『現代日本語書き言葉均衡コーパス』を用いた文末表現のバリエーションの分析

丸山 岳彦

国立国語研究所 言語資源研究系

1　はじめに

2006年度より国立国語研究所を中心として構築が進められてきた『現代日本語書き言葉均衡コーパス』(BCCWJ)が完成し、Web上の検索サイト「中納言[1]」「少納言[2]」、および「BCCWJ-DVD版」による一般公開が2011年より始まった。BCCWJは、日本語では初となる精密に設計された均衡コーパスであり、さまざまなメディア・ジャンル(これをレジスターと呼ぶことにする)から無作為抽出された約1億語分の書き言葉のサンプルが収録されている[3]。従来「たまたま入手できる電子資料を使っているに過ぎ[4]」なかった日本語のコーパス言語学的な研究は、BCCWJの完成により、詳細な書誌情報を利用した大規模データの位相的な研究が可能になった。

本稿では、BCCWJを利用した現代日本語文法研究および社会言語学的な研究の試みとして、レジスターごとに観察される文末表現のバリエーションについて分析を行なう。文末表現は文の表出にとって最も重要な要素であり、文末表現のバリエーションの豊富さは、そのレジスターが備える表現力の豊かさを示す手がかりとして捉えられることを述べる。

2　文末表現の豊かさとテキストの特性

ある文の末尾に現れる表現形式を、「文末表現」と呼ぶことにする。日本語の文末表現は、通常、述語となる要素(動詞、形容詞、名詞＋判定詞など)に、複数の文法カテゴリ(ヴォイス、アスペクト、肯否、テンス、モダリティ)を表わす助動詞や終助詞などの要素が後接することで構成される(図1)。

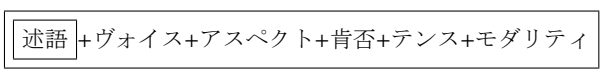

図1: 日本語の文末表現の構造

述語要素によって事態や状態が描写されたり、モダリティ要素によって書き手(話し手)による判断や読み手(聞き手)への伝達態度が表わされたりするという点で、文末表現は文の成立にとって重要な役割を果たす部分であり、テキストの特性に大きな影響を及ぼす要素であると言える。

ある主題に沿って書かれたひとまとまりのテキストを観察した時、そこに含まれる文末表現の種類が豊富であるほど、そのテキストは豊かな表現力・伝達力を持っていると見なすことができる。逆に、文末表現の種類が少なければ、そのテキストは単調な印象を与えやすく、定形性の高いテキストであると言える。この点において、文末表現のバリエーションを分析することは、そのテキストの表現力に関する特性を明らかにする手段と位置づけることができる。

そこで本稿では、BCCWJから文末表現を収集し、そのバリエーションの異なりをレジスターごとに集計することにより、テキストの定形性について分析する。

3　文末表現の収集

3.1　分析対象データ

BCCWJは、「出版サブコーパス」「図書館サブコーパス」「特定目的サブコーパス」と呼ばれる3つのサブコーパスから構成されており、さまざまなレジスターから無作為抽出されたテキストが、合計172,675サンプル収録されている。これらのテキストには、文章の構造をタグで表現した「文書構造タグ」が付与されており、いわゆる「文」の範囲が <sentence> タグによって囲まれている。文書構造タグが付与されたテキストの例を、一部抜粋して、図2に示す。

<title> タグはその文書のタイトル要素を示す。<paragraph> タグは段落範囲を示す。それぞれの内側に <sentence> タグがあり、「文」に相当する文書要素が囲まれる。なお、句点類(「。」「！」「？」)で終わる文は属性のない <sentence> タグで囲まれるが、句点類の区切り記号を持たない文は、<sentence type="quasi"> という属性をつけた形が使われる。図2では、タイトル部分に相当する「ままかり…【ママカリ】」は句点で終わっていないため、 <sentence type="quasi"> が用いられている。

```
<title>
<sentence type="quasi">ままかり…【ママカリ】</sentence><br/>
</title>
<paragraph>
<sentence>　瀬戸内海沿岸で、海魚サッパのことを、ママカリといいます。</sentence><sentence>その名の由来がおもしろいのです。</sentence><sentence>きくところによりますと、<quote>「飯＋借り」</quote>が、語源だといいます。</sentence><sentence>あまりおいしいので御飯が足りなくなり、借りにいったというのです。</sentence><br/>
</paragraph>
```

