Pronoun substitute annotation in seven Asian languages

Hiroki Nomoto[1] Ryuko Taniguchi[1] Shiori Nakamura[1] Yunjin Nam[1]
Sri Budi Lestari[2] Sunisa Wittayapanyanon (Saito)[1] Virach Sornlertlamvanich[3]
Atsushi Kasuga[4] Kenji Okano[1] Thuzar Hlaing[1]

[1]Tokyo University of Foreign Studies [2]Ritsumeikan Asia Pacific University
[3]Musashino University/Thammasat University [4]Kanda University of International Studies

{nomoto,ryukota,nakamura.shiori.v0,namyj,sunisa,okanok,thuzarhlaing}
@tufs.ac.jp tari0828@apu.ac.jp virach@musashino-u.ac.jp
kasugaat@kanda.kuis.ac.jp

Abstract

We have annotated first- and second-person expressions as well as address terms in the spoken language corpora of seven Asian languages. This annotation is important because these languages employ many non-pronominal expressions ('pronoun substitutes') in addition to personal pronouns when referring to the speaker/addressee. A total of more than four million words have been annotated. Furthermore, we built new (and first-ever annotated) spoken language corpora for five Southeast Asian languages. We also developed a web-based corpus annotation tool (ETA: Easy Text Annotator) that enables non-tech-savvy linguists to benefit from our annotation results. Both resources can be used for purposes other than the present study.

1 Introduction

Reference to the speaker and the addressee is achieved by personal pronouns such as *I* and *you*. However, some languages employ other expressions for the same purpose. Such expressions are called 'pronoun substitutes' (henceforth 'prosub'). For example, the word for 'teacher' can be used not only when referring to a third-person teacher but also when teachers refer to themselves (first person) and when students refer to their teachers (second person) in Japanese, Korean, Malay, Javanese, Thai, Vietnamese and Burmese [1]. Japanese examples are given in (1).[1)]

(1) a. *Sensei* ni choodai.
'Give it to *teacher* (= me).'
b. *Sensei* no heya wa doko desuka?
'Where is *teacher*'s (= your) room?'

Expressions such as this are thus ambiguous with regard to person. In language understanding, they require disambiguation. They can be prosubs or non-prosubs. If they are prosubs, they may be first person or second person. They pose a more difficult challenge to language generation. One needs to choose the most appropriate expressions for 'I' and 'you' according to the current context of conversation from a variety of candidates. Annotated corpora are indispensable in solving these problems automatically.

This study reports our project of annotating corpora with prosubs and related expressions in seven languages, i.e. Japanese, Korean, Malay, Indonesian, Thai, Vietnamese and Burmese.

2 Related work

The study of prosubs in linguistics has long been conducted at the level of specific languages and attempts at understanding their cross-linguistic properties have just begun recently. [2] is a summary of previous studies on Thai, Burmese, Malay, Indonesian, Javanese and Korean. In discussing these languages, it also makes reference to major previous studies on Japanese.

Regarding the use of large corpora, [3, 4] examined first- and second-person expressions, both personal pronouns and prosubs, in the monitor version of the Corpus of Everyday Japanese Conversation (CEJC) [5]. However, her annotation data are not openly available.

[1] created a multilingual dataset that provides information about whether or not a given expression can be used as

[1)] The examples in (1) and (2) were taken from the dataset reported in [1] available at https://github.com/matbahasa/ProSub/blob/main/data_all_v1.0.json.

(i) a first-person prosub, (ii) a second-person prosub, (iii) an address term and (iv) an honorific title in the seven target languages of the present study and Javanese. Examples are also given when the relevant use is available. The expressions are based on the common questionnaire for investigating pronoun substitutes and address terms developed by [6], which is a list of notions whose exponents are possible candidates of prosubs. The dataset and the questionnaire are available at `https://github.com/matbahasa/ProSub`.

3 Corpora

We used TUFS Asian Language Parallel Corpus (TALPCo) [7] and the corpora shown in Table 1.

TALPCo is an open parallel corpus consisting of 1,372 Japanese sentences and their translations into our target languages and English.[2] The sentences are given in a formal conversational register.

For Japanese and Korean, the national language institutes for the respective countries have developed spoken language corpora that we could use for our purpose. The two Korean corpora are part of the Modu Corpus published by the National Institute of Korean Language.[3] We selected 0.9% (228/25,696 files) of the Spoken Corpus and 26.9% (600/2,232 files) of the Dialogue Corpus.

