小説のセリフの分析を支援するツールと言語資源

佐藤理史
名古屋大学大学院工学研究科
sato.satoshi.g9@f.mail.nagoya-u.ac.jp

概要

本論文では、小説の会話文の文末形式を自動認定するシステム Kohaku、このシステムを利用して「BCCWJ 小説会話文」に文末形式を付与したデータ Kohaku-BCCWJ、および、小説の会話文において一定の頻度が観察される 1,923 種類の文末形式のリスト Kohaku-FFL を紹介する。さらに、これらを利用した「セリフの書き分け」の実態調査の実例と、データに基づく典型的な口調の探求について述べる。

1 はじめに

発話には話者の個性が投影され、それが言葉遣いに現れる。日本語の小説では、この性質を利用した「セリフの書き分け」が広く用いられている。

話者のキャラクタ性 (特徴・個性) は言葉遣いの色々な側面に現れるが、日本語では、終助詞を中心とした文末形式がその中心的な要素となる。よく知られているように、どのような終助詞を好んで使用するかには、性差が存在する [1, 2, 3]。現代社会では、使用される終助詞の性差は縮小傾向にあるが [4]、小説やマンガなどでは、キャラクタ性に基づく文末形式の選択的使用が積極的に利用されており、読者に登場人物の人物像を伝える一つの手段となっている。

我々日本語母語話者は、文末形式と想起されるキャラクタ性の関係を直感的に理解している。その関係は、役割語研究 [5, 6] 等により、定性的にも明らかにされている。しかしながら、話者のキャラクタ性と文末形式の関連性に焦点を当てた分析は、人手による小規模なもの [7, 8, 9, 10] が中心で、データによる裏付けは必ずしも十分とは言い難い。その理由のひとつは、コンピュータによる分析支援ツールが存在しないことにあると思われる。

このような背景より、筆者は、小説の会話文の文末形式を自動認定するシステム Kohaku[11] を作成した。同時に、このシステムを利用して「BCCWJ 小説会話文」[12] に文末形式を付与したデータ Kohaku-BCCWJ [13]、および、小説の会話文において一定の頻度が観察される 1,923 種類の文末形式のリスト Kohaku-FFL [14] を作成・公開した。本論文ではこれらのシステムと言語資源、および、これらを用いた小説セリフの分析について紹介する。

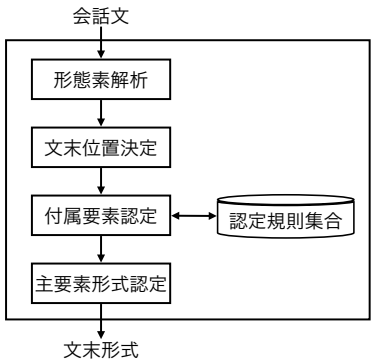

図 1 文末形式認定システム Kohaku の構成

2 文末形式認定システム Kohaku

Kohaku[11] の構成を図 1 に示す。

2.1 文末形式の認定処理

Kohaku は、与えられた会話文を形態素解析した後、以下の処理を経て、文末形式を決定する。形態素解析には Sudachi [15] を使用する。

1. 文末位置の決定
   会話文の末尾から先頭に向かって形態素列を走査し、品詞が「補助記号」以外の形態素を文末位置 (文末の形態素) とする。Sudachi が使用する辞書では、句点、疑問符、感嘆符などに、「補助記号」という品詞が割り当てられており、この処理は、文末のこれらの要素を無視することに相当する。
2. 付属要素の認定
   認定規則集合を適用し、付属要素を認定する。付属要素は、終助詞、接続助詞、丁寧表現、特

殊表現の 4 種類に分類されており、以下の出現順序制約を満たす場合に、付属要素を認定する。

特殊表現 < 丁寧表現 < 接続助詞 < 終助詞

3. 主要素形式の認定
   文末の付属要素以外の最も文末に近い形態素を主要素とみなし、この形式を認定する。

2.2 付属要素の認定

文末形式を自動認定するためには、以下の 2 点を定める必要がある。

- どのような範囲の形式を、文末形式に含めるか？
- 出現形をどこまで区別し、どこから正規化 (集約) するか？

付属要素の認定では、終助詞、接続助詞、丁寧表現、特殊表現のそれぞれに対して、これら 2 点を実態に則して定め、認定規則を定義した。認定規則は形態素列パターンで記述されており、その一覧は、文献 [11] に示されている。

