語彙は話題に従属する、文法も話題に従属する。

中俣尚己（京都教育大学）

1．はじめに

言語学習における「語彙」の重要性

本研究では……

- 「話題」に注目。
- 真正性のある接触場面会話コーパスを話題ごとに分割したうえで、話題別の特徴語を抽出した。
- 名詞だけでなく、種々の機能語も「話題別特徴語」だった。
- 教材開発のうえで重要な知見。

2.　先行研究

- 話題ごとの特徴語抽出の研究は長い歴史がある。
- 小宮(1995)などの専門語の研究も、一種の話題別特徴語研究。
- 昨今ではコーパスから対数尤度比(LLR)を用いて語を抽出する方法が効果的とされる。(内山ほか2004)
- 橋本(2016)では「食」の話題について特徴語を抽出し、構文ごとまた難易度ごとに特徴語を配置。これを100の話題に拡張したものが山内編(2013)。

これまでの研究は「書き言葉コーパス」から特徴語を抽出していた。

- 会話コーパスを元に特徴語を抽出した研究

- 中俣(2015b)　「食」
- 中俣(2016a)　「ポップ・カルチャー」

過去の研究で気になる点が……。

- 「ている」「た」など、従来の観点では特徴語とは考えられないような機能語が抽出された。
- 複数の話題の特徴語(松下2016)ということも考えられる。
- 本研究では複数の話題からなるサブコーパスに対して、
  全てLLRを計算し、
  本当に文法が話題に依存するのかを検証。

３．方法
3.1　使用したコーパス

「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。

「日中SKYPE会話コーパス」の四大特徴(中俣2016B)

1. 真正性がある。
2. 縦断的なデータである。
3. 一種の電話場面である。
4. 話題が指定されている。

３．２　サブコーパスの構築

- 話題が指定されているとはいえ、真正性を重視しているため、話題が逸れることもあり、これを考慮しないと特徴語の抽出はうまくいかない(中俣2015b)。
- 目視で実際に話題について話しているところだけを抜き出してサブコーパスを作ると、抽出された語の90％以上が実際に話題に関連していた(中俣2016a)。
- 発表者と調査協力者の２名でコーパス全文を目視し、話題ごとに分割。
  14のサブコーパスを構築した。

表１：サブコーパス一覧

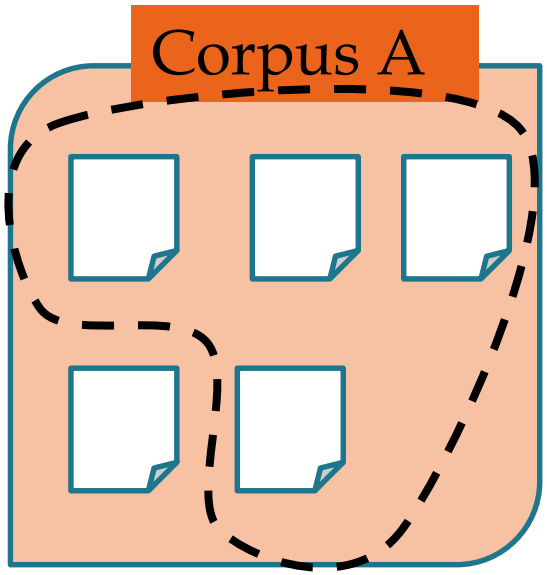
Topic	Token	Type	Topic	Token	Type
ポップ	25,675	2,102	大学	14,693	1,484
家庭	6,991	894	街	13,524	1,296
●開始	7,198	766	★天気	2,751	410
休暇	5,223	708	伝統	30,946	2,320
言語	37,412	2,095	食	29,443	2,025
●終結	10,398	869	★恋愛	2,163	388
★小中高	5,787	778	◆その他	18,429	1,763

★は当初話題としては指示していなかったが、その話題が頻繁に現れたため、サブコーパスにした。
●は電話場面ということを考慮し、サブコーパスにした。
◆はLLRの計算は行わない。

