A corpus-based study of *pro* drop in Standard Malay

Hiroki Nomoto Farhan Athirah binti Abdul Razak
Tokyo University of Foreign Studies
ISMIL 28, 9-10 June 2025

東京外国語大学
Tokyo University of Foreign Studies

Pro drop in Standard Malay

Standard Malay (henceforth “Malay”) allows the so-called *pro* drop, especially in its colloquial variety (Koh 1990: 141–146).

(1) Jangan bimbang pasal harga boleh berunding.
‘(You) Don’t worry because (regarding its) price (we) can negotiate.’

(2) Awak jangan bimbang pasal harga dia kita boleh berunding.

Previous studies on Malay *pro* drop

- All most all are qualitative.
  e.g. Koh (1990), Mashudi (2003), Ahmad Syafiq Amir et al. (2022), Nomoto (2022), Nomoto and Matsuura (2023)
- Nomoto and Kartini (2014) examine the frequency of *pro* occurring as *di-* passive agents.
- However, *pro* drop involving subjects and objects has not been investigated quantitatively.

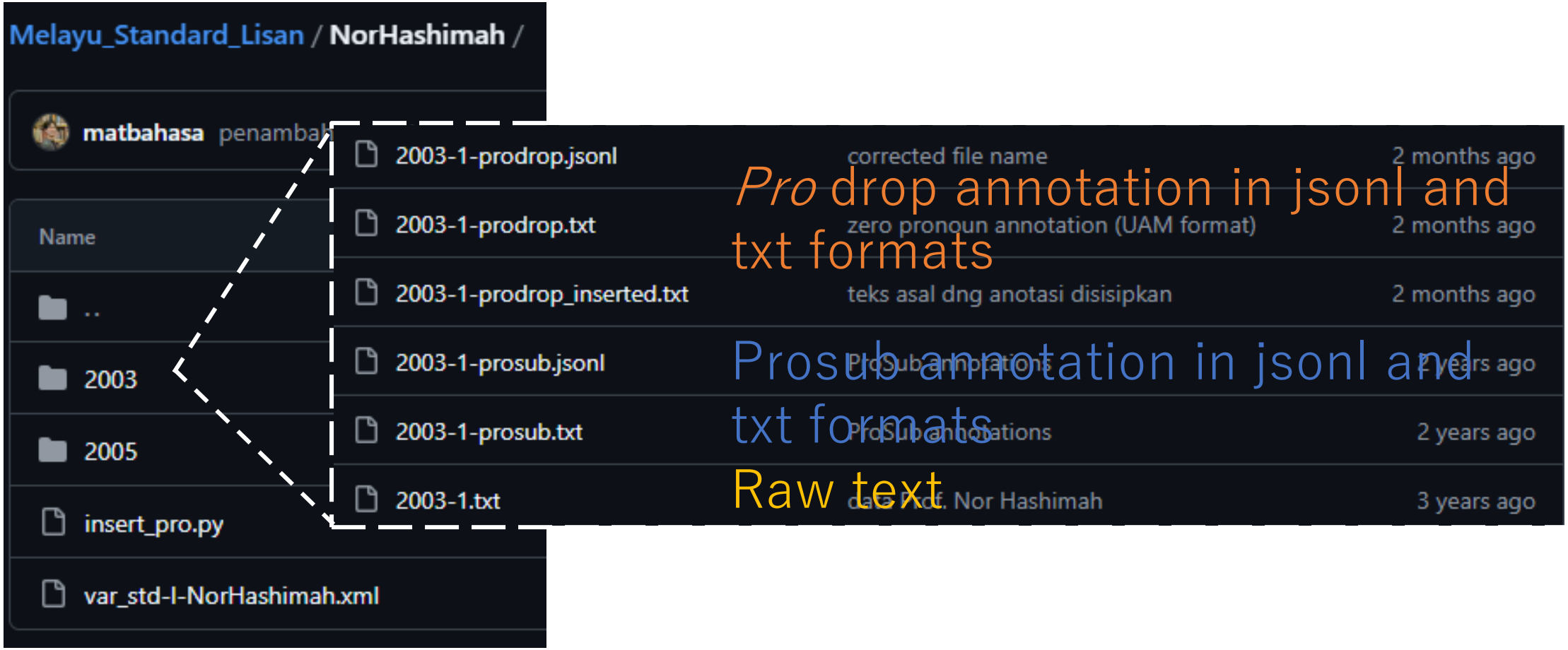
This study

- Fills this gap in research.
- Annotated a corpus for *pro* drop.
- Examined the frequency of *pro* dropped elements according to person and grammatical function.