2.2.1 終助詞

形態素解析結果の文末の終助詞列を、ひとまとまりの終助詞と認定する。文末および終助詞列の直前の「準体助詞の」は終助詞扱いとし、かつ、出現形の「の」と「ん」を区別する。いくつかの終助詞では前方文脈を区別する。たとえば、「終助詞ね」は、直前が判定詞の場合は〈 Dね 〉[1)]、直前が名詞・代名詞・形状詞の場合は〈 Xね 〉、直前がテ形の場合は〈 Tね 〉、のように区別する。認定規則は 254 規則存在し、427 種類の終助詞を認定する[2)]。

2.2.2 接続助詞

形態素解析で接続助詞と認定される形態素を、接続助詞と認定する。ただし、述語のテ形とみなせる「接続助詞て」と「接続助詞で」は、認定しない。この他に、〈 たら 〉、〈 なら 〉、〈 のなら 〉、〈 んなら 〉、〈 ので 〉、〈 んで 〉、〈 のに 〉、〈 って 〉等を接続助詞と認定する。認定規則は 38 規則存在し、34 種類の接続助詞を認定する。

2.2.3 丁寧表現

「です」と「ます」のバリエーションを丁寧表現と認定する。この他に、〈 ください 〉、〈 Tください 〉、〈 なさい 〉等を認定する。丁寧表現として認定するのは原則として 1 単位であるが、例外的に以下の並びを認め、たとえば、〈 ください. ます. でしょう 〉を丁寧表現と認定する。

$$\left\{ \begin{array}{l} \text{ください系} \\ \text{なさい系} \end{array} \right\} + \text{ます系}$$

$$\text{ます系} + \text{でしょう系}$$

認定規則は 40 規則存在し、158 種類の特殊表現を認定する。

2.2.4 特殊表現

「だろ (う)・じゃない」などは、会話文の末尾に頻出し、本来の意味を離れて、伝達や確認要求などの機能を果たすことが多い。このため、これらの表現を特殊表現として文末形式に含める。〈 だろう 〉等、〈 じゃない 〉等に加え、〈 のだ 〉、〈 んだ 〉、〈 じゃ 〉、〈 のじゃ 〉、〈 んじゃ 〉等を特殊表現として認定する。特殊表現として認定するのは原則として 1 単位であるが、例外的に以下の並びの 2 単位を認め、〈 じゃない. だろう 〉等を特殊表現として認定する。

$$\text{じゃない系} + \text{だろう系}$$

認定規則は 33 規則存在し、97 種類の特殊表現を認定する。

2.3 主要素形式の認定

主要素形式の認定は、付属要素をまったく伴わない場合の形式を区別するために行う。具体的には、

- 文末が述語とみなせるか (会話文が述語で終わる文とみなせるか)
- 述語の場合は、どんな活用形か。特に重要なのは、活用形が文の機能に直結する、命令形・意志推量形・仮定形である。

当該形態素が活用語の場合は、これを述語とみなし、活用型と活用形に応じた ID を付与する。ただし、「助動詞だ」と「助動詞や」は、出現形に対応する ID を付与する。形状詞・助動詞語幹の場合は、「助動詞だ」が省略されている述語とみなし、〈 状X 〉を付与する。それ以外の非活用語には品詞に対応する ID を付与する[3)]。

1) 本文中では、付属要素の ID を ⟨ID⟩ のように表記する。
2) ひとつの規則で複数の終助詞 (異形) を認定するため、規則数と種類数は一致しない。なお、本論文で示す種類数は、これまでに認定を確認した種類数である。
3) 主要素形式 ID は、デバッグ目的のため、付属要素を伴う場合にも付与する。

2.4 システムの出力

与えられた会話文に対して、Kohaku は、以下の情報を出力する。

F01 文末形式タイプ
F02 文末形式 ID
F03 終助詞 ID
F04 接続助詞 ID
F05 丁寧表現 ID
F06 特殊表現 ID
F07 主要素形式 ID