図 2: 文書構造タグの例（PB58_00012）

本稿では文末表現を分析するため、句点類のない <sentence type="quasi"> で囲まれた要素は除外し、属性のない <sentence> タグで囲まれた要素を分析対象とする。さらに、12 種類のレジスター間で分析対象データのサイズを揃えるため、各レジスターごとの層別情報（ジャンル、出版年など）による構成比率を反映させた形で、各レジスターからそれぞれ約 20,000 文ずつを無作為抽出した。ただし、データ量の不足などの事情により、20,000 文が取得できなかった部分もある。各レジスターの内訳を、表 1 に示す。

表 1: 分析対象データ

SC	レジスター	総文数	分析対象
出版 SC	書籍	1,087,715	20,000
	雑誌	197,069	19,999
	新聞	54,932	19,999
図書館 SC	書籍	1,276,651	20,002
特定目的 SC	白書	95,267	19,998
	教科書	39,966	18,766
	広報紙	97,454	19,976
	ベストセラー	170,940	20,000
	Yahoo!知恵袋	582,862	20,002
	Yahoo!ブログ	487,167	19,999
	法律	17,637	17,637
	国会会議録	116,022	19,619
合計		4,223,682	235,962

3.2　文末表現の収集方法

分析対象データの各文から、文末位置から左向きに 1 文字ずつを取得し、1〜20 文字（1〜20gram）までの範囲を、文末表現データとして抽出した（図 3）。

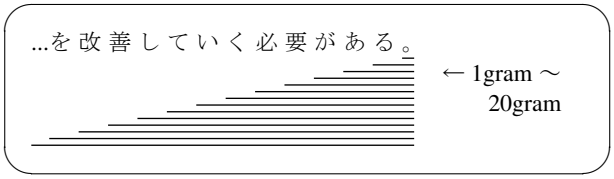

図 3: 文末表現データの収集

各レジスター・各 gram ごとに文末表現の生起頻度を集計し、当該のレジスターにおける全文数に占めるその表現の比率を算出した。白書から取得された文末表現データを集計した例を、一部抜粋して表 2 に示す。

表 2: 文末表現データの集計例（白書）

gram	文末表現	頻度	比率
1	。	19,996	99.99%
1	！	2	0.01%
2	る。	11,791	58.96%
2	た。	4,393	21.97%
2	)。	6,274	6.45%
	(中略)		
10	することとしている。	75	0.38%
10	ているところである。	67	0.34%
10	ることが必要である。	55	0.28%
10	していく必要がある。	49	0.25%

4　分析

4.1　文末表現の異なり率

収集した文末表現について、各 gram の異なり数をレジスターごとに集計し、それが全文数に占める比率（異なり率）を求めた。5gram 以降の結果を図 4 に示す。

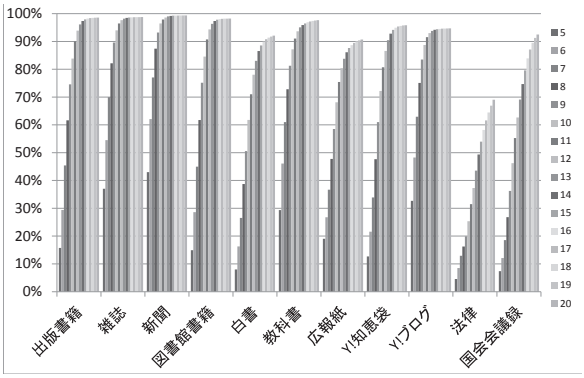

図 4: 5gram から 20gram までの異なり率

異なり率が高いほど文末表現の種類が豊富であり、逆に異なり率が低いほど文末表現の種類が少ないことを示す。図 4 では、特に法律の異なり率が低いことが目立つ。法規範の集合である法律では、文末表現のバリエーションが少なく、定形的な文末表現によって構成されていることがうかがえる。一方、書籍、雑誌、新聞などは異なり率が高いことから、これらのレジスターには多様な文末表現が含まれていると言える。

