自然談話コーパスに対する話題アノテーションの試み

中俣尚己[1]（京教大）・建石始（神戸女学院大）・堀内仁（国際教養大）・小西円（東京学芸大）・
山本和英（長岡技科大）

1　背景

全ての談話は話題を持つ。非母語話者のための日本語教育においても話題は切り離すことができない。従来の日本語教科書は、文法項目を易から難へと積み上げていく教授方式が主流であったが、昨今は一つの話題を中心として言語形式を扱う話題シラバスの教材も見られる。また、学習言語を使って、様々な問題を考える CLIL（内容言語統合型学習）あるいは TBL(タスク主導型学習）という学習法も注目を浴びている。仮に従来型の文法積み上げシラバスを採用するにしても、教科書には会話パートというものが必ず存在し、そこには必ず話題が存在する。

このように、話題は日本語の教材開発には欠かせない存在であるが、談話研究や日本語教育の文脈では、「話題転換」については多くの研究の蓄積があるものの、話題そのものについての研究は乏しい。唯一、話題別の語彙を提供しているリストが山内(編)(2013)であり、これは 100 の話題を用意し、それぞれについて単語と文型のリストを掲載している。さらに、山内(編)(2013)では「昨日」のような抽象的な意味を表す語は話題に依存していないとしているが、Nakamata(2019)は小規模な日本人と中国人の会話コーパスを分析し、「昨日」や「ている」のような時間を表す表現は「ポップ・カルチャー」の話題に多く出現することを明らかにしている。文法項目と特定の話題の結びつきが分かれば、より効果的な教材を作ることが可能になる。

しかしながら、山内(編)(2013)は『旧日本語教育能力検定試験　出題基準』と、『現代日本語書き言葉均衡コーパス』を元に作られたものであり、話し言葉に基づいているわけではない。また、Nakamata(2019)も日本人同士の会話を分析したわけではない。

さらに、このような研究をする時に問題となるのは、「話題をいくつ認定すべきか？」という問題である。山内(編) (2013)の 100 話題は十分な数値であるが、実際に教材開発をする際にはいささか多すぎる。中には「芸能界」と「メディア」のように明らかにまとめられそうな話題があるが、では、それをまとめていくと、いくつの話題にまで絞れるのかという問題がある。

そこで、本研究では、自然会話コーパスである『名大会話コーパス』(藤村ほか 2011)の全文を目視し、全ての発話に人手で話題タグをつけるというアノテーション作業を行った。この結果を元に、似ている話題を統合し、その後、話題別特徴語を抽出するという目論見である。本発表ではアノテーション作業の方針と、作業時に発覚した問題点ならびに対策について述べる。

[1] nakamata@kyokyo-u.ac.jp

２　アノテーション作業の方針

アノテーションはファイルを目視し、話題が変わっている箇所に行を入れ、その話題に該当すると思われる話題を山内(2013)から選び書き入れるという方法で行った。話題の変更を認定するためには、前後の話題を認定する必要があり、両者は不可分の作業と言える。

例　F144：＜笑い＞何も取り柄ないじゃないとか。

＠医療・健康

F148：昨日か何か、「あるある大辞典」で（ええ、ええ）亜鉛が大事とかっていうのをやってたんだね。

見なかった？

まず、予備調査として１つのファイルに対し日本語教育関係者５名で作業を行ったところ、分割箇所についてはほぼ揺れが見られなかった。そこで、以降は３名の作業者でアノテーションを進めた。『名大会話コーパス』は 129 のファイルからなる。これを４つのグループに等割して作業を進めた。うち２グループは筆頭発表者と２名の共同発表者で、残りの２グループはそれぞれ筆頭発表者と２名の文系の大学院生で作業を行った。

話題は山内(編) (2013)の 100 話題を基準にし、これでは不十分と思われる点については合議の際などに名称変更や追加を行った。利用した話題を別表に示す。

話題をつける単位としては５ターン以上継続したものを対象とし、４ターン未満のものは前後の話題に含めることにした。ある程度まとまった長さがなければ、文中に含まれる単語によってのみ話題に分割し、そこからまた各話題に含まれる語を抽出するという循環に陥るためである。

