国立国語研究所学術情報リポジトリ

‘Vocabulary Table of Japanese Topic O riented Conversation Corpus’ and a Consideration of Indices for Japanese Language Teaching

メタデータ	言語: jpn 出版者: 公開日: 2023-03-24 キーワード (Ja): キーワード (En): 作成者: 中俣, 尚己, 麻, 子軒, NAKAMATA, Naoki, MA, Tzuhsuan メールアドレス: 所属:
URL	https://doi.org/10.15084/00003723

『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　国際教育交流センター）†
麻　子軒（関西大学　国際教育センター）

'Vocabulary Table of Japanese Topic-Oriented Conversation Corpus' and a Consideration of Indices for Japanese Language Teaching

NAKAMATA　Naoki (Osaka University)
MA Tzuhsuan (Kansai University)

要旨

『日本語会話話題別コーパス:J-TOCC』の語彙表を公開する。表は２種類で、15 ある話題間での特徴度を比較するための粗頻度ならびに LLR の表と、各話題ごとに、240 名の調査協力者がそれぞれ何度その語を使用したかというデータを収めた表である。前者の表はどの話題に特徴的かという偏りを表し、後者の表は、ある話題を与えられた時に母語話者の何%がその語を使用するかという「使用者割合」を取り出せる。本プロジェクトの最終目標は日本語教育に役立つ「話題–語彙情報サイト」の構築であるが、現場に役立つ形で情報を整理するにはこの２種類の情報が必要であることを主張する。語の使用者の幅を見る指標としては tf-idf なども存在するが、検討の結果、本データでは使用総頻度の影響が大きすぎることがわかった。一方で、LLR は語の特徴語を効率よく抽出できるが、多義語など、他の話題の影響で値が低くなることもある。使用者割合はその点をカバーすることができる。

１．はじめに

筆者らは 2021 年 3 月に『日本語話題別会話コーパス：J-TOCC』(以下、「J-TOCC」)を公開した(中俣ほか 2021)。J-TOCC は、日本語教育において話題を中心とした学習活動や教材開発を支援するため構築されたコーパスであり、筆頭著者のウェブサイトで公開されている (http://nakamata.info/database/) 。 J-TOCC の名称は Japanese Topic-Oriented Conversation Corpus の略で、「ジェイトック」と読む。

J-TOCC では同じ会話協力者のペアが 15 の話題について、それぞれ５分間ずつ会話を行ってもらった。話題以外の条件が統制されていることが最大の特徴である。話題は表 1 に示した通り、身の回りの話題が 11、社会に関わる話題が 4 である。１つの話題につき 120 ペア、合計 10 時間の会話が収められている。語数は記号を除くと延べ語数ではおよそ 165 万語、異なり語数ではおよそ 4 万語である。話題による語数の差は大きくなく、１話題あたりおよそ延べ語数は 11 万語、異なり語数は 6 千語程度となっている。ペアは様々な話題について話せるよう、仲の良い関係の大学生どうしに限定し、性別の組み合わせや録音地の東西でバランスをとった。

表 1 J-TOCC の話題

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

† n.nakamata.ciee@osaka-u.ac.jp

また、話者がそれぞれの話題についてどのぐらい詳しいかという「話題精通度」についても 5 段階で尋ね、付属データとして添付している。

今回は、話題ごとの語の使用実態を明らかにするために、基礎資料として『日本語話題別会話コーパス：J-TOCC 語彙表』を公開する。本発表ではその概要について述べるとともに、日本語教育への応用のため、いくつかの指標についても検討する。

２　公開データの概要と作成手順

２.１　公開データの概要

公開するデータは２種類ある。１つは、「話題別特徴語表」であり、縦に語が並び、横に 15 の話題が並ぶ。それぞれの語がどの話題で多く使われているかを見るための表である。語の切り方として、短単位と長単位があり、また指標として粗頻度と対数尤度比(LLR)を用意し、４枚のシートからなる xlsx ファイルとして提供する。

