「語り性」を有する書きことばの典型例の分析

保田　祥[†]（国立国語研究所 コーパス開発センター）
柏野　和佳子（国立国語研究所 言語資源研究系）
立花　幸子（国立国語研究所 コーパス開発センター）
丸山　岳彦（国立国語研究所 言語資源研究系）

Analysis of Written Japanese Text with Addressing Expressions

Sachi Yasuda (Center for Corpus Development, NINJAL)
Wakako Kashino (Dept. Corpus Studies, NINJAL)
Sachiko Tachibana (Center for Corpus Development, NINJAL)
Takehiko Maruyama (Dept. Corpus Studies, NINJAL)

１．はじめに

我々のプロジェクトでは，『現代日本語書き言葉均衡コーパス』（BCCWJ）に収録されている図書館サブコーパスの書籍サンプルに，人手で文書分類の観点から情報を付与する作業を行っている（柏野・奥村 2012 予定）．

本稿では，人手によるアノテーション作業のうち，「語り性」の観点付与について取り上げる．書籍サンプルからランダムに選び出した約 500[1]のサンプルを 1 セットとし，「語り性あり」「語り性なし」のどちらかに 3 人の作業者の判断が一致したサンプル（本稿では「典型例」と呼ぶ）を分析することにより，どのような観点に基づいて「語り性」の有無が判断されているかを分析する．複数作業者の判断が一致したサンプルをもとに分析を行うことで，「語り性」の観点付与に有効な指標が抽出可能と考えられる．また，あわせて，「語り性」を持つと判断されたテキストを，別のセットの作業で「話しことば的」と判断されたテキストと比較し，両者の異同を確かめる．

２．「語り性」について

書籍テキストの中には，著者が読者に対して直接語りかけていると解釈できる文体がある．

以下の例は，読者に語りかけている文体のサンプルである．

例１）

> ときには一流ホテルのロビーでお茶を飲んで、ゴージャスな盛り花を見てきましょう。物ごとの上達は真似ることから始まるのです。真似るは学ぶの語源だそうです。真似を自分のものにしたいから、夢中になれるのではないでしょうか。
>
> 結婚式などフォーマルなお祝い事がつづくことがあります。そのたびにドレスを新調するのはたまったものではありません。かといって、いつもおなじものでは気がひけます。ある時期から黒い慶弔両用のドレスに、生花のコサージュをつけて出席することに決めました。
>
> コサージュは、お花屋さんでつくってくれます。あなたがすでにお花に詳しいなら、一〜二回の講習会で自分でつくれるようになれます。
>
> （LBo1_00017『ひとりって楽しい』）

[†] yasuda_s@ninjal.ac.jp

[1] セットごとに 456〜485 の幅があり，本稿で扱ったセットは 485 サンプルである．

例２）

> ですから、私たち親としてできることは、子どもの自律神経のバランスを整えるために、食事と睡眠を規則正しくとれるように注意することです。そのためには親子で食事を一緒にとるためにやりくりすることが大切です。親の都合を優先しておいて後から、基本的生活習慣への不安をもち、子どもに“早寝早起き”“三度の食事”を励行させても、それは一朝一夕にはできませんよね。　（LBo3_00103『ストレスから子どもを守る本』）

「あなた」や「みなさん」などの呼びかけ表現や，「でしょう」「ではないでしょうか」といった問いかけや相づちを求めるような文末表現など，「直接的な語り」表現と呼べるような表現が含まれるテキストを，本稿では「語り性を有するテキスト」と呼ぶ（柏野 2010）．

一方，書籍テキストの中には，「話しことば的」として解釈できる文体もある．会話文のみで地の文がない場合，戯曲調で地の文が卜書きの場合，講演会の書き起こし，一人称小説のようなサンプルである．本稿では，これらを「話しことば的」の典型例と呼ぶ．

以下，本稿は，「語り性」を有するテキストの文体的特徴を分析することにより，作業者が「語り性」を有すると判断する際の言語的な指標を抽出することを目的とする．

３．「語り性」に関する先行研究

文章を分類する試みの一つとして，石田（2003）の，量・構文・位置・表現・内容に関する様々な指標の提示がある．小磯ほか（2008）は，テキストに含まれる品詞率，語種率，異なり語率，文の長さなどの 12 の指標を選定してテキストの分類を行っている．また，前坊（2011）のように副詞と文末表現をとりあげて文書分類を行う研究も見られる．