Corpus

- The conversation data provided as appendices by Nor Hashimah (2003) and Nor Hashimah et al. (2005).
  - (2003) *Bahasa dalam Perniagaan: Satu Analisis Semantik dan Pragmatik*
  - (2005) *Sistem Panggilan dalam Keluarga Melayu: Satu Dokumentasi*
- Digitalized and published as part of *Korpus Variasi Bahasa Melayu: Standard Lisan*.
  https://github.com/matbahasa/Melayu_Standard_Lisan/tree/master/NorHashimah

Files


Melayu_Standard_Lisan / NorHashimah /
matbahasa penambah
Name
..
2003
2005
insert_pro.py
var_std-I-NorHashimah.xml
2003-1-prodrop.jsonl
corrected file name
2 months ago
2003-1-prodrop.txt
zero pronoun annotation (UAM format)
2 months ago
2003-1-prodrop_inserted.txt
teks asal dng anotasi disisipkan
2 months ago
2003-1-prosub.jsonl
ProSub annotations
2 years ago
2003-1-prosub.txt
ProSub annotations
2 years ago
2003-1.txt
data Prof. Nor Hashimah
3 years ago
Pro drop annotation in jsonl and txt formats
Prosub annotation in jsonl and txt formats
Raw text

Corpus details

- **Size**: 4,518 sentences, 34,724 tokens
- **Language**: Standard Malay (zsm) with dialectal mixing. [The data was been collected in various places, but the original authors normalized it (i.e. applied the standard orthography, e.g. *manɔ* → *mana*, not *mano* or *manor*).]
- **Content**:
  - 2003: Conversations between sellers and buyers at markets
  - 2005: (i) Conversations during cooking gatherings
    (ii) Interviews about the use of referring expressions

Annotation scheme

- No specific analysis of *pro* drop (e.g. Huang 1984; Barbosa 2019) was assumed.
- All null arguments in positions where an overt argument can occur were treated as *pro*.
- Exception: passive agents
- **Rule of thumb:** “Posit *pro* if you can insert a personal pronoun (e.g. *saya*, *awak*, *dia*).”
- The token immediately after *pro* was assigned an annotation tag (or tags).