No existing spoken language corpus suitable for our project was available for the other five languages. Hence, we decided to compile our own corpora. Some of the corpus data are already publicly available. The conversation data of Malay are from [9, 10].[4] The film data of Indonesian consist of the captions of two Indonesian films, i.e. *Gundala* and *5 cm*, taken from opensubtitles.org.[5] We also intend to make other data openly available in the future.

The word count of the Japanese corpus is based on the information provided on the corpus' website.[6] Those of the other languages were calculated based on the tokenization using the following tools: NLTK Tokenizer's `word_tokenize` [11] (Korean, Malay, Indonesian, Burmese), Trigram word segmentation and POS tagger [12] and Python Vietnamese Toolkit's ViTokenizer[7].

4 Annotation scheme and tools

The annotation was performed by undergraduate and graduate students as well as ourselves, with each language team consisting of 3–5 members. TALPCo was used for training before embarking on larger data.

The following three tags are used: `1st` (first person), `2nd` (second person) and `address` (address term). `1st` and `2nd` include not only prosubs but also personal pronouns. Personal pronouns are included because prosubs are extended personal pronouns, so to speak, and it is beneficial to study the two together. Address terms are also annotated because the expressions used as prosubs overlap considerably with those used as address terms, though they are not identical. An example of address terms in Japanese is given in (2). Recall that the word *sensei* can also be used as first- and second-person prosubs, as shown in (1).

(2) *Sensei*, bokura no heya wa doko desuka?
'Where is our room, *Teacher*?'

Such overlaps sometimes make it difficult to distinguish between a second-person prosub and an address term because, unlike in English, the subject does not have to be expressed overtly in our target languages. The rule of thumb when faced with uncertain cases is that second-person prosubs allow substitution by a second-person pronoun 'you', but address terms do not.

However, there are still cases that cannot be solved by such a rule. (3) is a case in point from CEJC. What makes this example complicated is the fact that the conversation involves more than two individuals, including Kaochan. Hence, *Kaochan* can be a third person when the sentence is not directed to her.

(3) *Kaochan* kore mo taberu?
(i) 'Does *Kaochan* (= you) wanna eat this too?'
→ `2nd`
(ii) 'Do Ø_{you} wanna eat this too, *Kaochan*?'
→ `address`
(iii) 'Does *Kaochan* wanna eat this too?'
→ no annotation

In the case of CEJC, the associated audio and video files enable us to choose from the three possibilities, based on the presence/absence of the prosody characteristic to address

[2] `https://github.com/matbahasa/TALPCo`
[3] `https://corpus.korean.go.kr`
[4] `https://github.com/matbahasa/Melayu_Standard_Lisan/tree/master/NorHashimah`
[5] `https://www.opensubtitles.org`
[6] `https://www2.ninjal.ac.jp/conversation/cejc/cejc-wc.html`
[7] `https://pypi.org/project/pyvi`

Table 1 Corpora used for prosub annotation

Language	Corpus	Content	Word count
Japanese	Corpus of Everyday Japanese Conversation [8]	conversations	2,421,162
Korean	NIKL Spoken Corpus, Dialogue Corpus 2020	monologues, conversations, drama scripts	1,484,527
Malay	original	conversations, play scripts	39,265
Indonesian	original	films, conversations	32,870
Thai	original	drama scripts, novels	272,342
Vietnamese	original	conversations, film scripts, stories	146,521
Burmese	original	play scripts	10,675

terms and the direction of the speaker's gaze. However, such cues are unavailable in other languages. The annotators were asked to make their best decision even for very difficult cases rather than leaving them unannotated.

We used doccano [13] (Thai, Vietnamese) and the UAM CorpusTool[8] (the other languages) for the annotation task. The annotation results can be exported in the jsonl format in the former and the tab-delimited text and XML formats in the latter. The examples of these formats are given in the Appendix.

5 Results

The annotation results will be made available to the public at `https://github.com/matbahasa/ProSub` either by uploading the files there directly or by providing links to other locations where the files are available.

5.1 TALPCo

Table 2 summarizes the number of annotated items in TALPCo in the seven target languages.

Table 2 The distribution of annotations in TALPCo

Language	`1st`	`2nd`	`address`	Total
Japanese	161	8	3	172
Korean	161	3	2	166
Malay	800	42	3	845
Indonesian	788	31	3	822
Thai	164	15	4	183
Vietnamese	749	45	1	795
Burmese	164	6	2	172

5.2 CEJC

Two groups conducted the annotation of CEJC independent of each other in order to examine the difficulty of the task. As of January 2023, 82% of the entire corpus have been completed by both groups. Table 3 summarizes the number of annotated items by the two groups. (4) shows some of the prosubs found in CEJC.

