2020 年 7 月 9 日

「自然会話コーパス話題アノテーション情報 ver1.0」について

中俣尚己(NAKAMATA Naoki)

１.「自然会話コーパス話題アノテーション情報」の概要

本データは大曾美恵子氏が作成された『名大会話コーパス』(藤村ほか 2009)の文字化ファイルの全ての行に対し､約 100 種類の話題アノテーション情報を与えるファイルである。

ファイルを解凍すると､『名大会話コーパス』の 129 の文字転記ファイルに対応する 129 の CSV ファイルが現れる。それぞれの CSV ファイルには行番号と、話題タグのみが含まれている。利用するためには、別途『名大会話コーパス』のファイル本体が必要となる。

『名大会話コーパス』の管理は現在は国立国語研究所に移管されており、本体ファイルは以下の URL で公開されている(URL は 2020 年 3 月時点)。そちらの利用条件なども読んだ上で利用されたい。

https://mmsrv.ninjal.ac.jp/nucc/

２．付与した話題タグセット

話題タグは日本語教育のために作成された山内(編)(2013)の 100 の話題タグセットを元に付与した。また、いくつかのタグは山内(編)(2013)をベースに日本語読解教材を話題ごとに分類するために作られた橋本(2018)のタグセットを利用している。実質的には同じであるものの、山内(編)(2013)から橋本(2018)にかけて名称の変更があったタグもあり、多くの場合で橋本(2018)に従っている。

その他、後述する合議の際に、いくつかのタグについては名称変更を行い、また､『名大会話コーパス』の録音調査そのものについて話している「名大会話」(種々の分析の対象外とすべき箇所)､「持ち物」「贈り物」という話題タグを追加した。

次ページに本データのタグの一覧を記載する。数字はその話題について話している談話上のまとまりが何例あるかというセッション数である。文の数や行の数ではないことに注意されたい。セッション数の多い順に並べた。

タグセットは全部で 104 であり、表 1 に示す通りである。うち、実際にコーパスに登場したのは 97 種類である。◆は採用したものの、名大会話コーパスには出現しなかった話題で、7 種類ある。97 種類のうち、表内で印のないものは、山内(編)(2013)から名称変更のないもので、71 種類ある。▲は山内(編)(2013)よりタグ名称を変更したものであり、変更前の山内(編)(2013)の名称を【】内に記す。全部で 18 種類ある。●は山内(編)(2013)に加えたもので、橋本(2019)の段階で加わったものと、前述の本研究が独自に加えたものがある。全部で 8 種類ある。

表 1　利用したタグセット（全 104）

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【家族】(80), ▲医療・健康【医療】(73), 衣(72), 人づきあい(71), ▲日常生活【季節・行事】(67), 通信(63), 町(63), ▲ヒト【人体】【容姿】(62), 芸能界(57), 就職活動(47), 写真(46), ▲お金【経済・財政・金融】(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【美容・健康】(38), ▲買い物・消費【買い物・家計】(38), パーティー(37), 映画・演劇(37), 結婚(36), 動物(35), 気象(34), ▲自動車【自動車産業】(34), 年中行事(33), ●贈り物(32), 家事(32), 試験(31), ▲文芸・漫画・アニメ【文芸・出版】(28), ●国際交流・異文化理解(28), 趣味(28), 学校（小中高）(28), ▲家電【家電・機械】(27), スポーツ(24), 事件・事故(24), 酒(23), ▲宗教・風習【宗教】(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【日曜大工】(17), 出産(15), 絵画(14), ▲農林業・畜産【農林業】(13), ▲外交・国際関係【外交】(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【社会運動】(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【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【国際経済・金融】(2), 災害(2), 宇宙(2), 税(2), 株(1), 文化一般(1),●若者論(1), 水産業(1), ▲法律・裁判【法律】(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)

