Zero pronoun annotation in Malay and beyond

Hiroki Nomoto[1] Farhan Athirah binti Abdul Razak[1] Kohei Fujita[2]
[1]Tokyo University of Foreign Studies [2]BFT Corporation
{nomoto, farhan.athirah.binti.abdul.razak.x0}@tufs.ac.jp k-fujita@bfts.co.jp

Abstract

This study proposes a zero pronoun annotation scheme that is easy to adopt cross-linguistically, regardless of language types, due to its reliance only on raw corpus data and the absence of prerequisites such as constituency trees or predicate-argument structures. A spoken language Malay corpus has been annotated using it. The results are compared to the distribution of zero pronouns in a Japanese corpus, namely the NAIST Text Corpus.

1 Introduction[1)]

Many languages in the world allow arguments that can otherwise be expressed overtly with pronouns or full noun phrases to be “omitted.” This phenomenon is known as “*pro* drop.”[2)] For example, the Malay example in (1) contains three instances of *pro* drop indicated by *e* (for “empty”), which we refer to as “zero pronoun.”[3)]

(1) Bila e_1 jumpa e_2, dia cakap e_3 dah penat.[4)]
when see she say already tired
‘When I$_1$ saw her$_2$, she said she$_3$ was already tired.’

According to Grambank (Feature GB522) [4, 5], 1,135 out of the 1,535 languages with the relevant data point (73.9%) allow subject *pro* drop. English is not a *pro* drop language and belongs to the minority. At least four types of *pro* drop languages have been identified: (i) languages with rich agreement (consistent/agreement-based *pro* drop; e.g. Italian), (ii) languages without agreement (radical/discourse-based *pro* drop; e.g. Japanese), (iii) languages with agreement and referential null subjects whose distribution is restricted (partial *pro* drop; e.g. Finnish) and (iv) languages with only impersonal and quasi-argumental null subjects (semi *pro* drop; e.g. Icelandic) [6]. *Pro* drop languages differ with regard to where *e* occurs and how *e* is interpreted. They also differ in the conditions under which *e* is chosen over its overt alternative.

Corpora annotated with zero pronouns are essential for investigating the linguistic properties of *pro* drop and solving NLP tasks involving *pro* drop languages such as zero anaphora resolution, machine translation and information extraction. However, as discussed in §2, such resources are available in only a handful of languages, despite the large number of *pro* drop languages. Moreover, there does not seem to exist a common scheme for zero pronoun annotation that can be utilized cross-linguistically. Therefore, this study proposes one such scheme (§3) and annotates a Malay corpus using it (§4). The annotation files are openly available at https://github.com/matbahasa/Melayu_Standard_Lisan/tree/master/NorHashimah/.

2 Existing methods of zero pronoun annotation

According to our survey, at least the following ten languages have publicly available corpora annotated with zero pronouns: Arabic, Catalan, Chinese, German, Indonesian, Japanese, Korean, Malay, Portuguese and Spanish. Table 1 summarizes the corpora we could find. Most of them depend on constituency trees. We regard it as a good feature because linguistic studies have shown that the position of *e* in the constituent tree and the grammatical function it determines are important. However, building a constituency treebank requires considerable effort, and hence presupposing it for the purpose of zero pronoun annotation is practically unrealistic for most languages. The ZAC corpus in Portuguese alters the corpus by inserting a tag in

[1)] A considerable part of this study is based on the second author’s Master’s thesis [1].

[2)] NLP practitioners should be more familiar with the related term “zero anaphora (resolution),” which is a kind of anaphora (resolution) that involves zero pronouns resulted from *pro* drop.

[3)] Other terms for *e* include “*pro*,” “null argument/pronoun” and “empty category/pronoun.”

[4)] http://aciklananovel.blogspot.com/2011/04/bab-22-kalau-memang-harus-begitu.html. This sentence was taken from the ZSM MXD2012 subcorpus of the Leipzig Corpora Collection [2] using MALINDO Conc [3].