Table 3 The distribution of annotations in CEJC

Group	`1st`	`2nd`	`address`	Total
A	9,298	3,522	2,634	15,457
B	9,447	4,231	1,286	14,964

(4) a. `1st`: *Sasagawa* (personal name), *baaba* 'grandma', *jibun* 'self', *kotchi* 'over here', *minna* 'everyone'

b. `2nd`: *Nomiyama san* (personal name + title), *mama* 'mama', *papa tachi* 'dad and his associates', *(go-)jibun* 'self',[9] *sensei* 'teacher', *minna* 'everyone'

The following two kinds of inter-annotator agreement scores were calculated:

1. Span match: The annotators identified an identical text span.
2. Value match: The annotators assigned the same value.

Value match is only relevant for matching text spans. Regarding Group B's annotations as the gold standard, the F-measure is calculated by the following equation in [14]. The F-scores thus calculated are 83.5 for span match and 95.5 for value match.

Span disagreements occur frequently with texts containing transcription symbols such as ')' and '/'. In fact, the annotators made guidelines to handle those symbols, but the presence of the symbols still disrupted their judgements. Most of value disagreements are concerned with the choice between `2nd` and `address`, as shown in the confusion matrix in Figure 1.

5.3 Facilitating use by linguists

The three kinds of output files created by the annotation tools, i.e. jsonl, tab-delimited text and XML, are not easy

[8] `http://www.corpustool.com`. We used version 3.3 because version 6 did not run on any of our Windows computers.

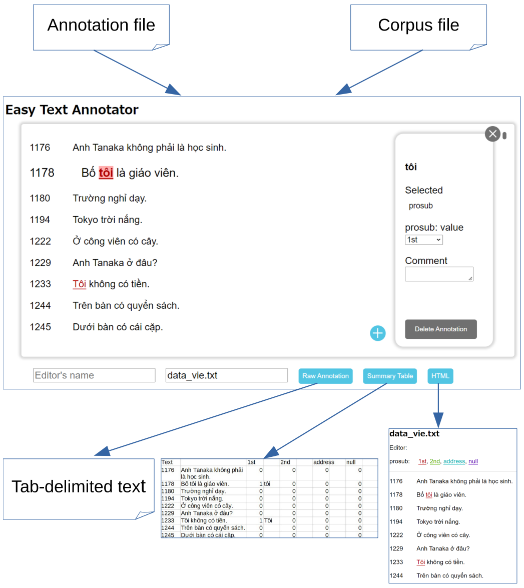
[9] *Go-* is an honorific prefix.

		Group B		
		1st	2nd	address
Group A	1st	8,046	25	8
	2nd	42	2,396	140
	address	11	847	1,005

Figure 1 Confusion matrix

to use for most linguists. Moreover, CEJC (Japanese) and NIKL corpora (Korean) prohibit redistribution of the original sentences, making it difficult for us to share the annotation results with the wider research community. As a solution to these issues, we developed a web-based tool called ETA (Easy Text Annotator) [15].

ETA takes the annotation and corpus files separately (Figure 2). For the CEJC and the NIKL corpora, we can only provide the annotation files. The corpus files need to be obtained/purchased from the institutes in charge. ETA can combine these files. One can modify the original annotation on ETA too. ETA supports three kinds of outputs. The summary table output is a text file formatted as a table and contains information about the frequencies of annotation values. It can be opened with a spreadsheet application such as Microsoft Excel. The HTML output allows one to examine an annotated text visually.



1176 Anh Tanaka không phải là học sinh.
1178 Bố tôi là giáo viên.
1180 Trường nghỉ dạy.
1194 Tokyo trời nắng.
1222 Ở công viên có cây.
1229 Anh Tanaka ở đâu?
1233 Tôi không có tiền.
1244 Trên bàn có quyển sách.
1245 Dưới bàn có cái cặp.