もう１つは「話者別使用頻度表」であり、こちらも縦に語が並び、横には J-TOCC の 240 名の参加協力者が並ぶ。そして各セルには各話者がその話題で何度その語を使用したかの数字が並ぶ。ここから 240 人の話者のうち何人が特定の話題で語を使用したかという使用率(UR)を計算することができるほか、話者の個人差も考慮した統計モデリングを行うこともできる(中俣 2023 予定)。話題ごとに表を作り、15 個の csv ファイルで提供する。また、こちらも短単位と長単位の両方を用意した。

２．２　形態素解析の方法及びデータの加工

形態素解析を行うのに用いた解析器、辞書といった解析環境は下記の通りである。

短単位解析器　MeCab 0.996
短単位解析辞書　UniDic-MeCab 2.1.2
長単位解析器　Comainu 0.72
長単位解析モデル　CRF++-0.58

手順は、まず J-TOCC の平文を Mecab と UniDic で短単位に分割し、次にそれを Comainu と長単位解析モデル CRF で長単位にまとめ上げる、という流れになるが、Comainu には Mecab を呼び出す機能が実装されているため、これら一連の処理は一度の実行で行えた。形態素解析の処理を行う前に、本文と見なさない話者記号及び丸括弧・山括弧で示されたメタタグとその内容（下記の例の下線部分）を削除した。

E-101-1M：お金かかるけど、（E-101-2M：うん）うまいんだよな。
E-101-2M：好きだ（笑）。
W-109-2M：いや、結構＜会話が重なり聞き取り不能＞

語彙表を作成する際に、形態素解析を終えた文字列に対し、以下の加工を行った。

外来語は UniDic では、「ファッション-fashion」のように語彙素の後ろに原語の綴りが併記されているが、検索の妨げとなるため、原語の綴りの部分を削除した。

また、LINE のような半角のアルファベットの文字列を短単位に分割すると、「LINE」のように１語になる場合と、「L」「I」「N」「E」のように４語になる場合とがある。後者は長単位への結合処理で自動的に LINE の１語になるが、アルファベット１文字の語彙素じゃ UniDic では全角で登録されているため、結合後は全角「ＬＩＮＥ」に変換される。一方、最初から短単位で「LINE」１語に切られた場合は半角であるため、作成した語彙表には同じ語にも関わらず、全角と半角が混在してしまう。これを解決するために、全角への一括置換を行った。

２.３　LLR の計算方法

LLR の計算には対象コーパスと参照コーパスが必要となるが、今回は１つの話題を対象コーパスとし、残り 14 の話題を合わせたものを参照コーパスとした。計算方法は下記の通りである(田中・近藤 2011)。

2(alna+blnb+clnc+dlnd-(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

３．データの分析

３.１　LLR の概要

以降、「01.食べること」をケーススタディーとして取り上げる。短単位のデータを使用する。まず、単純に LLR が高い語を取り上げると、表 2 のようになる。

表 2　「01.食べること」で LLR が高い語

語	LLR	語	LLR	語	LLR
食べる	6555	ラーメン	1411	肉	593
美味しい	2665	寿司	1106	焼き肉	575
好き	2089	食う	917	料理	487
外食	1520	【店名】	817	カレー	477

表 2 はどれも納得のいく語ではあるが、『J-TOCC』語彙表では、特徴語の基準となる LLR10.83 以上の語が「01.食べること」だけでも 750 語以上存在する。この豊かな情報は、一方で具体的な教育場面では、一体どれを選べばよいのかというネックともなりうる。そこで、以下では代表的な品詞に分けて分析を行うとともに、もう一つの指標として 240 名のうち何%が使用したかという「使用率」(UR)にも着目する。

３.２　動詞の場合

ここでは、UniDic において品詞が「動詞―一般」であるものを対象にする。表 3 は左側に LLR が高い語、右側に UR が高い語をそれぞれ 10 語ずつ並べたものである。