Table 1 Existing corpora with zero pronoun annotation

Corpus	Language	Dependency	Alter	Position	Function	Reference
OntoNotes [8]	Arabic, Chinese	constituency tree	no	yes	yes	no
Chinese Treebank [9]	Chinese	constituency tree	no	yes	yes	no
AnCora [10]	Catalan, Spanish	constituency tree	no	yes	yes	no
Tschick, AdT [11]	German	none	no	no	yes	yes
Penn Korean Treebank [12, 13]	Korean	constituency tree	no	yes	yes	no
NAIST Text Corpus [14]	Japanese	predicate-argument structure	no	no	(yes)	yes
Kainoki Treebank [15]	Japanese	constituency tree	no	yes	yes	no
TALPCo Treebank [16]	Indonesian, Malay	constituency tree	no	yes*	yes	no
ZAC [17]	Portuguese	none	yes	yes	no	yes

*The grammatical function is not explicitly annotated, but can be identified from the syntactic position.

Table 2 Tagset

Tag	Explanation	Example
PERSON		
1st	first person	***I** asked **my** mum to help **me**.*
2nd	second person	***You** asked **your** mum to help **you**.*
3rd	third person	***He** asked **his** mum to help **her**.*
GRAMMATICAL FUNCTION		
S	subject	***You** do it by yourself!*
DO	direct object	*Ken gave **it** to his friend.*
IO	indirect object	*Ken gave **her** a present.*
P	possessor	*I missed **my** train.*

the position of e. This method may be the easiest for ordinary linguists, for whom installing annotation tools such as doccano [7] is almost impossible. However, it is generally good to keep the raw corpus data separate from its annotations. The grammatical function value for the NAIST Text Corpus in Japanese is in parentheses because it employs morphological cases rather than grammatical functions. Morphological cases are a good indicator of grammatical functions, but the mapping is not perfect. Thus, although nominative case-marked noun phrases are usually subjects, they can also be objects. Moreover, many languages simply do not have morphological case.

3 Common scheme for zero pronoun annotation

We propose a common scheme for zero pronoun annotation that (i) does not presuppose another annotation, (ii) does not alter the corpus itself and (iii) can be used in any language.

```
Bila jumpa    ,    dia cakap  dah  penat.
     1st_S  3rd_DO            3rd_S
harga boleh berunding.
      3rd_P
      1st_S2
```

Figure 1 Sample sentences with zero pronoun annotations

3.1 Tagset

The proposed tagset consists of two categories: person and grammatical function. These two categories are frequently referred to in linguistic studies on *pro* drop. Table 2 summarizes the tags belonging to each category with examples in English, where the relevant items are indicated by **boldface**. These tags have 12 (= 3 × 4) possible combinations, which we will represent by joining the two categories with an underscore as in 1st_S.

3.2 What to annotate

Since zero pronouns, by definition, do not appear in any form in the text, **the token immediately after the position where *e* occurs** is the target of the annotation.[5] The next token includes punctuation marks. Figure 1 shows how sentences (1) and (3) are annotated.

One may wonder if it would be more intuitive to annotate white spaces. However, such a method is invalid for languages that do not use white spaces such as Chinese and Japanese. A special treatment is required in languages that lack punctuation marks to indicate a sentence boundary. Lao and Thai are the only such languages that we know,

[5] This makes it possible to formulate zero pronoun detection as a binary tagging problem (token preceded by e vs. token not preceded by e), as suggested by [18].

but there is a recent trend of not ending a sentence with a full stop or its equivalent in casual writing even in other languages such as English and Japanese. In such cases, when e occurs sentence-finally, the token *before* it can be annotated using a special symbol, such as 3rd_DO*, where * indicates that e occurs after the annotated token.

