話しことばに現れる文末表現の分類と文体的指標

丸山 岳彦

独立行政法人 国立国語研究所
maruyama@kokken.go.jp

1 話しことばの文体論

従来の文体論研究（計量文献学，計量文体論など）は，特定の作家・作品の文体的特性や，異なるジャンル間の言語的変異についての解明をその主な目的としており，研究対象として取り上げられるのは書きことばが中心であった．一方，話しことばに現れる文体という概念は，むしろ，「改まった口調」「くだけた話し方」のように聞き手側の印象の問題として把握されることが多かった．これらの印象は，実質的には，話しことばに現れる言語形式の特徴や発話の韻律的な特徴，話し手の属性や発話場面など，複数の要因の相互作用によって形成されると考えられる．話しことばの文体をコーパスを用いて明らかにしようとする試みは，LOB Corpus や London-Lund Corpus などを利用した Biber (1988) などがあるが，日本語の話しことばを対象とした文体論研究は，未開拓の状態であると言ってよい．

本稿では，日本語の話しことばを特徴づける文体的な指標について，特に言語の形式的な側面に着目して分析を行なう．これは，例えば，次の 2 つの話しことば（の書き起こし）の間に見られる文体的な差異を言語形式の観点から分析する，ということである[1]．

(1) 今日は私の住んでいる山室とはというお話をさしていただきます ‖ 私の住んでいる町は山室と言います ‖ 山室と言っても多分お分かりにならないと思いますので今から簡単ではございますが ‖ 山室というところの地域の説明をさしていただきます ‖ (F えー) 山室は埼玉県富士見市というところにあります ‖ ...

(2) (F えー) 私は (F あー) とても旅行が好きで (F えー) 一年に (D すー) 三回 (F えー) 以上は (D は) 旅行に行ってるんですが ‖ (F えー) 殆どが国内で (F えー) そうですね (F えー) その中でも特に (F いー) (F え) 国内で (F えー) 思い出に残ると言うか (F うー) 私の好きな (F えー) 場所として (F え) 神島というところがあります ‖ (F えー) この神島の (F おー) 場所なんですが ‖ ...

(1) と (2) の文体的な違いとしてまず目に付くのは，明示的な文末表現が現れるまでの長さ（文長）の違いや文末形式の違い，フィラーや語断片，言い直し表現の多寡などである．これらはいずれも，話しことばの文体的な特徴を捉えるための指標として考えることができる．

本稿では，話しことばの文体的な指標の中でも，特に文末表現に着目する．日本語の場合，文法的・意味的情報や待遇表現などの多くが文末表現によって表示されるため，そこには文体的な特徴が出やすいと考えられるからである．『日本語話し言葉コーパス』に含まれる 2 種類の異なる話しことばを対比させながら，文末表現が話しことばの文体の形成にどのように関わっているかについて分析を行なう．以下，2 節で使用するデータについて示し，3 節で文末表現の分類を行なう．4 節以降では，2 種類の異なる話しことばの間で見られる文体的な特徴や差異について，具体的に分析を行なう．

2 データ

本稿では，分析対象として『日本語話し言葉コーパス (Corpus of Spontaneous Japanese; 以下 CSJ と記す)』を用いる (前川, 2004)．CSJ は，約 661 時間（752 万語）分の日本語の自発音声を収録した音声コーパスであり，発話の書き起こしテキストを中心に，形態素，イントネーション，節境界，係り受け，談話構造などの研究用情報が付与されている．収録音声の中心は，「学会講演（Academic presentation speech; APS）」「模擬講演（Simulated public speaking; SPS）」と呼ばれる 2 種類の独話である．今回は，CSJ の「コア」と呼ばれるセット 177 講演分を使用する．内訳は，学会講演 70 講演，模擬講演 107 講演である．これらは人手により形態素が付与されているため，データとしての信頼性が極めて高い．

さて，本稿での主眼は話しことばに現れる文末表現に着目して分析を行なうことにあるが，そのためにはまず，文末の位置をあらかじめ特定しておく必要がある．しかしながら，句点の含まれない話しことばでは，「文」の境界が自動的・安定的に得られるとは限らない．一方，CSJ には日本語節境界解析の結果が格納されており，節の終端境界に「節境界ラベル」が付与されている．この節境界ラベルは，日本語節境界検出プログラム “CBAP-csj” に