本稿では文章の持つ「語り性」の観点を分析するため，「呼びかけ表現」や「文末表現」を抽出する．そこで，形態素解析を行い，品詞，活用形，語彙素の出現率を調べた．出現率の高い要素を抽出することにより，作業者が「語り性」の観点で分類を行う際に用いる指標を分析することができると考える．

また，小磯ほか（2011）は，調査者から得た評定語を指標として分析を行っている．そのとき，「書きことば的―話しことば的」という尺度に，「読み手に語りかける―語りかけの少ない」という尺度や，改まりの程度などの複数の観点が関与する可能性を示している．

そこで，本稿の「語り性」の観点分析にあたっては，「語り性」があるという分類と「話しことば的」であるという分類に差があるのかという点についても確かめることとした．「語り性」があると作業者が判定を行ったサンプルのセットと，「話しことば的」であると作業者が判定を行ったサンプルのセットを対照し，「語り性」と「話しことば的」の差を調べる．

４．調査データ

本稿で扱うのは，BCCWJ の図書館サブコーパスに含まれる書籍からランダムに選んだ 485 サンプルのセットである．このうち，398 サンプルが作業者 3 人全員に分類対象[2]として選ばれており，分析にあたっては，全作業者が分類を行ったこれらのサンプルを調査対象とする．

また，「語り性あり」「語り性なし」の判断が作業者全員で一致したサンプルは 80%であり，内訳は，「語り性あり」が 6.28%（25 サンプル），「語り性なし」が 73.87%（294 サンプル）である．作業者判断が一致していることから，これらのサンプルはそれぞれ「語り性」

[2] 対談，Q&A 形式，図解，用語解説など形式的に特徴のあるサンプルは，今回は分類対象外（非対象）とされた．作業者は，分類対象としたサンプルのみ観点付与を行っている．

の有無に関する典型例であると考える．「話しことば的」の典型例としては，「話しことば的」か「書きことば的」かの観点でアノテーションを行った 1,890 サンプル中，「話しことば的」の分類で作業者判断が全員一致した 12 サンプルを得ている．

典型例として選び出されたサンプルのセットについて，MeCab 0.98＋UniDic 1.3.12 を用いた形態素解析を行った．以下の分析結果に示す指標は，解析結果に基づく．但し，人手の観点付与作業では，「語り性」の判断に地の文のみが作業対象とされていることから，形態素解析にあたっては，会話文内の話しことばを解析対象としていない[3]．

本稿で扱うデータのセットは表１の通りである．

表１．分析対象データ

	語り性あり	語り性なし	非一致	話しことば的
サンプル数	25	294	79	12
語数	78,013	789,530	254,361	49,225

また，作業者の観点付与結果[4]の NDC 別分類を図１に表す．図１から，「語り性」があると判断されたサンプルは，NDC1 番台（哲学）と NDC4 番台（自然科学）に多い傾向が見られることがわかる．なお，図１の（）内数値は，分析対象としたセット 485 サンプルの NDC 別内訳を表す．