The position of e is determined based on the canonical word order. This rule is most relevant in languages whose word order is flexible such as Japanese and Ukrainian. For example, the canonical word order in Japanese is "S IO DO V" although other orders are also possible. The Japanese sentence in (2) illustrates this point.

(2) e_1 e_2 Kanshokukaado o o-watasi-itasimasu
(S) (IO) completion.card ACC POL-give-POL
node, e_3 rezi nite e_4
because (S) register at (DO)
go-teizi-kudasai.[6)]
POL-present-request
'We$_1$ will give you$_2$ a completion card, and you$_3$ are kindly requested to present it$_4$ at the register.'

4 Annotation of a Malay corpus

4.1 Methodology

Corpus We use the conversation data provided as appendices by [20, 21]. It consists of 4,518 sentences comprising 34,724 tokens. This data has been digitalized and made openly available as a part of Korpus Variasi Bahasa Melayu (Corpus of Malay Varieties).[7)] The conversations in [20] take place at markets and involve sellers and shoppers whilst those in [21] consist of two kinds, one being conversations during cooking events and the other being interviews about the use of person referring expressions. Although both contain elements presented as regional dialects, the entire data has been normalized, that is, converted to word forms of the standard variety. Hence, the corpus can be considered one of Standard Malay with occasional mixing of dialectal words.

Annotators and annotation tool The annotation was done by the second author and checked by the third author using doccano [7]. Since it does not support layers, we cannot separate the two annotation categories. We thus decided to create 12 tags by combining a person tag and a grammatical function tag (1st_S, 1st_DO, 1st_IO, . . .).

4.2 Language specific considerations

The actual annotation task requires various language specific considerations. Here we only note three of them that we think can affect the annotation results. Others are presented in Appendix A.

Secondary annotation tags The possessor follows the possessed noun it modifies in Malay. This word order and other Malay-specific phenomena can bring about a situation in which two zero pronouns occur in a row. An example is given in (3). e_1 is the possessor and part of the topic noun phrase whilst e_2 is the subject.

(3) harga e_1 e_2 boleh berunding.
price can negotiate
'(regarding) its$_1$ price, we$_2$ can negotiate.'

According to the common scheme proposed in §3 above, *boleh* will receive two tags, namely 3rd_P and 1st_S. The problem of simply assigning two tags is that the relative order information between the two is lost. To circumvent this problem, we introduced secondary tags with the suffix "2" to indicate a given tag follows the other unmarked tag. In this case, e_2 is represented as 1st_S2 (see Figure 1).

Dative alternation Malay has dative alternation between "S V DO *kepada* 'to'/*untuk* 'for' IO" (optional preposition phrase) and "S V IO DO" (double object construction). Hence, when IO is not overtly expressed, the sentence can be parsed as either pattern in principle. In such sentences, we chose the latter double object construction analysis. Thus, (4a) is parsed as (4b), which contains a zero indirect object.

(4) a. Kak bagi harga niaga dah ni.
elder.sister give price trade already this
'I've already given (you) the trade price.'
b. Kak bagi e harga niaga dah ni.

Bare definites vs. possessive definites Malay does not have a definite article like English *the*. Definite noun phrases can either be bare (bare definites) or involve a determiner such as a possessor (possessive definites) or a demonstrative. Consequently, some bare noun phrases can be parsed as either a bare definite or a possessive

[6)] This sentence is due to [19], who found it in an announcement about a food waste prevention campaign at an Otoya restaurant in March 2021. [19] posits e_2 in a different position, which we revise in this paper.

[7)] https://github.com/matbahasa/Melayu_Standard_Lisan

Table 3 Breakdown according to person

Person	1st	2nd	3rd	Total
Frequency	1,233	1,151	2,084	4,469
(%)	(27.6)	(25.8)	(46.6)	(100.0)

Table 4 Breakdown according to grammatical function

Function	S	DO	IO	P	Total
Frequency	3,044	480	628	317	4,469
(%)	(68.1)	(10.7)	(14.1)	(7.1)	(100.0)

definite with a zero possessive pronoun. In such cases, we chose the latter possessive definite analysis because possessive definites are not uncommon in Malay.[8] This is why we analysed (3) as involving a zero pronoun denoting the possessor (e_1) rather than analyze *ayah* 'father' as bare without any sort of omission.

