# Corpus Development for Pronoun Substitute and Address Term Study

Virach Sornlertlamvanich
*AAII, Department of Data Science*
*Musashino University*
Tokyo, Japan
*Faculty of Engineering*
*Thammasat University*
Pathumthani, Thailand
ORCID: 0000-0002-6918-8713

Sunisa Wittayapanyanon (Saito)
*World Language and Society Education Center*
*Tokyo University of Foreign Studies*
Tokyo, Japan
ORCID: 0000-0002-7892-3628

2023 8th International Conference on Business and Industrial Research (ICBIR) | 979-8-3503-9964-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICBIR57571.2023.10147532

***Abstract*—The development of a corpus of pronoun substitute and address term annotations is aimed at exploring the actual use of these language elements in dialogue sentences. To create the corpus, dialogue sentences were extracted from a collection of popular TV drama scripts and novels, representing both genres in order to provide a comprehensive overview of conversation settings. One key feature of the corpus is the inclusion of a detailed definition of kinship relations. This was done to gain a better understanding of the usage and meaning of kinship terms in dialogue. To ensure the accuracy of the corpus annotations, each sentence is annotated by two annotators, and the results are compared. This approach helps to minimize errors and personal biases, leading to a high-quality corpus that accurately reflects the usage of pronoun substitutes and address terms in natural language conversations. The comparison of annotations can also highlight any discrepancies, which can then be resolved through further discussion and collaboration between the annotators. This rigorous process helps to ensure that the corpus is a reliable and trustworthy resource. The resulting corpus will provide a wealth of information about these language elements and the way they are used in real-life dialogues, making it a useful resource for researchers in this field.**

***Keywords*—pronoun substitute, address term, corpus development, kinship term, dialogue**

## I. INTRODUCTION

Pronoun substitute is a word or a phrase that is intentionally used to refer to a speaker or an addressee of a conversation instead of using an usual personal pronoun. Pronoun substitutes are frequently and naturally used in Thai conversation. A preliminary study has been conducted by a collaborative collection of the pronoun substitute and address term in eight Asian languages under a designed encoding scheme to understand the context feature of the expression [1]. Resulting from the context feature study, the purpose of the corpus development in this research is to capture the expression in context of the pronoun substitute and address term, and their patterns of the possible word types in the actual dialogue sentences.

Pronoun substitutes and address terms are important linguistic cues that help us understand the intended referent of a noun in a sentence, such as the word "*girl*" in "That's a heavy load, *girl*." which is the second-person, "you" perspective, in the conversation. In natural language processing tasks such as machine translation and dialogue systems, accurate anaphora resolution is crucial for generating coherent and meaningful responses. By leveraging these linguistic cues, language models can improve their ability to accurately identify the intended referent of a pronoun and generate more accurate and contextually appropriate responses.

In the field of Thai linguistics, there is a lack of consistency in the meaning and definition of related terminology. To address this, a new term in Thai, "คำแทนบุรุษสรรพนาม" (*kham thɛɛn burùt sàpphanaam*) is proposed, which refers to words that can be used as substitutes for personal pronouns to indicate speakers and addressees [2]. This term includes all types of words such as kinship terms, occupation terms, and personal names, etc., which can be used as substitute for personal pronouns. In principle, the pronoun substitutes in Thai can be used for both of first-person expression and second-person expression, but there are words which can be used only to represent the speaker or the addressee of a conversation. In many cases, it is difficult to discriminate the type of pronoun substitute whether it refers to the speaker or the addressee. It is even more difficult when there is no sentential context or situation of the conversation. In general, the words which can indicate gender, social status, and attitude are possibly accounted for a pronoun substitute.

In the Thai language, there is a specific term used for addressing someone in a conversation or message, "คำเรียกบุคคล" (*kham rîak bùkkhon*) which is used to identify the person being spoken to and is a crucial element in the Thai language when it comes to communication [2]. The address term in Thai does not play the role of the subject or object in a sentence, but rather serves a function of calling attention to the interlocutor.

One of the unique features of this term is that it can be placed at different positions in a sentence. It can be found in the initial position, final position, or even in the medial position of a sentence, depending on the context and the speaker's intention. Additionally, the address term in Thai is often followed by particles, which add further emphasis and provide additional information about the person being addressed.

Similarly, in English, there is also a grammatical construct known as the vocative case, which is used to indicate the person being addressed [3]. The vocative case is typically used in spoken language and informal writing. Unlike some other languages, English does not use inflection to indicate the vocative case, but instead uses an optional noun phrase with a specific intonation [4]. This grammatical construct is known to express attitude, politeness, formality, status, intimacy, or role relationship, and is often used to characterize the speaker to the person being addressed [5]. The vocative can be indicated by using the word "you" in the sentence, or by using the person's name or title, such as "John," "Mom," or "Doctor." It is often found at the beginning of a sentence, but it can also be placed at the end or even at the middle of a sentence, depending on the context and the speaker's intention [6].