表 3　「01.食べること」における高 LLR 動詞と高 UR 動詞

高 LLR 語	LLR	UR	高 UR 語	UR	LLR
食べる	6555	93%	言う	94%	-44
食う	917	36%	食べる	93%	6555
飲む	260	19%	思う	80%	-57
頼む	235	18%	分かる	70%	5
太る	207	14%	出る	40%	13
作る	90	38%	違う	38%	2
焼く	78	10%	作る	38%	90
好く	45	10%	食う	36%	917
並ぶ	38	7%	入る	30%	-4
炙る	33	1%	知る	28%	-4

まず、左側を見るとLLRが高い語であっても、「炙る」のように使用者割合が1%というような場合も存在し、これはあまり重要な語とは認定できない。また、「食う」も教育という観点からはなかなか選びにくい語であろう。予想通り、これは男女によって使用割合に大きな差が見られ、男性の使用率は63%、女性の使用率は9%である。LLRの値は上位10語ほどではないが、使用率が比較的高い語としては売る(LLR20、 UR20%)や「痩せる」(LLR19、UR14%)がある。

次に、右側のURに注目すると「言う」のように必ずしも話題とは関係のない語も選ばれてしまう。しかしながら、「言う」「思う」は一種の機能語として、「分かる」「違う」は情報処理のマーカーとして頻繁に使用されていると考えられる。また、「入る」は「店に入る」「砂糖が入る」のような「01.食べること」と関係するコロケーションが多くあるものの、それ以上に「部活に入る」(04.スポーツ)、「バイトが入る」(09.アルバイト)など他の話題でより多くコロケーションが使用されたために、LLRが低く、特徴語とは判定されていない。このように、LLRは参照コーパスの性質によって、数値が変動する可能性がある。同様に、URが11番目に高い語として「買う」があるが、この語もファッションでの頻度が非常に高かったため、UR27%に対し、LLRは-5であった。

３.３　名詞の場合

ここでは、UniDicにおいて品詞が「名詞―普通名詞―一般」であるものを対象にする。表4は左側にLLRが高い語、右側にURが高い語をそれぞれ10語ずつ並べたものである。

表4　「01.食べること」における高LLR名詞と高Ratio名詞

高LLR語	LLR	UR	高UR語	UR	LLR
ラーメン	1411	39%	事	82%	27
寿司	1106	32%	奴	50%	0
肉	593	31%	人	46%	-106
焼き肉	575	23%	感じ	45%	-2
カレー	477	21%	方	40%	-1
御飯	452	36%	ラーメン	39%	1411
オムライス	391	14%	御飯	36%	452
チーズ	379	13%	家	36%	17
味	364	28%	本当	35%	1
食べ物	353	28%	物	33%	50

名詞の特徴として、LLRが高い語はURも軒並み高いことが挙げられる。これ以降も、20位まではURは10%を超え、38位まではURが5%を超えている。一方、右側のURに注目すると抽象的な語が多く抽出されてしまう[1]。名詞についてはLLRのみで十分と言えるかもしれない。

３.４　形容詞の場合

ここでは、UniDicにおいて品詞が「形容詞―一般」「形状詞―一般」であるものを対象にする。

表5　「01.食べること」における高LLR形容詞と高UR形容詞

[1] ただし、「もの」はLLRが50と高い。「もの」が食の特徴語であることは別コーパスを用いた研究(中俣2015)からも示唆されている。また、副詞可能な名詞含めると、「時」「後」「最近」などの語が使用率が高い。

高 LLR 語	LLR	UR	高 UR 語	UR	LLR
美味しい	2665	78%	好き	90%	2089
好き	2089	90%	美味しい	78%	2665
旨い	274	29%	凄い	48%	-10
甘い	251	17%	多い	43%	29
大好き	150	16%	確か	42%	-2
辛い	147	15%	まじ	30%	2
脂っこい	65	3%	旨い	29%	274
安い	42	20%	やばい	28%	2
しょっぱい	41	2%	そんな	27%	1
濃い	21	5%	嫌	27%	-32