Annotation tags

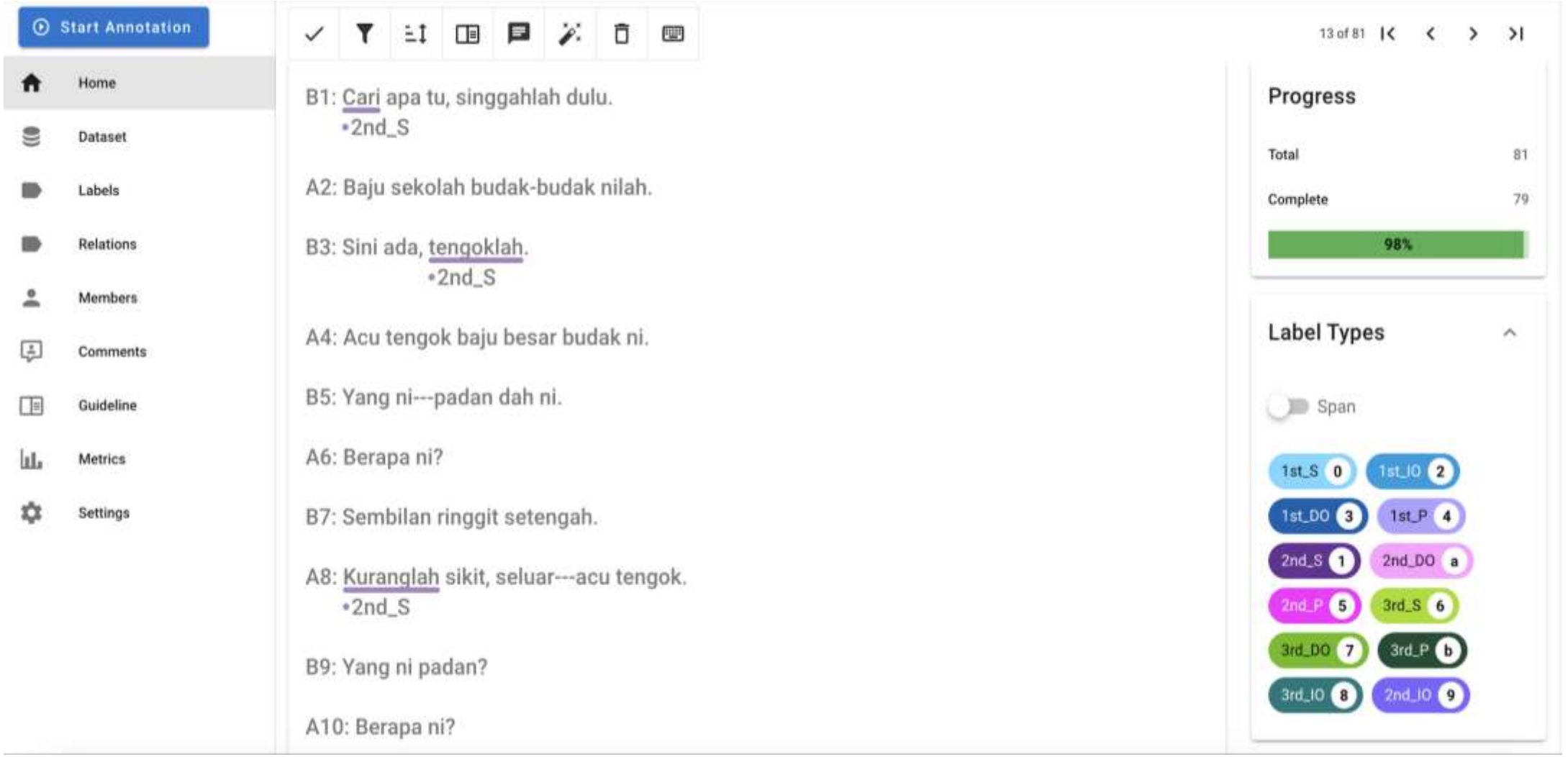
Person

- 1st
- 2nd
- 3rd

Grammatical function

- S
- DO
- IO
- P (possessor)

Doccano (Nakayama et al. 2018)



Examples

Rule: Annotate the token immediately after *pro*.

(3) a. Kenapa tak hidupkan?
2nd_S 3rd_DO

b. Kenapa awak tak hidupkan dia?

(4) a. Jangan bimbang pasal harga boleh
2nd_S 3rd_P
berunding. 1st_S

b. Awak jangan bimbang pasal harga dia kita boleh berunding.

Construction-specific rules: Dative alternation

<u>Adopt the double object analysis whenever possible.</u>

V DO (V = ditransitive/monotransitive)

☞ V ~~IO~~ DO (*pro* drop of IO)

V DO ~~*kepada/untuk IO*~~ (no *pro* drop)

(5) Kak **bagi harga niaga** dah ni.

a. ☞ Kak **bagi *pro*$_{awak}$ harga niaga** dah ni.

b. Kak **bagi harga niaga** dah ni.

Construction-specific rules: Bare definites vs. possessive definites

<u>Adopt the possessive definite analysis whenever possible.</u>

NP $Ø_D$ (no *pro* drop)

☞ NP Poss (*pro* drop of P)

(6) Jangan bimbang pasal **harga** boleh berunding.

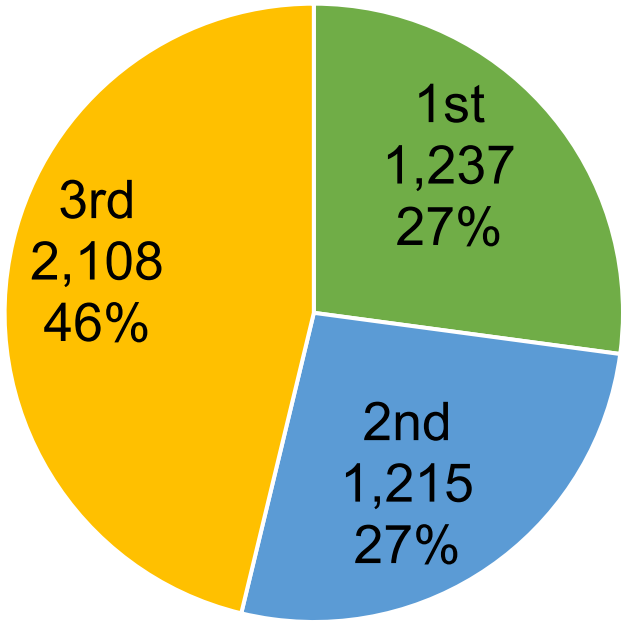
a. … **harga** boleh berunding.

