真正性のある接触場面会話コーパスを用いた話題特徴語の抽出

―ポップ・カルチャーの場合―

中俣尚己（京都教育大学）

1. はじめに

本研究は，トピックシラバスに基づいた教材を作る上で欠かせない語彙の選定を，実際にそのトピックについて話している会話コーパスのデータから半自動的に行うという試みである。

近年は日本語教育でも語彙に関する研究が盛んに行われているが(森(編)印刷中など)，特に山内(編)(2013)は「語彙は話題に従属する」という考えの元，100の話題に関して語彙表を整備した研究として特筆すべきものである。しかしながら，山内(編)(2013)は語の選定，難易度の判定の大部分は直感あるいは既存の教材に委ねており，学習者の接触場面の会話において必ずしもそれらの語が出現するかどうかは未検証である。本研究は山内(編)(2013)の枠組みに従いつつ，実際のデータから特定のトピックに出現する語彙表を構築する。

また，山内(編)(2013)は時間を表す表現などは「特定の話題に従属しない語」であるとしている。しかし，本研究の結果は，いわゆる機能語とされる語の中にも，特定の話題と相性が良い物が存在することを示唆する。これは，語彙学習だけでなく，いわゆる文法学習を目的とした教室活動も，特定の話題と結びつけて考えることができるということである。このことは，無論，経験的には教師に知られていたことであるが，それをデータで実証可能であることを示すことに本研究の意義がある。

2. 先行研究

特定の文書群と他の文書群を比較し，特定の文書群によく出現する語を特徴語として抽出する研究は広く行われており，指標としては対数尤度比(LLR)が効果的とされる(中條ほか2005)。日本語を対象にした大規模な研究としては田中・近藤(2011)の教科特徴語の研究があり，各教科の教科書に特徴的な語を，BCCWJの書籍コーパスと比較することで大量に抽出している。しかし，これまでの研究は異なる特徴を持つコーパスの比較が中心であり，同一人物の会話どうしを比較した研究はまだ少ない。

発表者は，中俣(2015b)において，自身が構築した『日中Skype会話コーパス』を分析した。このコーパスが話題が指定されていることを利用し，「料理」の話題の会話を分析したところ，244語が抽出され，80%以上が実際に「料理」と関係していることが確認された。これは十分に高い精度で，日本語教育に貢献できる情報であるが，誤抽出の語があることは問題である。しかし，誤抽出の語は，会話が「料理」の話題から逸れた箇所に集中していた。すなわち，中俣(2015b)は「料理」が話題の回のファイルまるごとを分析対象にしたことに問題があり，会話内容を精査し，実際にその話題について話している箇所だけを切り出してサブコーパスを構築すれば，100%に近い精度で抽出できることが予想された。

3. 方法

3.1 使用したコーパスについて

分析には筆者が構築した『日中 Skype 会話コーパス』を使用した。このコーパスは 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。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(2015a)を参照してほしい。

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 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。

『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

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。

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		

3.2 特徴語抽出の手順

まず、コーパス全体を目視し、「ポップ・カルチャー」が話題になっている箇所を抜き出し、その部分を特定コーパスとし、それ以外の部分を対照コーパスとした。この作業は調査協力者と発表者の２人で行い、意見が分かれた箇所は合議で決定した。ポップ・カルチャーにはドラマや音楽、アニメーションなどを含めるが、古典文学は含めない。語数は句読点などを除いて、特定コーパスが 24,496 語、対照コーパスが 180,589 語であった。

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３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(中俣 2015b)。

表３　『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(中俣 2015b)

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る(中俣　印刷中)。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。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12月26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。

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

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。

教科特徴語リストに合わせ0.1%水準で有意となる10.83よりも大きい語をポップ・カルチャーの話題特徴語と認定する。

4. 結果

発話の断片（「テレビ」と言おうとして「テレ」になったものなど）を誤解析したものを除くと、特徴語として251語が自動抽出された。これは特徴コーパスのうち、異なり語数の11.9%、延べ語数の28.7%をカバーする。以下、品詞ごとに代表的な語と語数を表4にまとめる。

表4　ポップ・カルチャー特徴語（品詞ごと）

品詞	語数	精度	代表的な語
一般名詞	103	91%	アニメ、映画、ドラマ、歌、題名、歌手、人気、曲、誰、主人公、番組、マンガ、グループ、テレビ番組、推理、音楽
固有名詞	92	100%	嵐、蛍の光、木村拓哉、SMAP、亮さん、ジェイ・チョウ、ハンガー・ゲーム、AKB、サザエさん、貞子、ナルト、陰陽師
動詞	19	95%	見る、聞く、知る、出る、読む、歌う、流れる、描く、参加する、見れる、おすすめする、はやる、捨てる、調べる、感動する
形容詞	17	100%	人気、この、好き、面白い、可愛い、かっこいい、有名、新しい、古い、大人気、無理、怖い、ソフト、真面目、爽やか
副詞	8	100%	ニコニコ、とっても、いろいろ、最近、去年、昔、今、さっき
感動詞	9	0%	あああ、ふうーん、へええ、ん、んー、うん、よーし、のう、え
機能語	3	？？	た、ている、の

