1D2-OS-29a-2

日本語小説の会話文タグ付コーパスの開発に向けて

Towards for development of tagged dialogue corpus of Japanese novel

村井 源[*1]
Hajime Murai

[*1] 東京工業大学
Tokyo Institute of Technology

In order to develop basis for computational analysis of fictional conversation in Japanese novel, format for tagged dialogue corpus of Japanese novel was developed based on 100 random samples of Japanese novel texts in BCCWJ. Developed corpus includes attributes about speaker's name, gender, listener's name, relationships between the speaker and the listener, field, profession, and age group.

1. はじめに

一般的な小説における会話文では，今話しているのが誰であるかを明示するような「○○は『……』と言った」のような地の文による説明的な情報は必ずしも記述されるわけではない．各登場人物の最初の発話では誰が話したかが明示されるが，その後は地の文を挟まずに会話文のみが交互に繰り返されるような形をとることは多くみられる．「と言った」のようなフレーズを各会話文の前後につけると文体が単調になり，なおかつ読者にはおおよそ誰が話しているかが前後の情報から推定可能であるため，重複してまどろっこしい印象を与えてしまうことがこの原因として考えられる．

読者は，会話文のおかれた文脈やすでに示された登場人物の個性や背景などの属性的な情報（例えば性別，年齢，立場，職業，出身地，教養，相手への評価……）を勘案し，明示されていない各会話の話し手と聞き手を解釈して読み進めることが前提とされている．また，会話文自体の中にも話し手の推定のために有用な要素が当然含まれている．会話の内容からその場においてこのような発話をするのが適切な登場人物を推定することや，「役割語」[金水 2003]のような登場人物ごとに設定された属性情報に応じたしゃべり方の特徴も話し手の推定には重要なカギとなるであろう．

一方で，文学作品などの会話文を含んだテキストに人工知能などを用いて機械的に処理する状況を考えると，人間の読者の場合には個々に推定が行われる話し手と聞き手の情報は機械の側で何らかのアルゴリズムを使って推定することが必要不可欠である．また，話し手や聞き手の推定に利用可能な様々な文脈情報や属性情報を何らかの手法で抽出できれば，それらの情報は例えば登場人物間の関係性の推定や主人公の感情状態の推定など他の様々なテキストの意味解釈にも利用が可能であると考えられる．

このため，文学作品における話し手の自動的な推定の研究が英語の作品などを対象として，ルールベースや機械学習などで進められてきているが，まだ十分に高い精度を得ているとは言い難いのが現状である[He 2013]．また日本語の文学作品の話し手の自動推定に関しては研究も少なく，学習用のデータとしたり推定アルゴリズムの精度を確認したりするために必要な，オープンな話し手情報付きのコーパスも存在していない．

そこで，機械的な会話文の話し手・聞き手の推定やその他の様々な意味処理，機械的な解釈に利用可能な話者やその属性情報を付与したタグ付コーパスを設計し，『現代日本語書き言葉均衡コーパス』(BCCWJ)をベースにデータ化を行った．

連絡先：村井源，東京工業大学大学院社会理工学研究科価値システム専攻，152-8552 東京都目黒区大岡山 2-12-1

2. 対象データ

『現代日本語書き言葉均衡コーパス』には販売された書籍の ISBN に基づくサンプリングで作られたコーパスと，図書館に所蔵された書籍のリストに基づくサンプリングで作られたコーパスが含まれているが，実際に流布している物語テキストにおける会話文の実態をとらえるということを考え，図書館の蔵書に基づくコーパスを用いた．また，日本語文学テキストを対象とするため十進分類で日本語の物語・小説自体にあたる 913 に分類されたサンプルのみを抽出した．結果として該当する 2189 サンプルが抽出されたが，それらの中に含まれる会話文への人手でのタグ付けと分類の労力を考慮し，2189 サンプルからランダムに 100 サンプルを抽出してタグ付けの対象とした．

タグ付けの対象となるのは対象テキスト中の時間軸で現在発話されている「」や『』でくくられた会話箇所とした．回想，置手紙，語り手や登場人物による想像上の会話などその場で行われていない発話的な描写は除外するが，電話の会話やファンタジーにおけるテレパシーでの会話は分析対象に含めた．結果として，日本語の物語文 100 サンプル中の 5632 発話を抽出してタグ付けを行った．会話単位でのタグ付けの例を表 1 に示す．

3. タグの要素

物語テキスト中での会話文において有用であると考えられる属性として下記を想定し，本文中から推定可能な範囲でタグ付けを行った．

話者：発話の話し手を記述した．登場人物に複数の呼び名がある場合は，他の登場人物と重複しない頻出の呼び名で統一的に記述した．話者は必ずしも単数ではなく，「人々」のような文字列であらわされる不特定多数の登場人物群も話者に含んだ．

性別：男性か女性か判断可能な情報が会話文や地の文にあった場合に性別を記述した．言葉づかいや職業からおそらく男性と推測されるが明確な記述のない場合などは空欄とした．

聴者：相手を前提とした発話の場合には，会話の場に居ることが明確な登場人物のうちで内容から他よりも聞き手にふさわしいと考えられえる登場人物がいた場合に聴者として記述した．複数名に向かっての発話である場合には複数の聴者を記述した．また相手を前提としない発話の場合にはそれぞれ，「独白」「祈り」「呪文」などのそれぞれの内容に該当するタグを付与した．