tôi
Selected
prosub
prosub: value
1st
Comment

Editor's name
data_vie.txt

Text		1st		2nd	address	null
1176	Anh Tanaka không phải là học sinh.	0		0	0	0
1178	Bố tôi là giáo viên.	1	tôi	0	0	0
1180	Trường nghỉ dạy.	0		0	0	0
1194	Tokyo trời nắng.	0		0	0	0
1222	Ở công viên có cây.	0		0	0	0
1229	Anh Tanaka ở đâu?	0		0	0	0
1233	Tôi không có tiền.	1	Tôi	0	0	0
1244	Trên bàn có quyển sách.	0		0	0	0
1245	Dưới bàn có cái cặp.	0		0	0	0

data_vie.txt
Editor:
prosub: 1st, 2nd, address, null

1176 Anh Tanaka không phải là học sinh.
1178 Bố tôi là giáo viên.
1180 Trường nghỉ dạy.
1194 Tokyo trời nắng.
1222 Ở công viên có cây.
1229 Anh Tanaka ở đâu?
1233 Tôi không có tiền.
1244 Trên bàn có quyển sách.

Figure 2 ETA: Easy Text Annotator

6 Conclusion

In this study, we annotated corpora in seven Asian languages with three prosub-related tags, i.e. 1st, 2nd and address. Moreover, we have built new spoken language corpora in five Southeast Asian languages. Compared to written language corpora, spoken language corpora, especially those involving naturalistic conversations, are limited to a handful of languages including Japanese and Korean. Our new corpora are thus valuable resources, even though their sizes are much smaller than those of the latter languages. We have also developed a corpus annotation and visualization tool ETA. It was originally developed for the project, but it can be used more generally for other projects.

The results of our corpus annotation will answer questions such as the following:

1. What are the expressions used to refer to the speaker and the addressee in language X?
2. What are the expressions that can be used as prosubs in language X? How are they similar to/different from those in language Y?
3. How are various speaker-/addressee-referring expressions used in a discourse?

[1] is an initial attempt to answer the second question. However, their dataset is based on a questionnaire rather than corpora. The present study can expand their list.

At least two improvements are conceivable in the present study. Firstly, 1st and 2nd can be supplied with subclass information. Expressions assigned these tags can be (i) personal pronouns, (ii) prosubs or (iii) apparent prosubs. Apparent prosubs differ from genuine ones in that the person information is not lexically encoded.[10] They just happen to refer to the speaker/addressee, given a particular context. English examples such as (5), discussed by [16] under the term of 'imposters', are thought to fall into this category.

(5) a. In this reply, *the present authors* attempt to defend {themselves/ourselves} against the scurrilous charges which have been made. → 1st
b. How is *my darling* tonight? → 2nd

Secondly, it is important to recognize the so-called '*pro* drop', i.e. a phenomenon in which the arguments of a predicate are not overtly expressed, in our target languages (cf. (3)). If our corpora were supplied with information about the presence of such empty items, they would also be subject to our annotation.

[10] We assume that the person information is lexically encoded in prosubs. For example, in Malay, although *cikgu* 'teacher' can be both first- and second-person prosubs, its synonym *guru* 'teacher' cannot be either even when the relevant teacher is the speaker/addressee in the context.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP20H01255.

References

[1] Ryuko Taniguchi, Wataru Okubo, Hiroki Nomoto, and Yunjin Nam. Daimeishidaiyou, yobikake hyougen no tagengo deetasetto [A multilingual dataset of pronoun substitutes and address terms]. In **The Proceedings of the 164th Meeting of the Linguistic Society of Japan**, pp. 307–313, 2022.

[2] Hiroki Nomoto, Sunisa Wittayapanyanon (Saito), Kenji Okano, Thuzar Hlaing, Yunjin Nam, and Sri Budi Lestari. Daimeishidaiyou, yobikake hyougen kenkyuu no genjou: Taigo, birumago, mareego, indoneshiago, jawago, chousengo [Current state of studies on pronoun substitutes and address terms: Thai, Burmese, Malay, Indonesian, Javanese and Korean]. **Gogaku Kenkyuujo Ronshuu**, Vol. 23, pp. 63–78, 2021. English version available at https://ling.auf.net/lingbuzz/005895.

[3] Ayami Kawano. **Nihongo Nichijoukaiwa Koopasu** no kazokukankaiwa niokeru koshou no shiyoujittai: Taishoudaimeishi o chuushin ni [Use of calling terms in family conversations in **The Corpus of Everyday Japanese Conversation**: Focusing on addressee-referring pronouns], 2019. Paper presented at Everyday Conversation Corpus Symposium IV.

[4] Ayami Kawano. **Nihongo Nichijoukaiwa Koopasu** niokeru jishou shiyou [Use of self-referring terms in **The Corpus of Everyday Japanese Conversation**], 2021. Paper presented at Everyday Conversation Corpus Symposium VI.