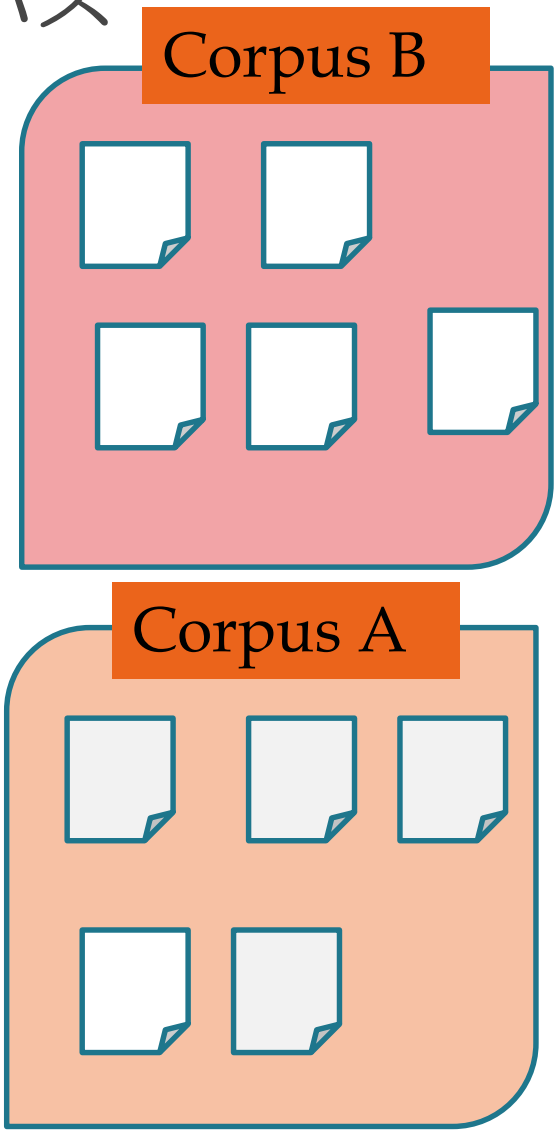
３．３　LLRの計算

- 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。(山本ほか2015)
- 「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2017年5月13日に行った。

対象コーパスと参照コーパス



今回はこちらを採用。
参照コーパスとして
適当な「接触場面会話
コーパス」が存在しな
いため。



LLRの計算式(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。
- 0.1％水準で有意となる**10.83**よりも大きい語を特徴語と認定する。

４．結果

- 以下、話題ごとに、
  左側に実質語、
  右側に機能語の特徴語を表示します。
- フィラーの類は除去しました。

ポップ・カルチャー

見る　(465.7)
人気　(419.1)
映画　(404.7)
アニメ(339.3)
ドラマ(187.4)
題名　(128.4)
好きな(127.5)
歌手　(118.8)
嵐　　(97.0)
マンガ(91.6)
聞く　(88.9)
曲　　(83.0)
知る　(81.0)

この　(144.4)
誰　　(97.2)
ている(38.7)
た　　(34.9)
の　　(24.6)　※格助

「今」「最近」「昔」
「去年」などの
時間表現も特徴語

家庭

子ども(95.9)
おじいさん　(92.9)
家庭　(84.8)
主婦　(77.0)
魚　　(70.3)
網・捕まえる　　(61.5)
ゴミ　(60.6)
家事　　(55.5)
一人っ子(54.3)
猫ちゃん　(47.4)
毛　　(45.6)
燃やす・倹約・かくれんぼ　(40.9)

頃　　(43.1)
時　　(26.5)
と　　(25.7)　格助
た　　(20.5)
も　　(17.2)

開始部

聞こえる(486.3)
こんにちは　(408.8)
もしもし　(359.8)
大丈夫　(281.5)
もしもーし　(162.4)
今日　(146.0)
悪い　(132.4)
すむ　(124.3)
映る　(95.0)
見える(69.5)
切れる(67.0)
こんばんは　(60.8)
声　(58.1)

ません(92.2)
ちょっと (91.1)
ます　(58.6)
た　(57.2)
か　(38.8)
です　(21.3)
ない　(18.9)　※助動

休暇

夏休み(353.1)
帰る　(181.9)
実家　(85.6)
休み　(77.9)
汽車　(72.6)
列車　(58.4)
１ヶ月(54.8)
西安　(52.3)
旅行　(51.6)
８月　(50.0)
こもる(45.0)
７月　(43.5)
ゴールデンウィーク(41.6)

まで　(47.9)
くらい(34.1)
から　(28.6)　※格助
に　(24.7)
うちに(22.2)
たい　(18.6)
です　(14.4)
ん　(10.9)　※んです

言語

敬語　(1061.9)
使う　(1058.0)
先生　(248.6)
尊敬語(207.8)
難しい(203.1)
言葉　(166.2)
謙譲語(152.4)
類義語(144.8)
丁寧　(138.5)
話す　(117.0)
目上・差し上げる(114.3)
意味　(108.4)
何々　(105.6)

に対する(180.2)
例えば(153.2)
場合　(138.7)
を　(122.0)
って　(81.3)
ても　(70.8)
時　(69.2)
と思う(58.6)
ば　(52.9)
と　(48.4)　※条件
とか　(33.5)
時に　(32.5)
れる　(19.4)

