アノテーターコメントを用いた「語りかけ性」分析の試み

—頻度情報から捉え難いテキスト性質の解明に向けて—

保田祥† 柏野和佳子‡ 立花幸子† 丸山岳彦‡
†国立国語研究所 コーパス開発センター ‡国立国語研究所 言語資源研究系

1. はじめに

『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(全10,551サンプル・28,892,944語)に,文書分類の観点から人手で情報を付与する作業を実施した(柏野・奥村2012).付与した情報は,対象とする読者(幼児・小学生～専門家:5段階)・文章の硬軟(とても軟らかい～とても硬い:4段階)・くだけているか(とても・どちらかといえば・くだけていない:3段階)・語りかけ性(とてもある・どちらかといえば・特にない:3段階)・小説の主人公あるいは語り手の人称・小説以外の文章の内容(とても客観的～とても主観的:4段階)である.

本稿では,このうちの「語りかけ性」と呼ぶ観点をとりあげ,人がどのように分類の判断をしているのか,その要因を調査する.「語りかけ性」は,テキストに現れる個別の要素の頻度から特徴が得にくいテキストの性質の一つである.

直感的に「語りかけ性」とは,書きことばではあるが,語りかけられている印象を受ける表現を含むと考えられる.但し,それらは「話しことば的」と作業者が判断する表現[1]とも異なる(保田ほか,2012a).どのような表現が「語りかけ性」の有無の分類根拠となるのか探るため,テキストに出現頻度の高い(特徴的な)表現を調査した.しかし,短単位レベルの頻度調査では,分類群に明らかな特徴が少なく,出現頻度の高い表現が含まれないテキストでも「語りかけ性」があると判断される場合がある(保田ほか,2012b).認知構造は素性や成分の束ではないことが言われており(e.g., Lakoff, 1987; Taylor, 1989),実際,テキストに出現する個々の要素からは指示物の全体像が得にくい(Yasuda, et.al., 2012).そこで,人の判断根拠を確かめるべく,アノテーターの「語りかけ性」を有すると分類した理由に関するコメント内容(保田ほか,2012c)の分析を試みた.

結果,「語りかけ性」は,特徴的な表現の多寡よりも,むしろテキスト全体から受け取られるものであり,いわゆるハウツー本のような教示的態度を強調するテキストでみとめられる傾向があるとわかった.

2. 「語りかけ性」とは何か

「語り(物語)」の特徴として,「歴史的現在形(historical-present)」の出現頻度が高いということが言われている(e.g., Shiffrin, 1981; Silva-Corvalan, 1983; 池上, 1986).この特徴は,もちろん語りかける表現に関係していると考えられるが,本稿でいうテキストの「語りかけ性」は,物語における「語り」に留まらない.

小磯ほか(2011)は,調査者から得た評定語を指標としてテキスト分析を行う際,「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」という尺度を含む複数の観点が関与する可能性があると示し,「語りかけ性」に関する尺度の有用性を考慮する.また,安藤(2012)は,小説における再現的提示の手法とは,二人称的世界が顕在しないことであると示し,読み手に語りかける言文一致の形がありえたならば,「言」に近い文体が創出されたかもしれないと述べる.すなわち「語りかけ性」は,既存の「言文一致」の範疇にない表現ということになるのだろう.

柏野(2010)は,「語りかけ性」を「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」と呼べるような表現が含まれるテキストを,「語りかけ性を有するテキスト」と呼ぶ.以下のような例が,「語りかけ性」があるとして,複数作業者の判断が一致した[2]テキストの典型例である.特徴的と考えられる表現に下線を引いた.

[1]「話しことば的」と判断されるサンプルは,調査を行った1,890サンプル中12サンプルに留まった.また,「語りかけ性」があるサンプルと対照すると,「話しことば的」サンプルにのみ出現率の高い要素として,感動詞・融合(「～じゃない」「～なきゃ」など)・「よ(終助詞)」などが突出する(詳細は,保田ほか,2012a).

[2] 約3,000サンプルについて,作業者3人全員の判断の一致率を確かめたところ,「語りかけ性」の有無についての判断は,80%で3人が一致し(保田ほか,2012a),判断に個人差は少ないと考えられる.

