文末表現の分布と文体
「現代日本語書き言葉均衡コーパス」を利用して

大学共同利用機関法人人間文化研究機構　国立国語研究所言語資源研究系准教授　**山崎　誠**

PROFILE

1957 年生。1980 年埼玉大学教養学部卒。1984 年筑波大学大学院文藝言語研究科単位取得退学。現在、大学共同利用機関法人人間文化研究機構国立国語研究所言語資源研究系准教授。1984 年より国立国語研究所で語彙調査、シソーラスの編纂、コーパスの構築に従事。産業日本語研究会世話人会。

yamazaki@ninjal.ac.jp

1 はじめに

コーパスを利用した言語研究の特徴のひとつにテクストの特性を明らかにする研究がある。特に、従来行われてきた文体に関する研究をより幅広く捉えられる環境がととのったことで、新たな研究が進んでいる。

本稿では文末表現の分布に現れるテクストの相違を観察する。文末表現は文体を端的に示す特徴的な言語要素であり、おもに「です・ます体」「だ体」「である体」などの名称とともに、文体の硬軟を表すものとして利用されてきた。

本稿では、市川 (1973) の文末表現の類別及び混合率・変移率という指標を用いて文末表現の状況を概観する。

2 データ

本稿で使用したデータは、『BCCWJ 領域内公開データ（2009 年度版）』である。このデータは、2011 年夏に公開した『現代日本語書き言葉均衡コーパス』の一部で、言語量は約 7,615 万語（短単位）である。2009 年 8 月時点での構築途中のものであるが、一定量のデータがあり、大体の傾向を把握するのには充分と思われる。データの内訳と語数（短単位）は表 1 のとおり。

表 1　使用したデータ

媒　体	語　数
出版書籍	2,333 万
雑誌	197 万
新聞	36 万
図書館書籍	2,800 万
白書	472 万
教科書	120 万
ベストセラー	364 万
Yahoo! 知恵袋	521 万
Yahoo! ブログ	282 万
国会会議録	490 万

3 句点の前に現れる語

まず、句点（。）の前に来る語（以下、文末語と称する）の出現状況を観察する。文末は必ずしも句点だけとは限らず、「？」などもその手がかりになるが、圧倒的に多いのが句点であるので、今回は句点のみを文の切れ目の判定に使った。

図 1 に媒体ごとに、文末語の平均使用度数を示した。

例えば、図書館書籍では、延べ語数で 1,139,370 個の文末語があるが、それらの異なりは 16,298 語であった。したがって、平均使用度数は 69.9 となる。平均使用度数は、文末語の種類の多寡を表し、値が高いほど種類が少ないことを意味する。

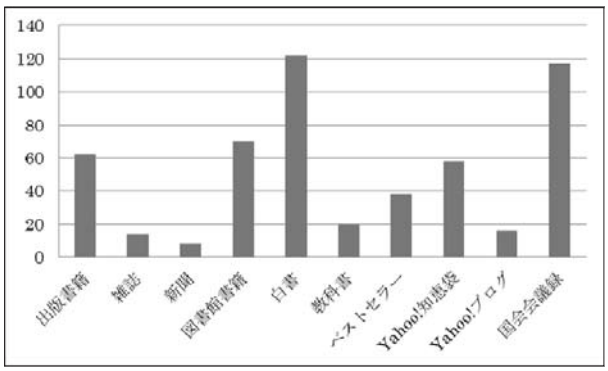

図 1 文末語の平均使用度数

図 1 からは、白書、国会会議録の値が相対的に高く、他の媒体と比べて同じような文末語が多いことが分かる。逆に値が相対的に低いのは、新聞、雑誌、Yahoo! ブログ、教科書である。新聞で文末語の種類が多い一つの理由は、「新学期になって 1 カ月。」のような名詞止めとくに「確実に実施。」のようなサ変名詞で終わる文が多いためであろう。後述の図 2 － 3 でもそのことが確かめられる。

どのような文末語が多いのか、媒体ごとに表 2 － 1 ～ 2 にまとめた。

表 2-1 ～ 2 は単純に文末語の属性を見ているだけであるため、実際には「有る」は「である」の形で用いられていることが多いであろうし、「居る」は、「～ている」での使用例が多いものと思われる。このような単純集計であっても、白書のみが「居る」がいちばん多いという特殊性を指摘することができる。

品詞で見ると、終助詞「か」「ね」「よ」のいずれかが上位 10 語に含まれている媒体が多いが、「白書」と「新聞」には終助詞は現れていない（新聞では 11 位に、白書では 26 位に「か」が現れる）。

Yahoo! ブログと Yahoo! 知恵袋に句点（。）が文末語として現れているが、これは、これらの媒体で文末に句点を重ねる表現法が多く取られていることを意味する。ちなみに句点が文末語になっている例は、新聞、白書では現れず、図書館書籍で 2 例、出版書籍で 14 例という低さだった。