本システムが認定する文末形式とは、文末形式タイプ (F01) と文末形式 ID(F02) の組である[4)]。

文末形式タイプ (F01) は、文末形式の大分類に相当する。具体的は、会話文に出現した付属要素の種別の最初の 1 文字を繋げたもの (ラベル) である。付属要素が全く出現しない場合、文末の形態素が述語の場合はタイプ「裸」、述語以外の場合はタイプ「なし ('-')」とする。

文末形式 ID (F02) は、構成要素の ID (F03–F07) から以下のように作成する。

1. 付属要素が存在する場合 — 存在する付属要素の ID (F03–F06) を出現順にドット ('.') で繋いだ表現
2. 付属要素が存在しない場合
   (a) 主要素が述語の場合 (文末形式タイプが「裸」の場合) — 主要素形式 ID (F07)
   (b) それ以外の場合 — なし ('-')

Kohaku の出力例を付録の表 A に示す。これまでに、5,247 種類の文末形式の存在を確認している。

3 文末形式リスト Kohaku-FFL

Kohaku システムの開発過程では、多くの会話文データに対してシステムを適用し、その出力を観察して認定規則集合の修正を繰り返した。使用した会話文データを以下に示す。

D1 「BCCWJ 小説会話文」[12] に含まれる 276,576 文
D2 ウェブサイト「小説家になろう」に掲載されている小説から収集した 5,297,896 セリフ [16] を文に分割した 8,879,279 文。

システム開発が完了した段階で、上記の会話文データの全てに文末形式を付与し、D1、D2 のいずれかで 0.1bp[5)] 以上の頻度で出現した 1,923 種類[6)] の文末形式を収録したリスト「小説会話文文末形式リスト Kohaku-FFL」[14] [7)] を作成した。その概要を付録の表 B に示す。

このリストには、以下の情報が収録されている。

- 1,065 種類の文末形式に対して、D1 における頻度 (回数と出現率)
- 933 種類に文末形式に対して、D2 における頻度 (回数と出現率)
- 806 種類の文末形式に対して、D1 と D2 の出現率の差 (対数比)
- 335 種類の文末形式に対して、話者の女性率[8)]

収録した 1,923 種類の文末形式は、上記 D1、D2 の会話文を 99%以上カバーする[9)]。

4 Kohaku-BCCWJ

Kohaku-BCCWJ [13] は、会話文データ D1 に対して文末形式を付与した全データである[10)]。このデータには、「BCCWJ 小説会話文」に含まれる、BCCWJ のサンプル ID、話者名、性別の情報が含まれている (付録の表 C) ので、それぞれの話者の文末形式の使用分布を取得することができる。

5 小説のセリフの分析

5.1 セリフの書き分けの調査

小説の登場人物のセリフを収集・電子化し、それぞれの会話文に Kohaku を用いて文末形式を付与すれば、その登場人物の文末形式の使用分布を取得することができる。同一作品に登場する複数の人物の使用分布を比較することにより、小説の作者がどのようにセリフを書き分けているかが明らかになる。

ここでは、2024 年の本屋大賞を受賞した『成瀬は天下を取りにいく』[17] の主人公「成瀬あかり」とその相方である「島崎みゆき」のセリフがどのように書き分けられているかを調べる。調査には、連作

4) 文末形式が F01 と F02 の組である理由は、文献 [11] を参照のこと。

5) bp は万分率を表す。1bp=0.01%
6) 文末形式タイプ「なし ('-')」を 1 種類として含む。
7) https://doi.org/10.18999/2013067 からダウンロードできる。
8) D1 に付与されている話者の性別情報を利用して算出した女性話者の比率 (%)。男性率 = 100 – 女性率
9) ここでのカバー率は、会話文の文末形式が 1,923 種類のいずれかとなる文の数を、総文数で割った値である。
10) Kohaku-BCCWJ は、国立国語研究所の中納言サイトで公開されているが、「BCCWJ 小説会話文」の一部を含むため、利用には『現代日本語書き言葉均衡コーパス (BCCWJ)』のオフライン版 (有償) のライセンスが必要である。