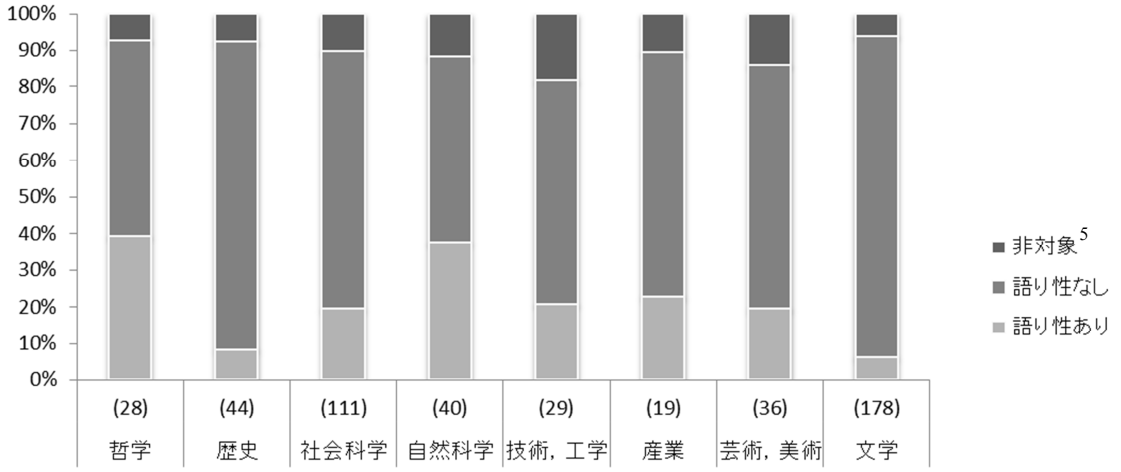

図１．NDC 別の「語り性」観点付与における作業者判断

５．語り性の観点付与結果に現れた指標

「語り性あり」「語り性なし」の典型例のセットを比較対照し，品詞と活用形，語彙素のそれぞれについて，観点付与の判断基準と考えられる指標の抽出を試みる．同時に，「語り性あり」の典型例については，「話しことば的」の典型例との対照も行い，「話しことば的」と「語り性」の差についても明らかにする．

---

[3] 「」内を会話文と見なした．固有表現や引用，強調などの部分に「」が用いられる場合も見られるが，「」は NDC9 番台に集中しており，概ね小説における登場人物の会話に用いられていることが確認された．

[4] 作業者 3 人が行った観点付与作業結果を足し合わせた．

[5] 注 2 を参照．作業者が作業対象外（非対象）と分類した割合を示す．

典型例のセットの比較にあたっては，カイ二乗検定を用いて要素ごとに有意差（有意水準 0.1%以下）の確認を行う．

５.１　語り性の有無と品詞

「語り性あり」「語り性なし」の典型例においては，出現頻度が上位である品詞では，表２のように差があまり見られない．しかし，全品詞を見ると，表３の通り，「語り性あり」の典型例は，「語り性なし」よりも終助詞，代名詞の出現率が高く，「語り性なし」では，「語り性あり」よりも固有名詞の出現率が高いという結果が現れた．

「語り性あり」との観点付与がなされた典型例は，終助詞と代名詞が多い．「語り性あり」の観点付与がなされた典型例は，相手に対する表現として，呼びかけや終助詞の付与が現れるためと考えられる．

また，「語り性なし」との観点付与がなされた典型例では，固有名詞の出現が多いといえる．NDC との関連（図１参照）を見ると，「語り性なし」の観点付与がなされた典型例は，NDC9 番台（文学：小説が多い）が 88%であり，登場人物の人名によって，固有名詞の人名が増えていることが考えられる．NDC2 番台（歴史）も「語り性なし」が 84%であり，固有名詞の地名等が増えることが考えられる．固有名詞の出現率は，ジャンルとの関連性が予測される．

なお，「語り性あり」と「話しことば的」の典型例を対照すると，終助詞，感動詞，固有名詞，代名詞のそれぞれが「話しことば的」に多いという有意差が現れていた．前述した固有名詞のほか，感動詞は，「語り性あり」の典型例よりも「語り性なし」の典型例に出現率が高い．「話しことば的」の判断には感動詞が関連していることが考えられるが，「語り性あり」の判断とは関わりが低い可能性がある．

表２．「語り性」の有無と品詞（頻度上位 10 位）

「語り性なし」典型例			「語り性なし」典型例		
助詞-格助詞	127,879	16.20%	助詞-格助詞	12,577	16.12%
名詞-普通名詞-一般	113,407	14.36%	名詞-普通名詞-一般	10,516	13.48%
助動詞	64,107	8.12%	助動詞	7,225	9.26%
動詞-一般	49,345	6.25%	動詞-非自立可能	5,526	7.08%
動詞-非自立可能	47,863	6.06%	動詞-一般	4,768	6.11%
補助記号-読点	41,238	5.22%	名詞-普通名詞-サ変可能	4,459	5.72%
名詞-普通名詞-サ変可能	35,938	4.55%	補助記号-読点	4,311	5.53%
助詞-係助詞	32,007	4.05%	助詞-接続助詞	3,614	4.63%
助詞-接続助詞	31,725	4.02%	助詞-係助詞	3,436	4.40%
補助記号-句点	30,750	3.89%	補助記号-句点	2,651	3.40%

