Q1-16

言語処理学会第31回年次大会　2025年3月11日

Zero pronoun annotation in Malay and beyond

［マレー語におけるゼロ代名詞アノテーション、そしてその先へ］

野元 裕樹[1]　ファルハン アティラ ビンティ アブドゥル ラザク[1]　藤田 航平[2]
[1]東京外国語大学　[2]BFT

概要

- 既存の言語資源の有無、言語のタイプを問わず使える、簡単なゼロ代名詞のアノテーション枠組みを提案
- マレー語のコーパスにゼロ代名詞のアノテーションを付与
- 結果をNAISTテキストコーパスと比較

pro脱落

- 有形の代名詞や名詞句で生起可能な項が「省略」される現象
- Grambank (Skirgård et al. 2023) によれば、1,535言語中1,135言語 (73.9%) で主語のpro脱落が可能
- pro脱落言語の例：日本語、朝鮮語、中国語、マレー語、タイ語、アラビア語、スワヒリ語、トルコ語、スペイン語、フィンランド語
- 英語はpro脱落言語ではなく、少数派

マレー語　　ゼロ代名詞

(1) Bila e_1 jumpa e_2, dia cakap e_3 dah penat.
when see she say already tired
'When I_1 saw her_2, she said she_3 was already tired.'

ゼロ代名詞アノテーション付きコーパス

- 10程度の言語でしか公開されていない
- ほとんどが構成素構造のアノテーションを前提とする
  - 利点：ゼロ代名詞の言語学的研究では構成素構造上の位置が重要であることが分かっている
  - 欠点：ほとんどの言語では構成素構造のアノテーションが付いたコーパスは存在せず、新規に作成することも非現実的

通言語的汎用性を考えた場合、
平文コーパスにアノテーションを行うのが望ましい

ゼロ代名詞アノテーションの共通枠組み

タグセット

- 人称：1st, 2nd, 3rd
- 文法機能：S（主語）、DO（直接目的語）、IO（間接目的語）、P（所有者）
- ゼロ代名詞が連続する場合は、2番目以降の要素に生起順を示す番号を付す（例：1st_S**2**）　cf. (2), (3)

アノテーションを付す場所

ゼロ代名詞が想定される位置の**直後のトークン**（句読点を含む）

- 句読点のない言語（タイ語、ラオ語）：
  直前のトークンに付し、特殊記号「*」を付す　cf. (4)
- 語順が比較的自由な言語（日本語、ウクライナ語など）：
  ゼロ代名詞が想定される位置は基本語順に基づき決定　cf. (3)

マレー語
(1) Bila jumpa , dia cakap dah penat .
1st_S 3rd_DO 3rd_S
(2) harga boleh berunding .
3rd_P （それの値段は
1st_S**2** 私達は交渉できます。）
日本語
(3) 完食カードをお渡しいたしますので、
1st_S （私達は
2nd_IO**2** あなたに完食カードをお渡しいたしますので、
レジにてご提示ください。
2nd_S 3rd_DO あなたはレジにてそれを…）
タイ語
(4) ผมสอบผ่านเพราะอาจารย์อบรมสั่งสอน
1st_DO*
（先生が私を教えてくれたおかげで私は試験に合格しました）

マレー語コーパスへのアノテーション付与

方法

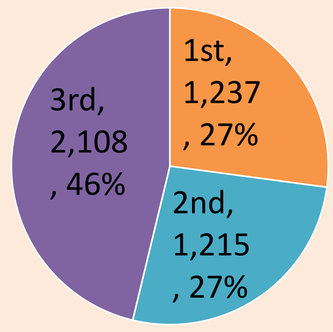
- Nor Hashimah (2003), Nor Hashimah et al. (2005)の付録
- 4,518文、34,724トークンから成る会話文
- 第二著者が付与した結果を第三著者が確認
- doccano (Nakayama et al. 2018) を使用

ゼロ代名詞の有無・位置の認定指針

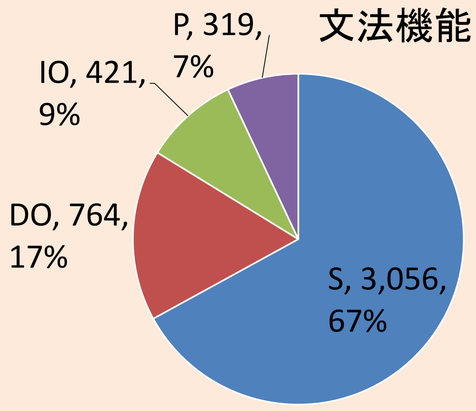
- 与格交替、定名詞、動詞＋小辞、動詞連続、目的語前置文、定型表現では複数の統語分析が可能
- 主にアノテーションの分かりやすさに従い1つの分析を選択

結果（予稿集提出後にデータの大幅修正あり）

人称



文法機能

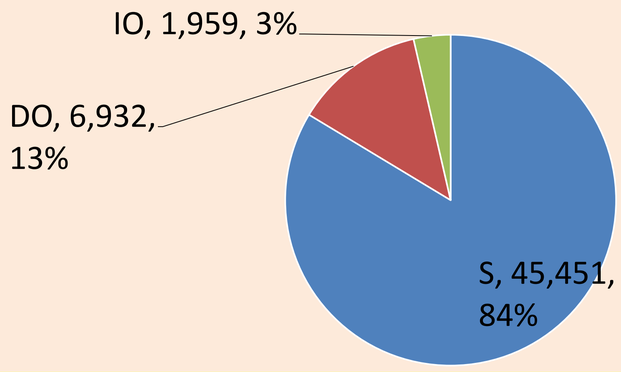


NAISTテキストコーパス（日本語）との比較

- NAISTテキストコーパスは
  - 会話文ではなく、新聞記事
  - 文法機能（統語情報）ではなく、格標示（形態情報）
    ガ格 ≈ S、ヲ格 ≈ DO、ニ格 ≈ IO
  - 人称とノ格（≈ P）はなし
- Iida et al. (2007) が報告する文内・文間ゼロ照応アノテーションのうち述語に関するものを比較の対象とする

比較から分かること

1. どちらのコーパスでも主語のpro脱落が圧倒的に多い
   ★通言語的に目的語より主語のpro脱落の方が一般的
2. NAISTテキストコーパスの方が主語の比率が高い
   ★マレー語は日本語よりも主語を言う



【謝辞】本研究はJSPS科研費JP23K25336の助成を受けたものです。タイ語の例文はスニサー・ウィッタヤーパンヤーノン氏にご教示いただきました。