例）お金を稼ぐために事業を始めるべきでないとしたら、なぜ事業を始めるのでしょうか。答えはあなたの情熱と夢にあります。お気に入りの趣味として事業を始めることを考えることができますか。それはほんの少数の人たちにしか理解できない夢です。なぜかって。まず第１に、たいていの人たちがそんなことが可能とさえ思っていないからです。(『世界一わかりやすいほんとうのお金持ちになる法』)

3. データ

BCCWJの図書館サブコーパスに含まれる書籍（10,551サンプル）をランダムに並べ替え，6人の作業者が文書分類を行った結果を用いた．調査にあたっては，作業結果から約半数をランダムに選び（5,652サンプル[3]），会話文を含む場合の多い小説を全て除いたサンプル(3,750サンプル・11,630,970語）を調査対象データとした．また，判断時に用いた表現や印象などが，備考欄へコメントとして記述されている場合がある．作業者によって記述量にはばらつきがあるが，「語りかけ性」に関しては作業者ごとにそれぞれの作業サンプル数の 2%～5%のコメントを得ている．

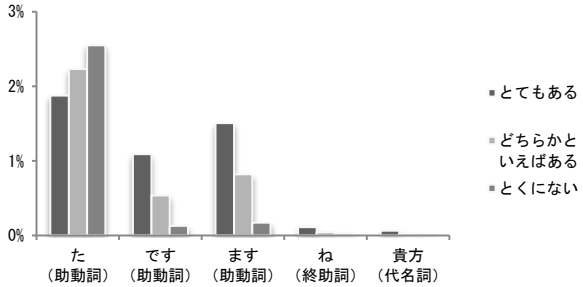
「語りかけ性」についてのアノテーションは，作業者が「とても（語りかけ性が）ある」「どちらかといえば（語りかけ性が）ある」「とくに（語りかけ性は）ない」の三種類の選択肢から該当すると判断した一つを選択する．作業の結果，「とてもある」は486サンプル（1,387,665語・本稿で扱うサンプルの13%），「どちらかといえばある」が805サンプル（2,347,671語・本稿で扱うサンプルの21.5%），「とくにない」が2,459サンプル（7,895,634語・本稿で扱うサンプルの65.5%）得られた．

サンプルの形態素解析には，MeCab 0.993＋UniDic2.1.0を用いた．分析結果に示す品詞情報や語彙素等の要素は，解析結果に基づく．

4. 「語りかけ性」を探る

4.1. 出現頻度に見る特徴的表現

作業者によって「語りかけ性」の有無で分類されたサンプル群のそれぞれに，高頻度に現れる[4]表現（保田ほか，2012a）を確認しておく．但し，「語りかけ性」の有無で分類されたサンプル群において，形態素解析結果から有意に差異の見られる個別の要素は僅少である．本稿で扱うデータについても，結果は図1のように現れた．しかし，新たに分類群毎の特徴的な要素などは得られていない．



【図1.「語りかけ性」の有無による分類群各々における有意差が期待される語の出現率】

「語りかけ性」がないサンプル群で「た」文末が多く見られることは，反対に「語り」として特徴的な非過去形が「語りかけ性」があるサンプル群に見られるためと推測され，「語りかけ性」は「語り」に類した表現も含むと考えられる．その他，「あなた」のように呼びかけと認識される代名詞や確認などの終助詞である「ね」，読み手に対する敬体としての「です」「ます」などで差異が見られることは当然であろう．むしろ，直感的に「語りかけ性」を受けると考えられる，先の例にあった「のでしょうか」「できますか」「なぜかって」のような問いかけなどは，「語りかけ性」の有無の分類で有意差が見られていない．

4.2. 人が判断根拠と認識する表現

短単位レベルの要素の頻度集計のみでは，得られる要素は直感的な判断と大差なく，「語りかけ性」を有するテキストに特徴的な要素が網羅できたとは言い難い．そこで，アノテーターのコメントを用い，実際に「語りかけ性」があるとの判断に用いたとする根拠の調査を試みた[5]（保田ほか，2012b）．

アノテーターのコメントに個別の要素として見られたのは，まず，読み手との一体感を生じる「私たち」「我々」のような人称がある．しかし，「単純な一人称複数ではない」と併記された場合もあり，

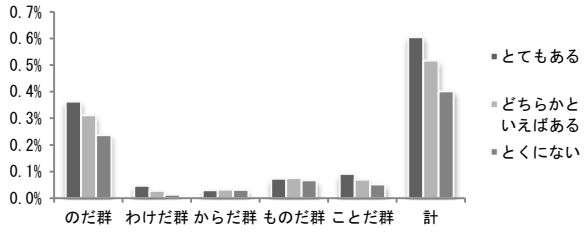
[3]対談，座談会をはじめ，Q&A形式，図解，用語解説など形式的に特徴のあるサンプルは，分類対象外（非対象）とされ，本サンプル数には含まない．アノテーション作業者は，分類対象としたサンプルのみ観点付与を行っている．

[4]図書館サブコーパスからランダムに選び出した約500のサンプルを1セットとし，「語りかけ性」があるとして3人の作業者の判断が一致したサンプル（約400）の分析を行った結果から，品詞・活用形・語彙素において，すべての要素の出現頻度について検定を行い，有意差の見られた表現を取得した．

