

BCCWJ小説会話文への話者情報の付与とその活用

| | |
|-----|---|
| 著者 | 山崎 誠, 柏野 和佳子, 宮寄 由美 |
| 雑誌名 | 言語資源活用ワークショップ発表論文集 |
| 巻 | 4 |
| ページ | 313-320 |
| 発行年 | 2019 |
| URL | http://doi.org/10.15084/00002582 |

BCCWJ 小説会話文への話者情報の付与とその活用

山崎 誠（国立国語研究所研究系言語変化研究領域）[†]

柏野 和佳子（国立国語研究所音声言語研究領域）

宮寄 由美（明治大学総合数理学部）

Annotation and Utilization of Speaker Information to Conversational Sentences in Novels Samples of BCCWJ

Makoto Yamazaki (National Institute for Japanese Language and Linguistics)

Wakako Kashino (National Institute for Japanese Language and Linguistics)

Yumi Miyazaki (Meiji University School of Interdisciplinary Mathematics)

要旨

本稿では「現代日本語書き言葉均衡コーパス」の図書館サブコーパスに含まれる小説（NDCで913, 923など）のサンプルにおける会話文に話者情報を付与した結果とそれを用いた分析について紹介する。付与したサンプル数は2,663サンプルである。付与した話者情報は「話者名、性別、年齢層」（これらは必須）のほか、「話者の社会的属性（職業など）、会話相手の情報、会話モード（電話での会話、方言での会話、外国人の会話等）」なども全てのサンプルにはではないが付けている。「話者名、性別、年齢層」については、「中納言」の検索結果に表示することを計画している。また、その他の話者情報は、中納言のサイトからBCCWJ所有者に限りダウンロードできるようにする予定である。分析から分かったこととして以下の4点を挙げる。(1)小説の全センテンスの約4割が会話文であること。(2)性別では女性の会話文が全体の約3割であること。(3)年齢層では約75%が成年層の会話であり、若年層は約20%、老年層は約5%であること。(4)会話モードでは、電話による会話全体の約4%程度あること。また、方言による会話文が約5,000あり、その多くは大阪を中心とした関西の方言であること。

1. はじめに

コーパスが単なるテキストの集合と異なるもっとも重要な違いは、アノテーション（付加情報）が付けられていることである。アノテーションが豊かであればあるほど分析の幅が広がる。近年公開されているコーパスにおいても、形態論情報を始めとしてさまざまなアノテーションが付与されている。例えば、話し言葉のコーパスにおいては話者の情報として、性別、年齢、社会的立場、出身地などがよく付けられている。書き言葉においても、小説の会話文を考えたとき、そこには、話者が存在し、話し言葉のデータと同じよう話者情報を付けることができる。もちろん、生の話し言葉ではないので、そこに書かれているのは擬似的な会話である。また、登場人物も架空のものがほとんどであるため、話者の属性の認定にも限界がある。しかし、大石（1987: 78-79）が指摘しているように、「会話の基本的性格をとらえようとするなど、話しことばの研究の一面として、会話文の研究が成り立つ」という考え方もある。実際の話し言葉と小説の会話文が違うことは調べてみるまでもなく明らかであろうが、どこがどう違うのかはやはり実際に調べてみないと分からないだろう。また、小説には役割語が使われる。それは、人間以外のキャラクター（妖精、人造人間、動物など）にも適用される。これらのキャラクターの会話は現実にはありえないため、どのような観点から言葉遣いが選ばれているか、研究の対象となろう。本稿は、書き言葉（小説）の話者に焦点をあて、その実態を報告するものである。

[†] yamazaki [AT] ninjal.ac.jp

2. データ

本稿で用いるデータは、「現代日本語書き言葉均衡コーパス」(以下、BCCWJ)の一部であるLB(図書館サブコーパス)に含まれる小説のサンプルである。サンプルが小説かどうかの認定は、BCCWJのメタ情報に含まれるNDC(日本十進分類法)の情報によった。具体的にはNDCが9x3の形式のものである。表1に今回の分析対象¹を示した。

表1 分析対象のファイル数

| NDC | 分類 | 作業済みサンプル数 |
|-----|-------------------|-----------|
| 913 | 日本文学・小説 物語 | 2,160 |
| 923 | 中国文学・小説 物語 | 27 |
| 933 | 英米文学・小説 物語 | 406 |
| 943 | ドイツ文学・小説 物語 | 8 |
| 953 | フランス文学・小説 物語 | 46 |
| 963 | スペイン文学・小説 物語 | 2 |
| 973 | イタリア文学・小説 物語 | 6 |
| 983 | ロシア・ソヴィエト文学・小説 物語 | 8 |
| 計 | | 2,663 |