表 2 － 1　文末語の上位10

順位	出版書籍	雑　誌	新　聞	図書館書籍	白　書
1	た（助動詞）	た（助動詞）	た（助動詞）	た（助動詞）	居る（動詞）
2	有る（動詞）	ます（助動詞）	居る（動詞）	有る（動詞）	た（助動詞）
3	ます（助動詞）	です（助動詞）	だ（助動詞）	だ（助動詞）	有る（動詞）
4	だ（助動詞）	だ（助動詞）	有る（動詞）	ます（助動詞）	）（記号）
5	です（助動詞）	居る（動詞）	為る（動詞）	です（助動詞）	為る（動詞）
6	居る（動詞）	有る（動詞）	ない（助動詞）	居る（動詞）	られる（助動詞）
7	ない（助動詞）	ね（終助詞）	」（記号）	ない（助動詞）	拠る（動詞）
8	為る（動詞）	ない（助動詞）	ます（助動詞）	無い（形容詞）	れる（助動詞）
9	無い（形容詞）	か（終助詞）	言う（動詞）	か（終助詞）	言う（動詞）
10	か（終助詞）	為る（動詞）	です（助動詞）	為る（動詞）	ない（助動詞）

表 2 － 2　文末語の上位10

順位	教科書	ベストセラー	Yahoo! 知恵袋	Yahoo! ブログ	国会会議録
1	た（助動詞）	た（助動詞）	ます（助動詞）	た（助動詞）	ます（助動詞）
2	ます（助動詞）	有る（動詞）	です（助動詞）	です（助動詞）	です（助動詞）
3	有る（動詞）	だ（助動詞）	。（記号）	ます（助動詞）	か（終助詞）
4	居る（動詞）	です（助動詞）	た（助動詞）	。（記号）	た（助動詞）
5	ね（終助詞）	ます（助動詞）	ず（助動詞）	ね（終助詞）	ね（終助詞）
6	為る（動詞）	居る（動詞）	ね（終助詞）	だ（助動詞）	ず（助動詞）
7	です（助動詞）	ない（助動詞）	下さる（動詞）	居る（動詞）	よ（終助詞）
8	見る（動詞）	無い（形容詞）	よ（終助詞）	・（記号）	ない（助動詞）
9	言う（動詞）	か（終助詞）	か（終助詞）	有る（動詞）	と（格助詞）
10	成る（動詞）	よ（終助詞）	・（記号）	ない（助動詞）	居る（動詞）

4 文末表現の類別

市川 (1973) の文末表現の類別は、常体のみの分類であったが、今回の調査は敬体（です・ます体）も多用されているため、敬体を追加して 14 分類とした。具体的な分類は以下のとおり。

「た」系列

①動詞＋た（終止形、以下同じ）〔「であった·ていた」以外の補助動詞の場合を含む〕

②形容詞・形容動詞＋た〔補助形容詞、および「形動＋あった」の場合を含む〕

③～だった·であった(体言以外の語に続く場合も含む)