The pronoun substitute and address term in Thai are also found in association with a title noun, describing someone's position or job. The title noun in Thai, “คำนำหน้านาม” (*kham namnâa naam*), such as “คุณ” (*khun*, Mr./Ms.) or “ท่าน” (*thâan*, Sir/Madam) is used as a prefix of a kinship term, occupation term, personal name. The title noun is used to indicate not only an honorific, but also an intimacy like “น้อง” (*nɔ́ɔŋ*, younger sibling) [2, 7].

This study collects dialogue sentences from TV drama scripts, which are rich in context, including conversational settings and participants. To provide a comparison, dialogue sentences with less clear context from novels are also included.

The remainder of the paper is organized as follows. Section II provides the information of the text corpus and the criteria for the annotation task. Section III exhibits the result of the corpus annotation. Section IV discusses the notable usages of pronoun substitute and address term picked up from the corpus. Finally, the conclusion and future work are remarked in Section V.

## II. Corpus Annotation

The goal of this study is to uncover the various types of words that can be used as pronoun substitutes or address terms in Thai popular drama scripts and novels, as listed in Table I.

TABLE I. THAI CONVERSATION CORPUS

| | Title | Type | Features | No. of Sentence | No. of Word |
|---|---|---|---|---|---|
| 1 | *nakii* (1 episode) [8]<br>นาคี | TV drama script | Contemporary spoken Thai, Fantasy | 255 | 3,016 |
| 2 | *phíphóphǐmmáphaan* (1 episode) [8]<br>พิภพหิมพานต์ | | Contemporary spoken Thai, Fantasy | 207 | 2,289 |
| 3 | *phləəŋnaakhaa* (1 episode) [8]<br>เพลิงนาคา | | Contemporary spoken Thai, Fantasy | 174 | 1,673 |
| 4 | Dare to Love (24 episodes) [9]<br>ให้รักพิพากษา | | Contemporary spoken Thai | 9,940 | 127,220 |
| 5 | *khwaamsùk khɔ̌ɔŋ kathíʔ* [10]<br>ความสุขของกะทิ | Novel | Contemporary spoken Thai | 92 | 1,459 |
| 6 | *tɛ̀ɛpaaŋkɔ̀ɔn* [11]<br>แต่ปางก่อน | | Spoken Thai from 1910 to present day including aristocratic words | 2,356 | 38,888 |
| 7 | *námsǎycayciŋ* [11]<br>น้ำใสใจจริง | | Contemporary spoken Thai | 3,676 | 41,609 |
| 8 | *phâathɔɔŋ* [11]<br>ผ้าทอง | | Contemporary spoken Thai | 3,906 | 56,188 |
| Total | | | | 20,606 | 272,342 |

The authors choose to focus on Thai popular drama scripts and novels because they provide a representation of natural conversation in Thai. This choice allows the study to identify the most common and prevalent forms of pronoun substitute and address terms used in Thai everyday conversation. By identifying these forms, the study aims to provide insight into the way Thai speakers use pronoun substitutes and address terms in their everyday language, which can be useful for language learners and researchers.