表３．「語り性」の有無，「話しことば的」の品詞（抜粋）

品詞	「語り性なし」典型例		「語り性あり」典型例		「話しことば的」典型例	
終助詞	1,667	0.21%	384	0.49%	1,145	2.28%
感動詞	506	0.06%	32	0.04%	404	0.80%
固有名詞	21,191	2.68%	436	0.56%	357	0.71%
代名詞	9,586	1.21%	1,133	1.45%	1,014	2.02%

５.２　語り性の有無と活用形

表４は，出現率に有意差の見られた活用形である．「語り性あり」「語り性なし」の典型例で，意志推量形（「～でしょう」「～だろう」など），命令形（「～ください」「～おけ」など）の出現率に差があった．「語り性あり」のテキストでは，読み手に対する表現が用いられているため，意志推量形と命令形の出現率が高いことが考えられる．但し，これらは「語り性あり」と「話しことば的」では差がない．

また，「語り性あり」の典型例には，「語り性なし」と比較すると，融合（例：「～じゃない」「～なきゃ」など）の出現率も高い．しかし，「語り性あり」と「話しことば的」を対照すると，融合は「話しことば的」の典型例で出現率が高いという結果が現れている．「語り性あり」は「話しことば的」とは同一の判断基準ではないことが示唆されよう．

表４．「語り性」の有無，「話しことば的」の活用形（抜粋）

活用形	「語り性なし」典型例		「語り性あり」典型例		「話しことば的」典型例	
意志推量形	1,895	0.24%	360	0.46%	239	0.48%
命令形	442	0.06%	79	0.10%	61	0.12%
撥音便	1,793	0.23%	325	0.42%	199	0.40%
融合	53	0.01%	29	0.04%	107	0.21%

５.３　語り性の有無と語彙素

図２の語種の割合を見ると，「語り性あり」の典型例と「語り性なし」の典型例には有意差がみられることがわかる．特に，「語り性なし」では前述の固有名詞の頻度が高いことから，固有語の割合が高く，また，会話文の補助記号（「」）をはじめとする記号の割合が高い．「語り性あり」の典型例では，「語り性なし」との違いとして，和語の多さがあげられる．

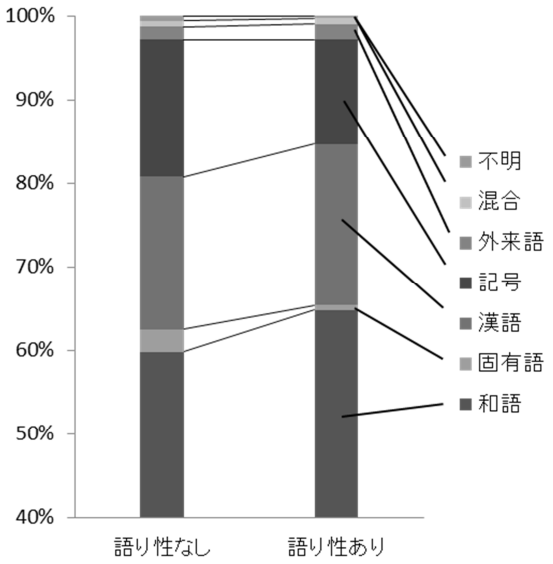

図２．「語り性」の有無と語種

「語り性なし」の典型例と対照すると，「語り性あり」の典型例にあきらかな語彙素は，以下であった．表５に，出現率と「語り性なし」「話しことば的」の典型例との対照を併せ

て示す.