4.2 1gram の分析

次に、1gram の文末表現を分析する。この場合の 1gram とは、文末に現れる句点類「。」「?」「!」を表わす。各レジスターの総文数に占める割合を集計し、「。」の比率で降順に並べ替えた結果を、表 3 に示す。

表 3: 1gram の集計結果

レジスター	。	?	!
法律	100.00%	0.00%	0.00%
白書	99.99%	0.00%	0.01%
国会会議録	99.98%	0.01%	0.01%
新聞	99.08%	0.50%	0.42%
教科書	98.37%	1.06%	0.58%
出版書籍	97.74%	1.40%	0.87%
図書館書籍	97.67%	1.48%	0.85%
ベストセラー	97.28%	1.64%	1.08%
広報紙	94.28%	1.76%	3.96%
雑誌	93.22%	2.54%	4.24%
Yahoo!ブログ	78.59%	7.85%	13.56%
Yahoo!知恵袋	72.23%	23.42%	4.35%

全レジスターにおいて、「。」で終わる文末表現が大半を占めていることが分かる。特に法律は、全ての文が「。」で終わっている。白書や国会会議録には「!」「?」の例も若干見られたが、実例を確認してみると、発言中に引用された記事や番組タイトルとして使用されている場合であった。以下、新聞、教科書などの規範性の高いレジスターが続き、雑誌では 93.2%となる。

これに対して、Yahoo!ブログでは 78.6%、Yahoo!知恵袋では 72.2%にまで「。」の比率が急落している。特に Yahoo!知恵袋における「?」の比率の高さからは、質問文の中で「?」で終わる文末表現が多用されている実態を見て取ることができる。

4.3 gram の増加と文末表現のバリエーションの推移

次に、1gram から 5gram までの範囲について、各 gram で最も多く出現する文末表現が全文数に占める比率について集計した。レジスターごとに比率の最高値がどのように推移したかを、図 5 に示す。

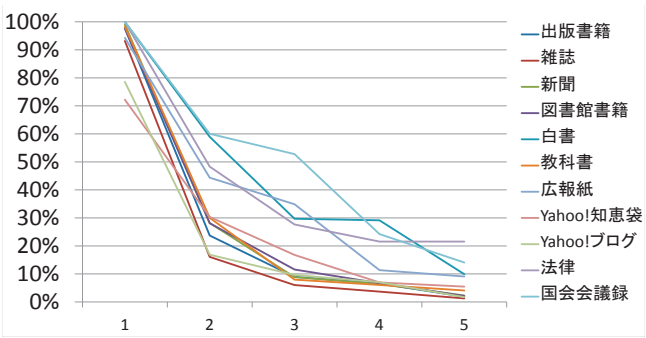

図 5: 文末表現のバリエーションの推移

法律、国会会議録、白書、広報紙、の 4 つは、5gram になってもほぼ 10%以上の値を示している。これらのレジスターでは文末表現のバリエーションが比較的少なく、定形的な文末表現が多用されていると言える。これら 4 つと他のレジスターは、2gram の時点ですでに 10%以上の開きがあることから、文末表現の定形性という点で両者には大きな隔たりがあると言える。

4.4 頻出する文末表現と出現率の比較

5gram 以上に現れた文末表現を集計し、レジスターごとに出現率の高い文末表現をリスト化した。結果を表 4 に示す（文末の「。」は省略してある）。ゴシック体は、表 4 中、そのレジスターにしか現れていない文末表現を表す。