b. ☞ … **harga** ***pro*** $_{dia}$ boleh berunding.

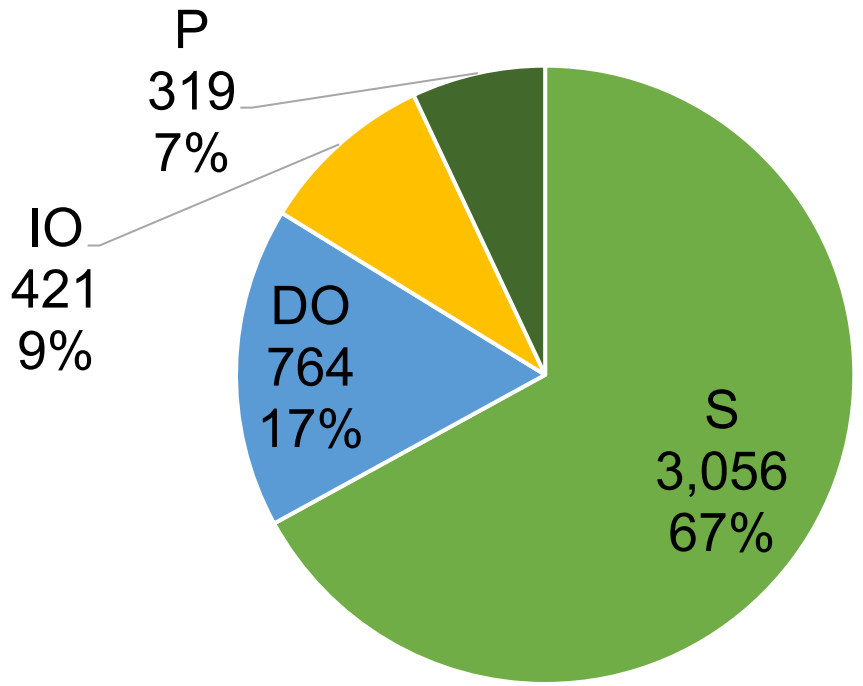
cf. Appendix in Nomoto et al. (2025) for other construction-specific rules.

Results (updated from Nomoto et al. 2025)