[1] 例文中の ‖ は，CSJ に付与された「節境界ラベル (丸山他, 2004)」のうち，「絶対境界」および「強境界」の位置を表す．また，(F) はフィラー要素を，(D) は語断片を表す．

よって自動的に付与されたラベルであり，統語的な観点から「絶対境界」「強境界」「弱境界」という 3 種のレベルが設定されている (丸山他, 2004)．本稿では，“CBAP-csj” によって「絶対境界」の [文末]，[文末候補] と認定された点を文末位置と見なし,「文」の境界とした．また，統語的に大きな切れ目である「強境界」の/並列節ケレドモ/[2]，/並列節ガ/と認定された点を発話の末尾と見なし，文境界と合わせて「発話」の境界とした．以降では，文境界・発話境界に現れる表現を合わせて文末表現と呼ぶ．

本稿で用いるデータの内訳を表 1 に示す．なお，以降で用いる「形態素」とは，CSJ の「短単位」を指す[3]．

表 1: データの内訳

	講演数	総形態素数	文境界	発話境界
APS	70	214,558	5,653	7,424
SPS	107	221,429	4,730	7,498

3 文末表現の分類

伝統的な日本語研究の中では，話しことばを対象とした文法研究・文体論研究が比較的早い時期から意識され，文末表現の類型化が行なわれてきた．例えば三尾 (1942) は，話しことばの文体に「だ體」「です體」「ございます體」という 3 つの「文體形」を設け,「文の終止部に用ひられた用言の文體形が，文の文體を決定し，文の内部（接續部，連體部）に用ひられた用言の文體形が，文の丁寧さを決定する (p.52)」と述べている．

ここでは，三尾の「文體形」を拡張して文末表現を類型化し，さらに形態的・文法的な情報を加味した上で，文末表現を素性の束として表現することを考える．本稿で考える文末表現の素性のうち，主なものを表 2 に挙げる．

表 2: 文末表現を表現するための素性

項目	素性
節境界 (CB_type)	文末，並列節ケレドモ，並列節ケレド，並列節ケドモ，並列節ケド，並列節ガ
文体表現 (Style)	ございます，であります，ていただきます いたします，ております，です，ます，φ
文タイプ (Sen_type)	コピュラ，サ変動詞，機能動詞，思考動詞，ノダ，モダリティ，イディオム，一般
モダリティタイプ (M_type)	かもしれない，そうだ，だろう，だろうか，はずだ，べきだ，みたいだ，ようだ，らしい，命令・依頼，終助詞，φ

項目「文体表現」は，三尾の「ございます體」に相当する「ございます，であります，ていただきます，いたします，ております」や,「です體」に相当する「です，ます」などの文末表現形式を，そのまま素性として採用したものである．また，項目「文タイプ」は，文末表現に現れる述語句をその形態によって 8 つのタイプに分類したものである.「コピュラ」は「結果です」など名詞+判定詞という構造を持つ述語句,「サ変動詞」は「勉強する」などサ変名詞+スルという構造を持つ述語句,「機能動詞」は「以上を結論とする」「夏になる」のようにスル・ナルが主動詞として働く述語句,「思考動詞」は「思う，考える」など思考に関わる動詞による述語句,「ノダ」は内部に「のだ」が含まれる述語句,「モダリティ」は「行くようです」「例外かもしれません」のように明示的なモダリティ表現を含む述語句,「イディオム」は「すみません」「ありがとうございます」のような定型句,「一般」は以上の分類に含まれない述語句を指す．

以上の分類を前提として，CSJ の形態素情報および節境界情報を入力すると文末表現の述語句を検出しその素性を出力するプログラムを作成した．(3) のような入力に対して，文末表現の述語句部分を検出した結果を (4) のように出力する．