精度については、その語が「ポップ・カルチャー」の文脈で語られているかどうかということを、発表者が特徴コーパスの原文を目視して確認した。感動詞と機能語を除いて、ポップ・カルチャーと関係のない語として判断したのは次ページの表5に示した10語である。

表5　誤判定と考えられる語

広い意味では関係するもの	日本、台湾、インターネット、無料、情報、ネット、集める
話題が微妙にずれたもの	校歌、スケジュール、卒業式

「日本」などは、「日本のテレビ」のように使われており、この語そのものはポップ・カルチャー関連語とは呼べないが、広い意味では関係しているというものである。そもそも、対数尤度比は特徴コーパスに多く出現する語という意味であり、語の意味に踏み込むものではない。ポップ・カルチャーについて話す時に「インターネット」「情報」などの語も出現しやすくなるということは考えられる話である。

他方、「校歌」などは全体としてポップ・カルチャーの話をしていても、挿話の形などで、やや話題が逸れたところに出現した語である。このようなものと、感動詞を除けば、抽出された251語は基本的にポップ・カルチャーに関連しており、効率よく抽出できたといえよう。

5. 考察

5.1　作品外語彙と作品内語彙

251語のうち、動詞や形容詞はそれぞれ10%にも満たず、少数の語彙が選ばれていた。反対に名詞は77%を占め、語彙学習における名詞の重要性を示している。ただし、ここで問題になるのは作品内の世界を表す語彙も含まれるということである。例えば、一般名詞「宇宙」は『宇宙兄弟』というマンガ原作の映画の内容を説明するのに使われている。固有名詞「イタチ」はマンガ『NARUTO』に登場する忍者である。このような語が多数を占めるのであれば、ポップ・カルチャー関連語とは言いがたい。名詞について、それが作品内の世界か作品外の世界を表すかについて分類を行ったところ、表6のように、作品外語彙が圧倒的に多くなった。よって、これらの語は「ポップ・カルチャーについて語る時によく使われる語」であるといえる。

表6　作品外語彙と作品内語彙

	作品外語彙	作品内語彙
一般名詞	「監督」「キャラクター」など 77例	「政治」「ロボット」など 16例
固有名詞	「福山雅治」「コクリコ坂」など 88例	「ピカチュウ」「ナルト」など 4例

5.2　抽出された機能語から考える教室活動

機械による特徴語抽出の利点の1つは、直感では気づかないような語を拾い上げることができる点である。その意味では、「誰」がかなり高い特徴度を示したことは注目に値する。このような疑問詞が話題に従属するとは一見思えないが、「食」の話題と比べれば、「誰が出演するのか」「誰が監督か」「誰が主人公か」ということで「誰」が話題になることは相対的に多いと言える。つまり、疑問詞「誰」を使う活動と「ポップ・カルチャー」という話題は相性が良いということになる。

同様に、疑問詞「いつ」こそ抽出されなかったが、時間を表す副詞的な表現として「最近」「去年」「今」「昔」「さっき」が抽出されたことも興味深い。そしてこれに呼応するかのように時間を表す「ている」と「た」も抽出されている。山内(編)(2013)では例えば時間を表す副詞などは「話題に従属しない」語の例とされている。しかし、「従属」とまでは言わなくても、(少なくとも教室活動を考えるレベルでは)「相性のよい話題」が存在すると考えられるのではないだろうか。

例えば、「駅前の鯛焼き屋はおいしいです。」と「駅前の鯛焼き屋はおいしかったです。」はもちろん、事実か体験かという意味では異なるが、それを受けた聞き手の行動がテンスによって変わるとは考えにくい。一方で、「「マッドマックス」は面白かったです。」と「「バットマンVSスーパーマン」は面白いです。」では、聞き手の行動が変わってくる。この話題ではテンスの違いはより重要な意味を持つ。テンスを使い分けなければならないという必然性が生まれるので、それを利用した教室活動(昔見た映画を紹介する/今やっているドラマを紹介する)が可能になるのである。

6. おわりに

本研究では、『日中 Skype 会話コーパス』から「ポップ・カルチャー」に特徴的な語を半自動的に 251 語抽出した。コーパスの規模から考えれば、量、質ともに高く、この手法は話題シラバスの教材作成に大変有益であると言える。さらに今回の分析では、これまで話題に従属しないと考えられていた機能語も、相性の良い話題が存在することが示唆された。これは話題から教室活動を考えるヒントとなる情報である。このことは無論経験的には知られていたことであるが、それをデータから定量的に示せたことに本研究の意義がある。今後は複数の話題を対象に同様の調査を行い、特徴語のリストを作成、相性の良い教室活動の提案を行っていきたい。

参考文献

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

中條清美・西垣知佳子・内山将夫・中村隆宏・山﨑淳史(2005)「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

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)

中俣尚己(2015a)「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(2015b)「「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』pp. 11-18

中俣尚己(印刷中)「学習者と母語話者の使用語彙の違い―『日中 Skype 会話コーパス』を用いて―」『日本語/日本語教育』7

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

森篤嗣(編)(印刷中)『ニーズを踏まえた語彙シラバス』くろしお出版

山内博之(編)(2013)『実践日本語教育スタンダード』ひつじ書房

謝辞 本研究は JSPS 科研費若手研究(B)26770180 の助成を受けた。