終結部

来週　(736.4)
日　(527.8)
じゃあ(503.5)
バイバイ(107.3)

「から」は抽出されなかった。→話題に関係なく因果関係を表している。
「ので」は相手への配慮が必要な約束場面に多い。

また　(242.7)
ございます(119.3)
から　(62.5)　※格助
にする(53.6)
まで　(25.2)
までに(25.0)
こそ　(20.1)
ので　(14.8)

小中高

小学生(180.4)
受験　(126.3)
制服　(107.6)
ルール(95.7)
光　　(74.4)
学校　(73.0)
保護者・セーター(72.0)
厳しい　(71.3)
高校　(70.3)
ブレザー(64.8)
蛍　　(60.0)
かばん(57.6)
セーラー服・スクールバス　(50.4)

とか　(15.6)
ちゃ　(11.7)　※ては
てく　(11.3)

大学

授業　(284.4)
大学　(238.6)
試験　(225.6)
学生　(124.2)
３年生(99.8)
レポート(98.6)
学校(93.2)
終わる(92.3)
就職　(83.5)
勉強する(70.6)
短大　(68.6)
夢中・サークル　(63.5)
多い　(63.0)

たり　(16.2)
くらい (15.3)
に比べ(14.7)
ます　(13.3)
の　　(11.1)　※格助
た

動詞文が多いことを
意味している。

街

故郷　(234.3)
東京　(180.0)
冬　(176.3)
長沙　(168.2)
北京　(162.4)
寒い　(135.9)
雪　(123.6)
田舎　(109.0)
きれい(96.4)
降る　(95.4)
桂林　(88.0)
所　(86.7)
山　(85.9)

が　(16.2)
たい　(1
存在文
し　(15.3)
の　(13.8)　※格助

LLR10.5に「や」。
並列表現と相性が
良い。

天気

暑い　(505.6)
雨　　(137.7)
天気　(110.9)
30度・大雨・蒸し暑い (69.5)
35度　(67.8)
涼しい(67.5)
最近　(62.9)
降る　(57.1)
気温　(51.3)
何度　(50.5)
今日　(47.5)
温度・20　(39.1)

です　(43.2)
くらい (38.7)
まだ　(28.7)
たびに(26.0)
は　　(14.2)
ね　　(13.4)

名詞文が多いことを意味している。

伝統

端午　(322.8)
節句　(230.6)
日　(218.3)
中秋節(169.1)
着物　(144.6)
こいのぼり　(134.5)
お盆　(119.5)
旧暦　(113.1)
月餅　(111.4)
着る　(106.9)
お祭り(106.8)
15日　(98.8)
５月　(96.7)

の　(52.6)　※格助
ながら(21.7)
では　(20.3)
を　(19.6)
になると(18.4)
と　(18.2)　※格助
よう　(15.9)　※助動詞
ん　(14.0)　※んです
たり　(13.7)
みたい(13.0)

食

食べる(868.0)
料理　(525.7)
おいしい　(312.0)
味　　(261.9)
肉　　(230.6)
入れる(205.4)
焼く　(161.0)
作る　(159.6)
甘い　(157.6)
食べ物(64.8)
おすし(134.0)
卵　　(116.2)
からい(114.3)

られる　(30.6)

※食べられる

ん　　(12.9)　※んです

かな　(12.4)

そう　(11.5)　※様態

恋愛

結婚する　(123.2)
姉　　(60.7)
チョコレート　(50.2)
人情　(45.8)
恋人　(45.7)
早い　(44.7)
29才・友チョコ・本命　(36.7)
背　　(36.0)
こそこそ　(35.3)
彼　　(34.2)
25才　(31.7)
彼氏・恋・ボーイフレンド　(29.1)

といい　(32.1)
です　(16.2)

5．おわりに

- 真正性のある接触場面会話コーパスを話題ごとに分割した上で、話題別の特徴語を抽出した。
- 結果、予想通り多くの話題に特有の実質語を抽出することができたが、先行研究の見解に反して、複数の機能語、すなわち文法項目が特徴語として抽出された。
- 「の」「を」などは複数の話題で抽出されたが、「ている」「し」など単一の話題で抽出されたものも多い。

- 文法も話題に従属する。

今後の課題

- OPIのような会話テストや，
  方言調査などへの応用可能性。

- 『日中Skype会話コーパス』は20万語規模。

- より大きな話題別会話コーパスを作り、
  研究を進めて行きたい。