無印　山内(編)(2013)と同じ……71 種類
▲　山内(編)(2013)より名称変更……18 種類
●　山内(編)(2013)に追加…… 8 種類
◆　『名大会話コーパス』に出現せず…… 7 種類
計……104 種類

山内(編)(2013)にあるが橋本(2019)で採用されていないため今回タグセットに採用しなかった話題タグを表２に示す。

表２　山内(編)(2013)にあるが利用しなかったタグ

手続き、感情、成績、手芸

また、以下のタグセットは橋本(2019)編の時点で統合・分割が行われていたのでそれにならった。

表３　タグの分割と統合

山内(編)(2013)	橋本(2019)ならびに本データ
【出産・育児】	「出産」と「育児」に分割
【容姿】【人体】	「ヒト」に統合

３．話題タグ付与の方針

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挿入し、その話題に該当すると思われる話題タグをタグセットから選び、＠に続けて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。

以下の例では、「＠食」がそれ以下の行の話題に該当する。また、この話題が継続する範囲、つまり次の＠が出現するまでがセッションである。CSV データでは、＠が入る行は取り除き、それ以外の全ての行について、行番号と話題の組を示している。

最終的なセッションの数は 3,461 になった。129 ファイルのそれぞれのセッション数の平均値は 26.8、中央値は 24、最小値は 4 (122.csv)、最大値は 107(87.csv)である。

例

（うん）それにあのー、今チケットってすごい全然取れないけど、（うん）あのー、取れるだけでもありがたいよねえ。

F148：そうそうそう、うん。

で、コンスタントに（うん）とりあえずは取ってもらえるわけだしね。

（うん）で、ほかにも見たいときにはそれを頼めるわけでしょう。

F144：うん。

＠食

F148 ちゃん、これ食べられる？

F148：からい。

F144：これかなりからいよねえ。

（中略）
好き嫌い、言われてみるとあるなあ。
F148：あるんだよ。
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？
F144：ああ、はいはい。
味覚障害がね、（そうそうそうそう）亜鉛がって、何か言われてるけど。

まず、予備調査として１つのファイルに対し、研究チームの既存コーパス分割班の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。なお、この５名はいずれも日本語教育を専門分野とする研究者である。

そこで、以降は１ファイルに対して、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グループは中俣、堀内、建石が、残りの２グループはそれぞれ中俣と２名の文系の大学院生で作業を行った。中俣が全てのファイルの作業を行うことで、一貫性を担保した。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

４．話題タグ付与のすり合わせと課題

その後、同じファイルに対して作業を行った３人の作業者が対面で合議を行い、一つの話題タグに決定した。１グループは約 32 のファイルからなるが、１ファイルにつきおよそ 30 分かかるため、合議は８時間×８日間を必要とした。

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。

方針では、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、今後は３ターンを

基準にして良い。

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。

５　今後の課題

話題ごとの語彙表を作成し、2020 年度末に公開する予定である。

参考文献

中俣尚己「自然談話コーパスに対する話題アノテーションの試み」『言語処理学会第 26 回年次大会発表論文集』
https://www.anlp.jp/proceedings/annual_meeting/2020/pdf_dir/P1-3.pdf
橋本直幸(2018)『話題別読解のための日本語教科書読み物リスト 2017』科研費報告書
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』p. 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

付記

本研究は JSPS 科研費基盤研究(B) 「話題が語彙・文法・談話ストラテジーに与える影響の解明」（課題番号 18H00676　研究代表者　中俣尚己）の成果物である。

本アノテーション作業には、科研メンバーのうち、特に以下のメンバーがかかわった。（※所属は 2020 年 3 月 1 日時点）

代表者：中俣尚己（京都教育大学）
既存コーパス分割班：建石始（神戸女学院大学）・堀内仁（国際教養大学）・小西円（東京学芸大学）・小口悠紀子（首都大学東京）
機械的分析班：山本和英（長岡技術科学大学）

また、アノテーション作業やタグ集計作業は学生アルバイトの皆様に協力頂いた。記して感謝申し上げる。