2019.08.30 国立国語研究所講堂

シンポジウム「話し言葉の多様性」

BCCWJ小説会話文の話者情報を利用した分析

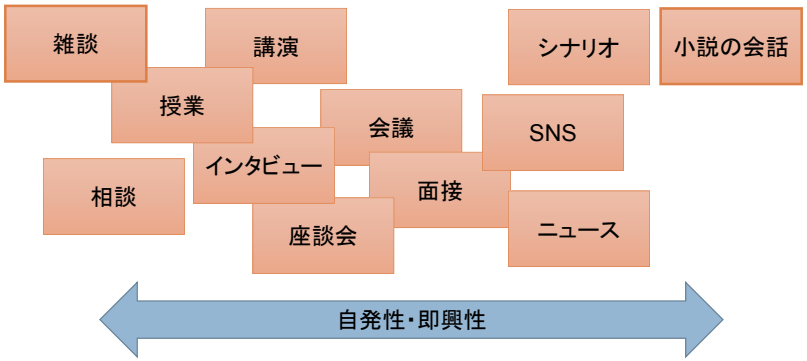
山崎誠（国立国語研究所）

概要

- 1.話し言葉の広がり
- 2.小説の会話
- 3.データと方法
- 4.結果
- 5.他のコーパスとの比較

1.話し言葉の広がり

- さまざまなタイプの話し言葉



2.小説の会話

2.1.小説の会話

- 高崎みどり(1981)
  - 小説の会話文とは、登場人物たちの言動の描写であるというよりは、さまざまの理由によって会話の形をとっている、地の文の延長であると考えた方がよさそうである。
  - 小説会話文は、話しことばの材料として、そのまま行生かすことは難かしく、取り扱いに慎重を要するように思われる。

2.2.小説の会話

- 大石初太郎(1987)
  - つまるところ、会話文は話しことばの資料として役立つものではないという結論を導き出すことになる。
  - たしかに、会話文は話しことばのイミテーションである。しかし，同時に、すぐれた作家の筆になる会話文は、話しことばの典型だとも考えられよう。
  - たとえば、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ことになろう。

2.3.小説の会話をとりあげる意義

- (1)実際の話し言葉とどのような点がどう違っているか。
- (2)過去の書き言葉資料(日本語歴史コーパスなど)との比較。
- (3)役割語の研究。
- (4)翻訳小説と日本語小説の文体の比較。
- (5)生の話し言葉の整形の自動化。

3.データと方法

3.1.データ

- 『現代日本語書き言葉均衡コーパス』(BCCWJ)の図書館サブコーパスに含まれる小説。2663サンプル。

3.2.分析対象とサンプル数

- 図書分類(NDC)により抽出。

NDC	分類	サンプル数
913	日本文学・小説 物語	2160
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		2663

3.3.話者情報の付与

- BCCWJには話者情報が付与されていないため，独自に話者情報を付与した。

3.4.タグの利用(<speech>)

- <quotation><speech><paragraph><sentence>
- 「兄弟って素敵？」</sentence>
- <br type= "automatic_original" /></paragraph></speech></quotation>
- <quotation><speech><paragraph><sentence>
- 「さあ、どうかな、もう七年も会ってないんですよ」
- </sentence>
- <br type="automatic_original" /></paragraph></speech></quotation>

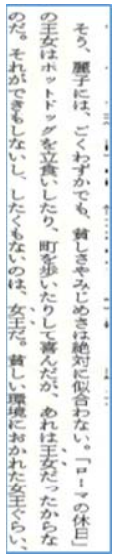
「兄弟って素敵？」
「さあ、どうかな、もう七年も会ってないんですよ」

- Lba9_00017「中国行きのスロー・ボート」(村上春樹)

3.5.タグの利用(<quote>)

- <sentence><quote>「ローマの休日」</quote>の
- 王女はホットドッグを立食いしたり、町を歩いたり
- して喜んだが、あれは王女だったからなのだ。
- </sentence>

  - Lba9_00051「死はやさしく奪う」(栗本薫)


そう、麗子には、ごくわずかでも、貧しさやみじめさは絶対に似合わない。「ローマの休日」の王女はホットドッグを立食いしたり、町を歩いたりして喜んだが、あれは王女だったからなのだ。それができもしないし、したくもないのは、女王だ。貧しい環境におかれた女王ぐらい、

3.6.会話箇所の特定

- 実際にはタグだけでは判断がむずかしいため，以下のような基準を設けた。
- (1)カギ括弧にくくられた，声に出したと想定される部分
- (2)カギ括弧にくくられた，(1)に準ずる部分･･･独話，沈黙，心内発話など。
- (4)カギ括弧にくくられていないが，声に出したと想定される部分･･･「と言った」などど明確に書かれている場合など。
- (5)「場面設定」を考慮した，(1)に準ずる部分･･･夢の中の会話，テレパシーなど。
- 宮嵜由美他(2017:39)をもとに作成