表 1　セリフの書き分けの実態

文末形式		女性率	成瀬あかり 頻度	%	島崎みゆき 頻度	%
終			35	17.2	65	41.1
	な	10.0	24	11.8	-	-
	のか	10.7	2	1.0	-	-
	Bな	11.0	2	1.0	-	-
	か	17.4	3	1.5	2	1.3
	Dよ	24.9	-	-	2	1.3
	Dね	26.9	-	-	3	1.9
	よ	39.9	1	0.5	15	9.5
	ね	57.4	-	-	5	3.2
	じゃん	72.5	-	-	6	3.8
	の	92.7	-	-	26	16.5
接終			4	2.0	-	-
	から. な	3.9	4	2.0	-	-
丁終			-	-	2	1.3
特終			13	6.4	5	3.2
	だろう. か	13.3	7	3.4	-	-
	じゃない. か	14.8	3	1.5	-	-
	んだ. Dよね	56.3	-	-	2	1.3
特接			2	1.0	1	0.6
	のだ. が	12.9	2	1.0	-	-
丁			-	-	6	3.8
	です	42.5	-	-	2	1.3
	でしょ	92.1	-	-	3	1.9
特丁			-	-	1	0.6
特			32	15.8	3	1.9
	だろう	8.0	7	3.4	-	-
	んだ	10.6	25	12.3	-	-
	じゃない	63.5	-	-	2	1.3
裸			99	48.8	40	25.3
	だ	7.1	14	6.9	1	0.6
	V意	19.9	13	6.4	1	0.6
	V	29.9	21	10.3	14	8.9
	Vタ	33.7	12	5.9	7	4.4
	A	33.9	28	13.8	7	4.4
	だっタ	37.9	2	1.0	1	0.6
	Aタ	43.3	7	3.4	2	1.3
	Vテ	78.2	1	0.5	4	2.5
	状X	79.8	-	-	2	1.3
-			15	7.4	20	12.7
		総会話文数	203	100	158	100

の最初の 2 編におけるセリフを使用する[11)]。

二人の文末形式の使用実態を表 1 に示す。この表では、文末形式タイプの各小計と、いずれかの人物が 2 回以上使用した文末形式を示した。文末形式の女性率は Kohaku-FFL に基づく。

二人の違いは、文末形式タイプ「裸」と「終」に顕著に現れている。成瀬は「裸」を多用する (48.8%) のに対し、島崎は「終」を多用する (41.1%)。島崎がよく使う終助詞「の・よ・じゃん・ね」を、成瀬は「よ」を除いて使わない。成瀬が使う終助詞「な」を島崎は使わない。作者が意図的に書き分けていることは明白である。成瀬の使用する文末形式は、あたかも成人男性のそれのようである。この時点の成瀬は中学生であるが、まったく女子中学生らしくない文末形式の使用は、常識の枠に収まらない稀有なキャラクタという成瀬の人物像の造形と密接に結びついていると思われる。

5.2　典型的な口調の探求

作品横断的に多くの登場人物の文末形式の使用分布を調べ、それを適切にクラスタリングすることによって、文末形式の使用分布に基づく典型的な口調 (多くの作品で観察される口調) を発見できる可能性がある。

Kohaku-BCCWJ には 16,625 名の人物が使用する 1,916 種類の文末形式が含まれている。ここから、出現率 5bp 以上の文末形式 158 種類と会話文数が 60 以上の 1,007 名を抽出し、文末形式の使用分布に基づいて人物を階層的にクラスタリングすると、最上位では、次の 5 つのクラスタが観察できた。

P　丁寧な文末形式を多用する人物群 (442 名)
M　男性的な文末形式を多用する人物群 (287 名)
F　女性的な文末形式を多用する人物群 (237 名)
O　古風な文末形式を多用する人物群 (21 名)
W　関西方言の文末形式を多用する人物群 (20 名)

これらのクラスタは、口調の大分類としては妥当を思われるが、我々が認識している「典型的な口調」の種類数は、もっと多いと思われる。

クラスタ P では、3 種類のサブクラスタ (特徴はそれぞれ：丁寧表現+終助詞、丁寧表現のみ、「ですわ・ますわ」)、クラスタ F でも、3 種類のサブクラスタ (丁寧表現多め、少なめ、中性的) の存在が観察できる。クラスタ M は一番混沌としているが、少なくとも 3 種類のサブクラスタ (柔らかめ、言い切り主体、粗野) の存在が観察できる。ただし、これらのクラスタは、それほど明快な形では現れない。ちなみに、この口調種別を当てはめると、成瀬は「M-言い切り主体」、島崎は「F-中性的」となる。