本稿で用いるデータは、センテンス単位(行単位)で作業を行った。この場合のセンテンスとは、BCCWJに付与されている<sentence>というタグで囲まれた範囲を表す。この範囲は実際の文に対応している場合もあるが、見出しなどにも対応している、また、基本的に句点(。)までをセンテンスの単位としているため、「」で囲まれた、一つの会話が複数のセンテンスに分かれている場合もある。全センテンスの合計は615,957行²であった。1サンプル平均231.3行となる。

3. 話者情報の付与

3.1 会話部分の認定

サンプル中のどの部分が会話に該当するかは、まず、BCCWJに付与されている、<speech>や<quote>というタグ³を作業上の目安とした。<speech>は、前後に改行をとめない、かぎ括弧(「」)で囲まれた部分に付与されており、また、<quote>は、1文中のかぎ括弧で囲まれた部分に付与されている。<speech>は会話の可能性が高いが、<quote>は、強調であったり、映画や本の名前などにも付与されており、必ずしも会話とは限らない。今回のデータでは、<speech>タグが付いているセンテンスは124,291行、そのうち、話者名が付いているものが119,443行あった。<sentence>タグが付与された箇所は、約96.1%の割合で会話であることが分かる。一方、<quote>タグが付いているセンテンスは、26,838行であり、そのうち、話者名が付いているものは21,282行であった。<quote>タグが付与された箇所が会話である割合は79.3%と相対的に低くなっている。

作業者はこれらのタグ以外の部分にも目を通し、上述のタグが無い部分の会話箇所も拾うようにしている。

なお、サンプル中のどの部分を会話とみなすかについては、宮寄他(2017)を参照されたいが、原則として、当該箇所「声に出したと想定される発話」を会話とみなし、話者

¹ 図書館サブコーパスの小説全体は2,707サンプルであり、表1はそれに44サンプル足りない。内訳は日本文学・小説物語29、英米文学・小説物語13、ドイツ文学・小説物語1、フランス文学・小説物語2である。

² センテンスを数える単位を便宜的に「行」とする。

³ 詳しくは西部他(2011: 240,274)を参照されたい。

情報を付与している。

3.2 例外的な会話

前節で会話部分を「声に出したと想定される発話」としたが、話者情報を付与した箇所には、若干その例外に当たるものもある。以下、例を3つ挙げる。1つ目は、伝達媒体が声でない場合である。具体的にはSFなどにおけるテレパシーによる会話や、コンピュータ画面上でのチャットなどが該当する。テレパシーが使われた会話は907行あり、25サンプルに出現している。

2つ目は、心内発話である。心内発話は実際に声には出していないが、会話に準ずるものとして話者情報を付与した。心内発話かどうかはデータ上区別できるようにしている⁴。心内発話は5,170行あり、800のサンプルに出現している。

3つめは沈黙である。沈黙は、多くの場合「…」で示され、発話自体は存在していないが、コミュニケーション上の役割を考慮して話者情報を付与している。沈黙は825行、344サンプルに出現している。

3.3 話者の属性

付与した話者の属性一覧と、それらの属性を付与した行数を表2に示す。このうち、話者名、性別、年齢層⁵を基本的な属性と考え、必須項目とした。それ以外の属性は分かる範囲で記入している。

表2 話者の属性

| 属性 | 説明 | 付与した行数 |
|----------|---|---------|
| 話者名 | (サンプル内での) 登場人物の呼び名 | 270,388 |
| 性別 | 男, 女, 不明 | 267,947 |
| 年齢層 | 若年層 (~19歳), 成年層 (20~59歳), 老年層 (60歳以上) | 263,510 |
| 年代の確信レベル | 年代の推定が極めて難しい場合に○ | 48,280 |
| 非人間 | 生物学的な人間以外のもの | 12,136 |
| 会話モード | 通常の対話場面でない場合。「電話, 方言, 外国人, テレパシー, 引用, 独話, 疑問, 沈黙, 驚き, 驚愕」など | 29,759 |
| 会話認定情報1 | 会話に準ずる場合, そのタイプを記入。独話, 心内発話など | 14,172 |
| 会話認定情報2 | 会話に準ずる場合に判定した根拠を記入 | 12,533 |
| 備考 | 注記 | 103,138 |
| 職業 | 発話者の職業, 社会的身分 | 18,073 |
| 相手 | 会話の相手 | 197,544 |

