

話題別会話コーパスに対する文体分析

ー地域差・性差・話題差ー

中俣尚己(大阪大学)

1. 研究目的

筆者は話題が固定された5分間の会話を集めたコーパス、『日本語話題別会話コーパス:J-TOCC』を構築、公開した(中俣ほか2021; 中俣(編)2023)。このコーパスに収められている会話の話題は表1の通り全部で15種類であり、これは「身の回りの話題」と「社会に関わることもある話題」に二分されている。しかしながら、この区分は会話録音前に日本語教育のプログラムなどを念頭に決定したものであり、実際の話し方(以下、話し方も含めて「文体」(style)と呼ぶ)がこの通りに分けられているかという保証はない。また、J-TOCCにおける形式と話題の関係については、「食べる事」「漫画・ゲーム」「アルバイト」では助詞なしが多く、「住環境」「家事」「日本の未来」では助詞なしが少ない(清水2023)、「まあ」は東日本・男性・「日本の未来」に多い(加藤2023)、「説明を逡巡する自問発話が「日本の未来」に多い(小西2023)」、「間投助詞「さ」は「旅行」に少なく「日本の未来」に多い(中俣2023)」など個別的に研究されているが、これらの現象は個別的ではなく連続しており、あり種の文体を形成していると考えられる。

表1『日本語話題別会話コーパス:J-TOCC』内の話題

身の回りの話題	社会に関わることもある話題
01. 食べる事, 02. ファッション, 03. 旅行, 04. スポーツ, 05. マンガ・ゲーム, 06. 家事, 07. 学校, 08. スマートフォン 09. アルバイト, 10. 動物, 11. 天気	12. 夢・将来設計, 13. マナー, 14. 住環境 15. 日本の未来

本研究の最終目的はJ-TOCC内の話題が文体の観点からどのように分けられるかを明らかにすることであるが、そのためには「かたい話し方」「やわらかい話し方」といった文体の評価軸を定めなければならない。そこで本研究ではコーパス内の語数という客観情報のみからまず文体の評価軸を抽出し、各軸の数値の多寡から話題の分類を試みる。なお、本分析ではテキストの語数のみを用い、事前に性別・地域・話題などの情報を与えない。そのため、性別や地域に対応する評価軸が「テキスト群の特性として」現れてくることも考えられ、それも考察の対象とする。

2. 計量文体論の先行研究

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。このように文体とは本来はテキストが持つ特性であるのだが、これまでの日本語の計量文体論は文章よりも「語の文体」(宮島1977)に注目したものが多かったと馬場(2022b)は指摘する。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、一つの到達点と言えよう。

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。現実の文章や談話は「硬い」「軟らかい」のような一次元的な尺度ではとらえきれない。

一方で、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し、さらに因子分析によって次元(評価軸)を取り出し、それによってジャンルの特徴を分析するという手法である。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。日本語では幅広いジャンルに対して同様の分析を行った研究は中俣(2024)までなかったが、この方法は日本語にも有効であると考えられるため、本研究でも因子分析による評価軸の抽出を行った。

3. 方法

本研究ではまず指標として文法形式の出現頻度に注目する。Biber(1988)ではアスペクト形式のようにまとめて集計を行っているが、日本語の場合「ている」と「てる」には文体差があるとみるべきである。そこで、形式の表層形に注目する。「構造格」として「が」「を」「に」、「意味格」として「と」「で」「から」のように細分化した文法カテゴリから代表的な形式を最大3つまで、J-TOCC全体での出現頻度が高い順に選んだ。形式の数は25カテゴリ73形式である。

まず、J-TOCC全文に対してcomainuを用いて形態素解析を行ったファイルを作成し、自作の検索プログラムで頻度を計算した。J-TOCCには15話題120ペア、合計1,800の会話が存在するが、これをさらに話者ごとに分け、3,600のファイルに対して、頻度をファイルの語数で割った調整頻度を算出した。

そして、73形式×3,600ファイルのデータに対しRを用いて5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性別、地域差、話題のデータは用いておらず、ただ調整頻度のみで因子を抽出しているということである。