３　アノテーション作業の実施と課題

話題が変更する箇所については作業者間の揺れはほぼ見られなかった。むしろ合議の話し合いで時間をとったのは、特に、複数のタグのどれを採用するかということが多かった。これは、１つの文に複数の話題タグをつけようとした作業者がいたこと、また、「話題の切り方」の捉え方が作業者によって異なっていたためである。そのため、「できるだけ細かく切る」「１つの箇所はできるだけ１つの話題にする」という作業方針を立てれば、効率よく話題タグを付与できるものと思われる。

もう、日本人には想像を絶する寒さなんじゃない。	気象	気象	旅行	気象
あの、ああいう、石の場所の村とかね（うん）都会の寒さってい	気象	気象	旅行	気象
何かあのね、えっと、イタリーに行ったときに、フィレンツェ、フィレンツェで、そこに住んでる日本人の女の子とたまたま飛行機	気象	気象	旅行	気象
（うん）そしたらね、フィレンツェは、あそこほんっとに石の都会だから、（うん）あそこの冬の底冷えといったら、日本、普通の	気象	気象	旅行	気象
F002：私は夏だわ、フィレンツェ、あのツアーで（ああ）行った	旅行	旅行	旅行	旅行
（うーん）よかったわよ。	旅行	旅行	旅行	旅行
あの、メディチ家の（そうそうそう）霊廟とかね。	旅行	旅行	旅行	旅行

図１　一番右が合議の結果

話題付与の最小単位を５ターンとしたが、実際に作業を行うと、３ターンのものでもまとまった話題として認定したくなる箇所が見つかった。今回は作業方針を堅持した。一方、それ未満の長さでまとまった話題と言える箇所は少なく、３ターンを基準にして良い。

合議後に修正・追加した話題は「贈り物」「持ち物」、そして分析から除外すべき「名大会話」の３つである。また、「自動車産業」は「自動車」に、「経済・財政・金融」は「お金」に名称変更し、身近な話題に付与しやすくした。それ以外には一切変更することなく、10 代～70 代の多種多様な話題を、山内(編)(2013)に従って分類することができ、「該当する話題がなくて困る」という状況もほぼなかった。

どの話題にするべきか悩んだ際にも、例えばディズニーランドに行く話題であれば、「遊園地」など関係がありそうな語を山内(編)(2013)の索引で調べ、「旅行」と「育児」に収録されているが文脈から考えて「旅行」とするなど、山内(編)(2013)を参照することで多くの場合は解決できた。また、それ以前の作業箇所で似たような話題の時はどのように行ったかを適宜検索して一貫性を持たせた。「日帰り旅行は「旅行」と見なす」「大学の授業の話題は「大学」にする」などのルールを適宜決めることもあった。

４　基礎統計

合議後の小話題のうち、主なものの基礎統計量を表１に示す。

表１　話題分割の結果

Topic	A.Token	B.Type	C.Variation	D.Session	E.Duration	F.DF	G.Ratio
食	114,882	4,406	13.0	313	367	85	66%
言葉	54,044	2,867	12.3	119	454	60	47%
労働	40,232	2,291	11.4	116	346	58	45%
旅行	52,497	2,998	13.1	147	357	55	43%
交通	42,021	2,339	11.4	128	328	54	42%
家庭	28,549	1,733	10.3	80	357	53	41%
人づきあい	23,652	1,680	10.9	71	333	53	41%
大学	45,871	2,320	10.8	107	429	51	40%
友達	21,624	1,562	10.6	93	233	49	38%
医療・健康	42,229	2,441	11.9	73	578	47	36%