④～ていた

⑤一般動詞＋た〔「だった」を除く〕

非「た」系列

⑥動詞（終止形、以下同じ）〔「てある・ている」以外の補助動詞を含む〕

⑦形容詞・形容動詞〔補助形容詞、および「形動＋ある」の場合を含む〕

⑧～だ·～である〔体言以外の語につづく場合を含む。ただし、形動に続く場合を除く〕

⑨～ている

⑩一般助動詞〔「だ」を除く〕

敬体「た」系列

⑪でした・ました

敬体「非「た」系列

⑫です・ます〔「でしょう」「ましょう」も含む〕

特殊

⑬言止め

⑭その他〔助詞止め、倒置的表現のほか、疑問文、命令文などの文末を含む〕

本稿で追加したのは⑪⑫である。また、データの制約のため、②の注記の「形動＋あった」の場合は③に含まれ、⑦の「形動＋ある」の場合は⑧に含まれている。

各媒体における文末表現の割合を図2−1～3に示す。

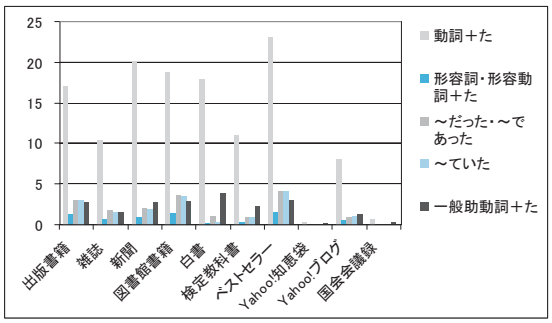

図２－１　文末表現の類別（常体・「た」系列）

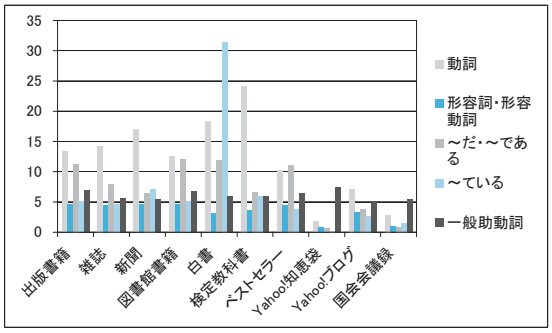

図２－２　文末表現の類別（常体・非「た」系列）

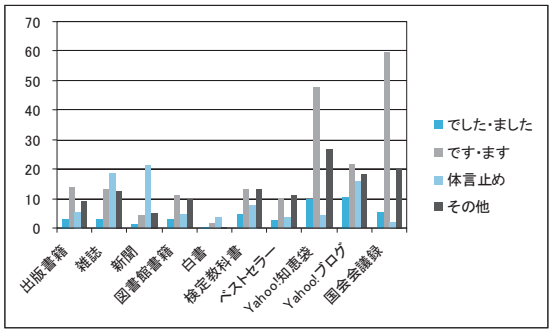

図２－３　文末表現の類別（敬体・特殊）

図２－１の常体の「た」系列で顕著なのは、Yahoo! 知恵袋と国会会議録の「動詞＋た」の割合が１%以下と低いことである。図２－２の常体の非「た」系列で顕著なのは、白書の「ている」が 31.5%と高い値であること、また、Yahoo! 知恵袋と国会会議録における「動詞」と「～だ・～である」の割合が低いことである。図２－３の敬体および特殊では、国会会議録における「です・ます」が約 60%と高いこと、新聞・雑誌における「体言止め」、Yahoo! 知恵袋における「その他」の割合も高い。以上の傾向はそれぞれの媒体の特徴を反映しているもの

であり、文体的な特徴であると言えよう。

市川 (1973) で紹介された混合率 (m)、変移率 (c) は以下の式で計算される。

混合率：

14種類の文末表現の混じり合いの度合いを表した指標

$$m = \left(1 - \frac{\sum_{i=1}^{14} fi^2}{\left(\sum_{i=1}^{14} fi\right)^2}\right) \times 100$$

変移率：

文末表現がまとまって出現するか、入り交じって出現するかの度合いを表す指標。例えば、文末表現 a と b があったとして、以下のような出現の仕方をした場合、それぞれの混合率は同じだが、文末表現の変移には大きな差がある。

(1)aaaaabbbbb

(2)abababababab

この差を表すのが変移率である。

$$c = p / n \times 100$$

ただし、p は変移箇所数；

n は連接箇所数（文の数－１）

各媒体の混合率と変移率を図３に示した。両者はほぼ同じような値を取るが、Yahoo! 知恵袋と Yahoo! ブログは、混合率が低い割りには変移率が高いという傾向がある。すなわち、文末表現の切り替えが相対的に多いということになる。

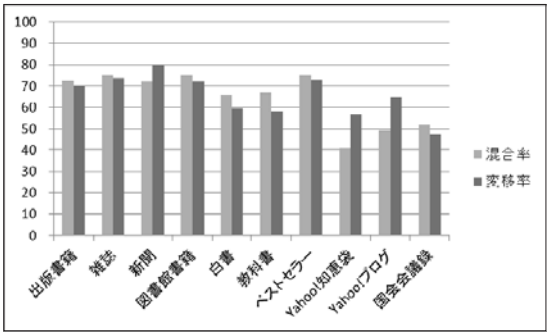

図３　混合率と変移率

ちなみに混合率、変移率はテクストの長さに影響されるのかどうかを確認する。テクストの長さと混合率・変移率が関係するのであれば、図３の傾向は媒体の差とばかりも言えなくなるからである。そこで、図書館書籍を対象にして、文の数を 10 刻みにして、混合率・変移率を算出したものが図４である。混合率、変移率ともに文の数が増えるにしたがって増加する傾向があるが、顕著とは言えない。文の数が 50 以上くらいから値は一定しているようである。

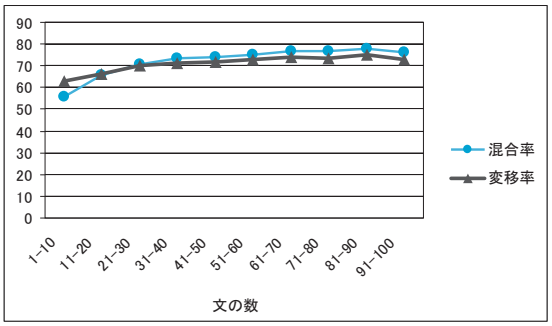

図４　文の数ごとの混合率・変移率

5 まとめと課題

本稿では、文末表現の分布を異なる媒体について観察し、特徴的な傾向をいくつか見いだすことができた。

文末表現の類別は本稿で採用したものだけではないので、今後どのような類別が媒体の特徴をよく捉えるのか検討する必要がある。また、本稿では混合率、変移率を扱う上で地の文と会話文との区別をしていないが、これを別々に扱うことも考えられよう。

その他、特徴的な副詞や接続詞、あるいは、口語的、文語的な表現などもカテゴリー化して類別に含めることもできよう。文体あるいはテクストのさまざまな特性を明らかにする研究はコーパスの利用で今後ますます精緻化していくことが期待される。

参考文献

市川孝 (1973)，文末表現の様相，「計量国語学」65，pp1-8