表2 指標として利用した73形式

カテゴリ	形態	カテゴリ	形態	カテゴリ	形態
構造格助詞	が, を, に	人称	俺, 私, お前	アスペクト縮約	てる, とる, ちゃう, てく
意味格助詞	で, から, より	疑問詞	どこ, どっち, 誰	アスペクト非縮約	ている, てしまう, ていく
並列助詞	NとN, NとかN	指示連体詞	そのN, このN, あのN	証拠性判断	そう様態, よう, みたい
連体助詞	NのN	固有名詞	一人称, 二人称, 三人称	推量性判断	だろう, やろう, でしょう, 思う, かもしれない系
提題助詞	Nは, Nも, Nって	ヴォイス	れるられる, せるさせる, 可能動詞	条件節	と, ば, たら
副助詞	だけ, くらい, しか	否定	ない, ず, へん	C類節	けど, し, から
指示代名詞	それ, あれ, これ	テンス	た(終止形のみ)	終助詞	な, よ, ね
陳述副詞	まあ, やっぱり, たぶん	程度副詞	ちょっと, 全然, 結構	時間副詞	もう, まだ, ずっと
接続助詞	て(補助動詞除く)	接続詞	で, じゃあ		

4. 抽出された因子およびその考察

表 3 に抽出された 5 因子をまとめる．説明力の強い順に D1 から D5 の順に並べ替えている．

表 3 因子分析の結果

	名称	＋の形式	－の形式	説明力
D1	東日本 VS 西日本	ね (.71), ない (.59), よ (.58), だろう (.41), でしょう (.33), 条件と (.27), 私 (.20)	やろう (-.56), ず (-.55), へん (-0.55), な (-0.52), たら (-.32), とる (-.27)	.372
D2	客観的伝達 VS 共感的対話	て (.58), で (.45), を (.39), その N (.38), に (.35), が (.30), N の N (.26), みたい (.23)	な (-0.25)	.188
D3	語り VS 非語り	た (.58), て (.37), てる (.34), 接続詞で (.28), ずっと (.23)	たら (-.27), 思う (-.24)	.164
D4	男性的 VS 女性的	おれ (.61), まあ (.31), N は (.28)	わたし (-.49), みたい (-.24)	.139
D5	配慮	まあ (.35), し (.34), けど (.33), やはり (.32), 条件と (.25), 結構 (.23)	N/A	.137

D1 は形式がもつ因子負荷量ならびにサンプルがもつ因子得点の双方から見て東日本と西日本の違いのみに対応している．説明力が他の因子よりも圧倒的に高いことから日本の大学生の話題をそろえた会話においては，地理的変異こそが最大の文体差であることが明らかになった．

D2 はこの結果だけからは判断しがたいが，中俣 (2024) の結果と重ねて考えると相手と共感しながら会話を行う文体と一方的に伝達する文体の違いと結論付けられる．客観的伝達の特性を持つジャンルは他に白書，新聞，科学論文，学会講演が挙げられる．J-TOCC 内では話題による差として現れ，「日本の未来」「マナー」「学校」は客観的伝達の特性が高い．前者 2 つは自身の意見を構築する必要があり，「学校」は相手が知らない情報を細かく正確に伝える必要性が生じるからだろう．反対に「食べること」「旅行」は共感的な箇所が多い．「雑談」は全体としては「共感的対話」の特性を持つため，特に負方向に高い形式は抽出されなかったと解釈できる．しかし，その中に「客観的伝達」の特性を持つ談話があるとそれが抽出される．

D3 は先行研究を踏まえて「語り VS 非語り」としたが，J-TOCC 内で見れば経験について語るか否かという区別と言えよう．これは話題による差が大きく，「社会にかかわることがある 4 話題」は全て「非語り」であった．「身の周りの 11 話題」では特に「学校」「漫画・アニメ」「スポーツ」が「語り」性が高い反面，「ファッション」と「家事」は負の値を示した．

D4 は形式がもつ因子負荷量ならびにサンプルがもつ因子得点の双方から見て男女の違いのみに対応している．説明力はそれほど高くない．また，男女の会話では同性どうしの会話と比べて中和される傾向がある．

