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

本研究の目的

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。

*語彙に関する研究　(森(編)2016など)

これまでの重要な成果

- 山内(編)(2013)
  『実践日本語教育スタンダード』
  (以下、実践S)
- 実質語は話題に従属するという考え方。

1.1.1.1.　食名詞：具体物の【料理名：個体】

- まず，100の話題を選定。
- 各話題ごとに文型を設定。
- その文型に入りうる名詞を
パラディグマティックに配列。
- 「A・B・C」３段階の難易度。

意味分類	A	B	C
【料理名：個体】	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

本研究が目指すもの

- 実践Sでの語の選定や難易度判定は大部分が執筆者の主観に基づくもの。
- 学習者の接触場面にその語が必要かは未検証。
- 会話コーパスから機械的にその話題に
  従属する語彙を抽出できれば,
  客観的かつ大規模な語彙表を作成できる。
- 実践Sの枠に，データから具体的な語を
  流し込む作業。
- 「話題に従属しない語」……本当に？
- 機能語と話題の関連性が分かれば、教室活動を考える手がかりに。

2．先行研究

特徴語抽出とは

- 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いる。
- 指標は（色々あるが）対数尤度比(Log-Likelihood Ratio)が効果的とされる(内山ほか2004, 中條ほか2005)。
- 田中・近藤(2011)　教科特徴語
- 中俣(2015b)　話し言葉「食」
- 山内・橋本(2016)　書き言葉「食」

中俣(2015b)

- 自身が構築した『日中Skype会話コーパス』について「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。
- 十分に高い数値であるが、誤抽出の語があることが問題。
- 「料理」回のファイルをまるまる分析対象にしたため。
- 誤抽出の語は、話題が逸れた箇所に集中。
- 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のでは？
- 本研究へ

３．方法

- ３．１　使用したコーパスについて
- ３．２　特徴語抽出の手順

3.1　使用したコーパスについて

「日中Skype会話コーパス」

- 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を利用した遠隔日本語会話活動(中俣ほか2013)を録音、文字化したもの。接触場面会話コーパス。

「日中Skype会話コーパス」とは

- 中国人学習者は全員２年生。
  日本人は３年生～M1。
- 9ペア。38会話。
- 総会話時間46:48:35。
  １会話あたり平均1:13:55。
- 語数は約20万語。
  （後述の「雪だるま」で解析）
- テキストファイル。
  笑いや発話の重なりといった簡単な
  記号を含む。

利用するには

中俣のホームページから無料でダウンロードできます。（「Skype　コーパス」で検索）

特徴１：真正性がある

- コーパスのための活動ではない。
- 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。
- →真正性のある接触場面の雑談。

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	Skype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

「明後日」や「木曜」は基本語であるのに，コーパスに出現しにくい(北村・冨岡・川村　2009)

特徴２：縦断的なデータ

- 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回分の会話がある。

特徴３：一種の電話場面

- 終結部などは電話場面そのものの展開が観察される(橋内1999)。
- コミュニケーション・ブレイクダウンや沈黙の研究にもどうぞ。

特徴４：話題が指定されている

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		

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も。

3.2　特徴語抽出の手順

- コーパス全体を目視し、「ポップ・カルチャー」が話題の特定コーパス(28,960語)とそれ以外が話題の対照コーパス(175,352語)に分割。
（調査協力者と発表者の２人で行った。）
- 「ポップ・カルチャー」にはドラマ・音楽・アニメーションを含めるが、文学は含めない。
- 学習者と母語話者は分割しない。
→なぜか？

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　（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**

接触場面において学習者と母語話者の語彙の違いは小さい。

- 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る。(中俣印刷中)
- 話者の違いによる特徴語＜話題の違いによる特徴語

日本語解析システム「雪だるま」

- 長岡技術科学大学の山本和英氏が開発。
- 形態素ではなく「単語」に分割することを目的とする。(森2016)
- 「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。
- 解析は2015年12月26日に行った。

対数尤度比を計算

- 田中・近藤(2011)の補正値
  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。
- 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」特徴語と認定する。

４．結果

３．３　結果

- 発話の断片（テレビと言おうとして「テレ」）を除いて，**251語**を抽出。

- 「ポップ・カルチャー」コーパスのうち……

異なり語数の11.9%
延べ語数の28.7%
(機能語・感動詞を含む)