- 「た」・「です」・「ます」（助動詞）が多い
- 「か」・「ね」・「よ」（終助詞）が多い
- 「私」（一人称代名詞）・「貴方」（二人称代名詞）・「自分」（名詞）が多い
- 「の」（準体助詞）・「事」（名詞）が多い
- 「「」・「」」（記号）が少ない

表５.「語り性」の有無,「話しことば的」の語彙素（抜粋）

語彙素	「語り性なし」典型例		「語り性あり」典型例		「話しことば的」典型例	
た（助動詞）	26,939	3.41%	1,285	1.65%	1,018	2.02%
です（助動詞）	388	0.05%	926	1.19%	591	1.17%
ます（助動詞）	698	0.09%	1,051	1.35%	406	0.81%
事（名詞）	4,368	0.55%	820	1.05%	316	0.63%
「（補助記号）	10,285	1.30%	406	0.52%	578	1.15%
の（準体助詞）	5,849	0.74%	825	1.06%	882	1.75%
自分（名詞）	819	0.10%	193	0.25%	125	0.25%
私（代名詞）	946	0.12%	146	0.19%	177	0.35%
貴方（代名詞）	55	0.01%	34	0.04%	14	0.03%
君（代名詞）	7	0.00%	5	0.01%	16	0.03%
か（終助詞）	1,192	0.15%	217	0.28%	245	0.49%
よ（終助詞）	107	0.01%	35	0.04%	280	0.56%
ね（終助詞）	95	0.01%	67	0.09%	286	0.57%

「語り性あり」の典型例には，助動詞の「です・ます」が高い出現率で見られている．しかし,「語り性あり」と「話しことば的」を対照すると,「です」には有意差がないが,「ます」は「語り性あり」の典型例に出現率が高いという違いがある．助動詞の「た」と各種終助詞は,「話しことば的」の典型例に出現率が高い.「語り性あり」の判断に「ます」が関わっており，また,「た」や終助詞は「話しことば的」の判断指標とされている可能性がある．

「語り性あり」における終助詞の多さは，５．１で見た通りであるが，語彙素レベルで見ると，終助詞「か」が「語り性なし」の典型例でも，0.15%の出現率と上位頻度語（55 位／異なり語彙素数 29,790）ながら,「語り性あり」の典型例では，0.28%の出現率となっており,「語り性あり」により多く見られる語彙素であるとわかる．

代名詞についても，５．１で見た通りといえるが,「語り性あり」の典型例と「話しことば的」の典型例では，一人称・二人称代名詞のそれぞれで,「私」のみ「話しことば的」に多く現れることを除き，ほぼ出現率に差がないという結果が現れている.「話しことば的」の判断指標として，一人称代名詞，二人称代名詞が用いられている可能性が考えられる．

また,「～なのだ」のように用いられる準体助詞の「の」や,「～という事」のように用いられる名詞の「事」が,「語り性あり」の典型例に出現率が高いこともわかった．上位頻度語という点では,「語り性なし」にも（「の」20 位：0.74%,「事」22 位：0.55%）多く現れる語彙素であるが,「語り性なし」の出現率を上回っている．とくに「語り性あり」では,「事」の出現率が高い.

なお,「語り性なし」の典型例には，NDC9 番台が多い（小説の会話文が多い）ため，補助記号も多く現れるが，前述の通り，語り性の判断は地の文で行われていることから，直接的な判断指標であるとは言い難い.

６．まとめ

既にアノテーション作業が完了しているサンプルを用い,「語り性」の観点付与に関して,複数作業者の判断が一致した典型例の分析を行った．品詞，活用形，語彙素の出現率を調べ，出現率の高い要素を抽出することで,「語り性」の有無について観点付与を行う作業者が，分類に用いている可能性の高い指標が整理された.

但し,「語り性」観点付与における作業者間の判断基準の差について,考慮しておきたい.ここまで，全作業者の判断が一致した典型例の分析を行って指標を得たが，全作業者の判断が一致しなかったサンプルもある．作業者の判断に個人差があり，作業者によっては判断基準とならない指標がある可能性が考えられる．よって，語り性の有無の観点で 3 人の作業者の判断が一致しなかったサンプル（79 サンプル）との対照を行い，作業者判断に揺れのある例のセットで，抽出した指標の出現率を確かめた.「語り性あり」の典型例に出現率が高くなければ，判断基準としての効果が低い可能性もあるといえよう．表６に対照結果を示す.