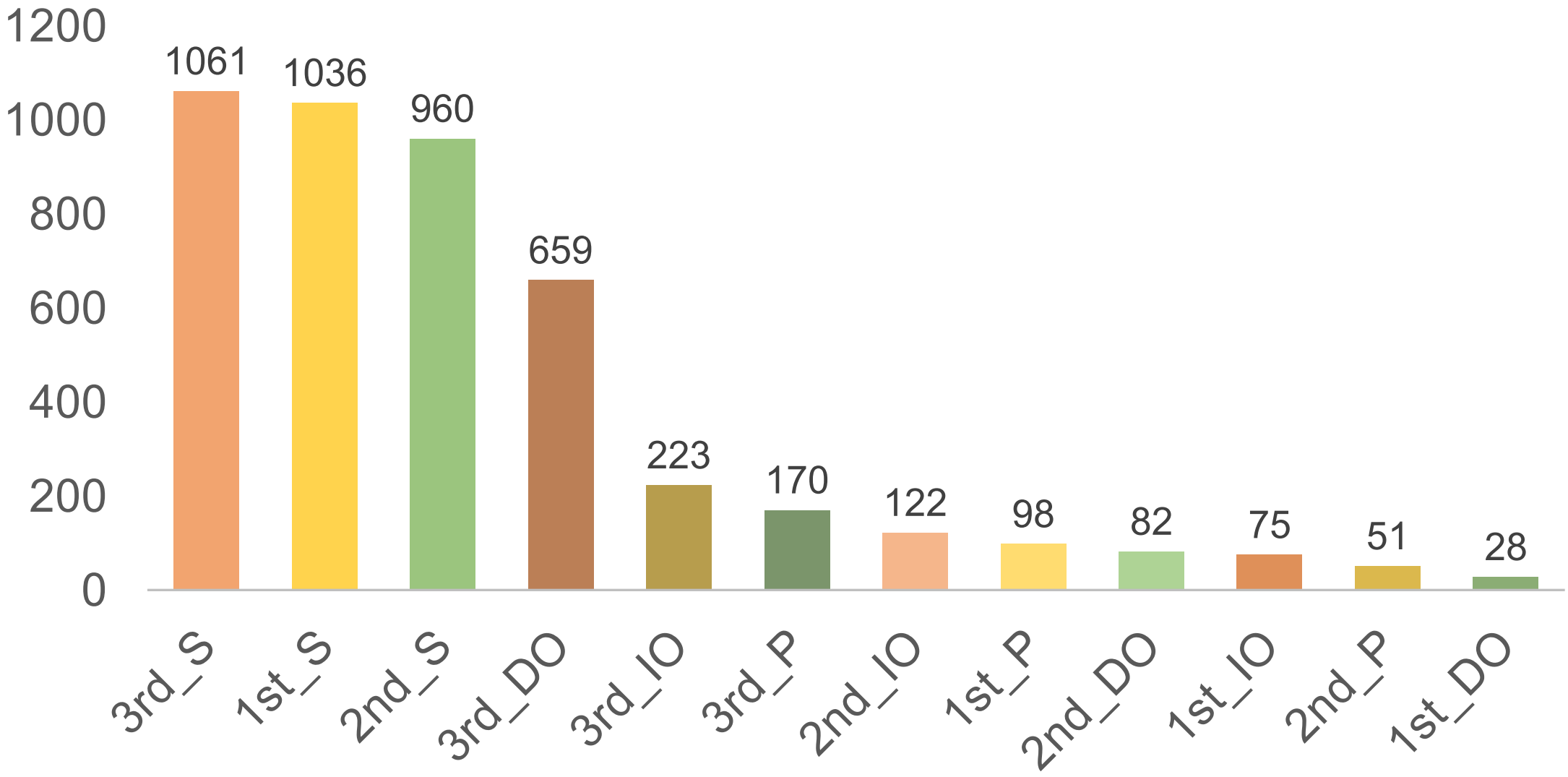
Person



Grammatical function



Person + function combined



On object *pro* drop

Nomoto and Matsuura (2023)

- Empty objects are not pro but a variable bound by a null topic (= Huang's (1984) analysis).
- **Prediction:** *MeN-* cannot occur before empty objects, given its movement blocking effect (Saddy 1991; Soh 1998; Cole and Hermon 1998)

(7) Siti membeli atau meminjam buku itu?
—Dia {pinjam/*meminjam} *e*. (N & M 2023)

(8) Nadiah menulis surat kepada Siti
dan Siti sudah {terima/*menerima} *e*. (N & M 2023)

Counterexamples in the corpus

(9) Abah, Mak **menanak *pro*$_{dia}$** kat rumah tu.

(10) Tak bubuh lagi, ni baru nak **menghidang *pro*$_{dia}$**.

(11) Kan Pak Long pergi **menyiasat *pro*$_{dia}$**, orang kata dia minta sedekah kat jejentas Johor?

(12) Sebab tu aku siap **merakam *pro*$_{dia}$** ni.

(13) Ingatkan nak pergi masa **menghantar *pro*$_{dia}$** nanti.

(14) Cangkerang menggunakan perkataan untuk **menerangkan *pro*$_{dia}$**… tak lepas dah tu. (unfinished sentence, so may not be problematic)

Possible reasons for *meN-*

1. The *meN-* verbs in (10)–(14) occur in embedded contexts.
   (10) … [$_{VP}$ nak [menghidang
   (11) … [$_{VP}$ pergi [menyiasat
   (12) … [$_{VP}$ siap [merakam
   (13) … [$_{CP}$ masa menghantar
   (14) … [$_{CP}$ untuk menerangkan
2. Some of these *meN-* verbs may have an intransitive use.