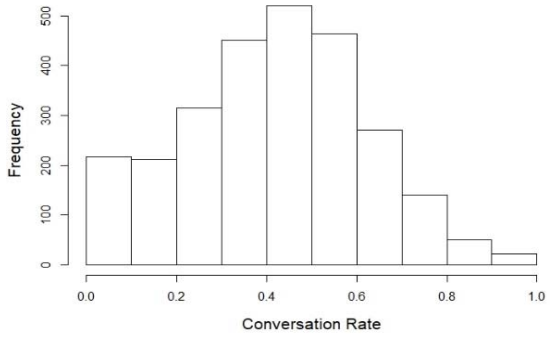
3.7.付与した属性

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	270388
性別	男，女，不明	267947
年齢層	若年層(～19歳)，成年層(20～59歳)，老年層(60歳以上)	263510
年代の確信レベル	年代の推定が極めて難しい場合に○	48280
非人間	生物学的な人間以外のもの	12136
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	29759
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	14172
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	12533
備考	注記	103138
職業	発話者の職業，社会的身分	18073
相手	会話の相手	197544

4.結果

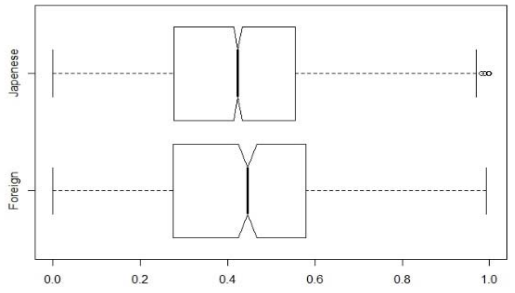
4.1.会話文の割合

- すべての文数に占める会話文の割合
- 平均は43.9%。



4.2.会話文の割合

- 日本の小説と翻訳小説
- 有意差はない（p=.0806，ウェルチのt検定）



4.3.性別

話者の性別	日本の小説	翻訳小説	計
女	62517 (30.2%)	17337 (31.9%)	79854 (30.5%)
男	144667 (69.8%)	37046 (68.1%)	181713 (69.5%)
計	207194 (100.0%)	54383 (100.0%)	261567 (100.0%)

- 複数人が同時に発している発話は除く。

4.4.年齢層

年齢層	日本の小説	翻訳小説	計
若年層	14510 (23.4%)	1716 (10.1%)	15866 (20.2%)
成年層	44531 (71.8%)	14414 (84.9%)	58945 (74.9%)
老年層	3023 (4.9%)	839 (4.9%)	3862 (4.9%)
計	62064 (100.1%)	16969 (99.9%)	78673 (100.0%)

- 日本の小説に若年層が多い。

4.5.年齢層×性別

年齢層	女	男	計
若年層	15866 (19.9%)	17640 (9.7%)	33506 (12.8%)
成年層	58945 (73.8%)	151926 (83.6%)	210871 (80.6%)
老年層	3862 (4.8%)	9687 (5.3%)	13549 (5.2%)
不明	1172 (1.5%)	2455 (1.4%)	3627 (1.4%)
計	79845	181708	261553

- 若年層の女性は男性の2倍近い。

4.6.会話モード

会話モード	会話文数
電話	11250
方言	5053
回想	3132
独話	2478
引用	1600
通信器	1356
関西	1319
テレパシー	907
大阪	740
外国人	409
憑依	374
叫び	363

会話モード	会話文数
録音音声	361
テレビ	246
インタビュー	224
江戸	204
鹿児島	176
インターホン	159
夢	154
京都	151
英語	125
歌	118
驚き	115
留守番電話	102

- 頻度100以上を挙げた。

4.7.会話モード(方言)

地域	会話文数
関西	1319
大阪	740
江戸	204
鹿児島	176
京都	151
名古屋	87
関東	57

地域	会話文数
佐賀	42
金沢	33
北海道	28
博多	23
九州	20
越後	3
熊本	1

- 関西方面の方言が圧倒的に多い。

4.8.職業

職業	会話文数
刑事	6846
高校生	6523
会社員	3663
警部	2969
探偵	2834
大学生	2818
武将	2448
武士	2251
弁護士	1913
警部補	1868
新聞記者	1546
中学生	1371
社長	1316

職業	会話文数
教師	1218
作家	1171
小学生	1159
主婦	1076
警察官	1069
小説家	1047
医師	900
学生	899
医者	857
検事	852
警視	839
記者	827
私立探偵	734

職業	会話文数
編集者	684
軍人	672
僧侶	666
カメラマン	653
秘書	648
将軍	632
王	602
銀行員	574
薬剤師	566
高校教師	544
俳優	526