TABLE II. TAGSET AND CRITERIA FOR LABELLING

| Main | Sub | Criteria |
|---|---|---|
| Speaker | | • Subject or object in a sentence |
| Addressee | | |
| Address Term | | • Not subject nor object in a sentence<br>• Found in either initial, medial, or final position of a sentence |
| | Kinship Term 1 | • Original meaning<br>• Family by blood including relations between parents<br>• Reference point = speaker<br>e.g. The speaker calls an elder brother as “พี่” (*phîi,* elder sibling). |
| | Kinship Term 2 | • Derived meaning<br>• Not family by blood<br>e.g. The speaker calls a senior or a clerk who look older than the speaker as “พี่” (*phîi,* elder sibling). |
| | Kinship Term 3 | • Original meaning<br>• Family by blood including relations between parents<br>• Reference point = One of family members of the speakers (mostly the youngest family member)<br>e.g. A mother calls her elder son as “พี่” (*phîi,* elder sibling). Reference point = younger child<br>e.g. A mother calls herself as “แม่” (*mɛ̂ɛ*, mother). Reference point = child<br>e.g. A father calls his wife as “แม่” (*mɛ̂ɛ*, mother). Reference point = child |
| | Kinship Term 4 | • Original meaning<br>• Not family by blood<br>• Reference point = One of non family members of the speakers (mostly the youngest person)<br>e.g. The speaker calls elder child of the speaker’s friend as “พี่” (*phîi,* elder sibling). Reference point = younger child of the speaker’s friend |
| | Kinship Term 5 | • Thai annotation only<br>• Derived meaning; not a family relation meaning<br>e.g. The speaker calls an idol regardless of their ages as “พี่” (*phîi,* elder sibling) / “น้อง” (*nɔ́ɔŋ,* younger sibling) + “เก้า” (*kâw*, personal name)”<br>e.g. The speaker calls a friend with the same age as “ไอ้น้อง” (*ʔây nɔ́ɔŋ,* little brother)<br>e.g. The speaker calls his/her mother as “พี่” (*phîi*, elder sibling) |
| | Title | • Equivalent to “คำนำหน้านาม” (*kham namnâa naam*) as described in Section II<br>• Found in front of a personal name<br>• Indicating honorific or intimacy<br>e.g. “คุณ” (*khun,* Mr./Ms.) / “บอส” (bɔ́ɔt, boss) + personal name<br>• Double titles are possible<br>e.g. “พี่” (*phîi,* elder sibling) + “ทนาย” (*thanaay,* advocate) + personal name<br>= kinship term + occupation term |
| | Particle | • A part of address term expression |
| | Pronoun | • Tag as a personal pronoun to mark that the word is not considered to be a pronoun substitute<br>• According to “The Royal Institute Dictionary” [12] |

This annotation system allows for multiple tags to be assigned to a single term in the text. The system uses two levels of tagging to identify and classify pronoun substitutes and address terms in the text. The main tag is used to identify the type of pronoun substitute (speaker for first-person pronoun, and addressee for second-person pronoun) and the address term, while the sub tag is used to provide additional syntactic and pragmatic information, such as title, particle, and kinship term. A separate pronoun tag is also used to distinguish these terms from the target pronoun substitute and address term. This allows for a more detailed and nuanced understanding of the use of pronoun substitutes and address terms in the text. The annotation criteria and guidelines, including label names and instructions for tagging, are outlined in Table II for reference. This allows annotators to understand and apply the guidelines consistently and accurately. Overall, the use of multiple tags and two levels of tagging makes the annotation process more comprehensive, and provides a more detailed understanding of the use of pronoun substitutes and address terms in the text.

In addition, when the speaker uses a common noun such as “ท่านประธาน” (*thâan prathaan*, president) to designate the addressee who is not actually a president, we note it as “alias” in comment box. Meanwhile, the other language teams in this project note “nickname” for the same case. This is because “nickname” has a special meaning in Thai. Most of the Thai people are given a nickname at birth, and we are very familiar with the term “nickname” in the sense of another personal name for acquaintanceship. In the case of “president”, as the speaker uses it to express the term according to the attitude, behavior and appearance of the addressee. To avoid the misunderstanding, we propose to note “alias” in the comment box.

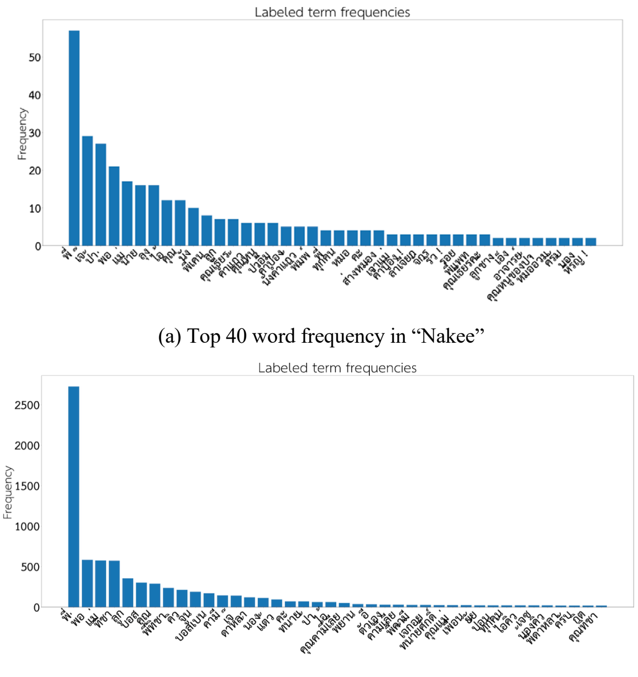
## III. Result of Corpus Annotation

A drama script is prepared in a form with a kind of structure which is helpful for extracting the dialogue and the speaker. In general, a script is composed of act, scene, setting, cast of characters, dialogue, and others. By separating the text into scene with setting information, an annotator can refer the circumstance for judging the intention of the speaker. The result of annotation of a script is more reliable comparing to the dialogue extracted from a novel which has no an apparent form for expressing a dialogue, and other circumstancing information. In this study, dialogues from drama scripts and novels are picked up to confirm the usages in both categories.