(3) a. 今回の実験の目的について説明いたします [文末]

b. こう滑らかに繋がっているようですけど/並列節ケド/

(4) a.
$$\left[\ |\text{説明いたします}| = \left(\begin{array}{l}\texttt{CB\_type} = [\text{文末}]\\ \texttt{Style} = [\text{いたします}]\\ \texttt{Sen\_type} = [\text{サ変}]\\ \texttt{HEAD} = [\text{説明}]\end{array}\right)\right]$$

b.
$$\left[\ \begin{array}{l}|\text{繋がっている}\\ \text{ようですけど}|\end{array} = \left(\begin{array}{l}\texttt{CB\_type} = [\text{並列節ケド}]\\ \texttt{Style} = [\text{です}]\\ \texttt{Sen\_type} = [\text{モダリティ}]\\ \texttt{M\_type} = [\text{ようだ}]\\ \texttt{HEAD} = [\text{繋がる}]\\ \texttt{ASP} = [\text{ている}]\end{array}\right)\right]$$

以上の手順で得られた文境界・発話境界の位置，および文末表現の素性を手がかりとして，次節以降では，学会講演（以下 APS）と模擬講演（以下 SPS）という異なる 2 つの話しことばを対象に，文末表現が両者の文体の形成にどのように関わっているかについて分析を行なう．

4 分析

4.1 文長・発話長の分布

まず，文長・発話長の分布について示す．1 文中・1 発話中に含まれる形態素数の平均値を，講演ごとに求めた．次に，APS・SPS ごとにそれらの平均値を求めた．これにより，APS・SPS に現れる文・発話の平均長を知ることができる．結果を図 1 に示す．

平均文長，つまり [文末]，[文末候補] が現れるまでに含まれる形態素数の平均値を比較すると，SPS の方が APS よりも顕著に長いことが分かる．一方，平均発話長，つまり文末以外に /並列節ケレドモ/，/並列節ガ/ の直後でも発話を分割した場合に得られる形態素数の平

[2] 「ケレド」「ケドモ」「ケド」という異形態を含む．

[3] 全ての短単位およびフィラーを含む．ただし語断片は含まない．

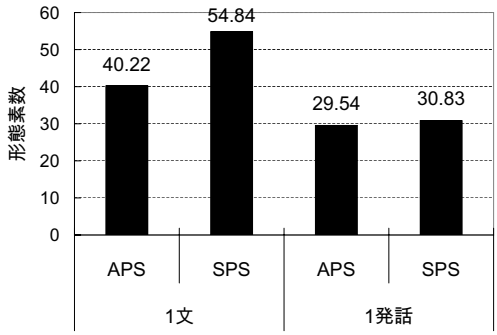

図 1: 平均文長・平均発話長の分布

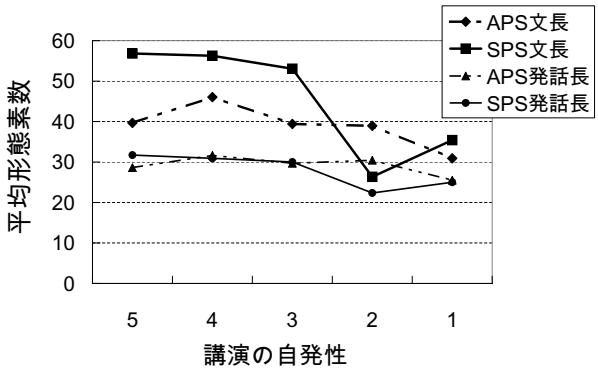

図 2: 平均文長・平均発話長と講演の自発性

均値は，APS・SPS の間でほぼ等しい．この結果から，SPS では /並列節ケレドモ/ /並列節ガ/ を頻繁に用いながら喋り続ける発話スタイルが比較的多く含まれ，結果として 1 文が長くなる傾向にあることが見て取れる．/並列節ケレドモ/ が頻出する発話の例を，(5) に示す．

(5) そこは (F まー) 現役高校生専門の予備校みたいなところだったんですけど ‖ そこは (F まー) 水道橋にあるところだったんですけど ‖ なかなかいい先生に会えまして ‖ で (F まー) 予備校ったってそんなに遊びに行くところじゃないんで (F まー) 楽しかったかって言うとそうではないかもしれないけども ‖ なかなかいい先生に会えて面白い授業が聞けて非常に価値はあったなと...