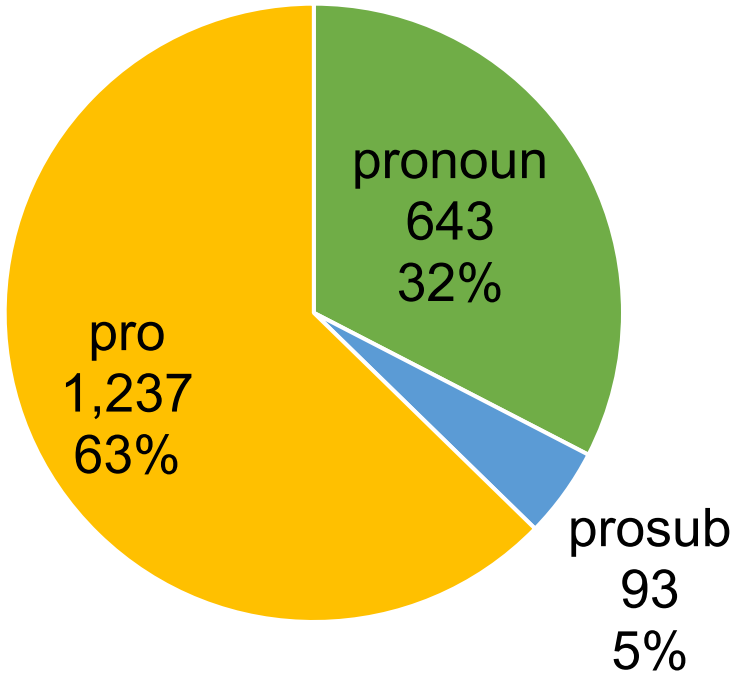
Overt vs. covert expressions

- The same corpus has been also annotated for overt first and second person expressions (Nomoto et al. 2023).
- So, we can now see the distribution of overt and covert expressions.
- Two types of overt fist and second person expressions:
  - Personal pronouns
    e.g. *saya*, *aku*, *I*, *awak*, *kamu*, *(eng)kau*, *you*
  - Pronoun substitutes (prosubs)
    e.g. *encik*, *puan*, *cik*, *pak cik*, *mak cik*, *mak*, *ayah*, *abang*, *kakak*, *adik*, *cikgu*, *Dato'*, *prof.*, *Dr.*, NAME, TITLE + NAME

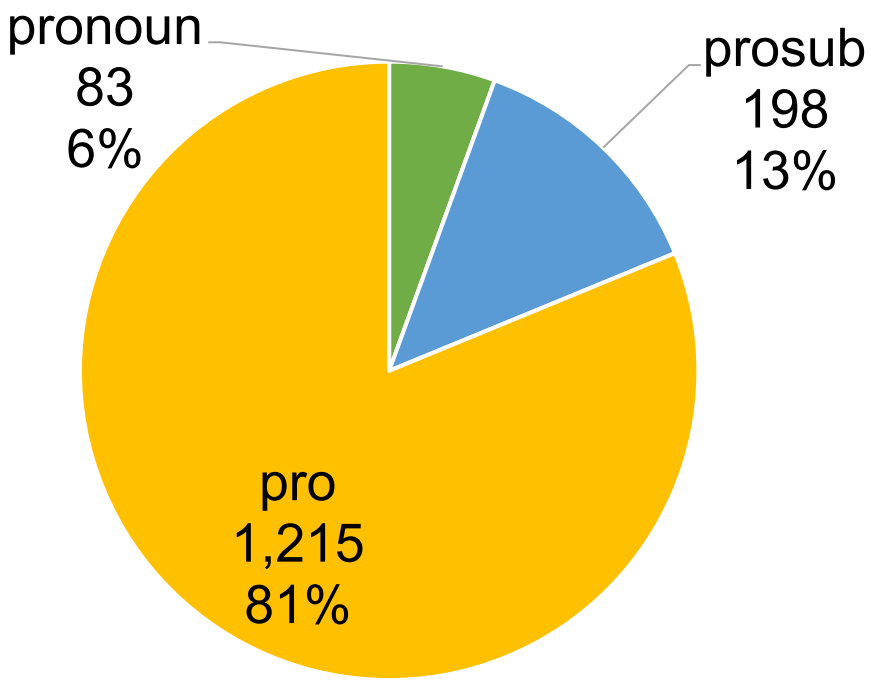
Distribution of 1st and 2nd person expressions in the corpus

1st

Personal pronouns are also frequent for 1st person.