もっとも情報が付与された数が多い、話者名を基準にすると、性別はそれに対して99.1%、年齢層は97.5%⁶、職業は66.8%、相手は73.1%の付与率である。しかし、情報が付与されていても、その内容が「不明」となっているものがある。「不明」が付けられたのは、話者名で108行、性別で6,163行、年齢層で453行であった。これらの数字のアンバランスは、

⁴ 表2の会話認定情報1に「心内発話」と記入されている。

⁵ 具体的な年齢が判明している場合は、備考欄「注記」に記入してある。

⁶ 話者名、性別、年齢層のすべてに属性(「不明」を含む)が付けられた数は263,503である。

属性のもつ性質による。話者名は、誰だか分からなくても、「子供」「村人」などと情報を記すことが出来るが、性別は手がかりがない場合、「不明」とせざるを得ない。年齢層は、状況的に成人であるとみなせる場合は、成年層となるため、まったく「不明」の場合は少なかったためである⁷。

4. 結果

4.1 地の文と会話文

便宜的に、話者名が付いているセンテンスを会話文、そうでない文を地の文とすると、会話文の数は 270,388 行となり、全体の 43.9%になる。また、各サンプルにおいて、会話文の数を全体の文数で割った値である、会話文率が求められる。図 1 に話者情報を付与した全 2,663 サンプルの会話文率の分布を示す。

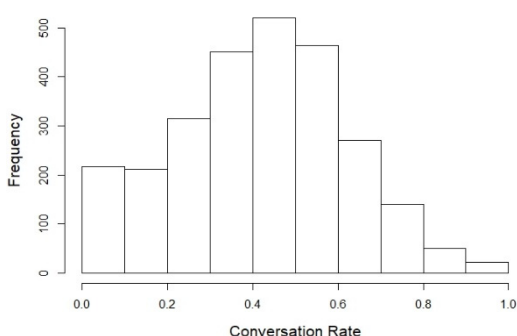


図 1 会話文率

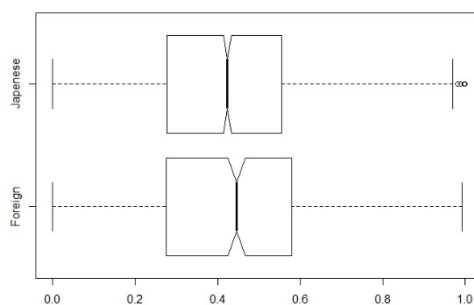


図 2 会話文率（日本の小説と外国の小説）

以下の分析では、日本の小説（NDCが 913）と外国の小説（NDCが 923, 933 など）とに分けた集計結果を示す。図 2 は、日本文学（913）とそれ以外の外国の小説の会話分率の比較である。外国の小説の方が少し値が高いようであるが⁸、ウェルチの t 検定の結果、5%水準で有意差はなかった ($p = .0806$)

4.2 性別

性別は単独の話者で発話しているもののみを対象とする。複数人の発話は 6,380 行あるがこれらは集計の対象外とした。表 3 は性別の会話文の数である。女性による会話が全体

⁷ 年齢層の確信度に不安がある場合は、「年代の確信レベル」欄に○が付与されている。

⁸ 外国の小説は平均 0.4319, 中央値 0.4460, 日本の小説は平均値 0.4131, 中央値 0.4240 である。

の約3割であった。この傾向は日本の小説でも外国の小説でもほとんど同じであった。

表3 会話文の数（性別）

| 話者の性別 | 日本の小説 | 外国の小説 | 計 |
|-------|---------------------|--------------------|---------------------|
| 女 | 62,517 (30.2%) | 17,337 (31.9%) | 79,854 (30.5%) |
| 男 | 14,4667 (69.8%) | 37,046 (68.1%) | 181,713 (69.5%) |
| 計 | 207,194 (100.0%) | 54,383 (100.0%) | 261,567 (100.0%) |