表 4 からは、各レジスターにおける文末表現の特徴を見て取ることができる。まずそれぞれの文末表現の出現率に着目すると、法律、国会会議録の出現率の値が全体的に高く、白書、広報紙がそれに続いていることが分かる。4.3 でも指摘したように、これらのレジスターでは、表 4 に挙げたような例を典型例とする定形的な文末表現が多用されていることが分かる。

一方、書籍、雑誌、新聞、Yahoo!ブログでは、出現率の値が全体的に低くなっている。これらには文末表現のバリエーションが豊富に存在するということであり、レジスターの特性として、豊かな表現力・伝達力を備えていると考えることができる。

教科書と Yahoo!知恵袋では出現率が若干高めだが、これは「～ましょう。」「～でしょうか?」などの定形的な文末表現が少数存在するためであると考えられる。

5 考察：テキストの機能と文末表現

ある事態や状態を描写したり、書き手（話し手）の判断や態度を表明したりする際、最も重要な役割を担うのが文末表現である。あるレジスターにおいて文末表現のバリエーションが豊富に観察されるということは、そのレジスターに含まれるテキストにはより豊かな表現力が備わっていると考えられる。逆に、文末表現のバリエーションが乏しく、定形的な文末表現が多く含まれる場合は、そのレジスターのテキストは斉一な文体によって成立していると見てよい。

文末表現のバリエーションが示す傾向は、当該のレジスターに含まれるテキストが担う機能と関係すると考えられる。そのテキストがどのような目的によって書かれたものかという点を考えることによって、文末表現のバリエーションの偏りや、そこに現れる文法カテゴリの傾向の違いを説明することができる。

その一例として、文末に現れるモダリティ要素の出現傾向について見てみたい。20gram までの文末表現データから、他者への働きかけを表す対人的モダリ

表 4: 頻出する文末表現（5gram 以上、上位 10 位）

出版書籍		図書館書籍		雑誌		新聞		白書		教科書	
のである	1.85%	のである	2.11%	ています	1.27%	している	2.32%	っている	9.91%	ましょう	4.08%
ています	1.73%	なかった	1.62%	している	1.00%	っている	1.43%	している	9.08%	**てみよう**	2.54%
っている	1.41%	っている	1.48%	っている	0.94%	れている	1.04%	**なっている**	6.28%	ています	2.09%
している	1.40%	している	1.18%	れている	0.90%	なかった	0.92%	れている	5.41%	れている	2.04%
なかった	1.34%	であった	1.18%	なかった	0.65%	**になった**	0.74%	**となっている**	5.09%	っている	1.57%
れている	1.21%	ています	1.17%	のである	0.65%	していた	0.67%	されている	2.73%	している	1.50%
あります	1.00%	っていた	0.97%	あります	0.61%	**たという**	0.58%	となった	2.29%	**みましょう**	1.37%
であった	0.96%	れている	0.95%	されている	0.46%	ています	0.54%	のである	2.03%	**てみましょう**	1.27%
りません	0.83%	たのである	0.78%	**でしょう**	0.46%	となった	0.49%	**めている**	1.99%	あります	1.25%
なります	0.73%	あります	0.74%	ください	0.45%	**るという**	0.47%	**ものである**	1.90%	りました	0.98%

広報紙		ベストセラー		Yahoo!知恵袋		Yahoo!ブログ		法律		国会会議録	
ください	9.07%	なかった	2.21%	**しょうか？**	5.46%	ています	1.72%	**ならない**	21.55%	**ございます**	14.08%
ています	8.46%	のである	1.94%	**でしょうか？**	5.44%	**いました**	1.48%	**なければならない**	18.79%	**でございます**	12.35%
しています	3.59%	であった	1.58%	**のでしょうか？**	2.79%	りました	1.21%	**ができる**	13.79%	**おります**	10.98%
てください	2.76%	っていた	1.43%	思います	2.77%	**きました**	1.15%	**ことができる**	13.78%	**ております**	10.84%
あります	2.66%	っている	1.23%	ています	2.77%	しました	1.11%	**ることができる**	11.95%	思います	8.47%
しました	2.49%	していた	0.80%	ください	2.75%	思います	0.82%	**しなければならない**	10.62%	あります	7.76%
れました	2.30%	している	0.80%	と思います	2.58%	と思います	0.73%	**ものとする**	8.12%	と思います	7.16%
なります	2.17%	たのである	0.70%	てください	2.44%	あります	0.64%	**することができる**	6.88%	**であります**	6.87%
してください	2.05%	ています	0.70%	りません	2.17%	**ていました**	0.61%	**るものとする**	6.36%	**いと思います**	4.18%
ましょう	1.94%	りません	0.60%	あります	1.60%	**っています**	0.57%	**準用する**	4.60%	**けでございます**	4.16%