### *A. Dialogues from TV Drama Scripts*

The TV drama scripts of "Nakee" and "Dare to Love" are used in this study, with a total of three episodes of "Nakee" and 24 episodes of "Dare to Love" as listed in Table I. The data consists of a total of 10,576 sentences and 134,198 words. The average sentence length is 12.69 words, which is a measure of how many words are used to express an idea.

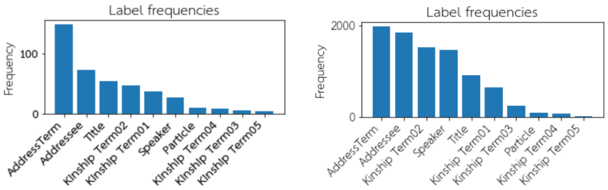
Only 106 words in “Nakee” are labeled, while 575 words in “Dare to Love” are labeled for the target of pronoun substitute and address term. As expected, the distribution of words and tags is similar, as shown in Fig. 1 and 2.



(a) Top 40 word frequency in “Nakee”

(b) Top 40 word frequency in “Dare to Love”

Fig. 1. Distribution of top 40 word frequency in TV drama scripts



(a) “Nakee”

(b) “Dare to Love”

Fig. 2. Distribution of tags in TV drama scripts

### *B. Dialogues from Novels*

In the four novels being studied, there are a total of 10,030 sentences and 138,144 words in the conversation. The average sentence length is 13.77 words. Additionally, 505 words have been labeled for the target of pronoun substitute and address term.

The analysis of the word distribution in novels suggests that it is similar to that of TV drama scripts. Both forms of text tend to use a lot of kinship terms when referring to people in conversation. This is likely because it is a way to express intimacy between conversation partners. The tag analysis apparently shows that terms referring to "Speaker" are used more frequently in novels than in TV drama scripts. The distributions of words and tags are illustrated in Fig. 3. This could be due to the nature of novels, which often have more internal monologues and self-reflection than a TV drama scripts, which tend to have more external dialogues and interactions.

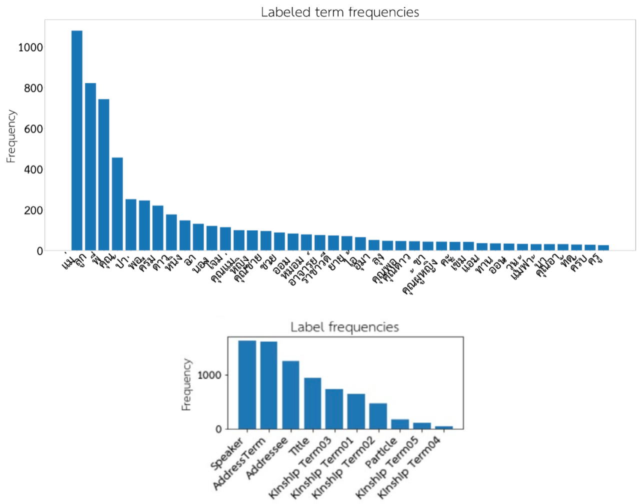


Fig. 3. Distribution of word frequency and tags in novels

### C. Dialogues from TV Drama Scripts and Novels

The distribution of word frequency and tags for the dialogues from TV drama scripts and the four novels is presented in Fig. 4. The dataset includes a total of 20,606 sentences and 272,342 words in the conversation. The average sentence length is 13.22 words, and a total of 1,106 words, out of a total of 16,736 counts, are labeled as the target for pronoun substitutes and address terms.