[5]作業者は，「語りかけ性」が「どちらかといえばある」という判断を行った際，コメントを記述する傾向がある．「とてもある」と判断されたサンプルにコメントがある場合は，「明らかにあなたに語りかけている体」などの記述にとどまっている．「とてもある」とまでは言い難いが，「語りかけ性」があると感じた場合に「どちらかといえばある」を選択し，その判断根拠を示すものと考えられる．

解析結果のみからは取得しにくい[6]要素といえる．また，着目されやすい表現でも，判断根拠に用いたとする例は単独で挙げられるのでなく，複数の種類（例：「のである」「からです」「ものだ」など）が並列的に例示され，この種の表現が「多い」ため，語りかけられている感じがした旨が記述されていた．テキストの総体から「語りかける」印象を得る可能性が指摘できる（保田ほか，2012c）．

そこで，「多い」として並列されていた表現の出現頻度を確かめた．結果を示した図 2 から，表現[7]によっては「語りかけ性」が「とてもある」サンプル群に出現割合が低いか他群との違いが少ないが，総計として「語りかけ性」があると判断されたサンプル群に，出現率が高くなっていることがわかる．



【図 2.「語りかけ性」の有無による分類群における
アノテーターが多いと感じた表現の出現率例】

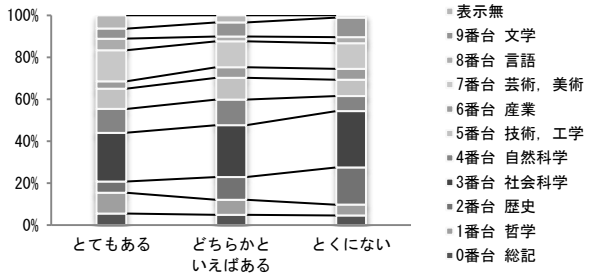
このほか，読み手の存在や判断を想定した表現（「いただく」「申し上げる」「あげる」「ください」など）が複数現れると，「語りかけ性」があるという印象になるようである．また，読み手にとって相手（書き手）の存在が認識されると推測されるような，評価に関する表現（「よい」「悪い」「大切」「便利」など）や可能（「できる」など），主観的かつ婉曲的な主張（「～と思う（見える／感じ）」「はず」など）が判断根拠とされている例も見られる．

しかし，アノテーターが「語りかけ性」があると判断するに用いたと認識する要素は，「語りかけ性」を形成する表現と言えるが，個別の出現頻度では影響が捉え難い．そもそも出現頻度を確認することも難しい．まとまった量のテキストにおいて，種々の表現の総体的な出現量と，文脈が要されるためである．以下に示すのは，直感的に，あるいは形態素解析結果でサンプル群間の出現頻度に有意差のあった表現を含まないが，「語りかけ性」があると判断されるテキスト例である．アノテーターのコメントに示される種類の要素（関連すると考えられる要素に下線を施した）が散見されることがわかる．

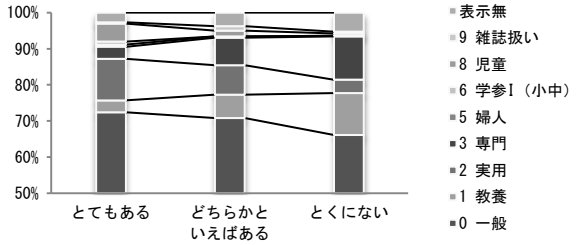
例）カップリングコンデンサが大きい場合、オレンジ色の側の配線が同じようにＩＣソケットの足にハンダ付け**できれ**ばどのように付けても**構わない**。完成図を見て**もらえれ**ば分かる**と思う**が、コンデンサの左の部分は大きくスペースが残してあるので、アキシャルリードのものも基板上に取り付け**可能**だ。また、大きすぎて基板からはみ出したとしても、特に**問題はない**。
なお、後で**説明する**が、このコンデンサは無しにも**出来る**。（『はじめてつくるプリアンプ』）

4.3. テキスト総体の情報

総体としてテキストの特徴は，書籍のジャンルやタイトルからも捉えることができると考えられる．そこで，「語りかけ性」があると判断されるテキストの現れる書籍の NDC 分類（ジャンル）と C コード（販売対象）及び，タイトルを確認した．



【図 3「語りかけ性」の有無と NDC 分布】



【図 4「語りかけ性」の有無と C コード分布】

図 3 で，「語りかけ性」による分類群ごとの NDC 分布を割合で示した．「語りかけ性」が「とてもある」群から「とくにない」群では，哲学（1 番台）・自然科学（4 番台）で分布割合が減少する．反対に，歴史（2 番台）・文学（9 番台）で増加する．「語りかけ性」は，自然科学分野の書籍で用いられやすい．

同様に，図 4 で，「語りかけ性」による分類群ごとのCコード分布を割合で示した．「語りかけ性」が「とてもある」群から「とくにない」群では，教養（1）・専門（3）の割合が増加しているのに対し，

