自然会話コーパスを元にした話題別語彙表の作成

中俣尚己（京都教育大学）

1. はじめに

すべての会話には話題が存在する。しかしながら、話題が言語の表現形式（語彙・文法・談話ストラテジー）にどのような影響を与えているかという点に着目した研究は多くない。山内(編)(2013)は話題ごとの語彙や文型を提案しているが、『旧日本語能力試験出題基準』という既存の語彙表にある語を分類したという側面が強い。Nakamata(2019)は実質語だけでなく、機能語も話題の影響を受けていることを明らかにしたが、利用したデータは小規模な接触場面会話コーパスであった。しかし、話題の統制がされていない既存の大規模な会話コーパスにも必ず話題が存在するはずである。そこで、『名大会話コーパス』（藤村ほか 2011）を人手で話題ごとのサブコーパスに分割し、話題ごとの特徴語を調査した。

2. 方法

2.1 話題ごとの分割の方法

『名大会話コーパス』の全ファイルを目視し、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。本研究では話者が交代するまでの発話の持続をターンと呼ぶ。話題タグ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ターン以下のものは前後の話題に含めることにした。

以下の例では、「＠食」がそれより下の行の話題タグである。また、この話題タグが継続する範囲、つまり次の＠が出現するまでをセッションと呼ぶ。

（１）＠食

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。

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。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、１つの話題タグに決定した。

2.2　話題タグセットについて

話題タグは山内(編)(2013)の 100 の話題タグセットを元に作成した。表２に示す。()内は名大会話コーパスに出現したセッションの数である。名大会話コーパスには 97 の話題が出現した。

表２　利用したタグセット（全 104 種）

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【人体】【容姿】(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【日曜大工】(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),●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)

無印　山内(編)(2013)と同じ　▲……山内(編)(2013)より名称変更

●……新たに追加　◆……コーパスに出現せず

2.3　特徴語の抽出

次に、comainu を用いて、BCCWJ における長単位と同じ単位で、各サブコーパスを語相当の単位に分割した。その後、話題 X を対象コーパス、話題 X 以外の全ての話題を参照コーパスとし、対数尤度比(LLR)を計算した。計算式は田中・近藤(2011)などの方法に拠った。

$$2(a\ln a+b\ln b+c\ln c+d\ln d-(a+b)\ln(a+b)-(a+c)\ln(a+c)-(b+d)\ln(b+d)-(c+d)\ln(c+d)+(a+b+c+d)\ln(a+b+c+d))$$

a：当該資料での当該語の度数　b：参照資料での当該語の度数
c：当該資料の延べ語数－a　d：参照資料の延べ語数－b
ln は自然対数を表す。a または b が 0 の場合、alna または blnb を 0 として計算する。
ad-bc<0 の場合の場合、-1 を乗じる補正を行う。

全てのサブコーパスの全ての語について計算を行った後、コーパスでの総頻度数 10 以上の語のみをまとめ、LLR の値に応じて色をつけた Excel ファイルを話題別特徴語リストとして出力した。

3. 結果

「話題別語彙表」は話題を選び、そこで使われる語彙を確認することにも、語を選びそれが使われる話題を確認することにも使えることが最大の特徴である。

3.1 話題から語彙を確認する

話題別語彙表の１行目には話題が並んでおり、ここから一つを選び Excel のフィルタ機能を使って「降順」で並べ替えるとその話題の特徴語を一覧することができる。

例えば「食」については LLR が 11 より大きい語は 241 語抽出される。「名大会話コーパス」を話題ごとに分割した結果としては、「食」の話題が一番テキスト量が多く（中俣ほか 2000）、そのため特徴語も多くなるのであるが、この中には「トマト」「おでん」「チーズ」などの名詞、「作る」「入れる」「食う」といった動詞以外にも、様態を表す「そう」や「食べれる」といった可能表現も特徴語であった。この結果は別のコーパスを用いた Nakamata(2019)の結果と一致しており、ある種の機能語が話題の影響を受けるということはほぼ確実であると言える。

