P9-4

本研究はJSPS科研費 JP20H01255の助成を受けたものです。

Pronoun substitute annotation in seven Asian languages

[アジア7言語における代名詞代用表現アノテーション]

野元 裕樹[1]　谷口 龍子[1]　中村 栞[1]　南 潤珍[1]　スリブディレスタリ[2]
スニサー ウィッタヤーパンヤーノン (齋藤)[1]　ウィラット ソンラートラムワニッチ[3]
春日 淳[4]　岡野 賢二[1]　トゥザライン[1]
[1]東京外国語大学　[2]立命館アジア太平洋大学　[3]武蔵野大学/タマサート大学　[4]神田外語大学

本研究で構築した言語資源

- アジア7言語のコーパスに対する話し手・聞き手・呼びかけ表現のアノテーション　☞ 計400万語以上！
- 東南アジア5言語の話し言葉コーパス
  ☞アノテーション付きのものは初の言語も！
- コーパス・アノテーションとアノテーション可視化のためのwebツール ETA: Easy Text Annotator
  ☞ふつうの文系研究者にも使いやすい！

本研究の重要性

- 日本語を含む対象言語では、話し手・聞き手を表す表現は人称代名詞に限らない。
  ☞文生成において、適切な表現を選ぶ必要

代名詞代用表現 (pronoun substitute)
話し手・聞き手を表す人称代名詞以外の表現

- 同じ表現が話し手（1人称）にも聞き手（2人称）にも他人（3人称）にもなる。
  ☞文理解において、人称についての曖昧性を解消する必要
- このような現象は西欧語や中国語では極めて限定的。

(1) a. 先生にちょうだい。　Give it to me.
　　b. 先生の部屋はどこですか？　Where is your room?

対象言語とコーパス

全言語共通：TUFSアジア言語パラレルコーパス（TALPCo; Nomoto et al. 2018）…1,372の日本語文とアジア言語への翻訳文

言語	コーパス	語数
日本語	日本語日常会話コーパス (Koiso et al. 2022)	2,421,162
朝鮮語	皆のコーパス（話し言葉コーパス、日常対話コーパス2020）	1,484,527
マレー語	新たに作成（会話、劇の台本）	39,265
インドネシア語	新たに作成（映画の字幕、会話）	32,870
タイ語	新たに作成（TVドラマの台本、小説）	272,342
ベトナム語	新たに作成（会話、映画の台本）	146,521
ビルマ語	新たに作成（劇の台本）	10,675

方法

- タグ：**1st**（1人称）、**2nd**（2人称）、**address**（呼びかけ）
- アノテーション対象：人称代名詞と代名詞代用表現

(1) a. 先生にちょうだい。　**1st**「私」で置換可能
　　b. 先生の部屋はどこですか？　**2nd**「あなた」で置換可能
(2) 先生、僕らの部屋はどこですか？　**address** 置換不可能

- 使用ツール：doccano (Nakayama et al. 2018; タイ語、ベトナム語), UAM CorpusTool（その他の言語）

結果と評価：『日本語日常会話コーパス』

- 2つのグループで独立してアノテーション付与作業を実施
- 全コーパスの85%が完了

アノテーションの分布とその例

グループ	1st	2nd	address	計
A	9,742	3,665	2,744	16,151
B	9,447	4,231	1,286	14,964

(3) a. 1st: 笹川, ばあば, 自分, こっち, みんな
　　b. 2nd: 野見山さん, ママ, パパたち, （ご）自分, 先生, みんな

- 会話参与者が3人以上の場合、判断が難しくなる。音声・動画ファイルが判断の助けになることも。

(4) かおちゃんこれ食べる？
　a. かおちゃんに対する発話、主語＝かおちゃん　→2nd
　b. かおちゃんに対する発話、主語＝ø　→address
　c. かおちゃん以外に対する発話　→アノテーションなし

アノテーター間一致率

1. 区間の一致（2人がテキストの同じ区間を選択）　$F = 84.4$
2. 値の一致（2人が同じタグを付与）　$F = 95.5$

混同行列

		グループB		
		1st	2nd	address
グループA	1st	8,460	25	8
	2nd	46	2,500	143
	address	12	886	1,040

不一致の原因

1. 区間
   「)」や「/」といった転記記号は、ガイドラインを作っても不一致を生みやすい。
   ☞転記とアノテーションの境界の問題
   - 転記を精密にするほど、アノテーションの精度が低下
   - 転記の一部をアノテーションとして切り離す可能性
2. 値
   格助詞が生起しないために、2ndとaddressの選択で多くの不一致が生じる。
   ☞話し言葉特有の問題。韻律情報のアノテーションも必要。

今後の展望

1. 1stと2ndについて、下位分類の情報を付与する。
   ① 人称代名詞
   ② 代名詞代用表現
   ③ 見かけ上の代名詞代用表現（なりすまし [imposter]）

(5) この問題について本学会の主催者は強く懸念しているのですが、ここにご出席の方々はどうお考えでしょうか？

2. Pro脱落（項の省略 cf. (4b)）による空項proを同定し、proに対して同様のアノテーションを付与する。
   ☞話し手・聞き手を表す表現のより包括的な理解へ

共同研究者募集！