4.3 年齢層

表4は年齢層別の会話文数である。全体の約3/4が成年層の会話である。また、外国の小説では若年層の会話が少ないことが分かるが、これは日本の小説にいわゆるライトノベルがあり、そこに高校生などの若年層が多く登場するためと思われる。表5は、年齢層と性別のクロス集計である。年齢層には「不明」あるいは、空欄のものがあり、それらも「不明」として扱った。女性は若年層の割合が男性の2倍近くあることが分かる。老年層の割合は男女でほぼ同じであるため、男性の成年層が女性より多いということになる。

表4 会話文の数（年齢層別）

| 話者の年齢層 | 日本の小説 | 外国の小説 | 計 |
|--------|--------------------|-------------------|--------------------|
| 若年層 | 14,510 (23.4%) | 1,716 (10.1%) | 15,866 (20.2%) |
| 成年層 | 44,531 (71.8%) | 14,414 (84.9%) | 58,945 (74.9%) |
| 老年層 | 3,023 (4.9%) | 839 (4.9%) | 3,862 (4.9%) |
| 計 | 62,064 (100.1%) | 16,969 (99.9%) | 78,673 (100.0%) |

表5 会話文の数（年齢層×性別）

| 年齢層 | 女 | 男 | 計 |
|-----|-------------------|--------------------|--------------------|
| 若年層 | 15,866 (19.9%) | 17,640 (9.7%) | 33,506 (12.8%) |
| 成年層 | 58,945 (73.8%) | 151,926 (83.6%) | 210,871 (80.6%) |
| 老年層 | 3,862 (4.8%) | 9,687 (5.3%) | 13,549 (5.2%) |
| 不明 | 1,172 (1.5%) | 2,455 (1.4%) | 3,627 (1.4%) |
| 計 | 79,845 | 181,708 | 261,553 |

4.4 会話モード

会話モードとは、通常の対面による、日本語での会話以外の場合にその場面を理解する

ために必要な情報として付与している。例えば、電話による会話や外国語（と想定される）会話などである。表 6 に会話モードのうち、頻度が 100 以上あったものを挙げた。なお、会話モードは、例えば、方言による電話での会話のように、ひとつの会話に複数の属性が付与されることがあるため、ひとつの会話を重複して集計している。

会話モードのうちもっとも多かったのは、電話による会話で全会話数⁹の約 4.2%を占める。それにつぐのが方言による会話で全会話数の約 1.9%となっている。方言による会話の地域別内訳を表 7 に示す。地域の情報が付いた方言による会話は 2,884 行あったが、そのうち、関西と大阪で約 71.4%を占め、小説における関西・大阪方言の使用が多いことが分かった¹⁰。山崎（2018: 9）で、日本語小説の特徴語として関西方言の助動詞「や」「じゃ」が挙げられていることもその現れであろう。

表 6 会話モードの数（頻度 100 以上）

| 会話モード | 会話文数 | 会話モード | 会話文数 |
|-------|--------|--------|------|
| 電話 | 11,250 | 録音音声 | 361 |
| 方言 | 5,053 | テレビ | 246 |
| 回想 | 3,132 | インタビュー | 224 |
| 独話 | 2,478 | 江戸 | 204 |
| 引用 | 1,600 | 鹿児島 | 176 |
| 通信器 | 1,356 | インターホン | 159 |
| 関西 | 1,319 | 夢 | 154 |
| テレパシー | 907 | 京都 | 151 |
| 大阪 | 740 | 英語 | 125 |
| 外国人 | 409 | 歌 | 118 |
| 憑依 | 374 | 驚き | 115 |
| 叫び | 363 | 留守番電話 | 102 |

表 7 方言による会話

| 地域 | 会話文数 | 地域 | 会話文数 |
|-----|-------|-----|------|
| 関西 | 1,319 | 佐賀 | 42 |
| 大阪 | 740 | 金沢 | 33 |
| 江戸 | 204 | 北海道 | 28 |
| 鹿児島 | 176 | 博多 | 23 |
| 京都 | 151 | 九州 | 20 |
| 名古屋 | 87 | 越後 | 3 |
| 関東 | 57 | 熊本 | 1 |

4. 5 職業

職業は、話者の社会的身分（話者名が付いた会話の約 66%の付与率であるが、暫定的な集計を行った¹¹。表 8 がその結果である。もっとも多かった職業は「刑事」で、6,846 の会話に登場している。全会話数の約 2.5%にあたる¹²。表 8 には、刑事以外にも、「警部、警部