4.2 文長・発話長と講演の自発性

次に，文長・発話長の分布と講演の自発性との関わりについて示す．講演の自発性に関する指標として，ここでは CSJ に付与された印象評定データ (籠宮他, 2004) のうち，単独評定の段階評定式印象評定項目「講演の自発性」を用いる．これは，その講演がどれだけ自発的に発話されているかを収録者が 5 段階で評定したものであり，数字が大きいほど自発性が高いと評定されたことを表す．

まず，APS・SPS に付された評定値全体の分布について，表 3 に示す．自発性が高いと評定された講演は，APS より SPS に顕著に高く見られることが分かる．

表 3: 印象評定値（講演の自発性）の分布

評定値	APS	SPS
5	19 (27.1%)	52 (48.6%)
4	20 (28.6%)	37 (34.6%)
3	11 (15.7%)	13 (12.1%)
2	11 (15.7%)	4 (3.7%)
1	9 (12.9%)	1 (0.9%)

次に，5 段階の評定値それぞれが付与された講演ごとに，文長・発話長の平均値を求めた．APS・SPS の別に，結果を図 2 に示す．

平均文長のグラフの傾きを見ると，APS・SPS ともに全体として右下がりとなっている．ここから，1 文が長くなるほど自発性が高く，逆に 1 文が短くなるほど自発性が低いと評定されている傾向が伺える．一方，平均発話長のグラフの傾きを見ると，文長の場合に比べて傾きが平坦になっている．ここから，発話長の長短は，自発性の評定に対して，文長の長短ほどには関わっていないと言うことができる．

4.3 節境界の分布

続いて，3 節で示した文末表現の分類結果に基づいて分析を行なう．まず，項目「節境界」の分布について，表 4 に示す．

表 4: 項目「節境界」の出現分布

CB_type 素性	APS		SPS	
文末	5,653	(76.1%)	4,730	(63.1%)
並列節ケレドモ	433	(5.8%)	882	(11.8%)
並列節ケレド	12	(0.2%)	51	(0.7%)
並列節ケドモ	126	(1.7%)	395	(5.3%)
並列節ケド	51	(0.7%)	707	(9.4%)
並列節ガ	1,149	(15.5%)	733	(9.8%)
合計	7,424	(100.0%)	7,498	(100.0%)

CB_type 素性のうち,「文末」の占める割合を見ると，APS で 76%，SPS で 63%となっており，先述の通り，並列節が頻出する発話スタイルが SPS に多く含まれていることを示している．次に,「並列節ガ」の占める割合は APS で 15.5%，SPS で 9.8%であるのに対して,「並列節ケレドモ・ケレド・ケドモ・ケド」を合わせた割合は APS で 8.4%，SPS で 27.2%となり，APS・SPS 間で異なる分布となっている．ここから,「並列節ガ」と「並列節ケレドモ」という 2 つの並列節が，APS・SPS という発話場面に応じて使い分けられている実態が伺える．改まった発話スタイルを持つ APS では「ガ」が多く用いられ，リラックスした発話スタイルを持つ SPS では「ケレドモ」が多用される，ということである．さらに,「並列節ケレドモ・ケレド・ケドモ・ケド」という各形態の分布を見ると，APS ではほとんど現れない「ケドモ・ケド」

という形態が，SPS では 5.3%，9.4%と比較的多く現れていることが分かる.「ケドモ・ケド」は「ケレドモ・ケレド」のくだけた言い方であり，そのような形式が SPS で多く現れるという傾向は，SPS が APS よりもくだけた発話スタイルを持っているということを裏付ける結果として考えることができる.

4.4 文体表現の分布

次に，項目「文体表現」の分布について，表 5 に示す.

表 5: 項目「文体表現」の出現分布

Style 素性	APS		SPS	
ございます	146	(2.0%)	50	(0.7%)
であります	68	(0.9%)	12	(0.2%)
ていただきます	88	(1.2%)	21	(0.3%)
いたします	195	(2.6%)	14	(0.2%)
ております	340	(4.6%)	151	(2.0%)
です	2,314	(31.2%)	4,192	(55.9%)
ます	3,622	(48.8%)	2,353	(31.4%)
その他	651	(8.8%)	705	(9.4%)
合計	7,424	(100%)	7,498	(100%)

Style 素性のうち，三尾の「ございます體」に相当する「ございます，であります，ていただきます，いたします，ております」は，いずれも SPS より 2 倍以上多く APS に現れている．これは，APS が SPS よりも改まった発話スタイルを持っていることの傍証として考えられる．また,「です，ます」の分布が APS・SPS 間で大きく異なっていることが分かる．これは，次に述べる項目「文タイプ」の素性「ノダ」の多寡に原因がある.