よりサイズの小さな話題として「コンピュータ」の特徴語を表２に示す。

表２：話題「コンピュータ」の特徴語

ウイルス	パソコン	マック	コンピュータ	ウィンドウズ
ノート	使う	エクセル	ソニー	開く
移す	のです	メール	人間	物理
古い	ソフト	へーえ	もの	せい
ゲーム	止まる	とりあえず	のだ	事務
2000	らしい	先生方	壊れる	問題

表２に出てくるカタカナ語は 2020 年現在から見れば少し古く感じられるものがあるかもしれない。『名大会話コーパス』の収録時にはまだ YouTube などは一般的でなかったため、固有名詞などには相違が見られる。しかし、「せい」「止まる」「とりあえず」「壊れる」からはトラブルについて語られている様子がわかる。「へーえ」「らしい」などからは情報の非対称性が読み取れる。説明のモダリティとされる「のです」「のだ」が特徴語として抽出されたことも興味深い。

（２）　でなんか、このアプリケーションがなかなか開かないんだって、それはウイルスのせいじゃあー。(data127)

（３）　ただ、あのー、古いので、インターネットをするのに問題があるんですよ。(data023)

（４）えっ、F005 ってさ、今さ、ウィンドウズ使ってんの、マック使ってんの。
両方使ってる。
両方使ってんの？へーえ。(data023)

3.2　語から話題を確認する

話題別語彙表の１列目の単語を検索すると、その語がどの話題でよく使われているか確認できる。類義の形容詞を例にあげると、「ハンサム」は「映画・演劇」の話題でのみ特徴語となる。対して「かっこいい」は「写真」「衣」「恋愛」「芸能界」などより広い範囲で使われ、「きれい」はそれらに加えて「旅行」「工芸」「建築・土木」「植物」にも使われることがわかる。このように類義語についても新たな情報を与えることが可能になる。

また、表２の「とりあえず」は「コンピュータ」と「喧嘩・トラブル」の話題のみで特徴語となっており、日常会話ではトラブル対応場面でよく使われていることがわかる。

機能語の「てほしい」は「恋愛」「出産」の特徴語であった。パートナーに対する要望が多いかと思いきや、（５）のように、第三者の恋愛を応援するような例もあった。一方で、使役の「せる」「させる」は「学校（小中高)」「教育・学び」「育児」「農林・畜産」の特徴語であり、被使役者は子どもか動物に限られるということが改めてわかった。

（５）　あ、ほんとに F114 ちゃんたちうまくいってほしい。(data066)

（６）　文法解説書を買わせても、いまいちわかんないんだ。(data106)

4.　おわりに

「話題別語彙表」はトピックに基づいた授業・会話例を作ること、あるいは実質語・機能語の解説に際して「その語がどのような話題の会話で使われているのか」を示す資料となる。類義語の違いではコロケーションの違いも重要であるが、コロケーションはリストという形で示すと情報量が膨大になる。話題の違いという観点を提供していきたい。

ひとまずは発表者の web サイトで Excel 形式で公開するが、より利便性のある形での提供を模索したい。97 の話題はいささか細かい分類かもしれないが、この中には互いに関係の深い話題もあり、プロジェクトではその関係についても調査を進めている。関係する話題についての特徴語も表示できるようになれば、さらに利便性は高まると考えられる。

参考文献

藤村逸子・大曽美恵子・大島ディヴィッド義和(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

Nakamata, Naoki(2019) Vocabulary Depends on Topic, and So Does Grammar, Journal of Japanese Linguistics: 35-2, 213-234.

中俣尚己・建石始・堀内仁・小西円・山本和英「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』

田中牧郎・近藤明日子(2011)「教科書コーパス語彙表」『言語政策に役立つ、コーパスを用いた語彙表・漢字表等の作成と活用』pp.55-63, 2011 文部科学省科学研究費特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：２１世紀の日本語研究の基盤整備」言語政策班

山内博之（編）（2013)『実践日本語教育スタンダード』東京: ひつじ書房.