「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（京都教育大学）†

Extraction of Topic-Specialized Vocabulary from "Skype Corpus" : A Case for the Topic of 'Eating'

Naoki Nakamata(Kyoto University of Education)

要旨

本発表では、発表者が構築した「日中 Skype 会話コーパス」を用い、会話で使用される語彙について分析する。このコーパスは日本の大学生と中国の大学生が Skype で会話交流活動を行ったのを継続的に録音、文字化したもので、真正な会話であるとともに、各回の話題が指定されていることに特色がある。今回は「食」がテーマの回とそれ以外のテーマの回に分け、日本語解析システム「雪だるま」を使って単語に分割した。その後、LLR を指標として「食」関連語が抽出できるかを検証した。結果、特徴度が高かった語は基本的に「食」に関連する語であり、高い精度で抽出できた。これは、会話コーパスにおいて話題の設定が重要であることを再確認できたと言える。

１．　はじめに

この発表の目的は２つある。１つは発表者が構築し、2015 年 4 月 1 日から公開している『日中 Skype 会話コーパス』の諸特性を紹介することである。もう１つは、その特性の１つである「会話の話題が決められている」点に着目し、話題別の語彙抽出を行った結果を示すことである。結果は高い精度を示しており、会話コーパスの構築においてはごく簡単にでも話題をあらかじめ決めておくことで、語彙表の作成に役に立つデータを得ることができると言える。

２．　『日中 Skype 会話コーパス』の紹介

２．１　『日中 Skype 会話コーパス』の概要

『日中 Skype 会話コーパス』は 2012 年５月～７月に、東京・実践女子大学と長沙・湖南大学の学生間で行った Skype を利用した遠隔会話活動(中俣ほか 2013)を録音、文字化したもので、接触場面の会話コーパスに分類される。中国側の学習者は全員２年生で、日本側の母語話者は学部３年～M１の学生で日本語教育を専攻したり、関連する授業を受講していた学生である。３ヶ月の間、ペアを固定し、１週間に１度のペースで Skype を用いた会話活動を行った。実際にはビデオ通話ではあるが、行ったのは録音のみで、現時点で公開しているのはその文字化資料のみとなる。

コーパスには延べ９ペア、３８の会話を収録している。総会話時間は 46:48:35 で、１会話あたり平均 1:13:55 とまとまった長さの会話と言える。後述する日本語解析システム「雪だるま」を使って分析した結果、総語数は 204,632 語であった(記号類を除く)。

コーパスはテキストファイルで提供され、笑いや発話の重なりといった簡単な記号を含んでいるが、これらは正規表現で簡単に取り除けるようになっている。コーパスの配布は

† nakamata[at]kyokyo-u.ac.jp

http://nakamata.info/database.html で行っている。氏名とメールアドレスを登録すればすぐにダウンロードできる。

会話活動の詳細な報告は中俣ほか(2013)、Skype コーパスそのものの説明については中俣(2015)にて詳しく説明している。

２．２　『日中 Skype 会話コーパス』の特性

『日中 Skype 会話コーパス』の言語資料としての特徴として、以下の４つを挙げる。

Ａ．　真正性がある。

このコーパスの設計はもともとコーパスを作ろうとしたものではなく、まずは Skype を用いた会話活動を通し、中国の学習者には学んだ日本語を使う機会を提供するとともに学習意欲を継続させること、日本の母語話者には外国人と文化交流をしたり日本語を教えたりしながら、日本語について考えてもらうことが第一の目的であり、それにあわせて計画がデザインされている。そのため、真正性のある接触場面コーパスになっている。以下、いくつかの語について、代表的な学習者コーパスである KY コーパスと比較したものが表１である。OPI という統制された会話である KY コーパスには出現しないような語が多数出現していることがわかる。

表１　KY コーパスと日中 Skype 会話コーパスの出現数の比較[1]

語	KY コーパス	日中Skype会話コーパス
明後日	0	7
木曜	6	41
すごい	77	211
すごく	190	86
すげえ	0	4

Ｂ．　縦断的なデータである。

会話活動は１週間に１回、継続的に行った。最も多いペアで７回分の会話があり、縦断的にデータを観察することができる。

Ｃ．　一種の電話場面である。

終結部には、例えば突然食事の話題をふって、会話を終結にもっていく前終結の段階が存在するなど、電話場面と同様の構造が観察される(橋内 1999)。　また、コミュニケーション・ブレイクダウンや沈黙も多く観察される。

