対応していると見ることもできる。

4.4 ジャンル内変異

最後に、ジャンル内の変異について触れると、白書は全ての評価軸で標準偏差が最も小さく、文体的に最も安定している。D2 の「語り」は自然科学で標準偏差が大きかった。D3「新情報の解説」は学会講演と模擬講演で標準偏差が大きかった。D4 の「聞き手配慮」は新聞、白書、雑談、学会講演以外では変異が大きかった。

5. 本研究のまとめと貢献

本研究では日本語を対象として Biber 流の多次元的文体分析を行った。４因子モデルを作ることができ、特に D1「共感的対話 VS 客観的伝達」と D2「語り VS 非語り」は他の研究とも通底する頑健な評価軸として抽出可能と言える。これにより、日本語の文体研究における評価軸を主観的かつ一次元的なものから、客観的かつ多次元的なものに更新したい。

参考文献

浅原正幸(2020)「Bayseian Linear Mixed Model による単語親密度推定と位相情報付与」『自然言語処理』27(1):133-150.

石黒圭(2015)「書き言葉・話し言葉」と「硬さ／軟らかさ」『日本語学』34-1:14-25.

井上次男(2009)「論説文における語の文体の適切性について」『日本語教育』141:57-67.

大江元貴・居關友里子・鈴木彩香(2020)「日本語の左方転移構文はいつ、どのように使われるか？」『社会言語科学』23(1):226-241.

柏野和佳子(2015)　『BCCWJ 図書館サブコーパスの文体情報』(2015 年公開第 1 版) https://doi.org/10.15084/00003109

佐野大樹(2016)「語彙密度から見た語彙シラバス」森篤嗣(編)『ニーズを踏まえた語彙シラバス』79-93, くろしお出版

陳志文 (2012)『現代日本語の計量文体論』くろしお出版.

中俣尚己(2020)「主成分分析を用いた副詞の文体分析」『計量国語学』32(7): 419-435.

馬場俊臣(2018)「接続詞の文体差の計量的分析の試み―『BCCWJ 図書館サブコーパスの文体情報』を用いて―」『北海道教育大学紀要　人文科学・社会科学編』69(1): 1-14.

馬場俊臣(2022a)　『語の文体値データ』(2022 年 2 月公開第 1 版) https://doi.org/10.15084/00003532

馬場俊臣(2022b)「「語の文体」と「文章の文体」―「語の文体値データ」を利用した「文章の文体」の確定―」『計量国語学』33(7):435-450

南雅彦(2009)『言語と文化』くろしお出版.

宮島達夫(1977)「単語の文体的特徴」松村明教授還暦記念会(編)『国語学と国語史』871-903, 明治書院.

山崎誠(編)(2014)『書き言葉コーパス―設計と構築―』朝倉書店.

Biber, D.(1988)*Variation across speech and writing*, Cambridge Univ. Press

Biber, D., Conrad, S. & Reppen, R.(1998) *Corpus Linguistics,* Cambridge Univ. Press