[5] Hanae Koiso, Haruka Amatani, Yuichi Ishimoto, Yuriko Iseki, Yasuyuki Usuda, Wakako Kashino, Yoshiko Kawabata, Yayoi Tanaka, Yasuharu Den, and Ken'ya Nishikawa. **Nihongo Nichijoukaiwa Koopasu monitaakoukaiban: Koopasu no sekkei to tokyuchou** [**Corpus of Everyday Japanese Conversation** monitor version: The architecture and features of the corpus]. National Institute for Japanese Language and Linguistics, Tokyo, 2019.

[6] Kenji Okano, Hiroki Nomoto, Sunisa Wittayapanyanon, Thuzar Hlaing, and Atsushi Kasuga. Ajia sangengo niokeru daimeishidaiyou, yobikakego no kyoutsuukoumoku chousa [An investigation of pronoun substitutes and address terms in three Asian languages based on a common questionnaire]. In **Proceedings of the Twenty-Eighth Annual Meeting of the Association for Natural Language Processing**, pp. 69–73, 2022.

[7] Hiroki Nomoto, Kenji Okano, David Moeljadi, and Hideo Sawada. TUFS Asian Language Parallel Corpus (TALPCo). In **Proceedings of the Twenty-Fourth Annual Meeting of the Association for Natural Language Processing**, pp. 436–439, 2018.

[8] Hanae Koiso, Haruka Amatani, Yasuharu Den, Yuriko Iseki, Yuichi Ishimoto, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, Yasuyuki Usuda, and Yuka Watanabe. Design and evaluation of the Corpus of Everyday Japanese Conversation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 5587–5594, Marseille, France, 2022. European Language Resources Association.

[9] 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.

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

[11] Steven Bird, Edward Loper, and Ewan Klein. **Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit**. O'Reilly Media Inc., Sebastopol, CA, 2009.

[12] Virach Sornlertlamvanich, Naoto Takahashi, and Hitoshi Isahara. Building a Thai part-of-speech tagged corpus (ORCHID). **The Journal of the Acoustical Society of Japan (E)**, Vol. 20, No. 3, pp. 189–198, 1999.

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

[14] George Hripcsak and Adam S. Rothschild. Agreement, the F-measure, and reliability in information retrieval. **Journal of the American Medical Informatics Association**, Vol. 12, No. 3, pp. 296–298, 2005.

[15] Hiroki Nomoto, Kaoru Kayukawa, and Yuta Takayasu. ETA: Easy Text Annotator. https://github.com/matbahasa/ETA, 2022.

[16] Chris Collins and Paul Postal. **Imposters: A Study of Pronominal Agreement**. MIT Press, Cambridge, MA, 2012.

Appendix: Output formats of doccano and UAM CorpusTool

A Jsonl (doccano)

```
{"id":265,"text":"2227\t わたしは きのう 大学へ 行きました。","label":[[5,8,"1st"]]}
{"id":266,"text":"2229\t たばこは やめました。","label":[]}
{"id":267,"text":"2232\t きょうは 誰も 来ませんでした。","label":[]}
{"id":268,"text":"2234\t あなたも よく 働きますね。","label":[[5,8,"1st"]]}
...
{"id":290,"text":"2744\t 先生、 こちらが 私の 母です。","label":[[5,7,"address"],[14,15,"1st"]]}
```

B Tab-delimited text (UAM CorpusTool)

Filename	ID	Start	End	Text	Comment	ParentID	prosub	1st	2nd	address
data_jpn.txt	39	6569	6572	わたし			1	1	0	0
data_jpn.txt	40	6636	6639	あなた			1	0	1	0
data_jpn.txt	41	7084	7086	先生			1	0	0	1
data_jpn.txt	42	7093	7094	私			1	1	0	0

C XML (UAM CorpusTool)

```
<?xml version='1.0' encoding='utf-8'?>
<document>
  <header>
    <textfile>Texts/data_jpn.txt</textfile>
    <lang>japanese</lang>
  </header>
    <body>
    2227 <segment id='39' features='prosub;1st' state='active'>わたし</segment>は きのう 大学へ 行きました。
    2229 たばこは やめました。
    2232 きょうは 誰も 来ませんでした。
    2234 <segment id='40' features='prosub;2nd' state='active'>あなた</segment>も よく 働きますね。
    ...
    2744 <segment id='41' features='prosub;address' state='active'>先生</segment>、こちらが <segment id='42' features='prosub;1st' state='active'>私</segment>の 母です。
    </body>
</document>
```