Ｄ．　話題が指定されている。

各回は次ページの表２のように話題が指定されており、数字はファイル名の末尾の数字

[1] 北村・冨岡・川村(2009)はコーパスの出現文書数から語の難易度を求める試みであるが、「あさって」「おととい」のような語は基本語であるものの、コーパスに出現しにくいという問題点を指摘している。また、CSJ と BCCWJ の調整頻度レベルでは一番頻度が少ない曜日は木曜である(Tono, Yamazaki and Maekawa 2013)。

に対応する。しかし、話題は必ずしも厳密に守られているわけではなく、話がそれたり日本語についての質問が行われることもある。これらの話題は事前に日中双方の学生から話してみたいことのアンケートを行い、決定した。

敬語に関しては張(2012)が、敬語について学習者で意義などについて話し合うことの効果を報告していることから採用した。

表２　日中 Skype 会話コーパスの話題

1	ポップカルチャー	6	伝統・行事
2	料理	7	夏休み・夏の予定
3	家庭・家族・子供	8	大学生活
4	故郷・今住んでいる場所	0	指定なし・トピック認定できず
5	敬語		

３．　「食」関連語彙の抽出

３．１　特徴語抽出の意義

日本語教育における教材作成において、語彙の選定は重要な作業である。中俣(2014)は文法積み上げ型シラバスを念頭に、特定の文法項目と共起する語彙をピックアップしているが、現在では話題シラバス・場面シラバスの教材も増えてきている。話題シラバス・場面シラバスの教材作成にあたっては、話題ごとにどのような語彙が用いられるかということが重要である。

話題ごとの語彙をまとめた重要な先行研究として山内(2013)『実践日本語教育スタンダード』（以下、実践 S）をあげることができる。実践 S はまず 100 の話題を選び、各話題ごとにまず文型を設定する。そしてその文型に入りうる名詞をパラディグマティックな形で提示したものであり、各名詞は難易度によって３段階に分けられている。実践 S の最初の話題は「食」であり、以下、「1.1.1.1.　食名詞：具体物」の【料理名：個体】の名詞を引用する。

表３　山内(2013)『実践日本語スタンダード』の一例

意味分類	A	B	C
【料理：固体】	カレー、パン、ごはん、サラダ、うどん、そば	サンドイッチ、ステーキ、ハンバーグ、刺身	ライス、粥、実、麺、漬物、～漬け

しかし、これらの語のピックアップや難易度判定は執筆者の主観に基づくものである。会話コーパスから機械的に話題関連語を抽出できれば、客観的かつ大規模な語彙表を作成することができ、さらに教材作成に活かすことができる言語資料となることが期待される。そこで本発表では、『日中 Skype 会話コーパス』から「食」関連語彙を機械的に抽出し、既存の語彙表である実践 S との比較を行う[2]。

2　ただし、実践Ｓの批判が目的ではない。山内(2013)は以下のように述べる。

このようなパラディグマティックに対立する語群を眺めると、語同士を直接比較できるようになるため、個々の語のレベル設定が非常に行ないやすくなる。（略）「同じ文の同じ位置に現れ得る語同士

３．２　手法

まず、コーパス全体を「料理」が話題の食コーパスとそれ以外が話題の対照コーパスに分割した(語数は食コーパスが 28,960 語、対照コーパスが 175,352 語)。一方で、学習者と母語話者の発話は分割しなかった。これは、表４に示す通り、接触場面においては学習者と母語話者の語彙に顕著な差は存在しないからである。

表４　『日中 Skype 会話コーパス』における話者別の異なり語数と延べ語数

話者	異なり語数	延べ語数	TTR
中国人学習者	5,374	103,883	0.0517
日本人母語話者	4,923	100,749	0.0489

細かく語彙を分析しても「母語話者はよく使うが、学習者はあまり使わない」あるいはその逆の語というものは一部の機能語的な語に限られていた[3]。実質語に絞って話者別に特徴語を抽出しても話題別の特徴語よりも少ない量しか抽出できない。特徴語を抽出する上では語数は多いほうが良いため、話者による語彙の違いは捨象して計算した。

次に、各コーパスを日本語解析システム「雪だるま」(http://snowman.jnlp.org/)にかけ、単語ごとに分割、品詞も付与した[4]。この「雪だるま」は長岡技術科学大学の山本和英氏が開発したシステムで、形態素ではなく「単語」に分割することを目的とし、「気が早い」のような慣用句、「かもしれない」のような複合辞、「勉強する」のようなサ変動詞、「無理だ」のような形容動詞をそれぞれ１語として出力することができる。解析は 2015 年 7 月 18 日に行った。