表６.「語り性あり」と作業者判断非一致サンプルとの対照

指標	語り性あり		非一致		高頻度
た(助動詞)	1,285	1.65%	6,889	2.71%	非一致
です(助動詞)	926	1.19%	1,416	0.56%	語り性あり
ます(助動詞)	1,051	1.35%	1,909	0.75%	語り性あり
事(名詞)	820	1.05%	2,150	0.85%	語り性あり
の(準体助詞)	825	1.06%	2,690	1.06%	有意差なし
自分(名詞)	193	0.25%	400	0.16%	語り性あり
貴方(代名詞)	34	0.04%	70	0.03%	語り性あり
君(代名詞)	5	0.01%	24	0.01%	有意差なし
か(終助詞)	217	0.28%	664	0.26%	有意差なし
よ(終助詞)	35	0.04%	170	0.07%	非一致
ね(終助詞)	67	0.09%	118	0.05%	語り性あり
意志推量形	360	0.46%	899	0.35%	語り性あり
命令形	79	0.10%	202	0.08%	有意差なし

結果として,「語り性」観点付与のための指標は，以下が得られた.

「語り性あり」
　活用形　　：意志推量形（「～でしょう」「～だろう」など）が多い
　語彙素　　：「です（助動詞)」「ます（助動詞)」「事（名詞)」が多い
　　「あなた（代名詞)」「自分（名詞)」「ね（終助詞)」が多い

「語り性なし」
　品詞・語種　：固有語が多い
　語彙素　　：「た（助動詞)」が多い

また，「語り性」の分類にあたり，作業者が「話しことば的」の観点とは異なる指標で判断を行っている可能性も示唆された．感動詞，融合（「〜じゃない」「〜なきゃ」など），終助詞「よ」のように，「話しことば的」で出現率が高くとも，「語り性あり」の典型例と比較すると，「語り性あり」では出現率が低いという要素があるためである．作業者の「語り性」の分類判断は，「話しことば的」との差異を含め，複雑な条件によって行われているものと考えられる．副詞や接続詞などの品詞毎に詳細な分析を行うことで，出現率が低くとも，判断に用いられる指標が得られる可能性もある．さらに，その他の観点の分析結果とあわせるなどの分析も必要であろう．

今後は，これらの指標とともに典型例を提示したマニュアルを作成し，マニュアルに沿ったアノテーション作業を進めることを予定している．観点付与にあたり，作業者が指標を参照することで，作業の効率化が見込めるとともに，有用なデータの作成が期待される．

謝　辞

本研究は，国立国語研究所の共同研究プロジェクト「テキストの多様性を捉える分類指標の策定」に基づくものです．また，BCCWJ の構築は，文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」（平成 18〜22 年度，領域代表者：前川喜久雄）による補助を得たものです．

文　献

石田栄美(2003)「テキストの自動分類に関わる諸要素」『日本図書館情報学会誌』49(2), pp65-78.

柏野和佳子(2010)「「直接的な語り」という表現スタイルをもつ書籍テキストの人手抽出の試み」『ことば工学研究会』35, pp.63-72.

柏野和佳子，奥村学(2012 予定)「書籍テキストへの分類指標人手付与の試み―『現代日本語書き言葉均衡コーパス』の収録書籍を対象に―」『言語処理学会第 18 回年次大会』B5-6.

小磯花絵，小木曽智信，小椋秀樹，冨士池優美，宮内佐夜香(2008)「『現代日本語書き言葉均衡コーパス』にもとづくジャンル間の文体差に関わる要因の分析」『社会言語科学会第 22 回研究大会発表論文集』pp.192-195.

小磯花絵，田中弥生，小木曽智信，近藤明日子(2011)「評定実験に基づくテキスト分類尺度の体系化の試み」『現代日本語書き言葉均衡コーパス』完成記念講演会予稿集，pp.47-52.

前坊香菜子(2011)「雑誌コラムに現れる語彙とモダリティ―副詞と文末表現を中心に―」『信学技報』，pp.55-60，電子情報通信学会

関連 URL

国立国語研究所の言語コーパス整備計画 KOTONOHA　http://www.ninjal.ac.jp/kotonoha/

特定領域研究「日本語コーパス」　http://www.tokuteicorpus.jp/