[6] 「我々」の出現頻度を見ると，「とてもある」群で 0.018% ，「どちらかといえばある」群で 0.018%，「とくにない」群で 0.029%と，「とくにない」群でむしろ出現頻度が高い．

[7] ここでは，「のである」「のです」「のだ」などをまとめて「のだ」群（他も同様）とする．

実用 (2)・児童 (8) で減少していることがわかる．すなわち，実用書と児童書[8]に「語りかけ性」が出現しやすいのだと考えられる．NDC分類における自然科学分野に現れやすいのも，実用書が多い[9]ためとの推測が可能である．

また，実用書とは，いわゆるハウツー本（啓蒙書・指導書）の類と推定され，書籍のタイトルは，その内容を代表して示すものと考えられる．そこで，「語りかけ性」が「とてもある」と判断されたサンプル（486 サンプル）の書籍タイトルを確認した．

内訳は，内容判別のできないタイトルが 200 サンプル(41%)，判別可能なタイトルが 286 サンプル(59%)であり，判別可能なタイトルのうち，物語が 6 サンプル，ハウツー本であることが明記[10]されたタイトルが 233 サンプル（「とてもある」サンプルの 48%．例：『目で見るパパとママの小児科入門』『リクガメが 100%喜ぶ飼い方遊ばせ方』『商標登録の実務がよくわかる本』など），ハウツー本であることが推測されるタイトル[11]が 47 サンプル（例：『乳酸菌パワーダイエット』『ひざの痛みをとる・治す』など）あった．「語りかけ性」が「とてもある」テキストは，約半数がハウツー本であるとタイトルにあきらかな書籍のサンプルだと言える．「語りかけ性」は，ハウツー本に用いられやすく，教示的な態度を示す表現手法の効果であると考えられる．

5. まとめと「語りかけ性」の再定義

テキストに出現する高頻度語からは判別し難くとも，読み手の認知するテキストの性質が存在する．

本稿は，人がどのような根拠をもとにテキストを分類するのか，「語りかけ性」があると判断した作業者のコメントから，ある種の表現群が文脈によって「語りかけ性」を与えていることを明らかにした．また，その性質は，テキストの総体に関わるため，書籍のジャンル分類や書籍タイトルに現れる傾向があることも確かめた．

「語りかけ性」のあるテキストとは，書きことばでありながら，読み手が「語りかけ」られていると感じるテキストである．現在形を多用するなど「語り」に特徴的な表現を含み，読み手への呼びかけや確認，敬体が頻出する．但し，特徴的な個別の要素（出現頻度）で捉えられるとは言い難く，文脈上，読み手の存在や判断を想定していると示す表現や，語り手の存在が推測される表現が複数現れることにより，総体的に生じるものでもある．実用書（ハウツー本）のように，教示的な態度を明らかにする際に用いられやすい傾向が見られる．

[8] 児童書は NDC 分類がほぼ全て不明（表示無）であるため論じないが，「とてもある」群で，後述する書籍タイトルに「物語」が含まれる例が 11%，「なぜ」「やさしい～」「おたすけ」などのハウツー本と推測される例が 22%見られる．

[9] NDC3・4 番台の 12%である．一般書が 72%と大部分であるため，C コード分類内では突出しているといえる．

[10] ～の本・～法・～方・入門などのほか，疑問文・命令文などがある．

[11] 明確な指標を含まないタイトル．

参考文献

安藤宏(2012)『近代小説の表現機構』岩波書店.

池上嘉彦(1986)「日本語の語りのテクストにおける時制の転換について」『記号学研究』 6(25), 61-74.

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)「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，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.

Lakoff, George. (1987) . *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago, University of Chicago Press.

Schiffrin, D. (1981) “Tense variation in narrative”. *Language*, 57(1), 45-62.

Silva-Corvalán, C. (1983) “Tense and aspect in oral Spanish narrative - context and meaning”. *Language*, 59(4), 760-780.

Taylor, John. R. 1989. *Linguistic categorization: Prototypes in linguistic theory*. Oxford: Clarendon Press.

保田祥，柏野和佳子，立花幸子，丸山岳彦(2012a)「「語り性」を有する書きことばの典型例の分析」『第 1 回コーパス日本語学ワークショップ』予稿集，pp. 139-146.

保田祥，柏野和佳子，立花幸子，丸山岳彦(2012b)「「語りかけ性」を有すると判断される書きことばの表現」『第２回コーパス日本語学ワークショップ』予稿集，pp. 43-50.

保田祥，柏野和佳子，立花幸子(2012c)「総体として印象を与える表現：「語りかけ性」を有すると判断する根拠」『ことば工学研究会』41, pp. 3-10.

Yasuda, S., Okamoto, M. & Aramaki, E. (2012) Ad hoc creature: Lost and added in translation from description to depiction, CogSci 2012.