最後に、解析結果を元に、特徴度の指数として、田中・近藤(2011)を参考に対数尤度比(LLR)を補正した値を計算した。計算式は下記の通りである。

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。

教科特徴語リストに合わせ 0.1%水準で有意となる 10.83 よりも大きい語を「食」特徴語と認定する。

の比較が可能」ということに大きな意味がある。（略）また、表９(発表者注：上記表３のこと)を見ると、「パスタ」と「ラーメン」が入っていないことに気づく。「パスタ」と「ラーメン」が入っていないことに気づくことができるのも、パラディグマティックに対立する語が集められていることの賜物である。従来よく見られた五十音順の配列による語彙表では、よほどのパスタフリーク、ラーメンマニアでない限り、「パスタ」や「ラーメン」がないことには気づかないものと思われる。(p.12)

つまり、実践Ｓは話題関連語がパラディグマティックに配列されるという「枠」を示したことに大きな価値がある。本発表はその「枠」の中にさらに実際のデータから具体的な語を入れ込むことができるか、という検証であり、両者は相補的な関係にあると考える。

[3] どのような語に差異が見られるのか、またなぜ実質語には差異が見られないのかといった考察は別稿(中俣　準備中)に譲る。

[4] 2015 年７月現在、限定公開となっている。興味をお持ちの方は山本和英氏まで。

３．３　結果

発話の断片（「レタス」と言おうとして「タス」になったものなど）を誤解析したものを除くと、244 語を抽出できた。これは食コーパスのうち、異なり語数の 11.9%、延べ語数の 16.0%をカバーする。表５に品詞ごとの数を示す。また、この数字はあくまでも機械的に抽出された語数である。そこで、実際に目視でそれぞれの語が食に関連する意味で使われているかを確認した。

表５　品詞ごとの「食」特徴語の語数

名詞 (複合名詞)	動詞 （非自立含む）	形容詞 （非自立含む）	その他 （副詞、感動詞、助詞、助動詞、複合辞）
190 語　83.7%	35 語 80.0%	11 語　90.9%	8 語

感動詞や助詞（「なあ」）が特徴語とは考えられないが、助動詞「られる」、複合辞「ないで」に関しては、食の場面でよく使用される可能性は考えられる。今後の課題としたい。

＜例１＞

Ｃ：うん。なぜ日本では、このチンジャオロースはとても有名です、か。

Ｊ：家庭一でよく食べます。中華料理の中でも、＜うん＞よく作られる。

＜例２＞

Ｊ：朝ごはん食べないで会社とか学校行って、お昼食べて夜食べて、の２食っていう生活の人、が多いですね。

以下、表６、７、８はそれぞれ名詞、動詞、形容詞・副詞の語彙リストであり、実践 S にならって提示してみる。

表６　「食」特徴語名詞リスト（190 語/83.7%）

【食べ物】料理、食べ物、もの
【食事】朝ごはん、弁当、給食、朝食、夕食、間食、昼食、懐石料理、昼
【料理名・固体】年越し、刺身、煮物、餃子、パン、寿司、餅、粥、ピータン、チンジャオロース、肉じゃが、麺類、ご飯、天ぷら、麺、ワンタン、焼き魚、チャーハン、回鍋肉、お好み焼き、カレー、ハンバーガー、きりたんぽ、ハンバーグ、ピザ、焼きそば、くさや、酢豚、ダック、卵焼き、サンドイッチ、スペアリブ、天津飯、水餃子、麻婆豆腐、関東煮、天津丼、中華丼、北京ダック、ピータン豆腐、チャオピン、親子丼、卵かけごはん、ジャージャー
【料理名・液体】スープ、味噌汁
【菓子・デザート】まんじゅう、肉まん、あんまん、クレープ、菓子、アイスクリーム、綿あめ、饅頭、ホットケーキ、綿、中華まん、チョコまん
【飲み物】梅酒、牛乳、紅茶、豆乳、酒、ジャスミン茶、日本酒、緑茶
【食材】肉、パスタ、アヒル、卵、なす、トマト、玉ねぎ、野菜、小麦、じゃがいも、犬、米、魚、ピーマン、レタス、生卵、納豆、いちご、中身、パプリカ、大根、食材、ネギ、にんじん、乾物、のり、小麦粉
【調味料】醤油、塩、わさび、あんこ、つゆ、山椒、油、めんつゆ、ティエン、調味料