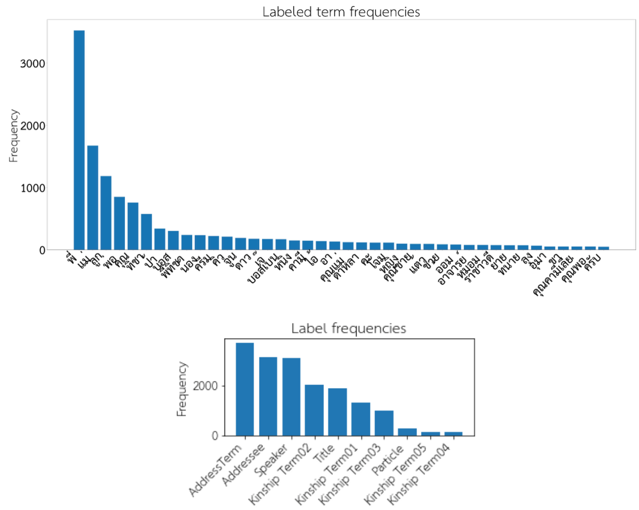
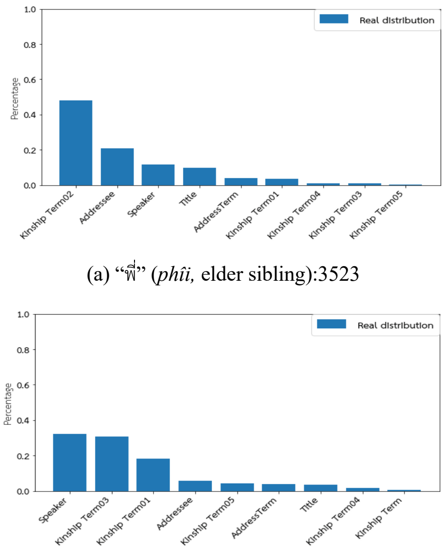


Fig. 4. Distribution of word frequency and tags in TV drama scripts and novels

From the total tagged data, it can be observed that kinship terms are frequently used in the dialogues. The use of kinship terms is apparently a way of expressing intimacy between the conversation partners. Moreover, when examining the types of kinship terms used, it can be seen that the use of "Kinship Term 2" as defined in Table II is particularly prevalent. This indicates that the speakers are drawing on a sense of familiarity with their conversation partners by referencing a close relationship, such as that of family by blood (Kinship Term 1). The use of such terms suggests that the speakers are attempting to create a comfortable and intimate atmosphere in their conversations.



(a) “พี่” (*phîi,* elder sibling):3523

(b) “แม่” (*mɛ̂ɛ,* mother):1674

Fig. 5. Distribution of tags for “พี่” (*phîi,* elder sibling):3523 and “แม่” (*mɛ̂ɛ,* mother):1674

Observing the kinship term closely in Fig. 5, we found that “พี่” (*phîi,* elder sibling):3523, “แม่” (*mɛ̂ɛ,* mother):1674, “ลูก” (*lûuk*, child):1182 ,“พ่อ” (*phɔ̂ɔ*, father):849 are the top four frequent words and they are all a kinship term of the first-degree relative.

“พี่” (*phîi,* elder sibling):3523 is the most frequently used in the tagged corpus. It shows the familiarity and honor to the conversation partners though there is no direct family relation. “แม่” (*mɛ̂ɛ,* mother):1674 is used for “Speaker” and “Kinship Term 3” in most cases.

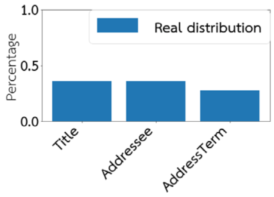


Fig. 6. Distribution of tags for “นาย” (*nāy*, Mr./boss/you):36

“นาย” (*naay,* Mr./boss/you):36 is one of special words which is popularly used in conversation. As expected, it is used in three possible meanings of Mr. (Title), boss (Address Term), and you (Addressee) as shown in Fig. 6.

The order of the most frequent use of the label is address term:3723, addressee:3162, and speaker:3107 as shown in Fig. 4. In terms of the variety of terms using to express the pronoun substitute and address term, the tagged data has shown in the average of number of label per term (i.e. no. of label divided by no. of term), namely, address term:4.25 (3723/875), addressee:9.88 (3162/320), and speaker:19.79 (3107/157). This means that the address term has more choice of words than addressee, and the addressee has more choice of words than the speaker. Accordingly, the words for speaker have the most specificity.

## IV. Pronoun Substitute and Address Term in Natural Conversation

One common challenge during annotation is determining if a term is a pronoun substitute or an address term when the conversational text lacks context. Without situational information, it can be difficult to distinguish between a pronoun substitute or common noun, and between an address term or proper noun.

Followings are ten representative cases that are commonly found with difficulties. The examples are itemized by the title with episode number and the order number as the sentence ID. In each sentence, the words that we have difficulty to identify are underlined, and followed by the discussion for a solution.

(1) Dare to Love 8: 9055
ชีวิตจะพังหรือไม่พัง ไม่ได้อยู่ที่คนอื่น แต่อยู่ที่ตัวเราเอง

| | | | | | |
|---|---|---|---|---|---|
| *chiiwít* | *càʔ* | *phaŋ* | *rɯ̌ɯ* | *mây* | *phaŋ* |
| life | will | break | or | NEG | break |
| *mây dây* | *yùu thîi* | *khon* | *ʔùɯn* | | |
| NEG | due to | person | other | | |
| *tὲɛ* | *yùu thîi* | *tua* | *raw* | *ʔeeŋ* | |
| but | due to | body | we | oneself | |

(NEG: negator)

"The life will be broken or not, it's not up to others, it's up to us.