4.3 Results

The total number of annotations made to the corpus is 4,469. Tables 3 and 4 show their breakdowns according to person and grammatical function, respectively. In terms of person, while third person is more frequent than first and second person, no striking difference exists between the latter two. As for grammatical function, about two thirds of zero pronouns in the corpus are subjects. Indirect objects turned out to be more prone to be zero than direct objects. It must be noted, however, that this result is partly due to our decision concerning dative alternation to choose the double object construction analysis (cf. §4.2), which will naturally increase the number of IOs.

Figure 2 shows the distribution of the combinations of person and grammatical function. The following two observations can be made. First, subject is most often realized as zero, regardless of person. Second, the other grammatical functions are realized as zero more often in third person than first and second person.

4.4 Comparison with Japanese

In this section, we compare the results above with Japanese, specifically the NAIST Text Corpus. Although it consists of formal writings unlike our corpus, [14] provide detailed statistics that enable an easy comparison. As noted in §2, its annotation scheme makes use of morphological case instead of grammatical function. A comparison of Tables 4 and 5 reveals two points. First, subject is far more often realized as zero than the other functions in both languages. Second, a substantial difference exists concerning the proportion of indirect object: rather big in Malay and very small in Japanese.

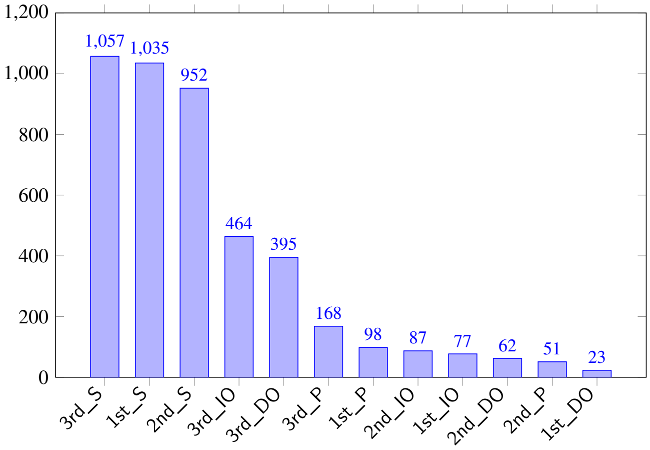

Figure 2 Combinations of person and grammatical function

Table 5 Distribution of cases in NAIST Text Corpus

Case	NOM (≈S)	ACC (≈DO)	DAT (≈IO)	Total
Frequency	45,451	6,932	1,959	54,342
(%)	(83.6)	(12.8)	(3.6)	(100.0)

5 Conclusion

This study has proposed a common scheme for zero pronoun annotation designed to be used cross-linguistically with no prerequisite annotations. We hope that it will contribute to increasing the number of corpora with zero pronoun annotation, which will enrich our understanding of the linguistic properties of *pro* drop and help improve the quality and quantity of related NLP research and development.

Regarding the Malay corpus, the same corpus has also been given other kinds of annotations, namely morphology, first and second person expressions, and address terms [22]. The zero pronoun annotation created by this study can be combined with these other annotations to gain new insights about the language. An obvious limitation of this study is the small size of the corpus, which is actually a problem of Malay linguistics in general. Since *pro* drop is a phenomena characteristic of spoken language in Malay, larger open spoken corpora are urgently needed.

[8] A similar zero pronoun annotation task should choose the bare definite analysis in languages in which possessive definites are not so common such as Japanese.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP23K25336.