典型的な口調を、小説のセリフデータに基づいて探求するためには、Kohaku-BCCWJ のデータ量では不十分である。より多くのデータを準備するためには、小説の登場人物アノテーション [18] の自動化が必須であろう。

11)　ただし、漫才のセリフは除く。

謝辞

「BCCWJ 小説会話文」データを提供してくださった国立国語研究所の山崎誠教授に感謝します。このデータは、国立国語研究所のプロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」（プロジェクトリーダー・小磯花絵）および日本学術振興会・科学研究費補助金「会話文への発話者情報の付与によるコーパスの拡張」（15H03212）の成果です。また、Kohaku-BCCWJ の公開に際しては「多世代会話コーパスに基づく話し言葉の総合的研究」（プロジェクトリーダー・小磯花絵）に支援いただきました。これらのプロジェクトに感謝します。

参考文献

[1] 益岡隆志, 田窪行則. 基礎日本語文法 第 3 版. くろしお出版, 2024.
[2] 上野智子, 定延利之, 佐藤和之, 野田春美（編）. 日本語のバラエティ. おうふう, 2025.
[3] 小川早百合. 話しことばの終助詞の男女差の実際と意識—日本語教育での活用に向けて—. 日本語ジェンダー学会（編）, 日本語とジェンダー, pp. 39–51. ひつじ書房, 2006.
[4] 鈴木睦. 言葉の男女差と日本語教育. 日本語教育, Vol. 134, pp. 48–57, 2007.
[5] 金水敏. ヴァーチャル日本語 役割語の謎. 岩波書店, 2003.
[6] 金水敏（編）. 〈役割語〉小辞典. 研究社, 2014.
[7] 遠藤織枝. ドラマのことば —NHK TV「レイコさんの歯医者さん」をめぐって—. 日本語学, Vol. 16, No. 1, pp. 67–79, 1997.
[8] 下條正純. 「マリヤ様がみてる」における女性文末辞と人物描写. コンテンツ文化史研究, Vol. 7, pp. 12–24, 2012.
[9] 朽方修一. ライトノベルにおける女性文末形式. ヨーロッパ日本語教育 (21), pp. 148–153, 2017.
[10] 安井寿枝. キャラクター言語に見るジェンダー意識—宮﨑駿作品の特徴とは—. 日本語学, Vol. 43, No. 1, 2024.
[11] 佐藤理史. 小説会話文の文末形式の自動認定, 2025. Jxiv, doi: https://doi.org/10.51094/jxiv.1401.
[12] 山崎誠, 宮嵜由美, 柏野和佳子. 小説会話文への話者情報付与. 国立国語研究所, 2022. https://www2.ninjal.ac.jp/conversation/report/report05.pdf
[13] 佐藤理史. BCCWJ 小説会話文に対する文末形式の自動付与. 名古屋大学大学院工学研究科, 2025. (Kohaku-BCCWJ に含まれる)
[14] 佐藤理史. 小説会話文の文末形式リストの作成. 言語資源ワークショップ 2025, 2025. https://clrd.ninjal.ac.jp/lrw/lrw2025/o08.pdf
[15] Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. Sudachi: a Japanese tokenizer for business. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, 2018.
[16] 川北雄大, 石川和樹, 夏目和子, 小川浩平, 佐藤理史. 口調弁別評価データセットの作成と口調エンコーダの評価. 情報処理学会研究報告, Vol.2024-NL-259 No.16, 2024.
[17] 宮島未奈. 成瀬は天下を取りにいく. 新潮社, 2023.
[18] 大島一海, 小川浩平, 佐藤理史. 小説テキストに対する登場人物アノテーション. 言語処理学会第 31 回年次大会発表論文集, pp. 340–344, 2025.