参考：　「食」コーパス
異なり11.9% 延べ16.0%

表４－１　代表的な特徴語

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、テレビドラマ、カラオケ、作品、漫画、人、一人、舞台、闘争、コント

※語数は機械的に抽出された語。精度はそのうち，実際に話題に関連している割合。

表４－２　代表的な特徴語

品詞	語数	精度	代表的な語
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、セーラームーン、福山、福山雅治、Shine、山口、ピカチュウ

※語数は機械的に抽出された語。精度はそのうち，実際に話題に関連している割合。

表４－３　代表的な特徴語

品詞	語数	精度	代表的な語
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、出演する、主演する

※語数は機械的に抽出された語。精度はそのうち，実際に話題に関連している割合。

表４－４　代表的な特徴語

品詞	語数	精度	代表的な語
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

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

５．考察

- ５．１　作品外語彙と作品内語彙
- ５．２　抽出された機能語から考える教室活動

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。

- 一般名詞「宇宙」→『宇宙兄弟』
- 固有名詞「イタチ」→『NARUTO』

- このような語が多数を
  占めているのではないか？

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」「バンド」など77例	「政治」「ロボット」「お見合い」など16例
固有名詞	「福山雅治」「コクリコ坂」「貞子」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

- 直感では抽出しにくい「誰」が抽出された。
- 「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」など「誰」が話題になることは相対的に多いと言える。
- 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。
- 他の疑問詞も？

「食」に関係しない語彙

- 疑問詞「いつ」は抽出されなかったが……
- 「最近」「去年」「今」「昔」「さっき」
- 「ている」「た」
- 山内(編)(2013)では「話題に従属しない」語。
- 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」が存在するのでは。

機能語が重要になるか？

- 「食」の場合
  - 「駅前の鯛焼き屋はおいしいです。」
  - 「駅前の鯛焼き屋はおいしかったです。」
    - 文の意味は異なるが、聞き手の行動は同じ。
- 「ポップ・カルチャー」の場合
  - 「昔「セーラームーン」を見ました。」
  - 「今「セーラームーン」を見ています。」
    - 聞き手の行動が異なってくる。
- 「ドラマ」や「映画」には時間的制約があるので、テンス・アスペクトが重要。
- 真正性のある活動になる。

教室活動の例

- 昔見た作品を紹介する。

  私は【時間の副詞】、
  ○○○○を見ました。
  ××××　が出ていました。

  ……
  ○○で見ることができます。
  ぜひ見て下さい。

教室活動の例

- 今放映中の作品を紹介する。

私は【時間の副詞】、
○○○○を見ています。
××××　が出ています。

……

○○チャンネルで、○○時から
やっています。
ぜひ見て下さい。

難易度による話題の分類
（山内・橋本2016:59より抜粋）

I-a	「町」「家族」「趣味」など
I-b	「食」「衣」「旅行」「交通」など
I-c	「住」「日常生活」「絵画」など
II-a	「ふるさと」「友達」「容姿」など
II-b	「音楽」「映画・演劇」「芸道」など
II-C	「文芸・出版」「家事」「祭り」など
III	「言葉」「思い出」「悩み」など
IV	「算数・数学」「サイエンス」など

機能語の観点を加えると

- 「衣」「食」「住」
  ……時間的変化がない。
- 「ふるさと」「ポップ・カルチャー」
  ……時間的制約がある。
  テンス・アスペクト
- 「サイエンス」
  ……ヴォイスなども重要。

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質・量ともに十分で、話題シラバスの教材作りに貢献可能。
- 機能語の中にも特定の話題と相性が良いものがある。逆に、機能語を練習する際に、その違いが真正な意味を持つ話題がある。
- これらは経験的には知られていたが、会話データから抽出できるという方法論を示し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、同様の調査を行いたい。

参考文献（追加）

- 内山将男・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
- 森篤嗣(2016)「旧JLPT語彙表に基づく形態素解析単位の考察」庵功雄・佐藤琢三・中俣尚己(編)『日本語文法研究のフロンティア』
- 山内博之・橋本直幸(2016)「第２章　教育語彙表への応用」　砂川有里子(編)『講座日本語コ–パス５．コーパスと日本語教育』

発表は以上です。
ご意見・ご質問よろしくお願いします。

利用希望者は
「Skypeコーパス」で検索！