【調理器具】鍋
【調理の場所】台所
【食器】椀、皿、箸
【飲食店】食堂、餅屋、回転ずし
【行列】満員
【味】味、舌、バニラ、味覚
【食欲】食欲
【団らんの場所】テーブル
【量】１杯、２杯
【調理法】生、生もの、固め
【未分類】茶道、赤、つば、系統、値段、１００、黄色、中国料理、日本料理、鍋料理、家庭料理、北京料理、四川料理、比較文化、食文化、１６元、広東料理、１００種類、福建省、東北人、湖南料理
【誤抽出】平成、子供、名刺、元号、字、みず、西暦、オン、メンツ、オッケー、体面、字幕、ビデオ、何、福山、キャンパス、テスト、比較、映像、テキスト、気晴らし、新暦、学期、皇暦、1つ、岳麓山、生田斗真、１時、はなみずき、新垣結衣、聴解、声優

表７　「食」特徴語動詞リスト（35 語/80.0%）

揚げる、切る、食べる、焼く、入れる、煮る、作る、潰れる、つける、煮込む、しびれる、かける、混ぜる、点てる、開ける、食べれる、盛る、冷やす、いためる、作る、たらす、さっぱりする、くさる、溶く、保つ、つつく、練る、かぐ
【誤抽出】数える、登る、参加する、主演する、通じる、延ばす、鍛える

表８　「食」特徴語形容詞（11 語/90.9%）

甘い、おいしい、辛い、臭い、薄い、辛い、苦い、酸っぱい、安い、簡単
【誤抽出】ふさわしい

３．４　考察

３．４．１　抽出精度とカバー率

まず、誤抽出の語について考えてみたい。ここを見ると、「平成」「元号」「西暦」「皇歴」といった暦に関する語群があることに気づく。これはある会話の終わりに、突然学習者が暦に関する質問をしたためである。その他の誤抽出の語も、会話の一部の個所で集中的に出現しており、別の話題についての個所であることが明白である。

このコーパスでの話題は、前もって表２のテーマについて話すように指示しただけであり、実際に会話参与者がそれを厳密に守っているわけではない。今回、分析対象をファイル丸ごとにしたため、このような語も「食」関連語として抽出されたが、内容を仔細に観察し、話題ごとに区切ってコーパスを作れば、誤抽出の語はほぼ全て排除できる。

つまり、話し言葉であれば、規模が数万語のコーパスであっても話題の特徴語は 100%に近い精度で機械的に抽出できるということである。この精度は子供話し言葉コーパスの特徴語分析(中條ほか 2005)、Facebook と Twitter の比較(石井 2011)、twitter を用いた時制関係語の抽出（赤崎ほか 2013）といった他分野の特徴語抽出の試みよりも明らかに高い。多くの実質語は話題に従属するという山内(2013)の方針が実証されたと言えよう。また、こ

の事実は会話コーパスを作る時、緩やかにでも話題を指定しておくと、日本語教育の教材作成に非常に有益な結果が得られるということを意味している。

その一方で、本当にすべての「食」関連語がカバーできているかという問題も残る。例えば、今回の調査では「食」コーパスにのみ、1例だけ出現した「味わう」のような低頻度語は抽出できない。これはコーパスサイズを大きくすることでしか対処できないかもしれない。

３．４．２　直感では気づきにくい特徴語

次に、個々の語について見ていく。もちろん、一見して食に関連する語が多く抽出されたわけであるが、機械的に抽出を行うメリットは直感では見逃してしまうような語も発見することができる点にある。例えば、【食べ物】に分類される名詞として「もの」が抽出されている。その理由は、以下のような例が「食」コーパスに多く見られたためである。

＜例３＞

Ｊ：えーと、ハンバーグというのは、あのー、お肉とか、あの、み、ミンチのお肉とか、あの、タマネギを刻んだものとかを、えーと、ね、練り合わせて、卵とか、小麦粉とかを練り合わせて焼いたもの。これがハンバーグで、ハンバーガーというのは、パンの間にそのハンバーグとか、レタスとか、チーズとかが挟んであるものがハンバーガーです。