References

[1] Farhan Athirah binti Abdul Razak. Mareego niokeru zero daimeishi no anoteeshon [Annotation of zero pronouns in Malay]. Master's thesis, Tokyo University of Foreign Studies, 2025.

[2] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In **Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)**, pp. 759–765, Istanbul, 2012. European Language Resources Association.

[3] Hiroki Nomoto, Shiro Akasegawa, and Asako Shiohara. Building an open online concordancer for Malay/Indonesian. Paper presented at the 22nd International Symposium on Malay/Indonesian Linguistics (ISMIL), 2018.

[4] Hedvig Skirgård et al. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. **Science Advances**, Vol. 9, 2023.

[5] Hedvig Skirgård et al. Grambank v1.0, mar 2023. Dataset.

[6] Pilar P. Barbosa. *pro* as a minimal nP: Toward a unified approach to pro-drop. **Linguistic Inquiry**, Vol. 50, No. 3, pp. 487–526, 2019.

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

[8] Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. OntoNotes release 5.0, 2013. LDC2013T19.

[9] Nianwen Xue, Xiuhong Zhang, Zixin Jiang, Martha Palmer, Fei Xia, Fu-Dong Chiou, and Meiyu Chang. Chinese Treebank 9.0, 2016. LDC2016T13.

[10] Mariona Taulé, M. Antònia Martí, and Marta Recasens. AnCora: Multilevel annotated corpora for Catalan and Spanish. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, **Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)**, pp. 96–101, Marrakech, Morocco, 2008. European Language Resources Association (ELRA).

[11] Magdalena Repp, Petra B. Schumacher, and Fahime Same. Multi-layered annotation of conversation-like narratives in German. In Jakob Prange and Annemarie Friedrich, editors, **Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)**, pp. 61–72, Toronto, Canada, 2023. Association for Computational Linguistics.

[12] Martha Palmer, Chung-Hye Han, Na-Rae Han, Eon-Suk Ko, Hee-Jong Yi, Alan Lee, Chris Walker, John Duda, and Nianwen Xue. Korean English Treebank annotations, 2002. LDC2002T26.

[13] Na-Rae Han, Shijong Ryu, Sook-Hee Chae, Seung yun Yang, Seunghun Lee, and Martha Palmer. Korean Treebank annotations version 2.0, 2006. LDC2006T09.

[14] Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. Annotating a Japanese text corpus with predicate-argument and coreference relations. In Branimir Boguraev, Nancy Ide, Adam Meyers, Shigeko Nariyama, Manfred Stede, Janyce Wiebe, and Graham Wilcock, editors, **Proceedings of the Linguistic Annotation Workshop**, pp. 132–139, Prague, Czech Republic, 2007. Association for Computational Linguistics.

[15] Ed Kainoki. The Kainoki Treebank – a parsed corpus of contemporary Japanese, 2022.

[16] Hiroki Nomoto. Kyokushoushugi ni motoduku heiretsu tsuriibanku no kouchiku [Building a parallel treebank based on minimalism]. In **Proceedings of the Twenty-Eighth Annual Meeting of the Association for Natural Language Processing**, pp. 103–107, 2022.

[17] Jorge Baptista, Simone Pereira, and Nuno Mamede. ZAC: Zero Anaphora Corpus (a corpus for zero anaphora resolution in Portuguese). In **Proceedings of Workshop on Corpora and Tools for Processing Corpora, PROPOR 2016**, pp. 38–45, 2016.

[18] Yaqin Yang and Nianwen Xue. Chasing the ghost: Recovering empty categories in the Chinese Treebank. In Chu-Ren Huang and Dan Jurafsky, editors, **Coling 2010: Posters**, pp. 1382–1390, Beijing, China, 2010. Coling 2010 Organizing Committee.

