Q1-6

言語処理学会第32回年次大会　2026年3月10日

Challenges in building a benchmark of linguistic minimal pairs for low resource languages: The case of Malay and Indonesian

野元 裕樹[1]　スリ ブディ レスタリ[2]　ダヴィド ムルヤディ[3]
ファルハン アティラ ビンティ アブドゥル ラザク[1]　稲垣 和也[4]　降幡 正志[1]
[1]東京外国語大学　[2]立命館アジア太平洋大学　[3]神田外語大学　[4]南山大学

【謝辞】本研究はJSPS科研費JP23K25336の助成を受けたものです。

概要

- 現在、マレー語・インドネシア語のBLiMP (Benchmark of Linguistic Minimal Pairs) MALINDO BLiMPを構築中
- https://github.com/matbahasa/MALINDO_BLiMP にて公開
- 可能な限り日本語のJBLiMP (Someya & Oseki 2023) に準拠
- 低資源言語ならではの困難と工夫を報告

【参考：BLiMPとは？】
- 言語モデルの文法知識を調べるためのベンチマークの1つ
- 対象文法現象においてのみ異なる正文・非文のペア（最小対）から成る

本研究のポイント

低資源言語でのBLiMP構築では
- 研究コミュニティが小さいため、データの収集・検証が大変
- 外国人のみの著者による研究が多く、データの信頼性が低い
- 関連言語間での翻訳によりデータ量不足を補うことが可能

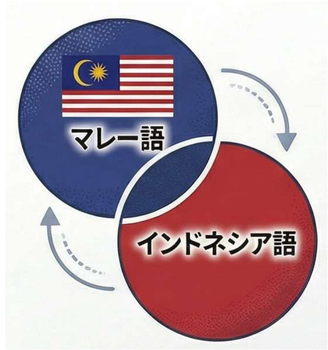
MALINDO BLiMPの構築手法

1. 非文の収集
研究コミュニティの小ささに起因する困難に直面
JBLiMP
1つの英文学術誌10年分
MALINDO BLiMP
70の文献（学術誌、書籍、博士論文）、4言語（日英馬尼）、1976–2023年

2. 最小対の作成
非文と対になる正文を特定または作成

○ 初めて会う人がけなすのは誰**を**ですか？
× 初めて会う人がけなすのは誰**に**ですか？
（JBLiMP、島の効果の例）

3. 翻訳によるデータ拡張
方言関係を活かし、翻訳によりデータ量を増加



4. 現象別の分類
45のパラダイムから成る12の現象に分類

母語話者に対する容認性判断実験によるコアデータの検証

JBLiMP
- クラウドソーシング「ランサーズ」を利用
- 240名の母語話者、367ペア（22~23ペア／人）
- 2択（どちらがより文法的？）
- 採択基準：
  原典と一致 ＞ 50%
- 採択率：90.9% (331/367)

MALINDO BLiMP
- 適切なクラウドソーシング提供会社なし→現地大学の協力＋Google Forms
- 308名の話者（母語話者280名）、300ペア（30ペア／人）
- 4択：Aのみ、Bのみ、both、neither
- 採択基準： both + neither (コントラストなし) ＜ 50%
  ∧ same (一致) + both, reverse (逆) + neither ≥ 55%
- 採択率：マレー語 58.0% (174/300)、インドネシア語 63.0% (189/300)

言語学での利用も視野に、高粒度の選択肢

著者の背景の違いが与える影響

JBLiMP
- 大半の原典が現地人（＋外国人）による執筆　23/28

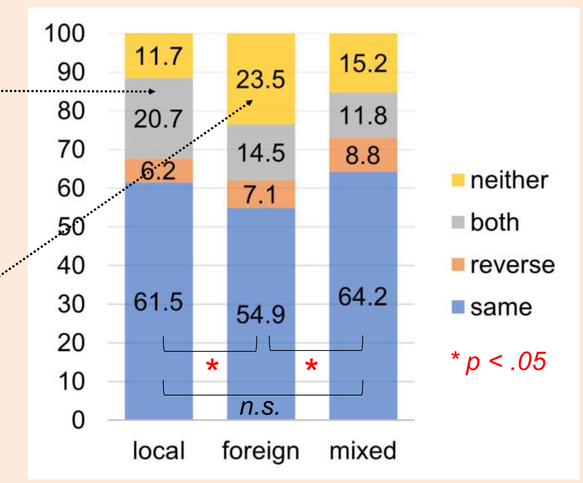
MALINDO BLiMP
- 半数以上の原典が外国人のみによる執筆　39/70
- 現地人のみ　21/70、現地人＋外国人　10/70

Q: 外国人が著者の場合、データの信頼性は下がらない？
A: 「現地人のみ」および「現地人＋外国人」に比べ、「外国人のみ」は一致率が有意に低くなる

現地研究者との協働は大切！

現地人のみ：
正文の間の微細な容認性の違いに着目する傾向

外国人のみ：
非文同士を比較する傾向



現象	マレー		インドネシア		現象	マレー		インドネシア	
項構造	33	(58.9%)	35	(62.5%)	島の効果	5	(18.5%)	10	(37.0%)
動詞の一致	3	(33.3%)	3	(33.3%)	埋語・空所	20	(48.8%)	24	(58.5%)
束縛	11	(100.0%)	9	(81.8%)	NPI認可	8	(88.9%)	8	(88.9%)
省略	6	(50.0%)	7	(58.3%)	名詞構造	7	(63.6%)	9	(81.8%)
形態論	31	(62.0%)	36	(72.0%)	制御・繰上げ	13	(54.2%)	15	(58.3%)
量化子	10	(100.0%)	8	(80.0%)	その他／不明	27	(67.5%)	26	(65.0%)

翻訳が与える影響

Q: 翻訳により拡張したデータの信頼性は下がらない？
A: 信頼性は下がらない

関連言語間での翻訳は有用なデータ拡張手法