2nd



Pro is prevalent across persons.

Conclusion and future direction

- Annotated a corpus of Standard Malay for empty arguments.
- Examined their distributions.
  - Person: 3rd >> 1st, 2nd
  - Grammatical function: S (approx. 2/3) >> DO > IO > P
- Utilizing the existing other annotation, showed how prevalent *pro* drop is.
- **Future:** Annotate overt expressions for grammatical functions (almost done) to enable a comparison/integration with the *pro* drop annotation.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP23K25336.

References (1)

Ahmad Syafiq Amir Abdullah Zawawi, Fazal Mohamed Mohamed Sultan, and Azhar Jaludin. 2022. Subjek nul *pro* bahasa Melayu: Suatu penilaian semula berpandukan kerangka minimalis. *Issues in Language Studies* 11:109–128.

Barbosa, Pilar P. 2019. *pro* as a minimal nP: Toward a unified approach to pro-drop. *Linguistic Inquiry* 50:487–526.

Cole, Peter and Gabriella Hermon. 1998. The typology of wh-movement: Wh-questions in Malay. *Syntax* 1:221–258.

Huang, C.-T. 1984. On the distribution and reference of empty pronouns. *Linguistic Inquiry* 15:531–574.

Koh, Ann Sweesun. 1990. *Topics in Colloquial Malay*. Doctoral Dissertation, University of Melbourne.

Mashudi Kader. 2003. Kategori kosong pro yang berfungsi frasa nama. *Jurnal Bahasa* 3:391–416.

Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. https://github.com/doccano/doccano.

References (2)

Nomoto, Hiroki. 2022. Aspek nahu dalam penterjemahan bahasa Jepun-bahasa Melayu: Ayat kewujudan dan pengguguran *pro*. In *Penterjemahan Struktur Bahasa Asing dalam Bahasa Melayu*, ed. Sang Seong Goh, 200–221. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Nomoto, Hiroki, Farhan Athirah binti Abdul Razak and Kohei Fujita. 2025. Zero pronoun annotation in Malay and beyond. *Proceedings of the Thirty-First Annual Meeting of the Association for Natural Language Processing*, 391-396.

Nomoto, Hiroki, and Kartini Abd. Wahab. 2014. Person restriction on passive agents in Malay: Information structure and syntax. In Current Trends in Malay Linguistics, ed. Siaw-Fong Chung and Hiroki Nomoto, volume 57 of *NUSA*, 31–50.

Nomoto, Hiroki, and Ai Matsuura. 2023. Pro drop in Standard Malay. Paper presented at the 26th International Symposium on Malay/Indonesian Linguistics (ISMIL).

Nomoto, Hiroki, Ryuko Taniguchi, Shiori Nakamura, Yunjin Nam, Sri Budi Lestari, Sunisa Wittayapanyanon (Saito), Virach Sornlertlamvanich, Atsushi Kasuga, Kenji Okano, and Thuzar Hlaing. 2023. Pronoun substitute annotation in seven Asian languages. In *Proceedings of the Twenty-Ninth Annual Meeting of the Association for Natural Language Processing*, 2242–2247.

References (3)

Nor Hashimah Jalaluddin. 2003. *Bahasa dalam Perniagaan: Satu Analisis Semantik dan Pragmatik*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Nor Hashimah Jalaluddin, Harishon Radzi, Maslida Yusof, Raja Masittah Raja Ariffin, and Sa'adiah Ma'alip. 2005. *Sistem Panggilan dalam Keluarga Melayu: Satu Dokumentasi*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Saddy, Douglas. 1991. WH scope mechanism in Bahasa Indonesia. In *MIT Working Papers in Linguistics 15: More Papers on Wh-Movement*, ed. Lisa L. S. Cheng and Hamida Demirdash, 183–218.

Soh, Hooi Ling. 1998. Certain restrictions on A-bar movement in Malay. In *Proceedings of the Third and Fourth Meetings of Austronesian Formal Linguistics Association 1996–1999*, ed. Matthew Pearson, 295–308. Department of Linguistics, University of California, Los Angeles.