D5 は形式から配慮表現と解釈できる．話題では「都会暮らしがいいか，田舎暮らしがいいか？」という文が指示された「住環境」で最も高く，「日本の未来」や「夢・将来設計」でも高かった．

D2 と D3 は英語の研究である Biber (1988)，日本語の様々なジャンルを対象にした中俣 (2024)，さらには BCCWJ に出現する副詞のみを分析した中俣 (2021) でも同様の結果が得られており，文体を論じるうえで頑健な評価軸と言えるのではないかと．また，形式は異なるものの「D5 配慮」の有無は中俣 (2021) でも抽出された．

5. 話題分類の結果とその考察

15 話題について，240 ファイルにおける各因子得点の平均点を算出した．その結果を用いて 15 話題に対し階層的クラスター分析を行ったところ，J-TOCC の 15 話題は「食べること・旅行・スポーツ・漫画・学校・スマートフォン・アルバイト・動物・天気」のクラスターと，「ファッション・家事・夢・マナー・住環境・日本の未来」のクラスターに二分されることがわかった．これはおおむね設計思想である「身の回りの話題」と「社会に関わることもある話題」に相当する．

しかし、予想外だったのは「ファッション」と「家事」の分類結果である。この2つは少なくとも大学生にとっては「身の回りの話題」とは呼べないことを示している。表4に示した通り、因子得点の平均値の中では「D3 語り・非語り」が最も話題による分散が大きく、この指標がクラスター分析に大きく寄与したと考えられる。ファッションは15話題で最も話題精通度が低かった。「家事」もその内容は「家事ができない、できていない」というものが多く、経験を語る談話は少ない。文体という観点から見れば、単に語る事物が日常的か否かということではなく、「自身の経験を語れるか否か」によって話題は区分されるということを示している。言語教育において話題を選ぶ時も、この観点が最も重要であるということになる。

表4 話題と因子得点

	話題	D1 東西	D2 客観 VS 共感	D3 語り	D4 男女	D5 配慮
身近で経験を語りやすい話題	11 天気	0.004	-0.146	0.185	-0.095	-0.008
	3 旅行	-0.082	-0.283	0.208	-0.232	-0.078
	1 食	0.027	-0.329	-0.091	-0.056	-0.182
	10 動物	-0.039	0.031	0.128	-0.169	-0.152
	9 バイト	-0.019	0.037	0.190	-0.032	-0.033
	8 スマホ	0.021	-0.044	0.064	0.061	-0.206
	4 スポ	-0.015	-0.070	0.458	-0.060	-0.077
	5 漫画	-0.026	-0.151	0.538	-0.032	-0.311
	7 学校	-0.162	0.288	0.938	-0.102	-0.202
経験を語りにくい話題	15 未来	0.151	0.292	-0.767	0.230	0.315
	14 住	-0.056	-0.069	-0.651	-0.022	0.823
	13 マナー	0.002	0.234	-0.339	0.085	0.027
	12 夢	0.060	0.121	-0.440	0.184	0.256
	2 衣	0.072	-0.035	-0.238	0.075	-0.025
	6 家事	0.060	0.124	-0.183	0.164	-0.146

謝辞

本研究の遂行には JSPS 科研費 23H00668 の助成を得た。

参考文献

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.

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>

加藤絵梨(2023)「話題による「まあ」の使用傾向」中俣(編)(2023)に収録

小西円(2023)「話題・地域による自問発話の使用傾向」中俣(編)(2023)に収録

清水由貴子(2023)「話題と無助詞現象」中俣(編)(2023)に収録

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.

中俣尚己ほか(2021a)「日本語話題別会話コーパス: J-TOCC」『計量国語学』33-1, pp. 205-213.

中俣尚己(2023)「助詞の出現頻度と話題」中俣(編)(2023)に収録

中俣尚己(編)(2023)『話題別コーパスが拓く日本語・日本語教育研究』ひつじ書房.

中俣尚己(2024)「文法形式に基づいた日本語文体の多次元的抽出」『日本語文法学会第 25 回大会発表予稿集』.

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」『計量国語学』33(7):435-450.

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.

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

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