[19] Hiroki Nomoto. Aspek nahu dalam penterjemahan bahasa Jepun-bahasa Melayu: Ayat kewujudan dan pengguguran *pro* [Grammatical aspects in Japanese-Malay translation: Existential sentences and *pro* drop]. In Sang Seong Goh, editor, **Penterjemahan Struktur Bahasa Asing dalam Bahasa Melayu**, pp. 200–221. Dewan Bahasa dan Pustaka, Kuala Lumpur, 2022.

[20] Nor Hashimah Jalaluddin. **Bahasa dalam Perniagaan: Satu Analisis Semantik dan Pragmatik [Language in Commerce: A Semantic and Pragmatic Analysis]**. Dewan Bahasa dan Pustaka, Kuala Lumpur, 2003.

[21] Nor Hashimah Jalaluddin, Harishon Radzi, Maslida Yusof, Raja Masittah Raja Ariffin, and Sa'adiah Ma'alip. **Sistem Panggilan dalam Keluarga Melayu: Satu Dokumentasi [Address System in Malay Families: A Documentation]**. Dewan Bahasa dan Pustaka, Kuala Lumpur, 2005.

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

A Other language specific considerations

A.1 Verb + particle

The direct object of a verb accompanied by particles such as *lagi* 'more', *balik* 'back', *kembali* 'back', *semula* 'again' can occur either after the verb (6a) or between the verb and the particle (6b). We chose the former analysis in our annotation scheme.

(5) Tak boleh *kurang* *lagi* kak?
not can discount more elder.sister
'Can't you discount more, sis?' [20]

(6) a. ✓ Tak boleh kurang *e* lagi kak?
b. Tak boleh kurang lagi *e* kak?

A.2 Serial verbs

Some serial verbs (V_1 V_2) allow the object of V_1 to occur either after V_1, as in (8a), or after V_2, as in (8b). We chose the former analysis.

(7) Saya baru *bawa* *keluar* kejap tadi.
I just carry go.out for.a.moment just.now
'I just took it out a moment ago.' [20]

(8) a. ✓ Saya baru bawa *e* keluar kejap tadi.
b. Saya baru bawa keluar *e* kejap tadi.

A.3 Left dislocation vs. topicalization

Left dislocation refers to a construction in which the sentence-initial topic is repeated by a resumptive pronoun, as in (9a). It differs from topicalization in that the latter does not involve a resumptive pronoun, as in (9b).

(9) a. Left dislocation

Yang merah itu aku nak *dia*.
REL red that I want it
'The red one, I want it.'

b. Topicalization

Yang merah itu aku nak.
REL red that it want
'As for the red one, I want.'

Another difference is that only topicalization is subject to the so-called island conditions. Therefore, when the pronoun is absent, only the left dislocation analysis is available if islands are involved, as in (10).

(10) a. Yang merah itu ada orang beli.
REL red that be person buy
'The red one, there's a person who bought it.' [20]
b. Yang merah itu ada orang beli *e*.

However, if not, both left dislocation and topicalization analyses are possible. In this case, we chose the left dislocation analysis.

(11) a. ✓ Yang merah itu aku nak *e*.
b. Yang merah itu aku nak.

A.4 Fixed expressions

Some fixed expressions can be analysed as resulting from omitting arguments. However, we treat fixed expressions as not involving *pro* drop.

(12) (Saya) Tak apa.
I not what
'It's OK.'

(13) a. ✓ Tak apa.
b. *e* Tak apa.

(14) Apa khabar (awak)?
what news your
'How are you?'

(15) a. ✓ Apa khabar?
b. Apa khabar *e*?

(16) (Saya) Terima kasih (awak).
I receive love your
'Thank you.'

(17) a. ✓ Terima kasih.
b. *e* Terima kasih *e*.

(18) (Saya) Biar-lah (awak).
I let-PART you
'Let it be.'

(19) a. ✓ Biarlah.
b. *e* Biarlah *e*.

B Abbreviations

ACC: accusative; DAT: dative; NOM: nominative; PART: particle; POL: polite; REL: relativizer