(2) Dare to Love 8: 9008
คุณจะหลอกใครก็ได้แต่หลอกตัวเองไม่ได้

| | | | | | |
|---|---|---|---|---|---|
| *khun* | *càʔ* | *lɔ̀ɔk* | *khray* | *kɔ̂ɔ* | *dây* |
| you | will | cheat | someone | also | can |
| *tὲɛ* | *lɔ̀ɔk* | *tuaʔeeŋ* | *mây* | *dây* | |
| but | cheat | oneself | NEG | can | |

(NEG: negator)

"You can cheat someone, but you can't cheat yourself."

(1) and (2) show the cases that it is possible to identify the underlined words as common nouns (anaphor) or first-person expression or second-person expression. As for (1), we need to consider it carefully by the context it has. As เรา (*raw*, we/you) in Thai can be used to refer to either first-person or second-person. In addition, "ตัว" (*tua*, body) and "เอง" (*ʔeeŋ*, oneself) are added to "เรา" (*raw*, we/you), it can also be regarded as a common noun to mean "our body". Through the discussion among the annotation team members, this word is identified as a first-person expression following the general perception. Although "ตัวเอง" (*tuaʔeeŋ*, oneself) in (2) could be also considered as a common noun or a second-person expression. The addressee is tagged by the explicit context that this word clearly designates the addressee "คุณ" (*khun*, you) in the beginning of the sentence.

(3) Dare to Love 5: 7902
คนทำผิดก็ต้องขอโทษสิ

| | | | | |
|---|---|---|---|---|
| *khon* | *tham* | *phìt* | *kɔ̂ɔ* | *tɔ̂ɔŋ* |
| person | do | wrong | also | must |
| *khɔ̌ɔthôot* | | *sìʔ* | | |
| apologize | | PTCL | | |

(PTCL: particle)

"The person who did (something) wrong need to apologize."

The noun phrase "คนทำผิด" (*khon tham phìt*, person who did (something) wrong) is possibly interpreted as a common noun. However, we tagged it as an addressee because it is confirmed by the context that the noun phrase designates the person who is talked to.

(4) Dare to Love 2: 6533
ผมแค่ไม่คิดว่า ระดับเอ็มดีจะมาขอบคุณทนายฝึกหัดด้วยตัวเอง

| | | | | | |
|---|---|---|---|---|---|
| *phǒm* | *khɛ̂ɛ* | *mây* | *khít* | *wâa* | *radàp* |
| I | just | NEG | think | that | level |
| *ʔemdii* | *càʔ* | *maa* | *khɔ̀ɔpkhun* | | |
| MD | would | come | thank you | | |
| *thanaay fùkhàt* | | *dûay tuaʔeeŋ* | | | |
| junior lawyer | | by yourself | | | |

(NEG: negator)

"I didn't think that the person of MD level (= addressee) would come to say thank you to the junior lawyer (= speaker)."

It is possible to identify both "ระดับเอ็มดี" (*radàp ʔemdii*, person of MD level) and "ทนายฝึกหัด" (*thanaay fùkhàt*, junior lawyer) as a common noun. In this case, we can interpret that "ระดับเอ็มดี" (*radàp ʔemdii*, person of MD level) designates the addressee and "ทนายฝึกหัด" (*thanaay fùkhàt*, junior lawyer) designates the speaker based on the situational context of the story.

(5) Dare to Love 7: 8501
ในฐานะบอส

| | | |
|---|---|---|
| *nay* | *thǎanáʔ* | *bɔ́ɔt…* |
| as | position | boss… |

"As a boss…"

Even though "บอส" (*bɔ́ɔt*, boss) is an actual position of the speaker, we do not tag this word in this case because the speaker has an intention to mean the position in general, not the speaker individually.

(6) Dare to Love 8: 9122
ไม่เรียกว่าพี่เบนแล้วเหรอครับ

| | | | | |
|---|---|---|---|---|
| *mây* | *riak* | *wâa* | *phîi* | *ben* |
| NEG | call | that | elder brother | Ben |
| *lɛ́ɛw* | *rə̌ə* | *khráp* | | |
| PRF | Q | PTCL | | |

(NEG: negator, PRF: perfect, Q: question particle, PTCL: particle)

"Don't you call (me) as big brother Ben anymore?"

In case of (6), "พี่เบน" (*phîi ben*, big brother Ben) is not tagged because it is a name to call the speaker, and it does not yet designate the speaker of the sentence.

