『日本語話題別会話コーパス:J-TOCC語彙表』の公開と日本語教育むけ情報サイトにむけた指標の検討

中俣尚己(大阪大学)・麻子軒(関西大学)

1. はじめに

日本語話題別会話コーパス: J-TOCC

120ペアの大学生に15の話題について5分ずつ会話してもらったコーパス(中俣ほか2021)
・延べ10時間、165万語
・話題/性別の組み合わせ/録音地(東西)を比較可能
・日本語教育への応用が目的

身の回りの話題	社会にもかかわる話題
01.食べること 02.ファッション 03.旅行 04.スポーツ 05.マンガ・ゲーム 06.家事 07.学校 08スマートフォン 09.アルバイト 10.動物11.天気	12. 夢・将来設計 13.マナー 14. 住環境 15.日本の未来

2. 公開データの概要と作成手順

「話題別特徴語彙表」 対数尤度比(LLR)&頻度
縦に語、横に15話題が並んだxlsxファイル(4シート)。
LLRは特徴度の頑健な指標(内山ほか2004)
「話者別使用頻度表」 頻度
縦に語、横に240人の参加者が並んだ15のcsvファイル。
集計することで、「何%の母語話者が該当の話題でその語を使用したか」という話者使用率(UR)を計算可能。
※使用ソフト:Comainu 0.72
全てのデータに対し短単位と長単位を用意
※LLRは1つの話題を該当コーパス、
他14話題を参照コーパスとして計算(中俣2015)

最終的にはファイルではなく現場で使いやすい情報サイトにまとめるのが目標。そこで利用する指標として、LLRだけでなくURも有効ではないか?

3. データの分析

表1.「01.食べること」のLLRが高い語とそのUR

V	LLR	UR	N	LLR	UR	A	LLR	UR
食べる	6555	93%	ラーメン	1411	39%	美味しい	2665	78%
食う	917	36%	寿司	1106	32%	好き	2089	90%
飲む	260	19%	肉	593	31%	旨い	274	29%
頼む	235	18%	焼き肉	575	23%	甘い	251	17%
太る	207	14%	カレー	477	21%	大好き	150	16%
作る	90	38%	御飯	452	36%	辛い	147	15%
焼く	78	10%	オムライス	391	14%	脂っこい	65	3%
好く	45	10%	チーズ	379	13%	安い	42	20%
並ぶ	38	7%	味	364	28%	しょっぱい	41	2%
炙る	33	1%	食べ物	353	28%	濃い	34	5%

表2.「01.食べること」のURが高い語とそのLLR

V	UR	LLR	N	UR	LLR	A	UR	LLR
言う	94%	-44	事	82%	27	好き	90%	2089
食べる	93%	6555	奴	50%	0	美味しい	78%	2665
思う	80%	-57	人	46%	-106	凄い	48%	-10
分かる	70%	5	感じ	45%	-2	多い	43%	29
出る	40%	13	方	40%	-1	確か	42%	-2
違う	38%	2	ラーメン	39%	1411	まじ	30%	2
作る	38%	90	御飯	36%	452	旨い	29%	274
食う	36%	917	家	36%	17	やばい	28%	2
入る	30%	-4	本当	35%	1	そんな	27%	1
知る	28%	-4	物	33%	50	嫌	27%	-32

●動詞・形容詞には、LLRが高くてもURが低い語が存在する。

●対照的に、名詞でLLRが高い語はURも高い。これ以降も上位20語はURが10%を超え、38語まではURが5%を超えている。

●URが高いが、LLRが低い語には一種の機能語や、情報処理のマーカー、口語的表現が含まれる。特徴語でなくても重要と考えられる。

●「入る」は「店に入る」「砂糖が入っている」のように「食」のコロケーションで使われているが、「部活に入る」「バイトが入る」など他話題でより多く使われているために、LLRが低くなる。

●「もの」は「食」のLLRが高い。(中俣2015)

●表は「01.食べること」についてのデータであるが、「06.家事」と「15.日本の未来」でも同様の傾向を確認。

4. 指標の検討

●UR, 頻度、tf-idfの相関はどれも0.8以上で高い。
●URは母語話者のX%が使用するという意味で理解しやすい。会話コーパスの観察単位を人にするという点(森2017,2019)からも支持される。
●URとLLRの相関は下記の通り。

All	V	N	A
.21	.48	.48	.74

5. おわりに

●話題からそこで使う語彙を探せる情報サイトの構築。トピックベースやタスクベースの授業構築支援。
●「炙る」は確かに「食」だが、重要度の観点から、複数の指標が必要ではないか?

●名詞の選定にはLLRのみで十分である。
●動詞・形容詞の選定にはLLRに加えてURを加味し、URが低い語の重要度を下げ、URが高い語を加えるべきである。

参考文献
内山将夫・中條清美・山本英子・井佐原均(2004)「英語教育のための分野特徴単語の選定尺度の比較」『自然言語処理』11:3, 165-197.
中俣尚己(2015)「「日中Skype会話コーパス」を用いた話題別語彙の抽出—「食」の場合—」
中俣尚己・太田陽子・加藤恵梨・澤田浩子・清水由貴子・森篤嗣(2021)「「日本語話題別会話コーパス:J-TOCC」」『計量国語学』33:1, 205-213.
森秀明(2017)「コーパス間における単語使用率の比較—観察単位(ケース)は単語か文書か—」『計量国語学』31:3, 205-221.
森秀明(2019)『コーパスの計量的分析法再考』東北大学博士論文.

本研究はJSPS科研費18H00676、22H00668の助成を受けた。