「辛いもの」といった単純な例も「食」コーパスに見られたが、＜例３＞のような「～を～したもの」という構文は「食」コーパスにのみ出現した。これは料理の説明をする時に頻用され、また使えると説明がスムーズにいく項目であると言える。

また、動詞では「潰れる」「しびれる」「保つ」などが出現しているが、これらはそれぞれ「お酒を飲んで潰れる」「四川の本格マーボーは舌がしびれる」「調理時、一定温度に保つ」といった文脈で使われている。

これらの構文や語は実践Ｓには収録されていない。

３．４．３　難易度をどう考えるか

実践ＳではＡ、Ｂ、Ｃの三段階で難易度が表示されているが、コーパスの出現頻度から再考できる余地がある。

表９　実践Ｓと「食」コーパスの比較

	焼く	煮る	炒める
実践Ｓ	Ａ	Ｂ	Ｂ
「食」コーパス	48 回	26 回	9 回

また、実践 S では「焼ける」「グラム」「センチメートル」といった語が難易度Ａになっていたが、これらは『日中 Skype 会話コーパス』全体を通しても出現しない。

さらに、「鍋」と「包丁」はどちらも実践ＳではＡランクにあたり、直感的にもどちらも調理に不可欠の道具であるように思えるが、『日中 Skype 会話コーパス』全体での出現数は鍋が 34 回に対し、包丁は 0 回である（フライパンは 4 回）。つまり、実際の重要度と「会話で使用するか」ということは全く別の次元の尺度であり、コーパスからわかる「会話でどのぐらい使うか」という情報が会話教材において重要になると考える。

４．　おわりに

この発表では、『日中 Skype 会話コーパス』の特性について紹介し、「食」特徴語の抽出を行った結果を発表した。会話コーパスの特徴語抽出において話題が果たす役割を極めて大きいと言える。また、機械による特徴語抽出は、直感では気づきにくい語を抽出したり、難易度を考慮することにより、日本語教材作成に貢献できることを示した。

謝　辞

本研究は、JSPS 科研費(26770180)による補助を得ました。また、LLR の計算方法については帝塚山大学の森篤嗣氏、コーパスに出現しにくい語については東京国際大学の川村よし子氏に助言を頂きました。また、単語解析は山本和英氏と長岡技術科学大学自然言語処理研究室のメンバーが作成した「雪だるま」を利用させて頂きました。お世話になった皆様に感謝申し上げます。

文　献

赤﨑優介・森田和宏・泓田正雄・青江順一(2013)「Twitter を用いた時制を表す特徴語の自動収集に関する研究」『言語処理学会第 19 回年次大会発表論文集』

石井健一(2011)「Facebook と Twitter の発言における特徴語の比較 (http://hdl.handle.net/2241/115339 よりダウンロード可能)

北村達也・冨岡洋介・川村よし子(2009)「IDF を用いた単語レベル判定システムの構築と検証」『日本語教育方法研究会誌』16(1), pp.52-53

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

中條清美・西垣知佳子・内山将夫・中村隆宏・山﨑淳史(2006)「子供話し言葉コーパスの特徴語抽出に関する研究」『日本大学生産工学部研究報告 B 文系』39,pp. 65-78,日本大学生産工学部.

張贇(2012)「敬語コミュニケーション学習における「変容」に関する考察 : 上級学習者の事例分析から」『待遇コミュニケーション研究』9, 待遇コミュニケーション学会

中俣尚己・漆田彩・小野真依子・北見友香・竹原英里(2013)「Skype を活用した日中会話交流プログラム」『実践国文学』83, pp.132(25)-109(48), 実践国文学会

中俣尚己(2014)『日本語教育のための文法コロケーションハンドブック』くろしお出版

中俣尚己(2015)「日中 Skype 会話コーパスについて」(http://nakamata.info/about_skype_corpus.pdf よりダウンロード可能)

中俣尚己(準備中)「接触場面における学習者と母語話者の語彙はどこが異なるのか？―「日中 Skype 会話コーパス」の分析―」『日本語／日本語教育研究会第 7 回大会予稿集』

橋内武(1999)『ディスコース　談話の織りなす世界』くろしお出版

山内博之(2013)『実践日本語教育スタンダード』ひつじ書房

Tono, Y., Yamazaki, M., Maekawa, K. (2013) *A Frequency Dictionary of Japanese* Routledge.

関連 URL

中俣尚己のウェブサイト http://nakamata.info/

雪だるまプロジェクト http://snowman.jnlp.org/