(7) Dare to Love 16: 41847
เดี๋ยวแฟนคุณไปเอาอาหารเช้ามาให้นะครับ เตรียมไว้ให้แล้ว

| | | | | |
|---|---|---|---|---|
| *dǐaw* | *fɛɛn* | *khun* | *pay* | *ʔaw* |
| later | lover | you | go | bring |
| *ʔaahǎan cháw* | | *maa* | *hây* | *náʔ khráp* |
| breakfast | | come | give | PTCL |
| *triam* | *wáy* | *hây* | *lɛ́ɛw* | |
| prepare | STAT | for | PRF | |

(PTCL: particle, STAT: stative, PRF: perfect)

"Your lover (= the speaker) will bring breakfast to you later. It's ready."

“แฟนคุณ” (*fɛɛn khun*, your lover) can be regarded as a third-person expression when we read the sentence singly. But situational context of the story shows the intention of the speaker to express the intimacy. So, the word is tagged as the speaker of the sentence.

(8) Dare to Love 6: 7964 8236
คนบางคนไม่เอ่ยปากขอให้ใครช่วยหรอกครับ ที่เราทำได้คือสังเกตให้ดีๆ แล้วเราจะรู้ว่าเขากำลังต้องการความช่วยเหลืออยู่

| | | | | | |
|---|---|---|---|---|---|
| *khon* | *baaŋ* | *khon* | *mây* | *ʔə̀əy pàak* | |
| person | some | CLF | NEG | say | |
| *khɔ̌ɔ hây* | *khray* | *chûay* | *rɔ̀ɔk khráp* | | |
| ask | someone | support | PTCL | | |
| *thîi* | *raw* | *tham* | *dây* | *khɯɯ* | |
| what | we | do | can | COP | |
| *sǎŋkèet* | *hây diidii* | | | | |
| observe | well | | | | |
| *lɛ́ɛw* | *raw* | *càʔ* | *rúu* | *wâa* | *kháw* |
| then | we | will | know | that | they |
| *kamlaŋ* | *tɔ̂ŋkaan* | *khaam chûaylʉ̌a* | | *yùu* | |
| PROG | need | support | | PROG | |

(CLF: classifier, NEG: negator, PTCL: particle, COP: copula, PROG: progressive)

“There is someone who don’t ask support to the others. What we can do is to observe well, then we will see what they need now.”

The cases (8) to (10) which concern personal pronouns, not pronoun substitutes, are selected to demonstrate the challenge in determining a person's identity without the context.

In Thai, “เรา” (*raw*) can be used to indicate both singular/plural first-person and singular/plural second-person [3]. In general, it is difficult to understand what “เรา” (*raw*) designates among the four possible interpretations. In addition, “เรา” (*raw*) can also be interpreted as a noun to mean persons in general, as shown in the case of (8). Considering the context in (8), we can tag it as the speaker. Furthermore, it also has a possibility to consider whether “เรา” (*raw*) has a meaning to include the addressee or not. Therefore, the context and the situational context are crucial in identifying the scope of the meaning of personal pronoun in Thai.

(9) Dare to Love 6: 7968
บอสเบนส่งมาว่า “ถึงบ้านหรือยังครับ ผมเห็นฝนตก ถ้าตากฝน รีบอาบน้ำสระผมนะครับ เดี๋ยวไม่สบาย” บอสเบนเป็นห่วงฉันด้วย

| | | | | | |
|---|---|---|---|---|---|
| *bɔ́ɔt* | *ben* | *sòŋ* | *maa* | *wâa* | *“thʉ̌ŋ* |
| boss | Ben | send | come | that | arrive |
| *bâan* | *rʉ̌ʉyaŋ* | *khráp* | | | |
| home | Q | PTCL | | | |
| *phǒm* | *hěn* | *fǒn* | *tòk* | | |
| I | look | rain | fall | | |
| *thâa* | *tàak* | | *fǒn* | *rîip* | |
| if | encounter | | rain | hurry | |
| *ʔàapnám* | | *sàʔ* | *phǒm* | *náʔ khráp* | |
| take shower | | wash | hair | PTCL | |
| *dǐaw* | *mây* | *sabaay”* | *bɔ́ɔt* | *ben* | |
| soon | NEG | fine | boss | Ben | |
| *penhùaŋ* | *chǎn* | *dûay* | | | |
| worry | I | too | | | |

(Q: question particle, PTCL: particle, NEG: negator)

“Boss Ben sent me a message “Have you arrived home? I feel it rains. If you encounter the rain, take shower and wash your hair suddenly. If not, you will catch a cold.” Boss Ben worried about me.”

The case (9) is a quoted sentence in which the first or second person pronoun can either be a speaker or an addressee, based on the speaker's intent. In (9), “ผม” (*phǒm*, I) is labeled as the speaker, following the same criteria as other languages in the project.