⁹ 話者名が付いた 270,388 行を全会話数とみなす。

¹⁰ 方言の情報は今後整備予定であり、以下は現時点での入力情報から算出したものである。

¹¹ この欄は自由記述であるため、3,638 種類の属性が用いられている。今後同じまとめるなどして整理する必要がある。

¹² 「刑事部長」「部長刑事」など、「刑事」を含む会話は 8,150 あり、全会話数の約 3.0%に当たる。

補, 警察官, 警視」があり, これらの合計は 13,591 であり, 全会話数の約 5%となる。同じく, カテゴリとして多いのは「学生」であり, 表 8 の「高校生, 大学生, 中学生, 小学生, 学生」の合計は 11,871 であり, 全会話数の約 4.4%を占める。警察や学生が関係するサンプルが多いことが分かるが, 「刑事」という文字列を職業欄に含むサンプル数の異なりが 198 であるのに対し, 「学生」または「高校生」を職業欄に含むサンプル数の異なりは 318 であり, 学生の会話に比べて警察関係の会話は相対的に特定のサンプルに集中して出現していることが分かる。

他に特徴的な職業として「武士」「武将」がある。これらはすべて日本の時代小説に現れるものである。

表 8 職業別の会話数 (頻度 500 以上)

| 職業 | 会話文数 | 職業 | 会話文数 | 職業 | 会話文数 |
|------|-------|------|-------|-------|------|
| 刑事 | 6,846 | 教師 | 1,218 | 編集者 | 684 |
| 高校生 | 6,523 | 作家 | 1,171 | 軍人 | 672 |
| 会社員 | 3,663 | 小学生 | 1,159 | 僧侶 | 666 |
| 警部 | 2,969 | 主婦 | 1,076 | カメラマン | 653 |
| 探偵 | 2,834 | 警察官 | 1,069 | 秘書 | 648 |
| 大学生 | 2,818 | 小説家 | 1,047 | 将軍 | 632 |
| 武将 | 2,448 | 医師 | 900 | 王 | 602 |
| 武士 | 2,251 | 学生 | 899 | 銀行員 | 574 |
| 弁護士 | 1,913 | 医者 | 857 | 薬剤師 | 566 |
| 警部補 | 1,868 | 検事 | 852 | 高校教師 | 544 |
| 新聞記者 | 1,546 | 警視 | 839 | 俳優 | 526 |
| 中学生 | 1,371 | 記者 | 827 | | |
| 社長 | 1,316 | 私立探偵 | 734 | | |

また, 職業の欄では, 表 2 の話者の属性で「非人間」となっているものの実体を確認することができる。表 9 に頻度 100 以上のものを示す。多くは特定のサンプルで集中的に用いられたものが多い。

表 9 非人間の属性が付いた職業の会話数 (頻度 100 以上)

| 職業欄の表示 | 会話数 |
|------------|-----|
| 列車 | 412 |
| 小人 | 246 |
| 異星人/植物系生命体 | 206 |
| 修験者 | 150 |
| 天使 | 145 |
| ドブネズミ | 141 |
| 整体師 | 141 |
| 妖 | 132 |
| ヒューマノイド | 123 |
| なめくじ | 104 |
| 悪神の血をひくもの | 104 |

5. まとめと今後の課題

本稿では、BCCWJの図書館サブコーパスにおける小説の会話文に対して、話者情報（話者名、性別、年齢層等）を付与し、現整備状況から算出した結果に基づいた、話者の属性からみた会話文の量的傾向を概観した。今後は形態論情報を用いた語彙的な分析に進む予定である。

謝 辞

本研究は、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」（JP15H03212）による成果である。

話者情報の付与は、論文執筆者に加え、河野礼実氏、田嶋明日香氏、立花幸子氏、平本智弥氏、牟田浩子氏、山縣智子氏が作業に当たった。

文 献

- 大石初太郎（1987）近代・現代小説会話文の資料性，「国文学解釈と鑑賞」52（7），72-79.
西部みちる，大島一，間淵洋子，小林正行，田島孝治，高田智和，山口昌也（2011）『現代日本語書き言葉均衡コーパス』における電子化テキストの構築，国立国語研究所.
宮寄由美，柏野和佳子，山崎誠（2017）発話文への発話者情報付与の基本設計：『現代日本語書き言葉均衡コーパス』収録の小説を対象に，「言語資源活用ワークショップ発表論文集 2016」，pp.38-48. <http://doi.org/10.15084/00001456>
山崎誠（2018）翻訳小説と日本語小説における会話文の計量語彙論的比較，「語彙研究」，15，pp.1-15.