- 頻度500以上を挙げた。警察関係が全体の約5%。

5.他のコーパスとの比較

5.1.比較したコーパス

- 『日本語話し言葉コーパス』
  学会講演
  模擬講演
- 『日本語日常会話コーパス』
- 『名大会話コーパス』
- 『女性のことば・男性のことば(職場編)』

5.2.品詞構成比

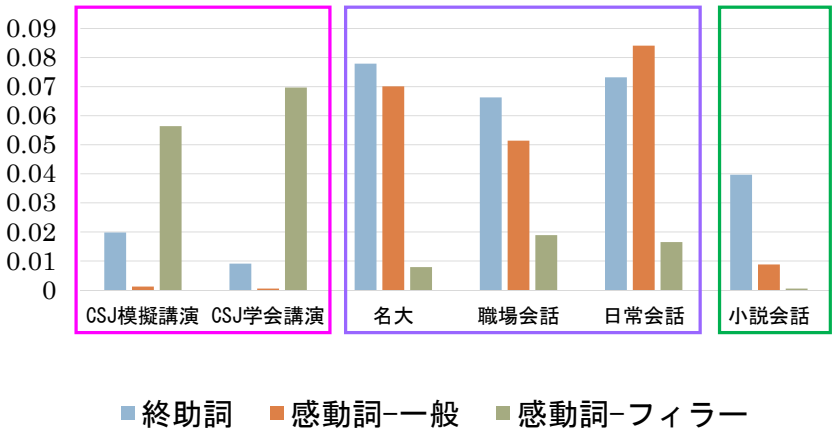
	模擬講演	学会講演	名大	職場会話	日常会話	小説会話
名詞	0.217	0.277	0.172	0.198	0.177	0.224
代名詞	0.024	0.018	0.042	0.035	0.038	0.037
形容詞	0.018	0.010	0.031	0.025	0.032	0.022
形状詞	0.012	0.013	0.011	0.008	0.010	0.011
動詞	0.134	0.133	0.115	0.116	0.110	0.143
副詞	0.037	0.020	0.051	0.046	0.061	0.030
連体詞	0.011	0.014	0.014	0.011	0.008	0.013
接続詞	0.011	0.012	0.007	0.008	0.008	0.003
感動詞	0.058	0.070	0.078	0.070	0.101	0.009
助詞	0.326	0.295	0.332	0.320	0.306	0.338
助動詞	0.127	0.108	0.127	0.137	0.126	0.136
接頭辞	0.006	0.006	0.005	0.005	0.007	0.009
接尾辞	0.019	0.024	0.016	0.019	0.017	0.025

山崎(2017:281)より

5.3.品詞構成比

- 名詞は，学会講演で多い。
- 代名詞は，学会講演で少ない。
- 形容詞は，日常会話，名大で多く，学会講演で少ない。
- 副詞は，日常会話で多く，学会講演で少ない。
- 連体詞は，日常会話で少ない。
- 接続詞は，学会講演で多く，小説会話で少ない。
- 感動詞は，日常会話で多く，小説会話で少ない。

5.4.品詞構成比



山崎(2017:281)より

5.5.特徴語

- 対数尤度比による特徴語(上位20語)

- あなた(代名詞), わ(終助詞), よ(終助詞)
- 俺(代名詞), さん(接尾辞), ず(助動詞)
- 私(代名詞), は(係助詞), きみ(代名詞)
- た(助動詞), いる(動詞), おまえ(代名詞)
- さま(接尾辞), お(接頭辞), ぞ(終助詞)
- 男(名詞), 殺す(動詞), へ(格助詞)
- だ(助動詞), ない(形容詞)

5.6.特徴語(日本の小説と翻訳小説)

- 上位10語(固有名詞を除く)
- 日本の小説
  - さん(接尾辞), や(助動詞), な(終助詞), お(接頭辞),
  - さま(接尾辞), 先生(名詞), ちゃん(接尾辞), か(終助詞),
  - くん(接尾辞), あっ(感動詞)

- 翻訳小説
  - 彼(代名詞), 私(代名詞), きみ(代名詞), あなた(代名詞),
  - を(格助詞), 作品(名詞), 僕(代名詞), 炉辺(名詞),
  - 居る(動詞), 彼女(代名詞)

山崎(2018:9-10)より

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。
- 話者情報の付与作業には, 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, 山縣智子氏の協力を得た。

参考文献

- 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.
- 高崎みどり(1981)小説の中の会話文について,「ことば」2, 86-97.
- 宮嵜由美, 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(2017)レジスター・位相の違いによる会話文の語彙的多様性,「言語資源活用ワークショップ発表論文集2017」, pp.278-289. 2017 http://doi.org/10.15084/00001529

参考文献

- 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.

- ご清聴ありがとうございました。