付録

表 A Kohaku の出力例
会話文の出典はすべて『成瀬は天下を取りにいく』[17]。上半分の話者は成瀬、下半分は島崎。

F01	F02	F03	F04	F05	F06	F07	会話文
裸	V	-	-	-	-	V	島崎、わたしはこの夏を西武に捧げようと思う
終	な	な	-	-	-	だ	みうらじゅんみたいだな
接終	から. な	な	から	-	-	A	万が一のことがあるといけないからな
特終	だろう. か	か	-	-	だろう	A	ところで、マスクに何か書けないだろうか。
特	んだ	-	-	-	んだ	な	あくまでマスクを有効活用するのが重要なんだ。
特接	のだ. が	-	が	-	のだ	A	最後まで出られたらいいのだが
裸	だ	-	-	-	-	だ	これがM－１グランプリのエントリー用紙だ
裸	V意	-	-	-	-	V意	まだ間に合うから、とりあえずエントリーしておこう
裸	A	-	-	-	-	A	心配しなくていい。
終	ねん	ねん	-	-	-	V	その気になったらしゃべれんねん
特終	じゃない. か	か	-	-	じゃない	名	でも結局そういうことじゃないか？
終	よ	よ	-	-	-	A	何もないよ
終	よ	よ	-	-	-	Vタ	ちゃんと映ってたよ。
裸	V	-	-	-	-	V	行けたら行く
終	の	の	-	-	-	Vタ	わざわざ二枚買ったの？
丁終	んです. よね	よね	-	んです	-	な	あの、あかりちゃんのお母さんは滋賀の出身なんですよね？
裸	Vテ	-	-	-	-	Vテ	成瀬に渡すものがあって
接	けど	-	けど	-	-	A	まぁ、わたしは行くかどうかわからないけど
終	じゃん	じゃん	-	-	-	状X	めっちゃライオンズ好きな人みたいじゃん
丁	でしょ	-	-	でしょ	-	名	どう考えても成瀬がボケでしょ
特終	じゃない. じゃん	じゃん	-	-	じゃない	名	成瀬だって普段関西弁じゃないじゃん
特終	んだ. Dよね	Dよね	-	-	んだ	A	あと、ボケに意外性がないんだよね。
特終	んじゃない. かなあ	かなあ	-	-	んじゃない	A	もっと身近なテーマがいいんじゃないかなぁ

表 B 文末形式リスト Kohaku-FFL の概要
主要なフィールドは、F01, F02, F08, F10, F11, F15

ファイル名		KohakuFFL.20250716.tsv
ファイル形式		TSV 形式
行数		1,923 行
各行のフィールド数		16 フィールド (F00–F15)
F00	通し番号	0000 – 1922
F01	文末形式タイプ	文末形式の大分類
F02	文末形式 ID	
F03	終助詞 ID	終助詞等
F04	接続助詞 ID	接続助詞等
F05	丁寧表現 ID	「ます・です」など
F06	特殊表現 ID	「のだ・だろう・じゃない」など
F07	BCCWJ 頻度	D1 における頻度 (回数)
F08	BCCWJ 出現率	D1 における出現率 (万分率)
F09	なろう頻度	D2 における頻度 (回数)
F10	なろう出現率	D2 における出現率 (万分率)
F11	$\log_2$(F10/F08)	D1 と D2 の出現率の対数比
F12	女性頻度	D1 における女性話者の頻度 (回数)
F13	男性頻度	D1 における男性話者の頻度 (回数)
F14	その他頻度	D1 における女性・男性以外の頻度 (回数)
F15	女性率	D1 に基づく女性率 (%)

表 C Kohaku-BCCWJ の概要
F08–F13 は「BCCWJ 小説会話文」から抽出

ファイル名		KohakuBCCWJ.20250716.tsv
ファイル形式		TSV 形式
行数		276,576 行
各行のフィールド数		16 フィールド (F00–F15)
F00	通し番号 (000001 – 276576)	
F01	文末形式タイプ	
F02	文末形式 ID	
F03	終助詞 ID	
F04	接続助詞 ID	
F05	丁寧表現 ID	
F06	特殊表現 ID	
F07	主要素形式 ID	
F08	会話文	
F09	元データの No.	
F10	BCCWJ のサンプル ID	
F11	話者名	
F12	性別	
F13	年齢層	
F14	サンプル ID における会話文開始位置	
F15	サンプル ID における文末形式開始位置	