表 1 タグ付きコーパスのサンプル（BCCWJ サンプル ID:LBt9_00252）

話者	性別	聴者	関係	場	職種等	年齢等	発話
藤岡哲	男	羊	友人			高校生	「羊っちゃん、まだ決めつけるのは早いんじゃあ…」
スズキ（姉）	女	羊		家庭	教師		『はーい』
羊	女	スズキ（姉）		家庭		高校生	「パパいますか」
スズキ	男	羊	養子	家庭	教師		『なんだ？』
羊	女	スズキ	義父	家庭		高校生	「荷物」
羊	女	スズキ	義父	家庭		高校生	「もう、帰ってこなくていーですから」
スズキ	男	羊	養子	家庭	教師		「だろうな」
藤岡哲	男					高校生	「ほげぇ…」

相手の関係：上司・部下・同僚や親子・恋人など登場人物間の関係性が文章中の情報から推定可能な場合は相手との関係として記述した．刑事ものなどで警察と容疑者の場合は，警察は職業名でもあるので後述の「職種」欄に，容疑者は職業ではないので「相手との関係」欄に記述している．

場：物語の舞台となる場が推定可能な場合は，時代（古代，平安時代，江戸時代）や，ジャンル（ファンタジー，西部劇）などを記述している．推定される時代が「現代」，場所が「日本」の場合には省略して記述していない．また時空間的な場だけでなく，「裁判」や「家庭」「職場」などの会話に影響を与えると考えられる状況設定も合わせて記述している．

職種等：話し手の職業などが警察や裁判官など明確な場合は「職種等」の欄に記述している．ファンタジー等での人間外の種族（妖怪・悪魔等）も記述している．

年齢等：物語中で年齢そのものが明記されることはほとんどないが，話し手が高齢者や幼児であることが明確な場合には年齢等に記述している．中学生や高校生などの情報も年齢等．

4. コーパスの特徴と問題点

得られたコーパスの属性を集計したところいくつかの属性に偏りがみられる結果となっている．例えば性別であるが，表 2 のように倍近く男性の発話者が多い．サンプルテキスト中に事件の推理や戦争を扱ったものが多く含まれ，警察や捜査関係者，武士や軍人が多くなることが一つの原因として考えられる．

表 2 性別の集計

男	3296
女	1678
その他	46
無（機械）	3
不明	606

表 3 聴者の関係の集計上位

職場関係	798
友人	660
家族・親戚	408
恋愛（夫婦除く）	332
店（客・店員，患者・医者等）	116

また話者と聴者の関係であるが，職場における上司や部下，同僚などの関係が多く，それに続いて友人や家族関係も多いが必ずしも特定の関係が読み取れない発話が多かった（5629 中 2464 と半数弱）．関係特定不能な発話の中には，関係性に関する情報が会話の前後の地の文や会話文中に記述されていないもの，その場にいる不特定の何人かを相手にしているような発話，単なる顔見知りらしき関係，職場外での仕事上の関係者，捜査関係者と捜査対象の人物などが含まれている．なお捜査関係のうちで警察内の同僚同士や上司と部下の会話は表 3 中の「職場関係」中に含んでいる．

職業等に関しては明確な記述が多いのは警察・検事・弁護士・探偵等推理物のジャンルに頻出する職種であり，その他には戦記物や歴史もの関係に頻出する武士・軍人・王族関連や学校等を舞台とした場合の教授や教師等ごく限られた職種が抽出された．また年齢等に関しても関連記述のある登場人物はごくわずかであり，約 90%が年齢特定不能であった．

5. 結論と今後の課題

BCCWJ の図書館所蔵サンプリングは所蔵テキストのジャンルの偏りを反映して人気のある特定ジャンルが多量に含まれると考えられる．また物語本文からのランダム部分サンプリングであるため前後関係の把握が困難で登場人物の関係性が記述しにくいという側面もある．今後属性情報の明確な会話文のみを対象として分析を行うという方法もありうるが，より一般的な会話文の特徴を抽出するためには，サンプリングされていないテキスト個所や関連する書誌情報等から登場人物の属性を抽出するなどの工夫が必要になると考えられる．

本稿にて提案したタグ付きコーパスは個人による開発のため均衡コーパスからのランダムサンプリングに対してタグ付けを行っているが，コーパスのサイズは十分とは言い難い．現在本稿と非常に近い設計理念で国立国語研内の研究グループ「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」内のレジスター班において均衡コーパス内の物語テキストの会話文へのタグ付け作業が進められており[国立国語研究所 2017]，筆者も今後当研究グループと協力してより大規模で精細なコーパスの開発と公開を目指すことを検討している．

参考文献

[金水 2003] 金水敏: ヴァーチャル日本語 役割語の謎，岩波書店，2003.

[He 2013] Hua He, Denilson Barbosa, and Grzegorz Kondrak: Identification of speakers in novels, In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 1312–1320, 2013.

[国立国語研究所 2017] 国立国語研究所，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，http://pj.ninjal.ac.jp/conversation/corpus.html，2017/2/27 参照，2017.