状況としては動詞に似ており、LLR が高くても UR が低い語が見られる。反対に UR が高いが LLR が低い語には「凄い」「まじ」のような程度表現や「やばい」「多い」「嫌」のように「01.食べること」に関する会話でよく使用される語が含まれており、使用率を LLR と併用することで、より効果的に語彙を選定できるようになると考えられる。

４．指標の検討

ここまで述べてきたようなこと、例えば「炙る」は食の特徴語と言えるが重要度は低いということは経験のある教師ならば直観でわかることである。しかしながら、筆者らは話題を選ぶと必要な語彙を提供したり、語を選ぶと相性の良い話題を提供する「話題―語彙情報サイト」の構築を進めており、そのためには上記のような直観も何らかの形で指標化する必要がある。ある話題に特徴的かという指標としては LLR が頑健で一貫性のある指標であることがわかっている（内山ほか 2004）。一方で、重要度の指標としては今回用いた使用率(UR)のほか、使用頻度(frequency)や、その２つを組み合わせた tf-idf などの指標も用いられている。そこで、以上の４指標についての相関係数の計算を行った。記号を除いた全語、動詞、名詞、形容詞についてそれぞれ計算を行った結果が表 6 である。

表 6　指標間の相関係数

全語	UR	freq.	tf-idf	LLR	動詞	UR	freq.	tf-idf	LLR
UR	1				UR	1			
freq.	0.79	1			freq.	0.85	1		
tf-idf	0.82	1.00	1		tf-idf	0.89	0.99	1	
LLR	**0.21**	0.18	0.20	1	LLR	**0.48**	0.81	0.79	1
名詞	UR	freq.	tf-idf	LLR	形容詞	UR	freq.	tf-idf	LLR
UR	1				UR	1			
freq.	0.96	1			freq.	0.91	1		
tf-idf	0.95	0.98	1		tf-idf	0.95	0.99	1	
LLR	**0.48**	0.61	0.67	1	LLR	**0.74**	0.90	0.87	1

まず、UR、frequency、tf-idf の３指標はどの品詞でも非常に相関が高い。このうち、教育目的としては使用率が有効な指標と言える。頻度では 258 回という数字の意味は単独ではわからないが、使用率は「その話題においては、母語話者の 80%が使用する」というように単独で意味を解釈できる指標だからである。また、森(2017、2019)におけるコーパスの単位を文書（人）にすべきであるという主張からも正当性があると言える。

次に、LLR と UR の相関に着目すると、全語では 0.21 で低く、動詞と名詞では 0.48 と

中程度の相関を示し、形容詞では 0.74 と高くなった。特に形容詞の相関がこれだけ強くなったのは不明である。LLR は他の 3 つとは異なるタイプの指標であるとは言える。

また、３．における高 LLR 語と高 UR 語の比較から、指標の選定については以下のように言えそうである。

●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。

上述の提言は他の話題のデータにも成り立つであろうか。「06.家事」と「15.日本の未来」について調査を行ったところ、同様の傾向を示すことが確認された。

５．おわりに

本研究では『日本語話題別会話コーパス：J-TOCC 語彙表』を作成・公開した。語彙表は特徴度を示す LLR の表と、重要度を示す「話者別使用頻度表」からなる。名詞においては特徴度が高い語はたいてい重要度が高かったが、動詞や形容詞においては両者は必ずしも一致しないことがわかった。

謝　辞

本研究は JSPS 科研費 18H00676、22H00668 の助成を受けた。

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.

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。

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―」『第 8 回コーパス日本語学ワークショップ予稿集』11-18. http://doi.org/10.15084/00003427

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.

中俣尚己(2023 予定)「間投助詞「さ」の使用に話題が与える影響」中俣尚己(編)『話題別コーパスが拓く日本語教育と日本語学』ひつじ書房.

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.

関連 URL

データベース｜中俣尚己ウェブサイト　http://nakamata.info/database/