A 列は延べ語数、B 列は異なり語数である。形態素解析には Sudachi の分割モード B を用いた。C 列は語彙の豊かさを意味する語彙多様性(石川 2012)という指標である[2]。この数値が高いほど様々な語が出現していると言える。固有名詞の多い「食」と「旅行」で高く、この２つは全話題の中で最も高い。D 列は３ターン以上持続した単位の数である。E 列は話題持続度であり、A÷D で計算した。これが高い話題は「話し始めると長くなる話題」と

[2] 通例、Type÷Token の値である TTR が使われるが、コーパスサイズが大幅に異なるため、Type ÷√Token である Guiraud 値を利用した。

言えよう。専門的な話題や経験談に近いものほど高い数値を示す傾向にある。「医療・健康」は高い数値を示しているが、全体では 11 位であり、上位５話題は「工芸」「映画・演劇」「出産」「育児」「習い事」である。F 列は出現文書数（ここでは会話数）であり、G はその割合である。

「食」が話題の基本であることがわかった。「言葉」「大学」が多いのは日本語教育関係者が多い名大会話コーパスの特徴を反映している。

５　今後の課題

今後はこの分割した小話題をまとめる必要がある。方法としては「出現する単語が似ている話題は似ている」という仮説に基づいた、単語文書行列を作成し多変量解析を行う方法のほか、「連接しやすい話題は似ている」という仮説に基づいた、話題バイグラムを作成し、グラフを作成する方法も試す予定である。

また、『名大会話コーパス』の各行に対応した小話題のファイルデータを公開する予定である。

参考文献

石川慎一郎(2012)『ベーシックコーパス言語学』ひつじ書房

藤村逸子ほか(2011)「会話コーパスの構築によるコミュニケーション研究」藤村逸子・滝沢直宏編『言語研究の技法：データの収集と分析』pp. 43-72、ひつじ書房

山内博之(編)(2013)『実践日本語教育スタンダード』ひつじ書房

Nakamata, Naoki (2019) Vocabulary Depends on Topic, and So Does Grammar

Journal of Japanese Linguistics35-1

別表　利用した話題タグ　（数字は出現セッション数）

食(313), 名大会話(171), 旅行(147), 交通(128), 言葉(119), 労働(116), 大学(107), 教育・学び(96), 友達(93), 調査・研究(85), 家庭(80), 医療・健康(73), 衣(72), 人づきあい(71), 日常生活(67), 通信(63), 町(63), ヒト(62), 芸能界(57), 就職活動(47), 写真(46), お金(44), 住(43), メディア(42), 人生・生き方(42), 恋愛(42), 性格(41), 思い出(41), 喧嘩・トラブル(40), 音楽(39), 美容(38), 買い物・消費(38), パーティー(37), 映画・演劇(37), 結婚(36), 動物(35), 気象(34), 自動車(34), 年中行事(33), 贈り物(32), 家事(32), 試験(31), 文芸･漫画･アニメ(28), 国際交流･異文化理解(28), 趣味(28), 学校（小中高）(28), 家電(27), スポーツ(24), 事件・事故(24), 酒(23), 宗教・風習(22), 遊び・ゲーム(23), 育児(21), 習い事(19), コンピュータ(20), ものづくり(17), 出産(15), 絵画(14), 農林業・畜産(13), 外交・国際関係(12), ふるさと(12), 死(12), 植物(11), 夢・目標(11), マナー・習慣(10), 戦争(10), 持ち物(10), 引っ越し(9), 工芸(8), 悩み(8), 建設・土木(7), テクノロジー(7), 少子高齢化(7), 歴史(7), 社会活動(6), ギャンブル(6), 自然・地勢(6), ビジネス(6), コレクション(6), 政治(5), ジェンダー(5), 会議(4), 伝統文化・芸道(4), 社会保障・福祉(4), 環境問題(4), サイエンス(3), 芸術一般(3), 祭り(3), 国際経済・貿易(2), 災害(2), 宇宙(2), 税(2), 株(1), 文化一般(1), 不明(1), 若者論(1), 水産業(1), 法律・裁判(1)
工業一般(0), 重工業(0), 軽工業・機械工業(0), エネルギー(0), 差別(0), 選挙(0), 算数・数学(0)