ティ表現のうち、「依頼」「質問」を表す「ください。」「[でま] すか [。？]」「[でま] しょうか [。？]」という形の文末表現を抽出した。また、書き手の主観的な判断を表す対事的モダリティ表現のうち、「禁止」「義務」「許可」を表す「[はば] ならない。」「[てで] もよい。」「ことができる。」という形の文末表現を抽出した。出現率で逆順に並べた結果を表 5 に示す。

表 5: モダリティ表現の出現率

対人（依頼・質問）		対事（禁止・義務・許可）	
Yahoo!知恵袋	18.93%	法律	35.34%
広報紙	9.89%	教科書	1.07%
国会会議録	5.86%	出版書籍	1.03%
教科書	2.36%	白書	0.68%
Yahoo!ブログ	1.44%	図書館書籍	0.53%
雑誌	1.30%	ベストセラー	0.48%
出版書籍	0.92%	雑誌	0.42%
図書館書籍	0.84%	新聞	0.27%
ベストセラー	0.78%	国会会議録	0.17%
新聞	0.29%	Y!ブログ	0.11%
白書	0.07%	広報紙	0.02%
法律	0.00%	Y!知恵袋	0.01%

Yahoo!知恵袋では全文末表現のうち約 20%が、広報紙では約 10%が、「依頼」「質問」のモダリティ表現を含んでいることになる。Yahoo!知恵袋の質問部分は、参加者に質問をしたり情報提供を依頼したりすることを目的に書かれたテキストであることから、「依頼」「質問」の比率が高いことを説明できる。また、広報紙は区市町村の住民に情報を提供するための媒体であり、「お問い合わせください。」「ご利用ください。」のように直接呼びかけるような文末表現が多用されている。いずれの場合も、テキストが読み手に直接訴えかける機能を担っているということが、対人的モダリティを伴う文末表現が頻出する理由と考えられる。

一方、「禁止」「義務」「許可」のモダリティ表現を伴う文末表現の出現率は、法律が約 35%と、他のレジスターに比べて圧倒的に高い。表 4 においても、上位 10 位の大半が「禁止」「義務」「可能」のモダリティ表現を伴っている。法律とは、ある行為を命令・禁止したり、ある権利を保障したりすることを明示的かつ曖昧性のないように述べるためのテキストであり、そのことを示すための文末表現に特化していると考えることができる。表 4 に挙げた上位 10 位のすべてが法律にしか現れていない文末表現となっていることからも、法律の文末表現が極めて特殊な振る舞いをしていることが見て取れる。

6 まとめ

BCCWJ に含まれる様々なタイプのテキストから文末表現を収集し、そのバリエーションの分布について分析を行なった。レジスターによって文末表現のバリエーションが異なること、文末表現に現れる要素はテキストが担う機能と関係があることを述べた。

参考文献

[1] https://chunagon.ninjal.ac.jp/.

[2] http://www.kotonoha.gr.jp/shonagon/.

[3] 国立国語研究所. 『現代日本語書き言葉均衡コーパス』利用の手引. （BCCWJ-DVD 版に収録）.

[4] 丸山岳彦, 田野村忠温. コーパス日本語学の射程. 日本語科学, Vol. 22, pp. 5–12, 2007. 国書刊行会.