(10) Dare to Love 9: 9218
คามี่ ๆ เธอ ไม่ได้สนใจเรื่องบอสเบนซะหน่อย

| | | | | |
|---|---|---|---|---|
| *khaamîi* | *khaamîi* | *thəə* | *mây dây* | *sǒncay* |
| Kami | Kami | you | NEG | be interested |
| *rʉ̂aŋ* | *bɔ́ɔt* | *ben* | *sáʔ nɔ̀y* | |
| story | boss | Ben | PTCL | |

(NEG: negator, PTCL: particle)

“Kami, Kami, you are not interested in boss Ben.”

Since the corpus we are considering are TV drama scripts and novels, there are some sentences of soliloquy. Some language specific difficulties can be found in Thai soliloquy. In Thai, it is very common that a personal name is used as a first-person expression, especially by female speakers. Furthermore, “เธอ” (*thəə*) which is used in (10) has a meaning of second-person (you) as well as female third-person (she). Even more, “เธอ” (*thəə*) which does not have the meaning of speaker in general, but “เธอ” (*thəə*) is tagged as the speaker in (10). This follows the same principle as in the quotation sentence of (9). Going through the annotation process, we can observe the very special cases of usage that differ from the original meaning interpretation, as described in the case of (10) and Kinship Term 5 in Table II.

## V. Conclusion

A large number of dialogue sentences have been annotated using a set of criteria. While some sentences may still have room for interpretation, multiple annotators have been assigned to reach a consensus based on the provided context information in the scripts. In fact, some criteria are added after encountering multiple interpretations during annotation. As a result, the pronoun substitutes and address terms can be confirmed through actual conversation. The presence of both pronoun substitutes and address terms can serve as a crucial factor in understanding the speakers' intentions, which can aid in the process of resolving anaphora in language understanding. This is crucial for accurate comprehension of the dialogue, as well as for natural language processing tasks such as machine translation and dialogue systems.

## Acknowledgment

This work was supported by the Thailand Science Research and Innovation Fundamental Fund, Contract Number TUFF19/2564 and TUFF24/2565, and by JSPS KAKENHI Grant Number JP20H01255. We would like to extend our appreciation to Dr. Sorarat Jirabawornvisut, Khunying Vinita Diteeyont (aka V. Vinichaikul), and Tipthida Satthathip for the valuable contributions of their writings for this study. Additionally, we would like to acknowledge the publishers, Amarin Printing and Publishing PCL. and BEC World PCL., who have graciously allowed us to use their works in our research.

## REFERENCES

[1] V. Sornlertlamvanich, et al. "Collaborative Collection of Multilingual Pronoun Substitutes and Address Terms," In *Proc. 7th Int. Conf. on Business and Industrial Research (ICBIR2022)*, Bangkok, Thailand, May 19-20, 2022, pp. 36-40.

[2] S. Wittayapanyanon. "A Review of Studies of Pronoun Substitute and Address Term," In *Southeast Asian Studies Tokyo University of Foreign Studies*, No. 26, Dec 2020, pp.1-23.

[3] R. Quirk, et al., *A Comprehensive Grammar of the English Language*, USA: Longman Group Ltd., 1985, pp. 773.

[4] D. Crystal, *A Dictionary of Linguistics and Phonetics*, 5th edition, Oxford: Blackwell Publishers Ltd., 2003.

[5] A. M. Zwicky, "Hey, What's your name!," In *10th Regional Meeting of the Chicago Linguistic Society*, 1974, pp. 787-801.

[6] D. Biber, et al., *Longman Grammar of Spoken and Written English*, London: Longman, 1999.

[7] V. Sornlertlamvanich, N. Takahashi, and H. Isahara. "Building a Thai Part-Of-Speech Tagged Corpus (ORCHID)," In *The Journal of the Acoustical Society of Japan (E)*, Vol.20, No.3, May 1999, pp 189-198.

[8] S. Jirabawornvisut. *Nakee(2015), Pleang Naka (2017), Phiphop Himmapan (2018)*, BEC World PCL., 2015.

[9] T. Satthathip and S. Panyaopas. *Dare to Love*, BEC World PCL., 2021.

[10] N. Vejjajiva. *The Happiness of Kati*, Amarin Printing and Publishing PCL., 2019.

[11] V. Diteeyont. *Tae Pang Korn (2005), Nam Sai Jai Jing (1994), Pha Tong (2020)*, Amarin Printing and Publishing PCL., 1994.

[12] Royal Society of Thailand, *The Royal Institute Dictionary*, Bangkok: Royal Society of Thailand, 2011.