4.5 文タイプの分布

最後に，項目「文タイプ」の分布について，表 6 に示す.

表 6: 項目「文タイプ」の出現分布

Sen_type 素性	APS		SPS	
コピュラ	1,625	(21.89%)	1,430	(19.07%)
サ変動詞	807	(10.87%)	179	(2.39%)
機能動詞	556	(7.49%)	209	(2.79%)
思考動詞	469	(6.32%)	345	(4.60%)
ノダ	833	(11.22%)	2,794	(37.26%)
モダリティ	124	(1.67%)	172	(2.29%)
イディオム	36	(0.48%)	20	(0.27%)
一般	2,974	(40.06%)	2,349	(31.33%)
合計	7,424	(100%)	7,498	(100%)

ここでは，SPS の「ノダ」の割合が APS よりも顕著に高い点が特徴的である．表 5 で「です」の割合が SPS で高かったのは，(6) のように「のだ」という形式の多用によるところが大きいと考えられる．一方,「サ変動詞, 機能動詞」の割合は APS において顕著に高い．これは，学術講演という APS の性格によるところが大きいと考えられる.

5 まとめと課題

日本語の話しことばを特徴づける文体的な指標について，文末表現に着目して分析を行なった．話しことばの中から文境界・発話境界を設定することによって文長・発話長の分析を，また，文末表現を形態的・文法的な素性の束として捉えることによって文末表現のバリエーションの分析を，それぞれ行なうことができた．APS・SPS という 2 種類の講演を対象として分析を行なった結果，両講演の持つ発話スタイルに関して，複数の観点から違いを明らかにすることができた．4 節で示した分析の観点は，いずれも，話しことばの文体的な特徴を示す指標として有効であると言える.

話しことばの文体を捉えるための指標には，本稿で挙げたもの以外にも，さまざまな観点を考えることができる．発話スタイルに関係する指標としては，冒頭で指摘したように，フィラーや語断片の多寡，言い誤りや言い直しの多寡などが挙げられる．さらに，語種の比率や使用語彙の難易度などの語彙論的な指標，テンス・アスペクト・モダリティなどの文法形式に関わる文法的な指標,「観察できます」「観察することができます」「観察が可能です」などパラフレーズ可能な言い回しがどのように分布しているかという語用論的な指標，そして，文境界や節境界のイントネーション・ポーズなどに着目した音韻論的（パラ言語的）な指標もまた，話しことばを特徴づける文体的な指標として見ることができる.

そもそも，書きことば・話しことばに限らず，ある言語表現の文体的な特徴を捉えるためにどのような項目（変数）をいくつ立てるべきかについては，さらに深く検討される必要がある (Nakamura, 1995)．多様なタイプのテキストを比較するために適切かつ有効な変数を設計するには，さまざまな角度からの観察と記述をケーススタディ的に積み上げていくことが有効であると思われる.

謝辞: 本研究の一部は、日本学術振興会科学研究費補助金（萌芽研究）課題番号 16652031 を受けている。

参考文献

Biber, D. (1988). *Variations across Speech and Writing*. Cambridge: Cambridge University Press.

籠宮隆之・山住賢司・槙洋一・前川喜久雄 (2004).「講演音声に対する印象評定尺度の作成と分析」.『第 3 回話し言葉の科学と工学ワークショップ講演予稿集』, pp. 47–52.

前川喜久雄 (2004).「『日本語話し言葉コーパス』の概要」.『日本語科学』, **15**, 111–133.

丸山岳彦・柏岡秀紀・熊野正・田中英輝 (2004).「日本語節境界検出プログラム CBAP の開発と評価」.『自然言語処理』, **11** (3), 39–68.

三尾砂 (1942).『話言葉の文法 言葉遣篇』.（くろしお出版 1995 年）.

Nakamura, J. (1995). Text Typology and Corpus: A Critical Review of Biber's Methodology. *English Corpus Studies*, **2**, 75–90.