

Collaborative Collection of Multilingual Pronoun Substitutes and Address Terms

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

Hiroki Nomoto
Department of Language and Culture
Studies
Tokyo University of Foreign Studies
Tokyo, Japan
ORCID: 0000-0003-3982-4368

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

Atsushi Kasuga
Department of Asian Languages
Kanda University of International
Studies
Chiba, Japan
kasugaat@kanda.kuis.ac.jp

Kenji Okano
Institute of Global Studies
Tokyo University of Foreign Studies
Tokyo, Japan
ORCID: 0000-0001-8494-9771

Wataru Okubo
Graduate School of Global Studies
Tokyo University of Foreign Studies
Tokyo, Japan
ORCID: 0000-0001-9320-3787

Yunjin Nam
Institute of Global Studies
Tokyo University of Foreign Studies
Tokyo, Japan
namyj@tufs.ac.jp

Yoshimi Miyake
Department of Resource Policy and
Management
Akita University
Akita City, Japan
miyake@ed.akita-u.ac.jp

Thuzar Hlaing
World Language and Society Education
Center
Tokyo University of Foreign Studies
Tokyo, Japan
thuzarhlaing@tufs.ac.jp

Ryuko Taniguchi
Institute of Japan Studies
Tokyo University of Foreign Studies
Tokyo, Japan
ryukota@tufs.ac.jp

Sri Budi Lestari
Center for Language Education
Ritsumeikan Asia Pacific University
Oita, Japan
tari0828@apu.ac.jp

Abstract—This paper describes the encoding scheme for pronoun substitutes and address terms in eight Asian languages based on the vocative studies. The target languages are selected according to the availability of the language experts and resources. The nature of pronoun substitutes and address terms expression across the languages can be confirmed by the concepts defined in the WordNet. In this study, a workbench for text data collection (WordList) has been carefully designed to maintain the input data consistency and the semantic linkage between the target languages. The WordList is a web-based application facilitating an online collaborative data input. It maintains the data in MongoDB, and supports JSON and CSV format file exporting for database backup and batch data cleansing for further expression pattern study.

Keywords—pronoun substitutes, address terms, vocatives, workbench, Asian languages

I. INTRODUCTION

Wikipedia is one of the great successful collaboration in user content creation for a huge text collection. It contains a large amount of user generated contents. It significantly gains a high potential in content creation and maintaining its consistency. In the same time, many language processing models can be efficiently generated from a large amount of data due to the advancement in the recent machine learning research. As a result, the Wikipedia text data becomes one of the important text resources for training the language models to use in various kinds of language processing tasks i.e. machine translation, summarization, question generation, named entity recognition and so on. It is also applicable for

transfer learning to save the language core training efforts and to apply to low-resource languages. Online and collaborative editing features of Wikipedia are the two main concerns in preparing a workbench for our multilingual pronoun substitutes and address terms study. Pronoun substitutes and address terms are actually a kind of noun or noun phrase that functions as a pronoun which indicates a noun in a conversation. The expression becomes redundant when it is used for calling for attention, such as “Thank you, John”, or “Mom, I am hungry.” The sentence structure becomes more complicated, especially for most of the Asian languages which are non-segmented and rich of honorific expression. To understand the characteristics of the expression, the pronoun substitutes and address terms are collected by language and connected via WordNet [1] synset ID to capture their interchangeability among languages. Therefore, the contributors of each language have no need to be concerned about the consistency with other languages. The remarkable features of pronoun substitutes and address terms are listed up in a hierarchical manner [2].

The target languages in this study are Burmese, Indonesian, Japanese, Javanese, Korean, Malaysian, Thai, and Vietnamese. Organizing the text data collection of low resource languages is not trivial. There are difficulties in managing the interfacing and supporting sufficient information for collaborators to consider just only their language in concern. WordList workbench is currently designed to provide at least English and Japanese menu since they are both common languages among the current contributors.

WordNet synset ID is introduced to be a semantic identifier for connecting terms from different languages. Since pronoun is the main word category we are focusing on, the Extended Open Multilingual WordNet (OMW) [3] is introduced to fulfill the absence of pronoun and multilingual features of the original WordNet.

Pronoun substitutes and address terms are a part of vocative study by limiting the sophistication in determining the form of expression. To be able to cover the necessary information, the encoding scheme is grouped up into four categories, i.e. morphological, syntactic, semantic and pragmatic information. The provided sample sentences are crucial for making the interpretation clear and learning the word context.

WordList workbench exports the data in both JSON and CSV formats for batch data cleansing and further analysis. Database backup is also conducted regularly in the formats of JSON and CSV.

The remainder of the paper is organized as follows. Section II discusses some related works in the development of collecting text data via Internet collaboration. Section III explains the design of encoding scheme for pronoun substitutes and address terms collection focusing on the necessary information to establish the forms for multilingual text data analysis. Section IV elaborates the development of WordList workbench and reports the past one year of text data collection.

II. LITERATURE REVIEW

Various kinds of editor for corpus and dictionary development are available in both opensource and commercial software. To support a wide collaboration between collaborators in both online and offline manner, a web application frontend framework with database and authentication backend is commonly implemented.

LEXiTRON is a corpus based dictionary that collects sample sentences of a word to describe its syntactic context and semantic interpretation by its usages. A word in LEXiTRON can be comfortably identified by its surrounding context that becomes one of the main streams today for generating word embeddings in n-gram, skip-gram or bag-of-word in word to vector, and sentence piece algorithms. The concept has been extended to build a dictionary network for Asian languages [4].

WNMS (WordNet Management System) is developed as a platform for editing the result of WordNet conversion to target languages in Asian WordNet (AWN) development initiative [5]. In addition to the basic functions of editing and reporting, WNMS provides voting function to collect the supporting score from collaborators in an open collaboration manner. Synset ID assignment is the main issue in aligning the meaning across languages. Many efforts have been spent in the synset ID assignment initially from automatic assignment by similarity calculation between the synset and existing bilingual dictionaries. WNMS provides an editing function to keep the consistency of the linkage through its web service API [6].

In the study of pronoun substitutes and address terms, we prepare a collaborative web application for collecting the possible forms of occurrence. The template for collecting the crucial word information have been designed in the encoding scheme based on the vocative studies. It is reported that the

vocative form can be divided into five distinct classes: unbound pronouns, names, kinships terms, titles and descriptors [7]. Also, vocatives can indicate the nature of relationships between people of primary importance whether the terms are used reciprocally or non-reciprocally. The former indicates equality and are common within a status group like, children, students, and fellow workers. The latter, on the other hand, indicates an imbalance in power or prestige; an example of this is teacher-student relationship [8]. The encoding scheme in this study is prepared in four groups of information as described in Section III.

To handle the multilingual issue, we introduce WordNet synset ID which can be retrieved from the open API of OWM and WordNet database. The WordNet synset ID provides the synset and description for collaborators to assign the semantic information to the target word. By this, it enables the semantic linkage between the languages.

III. ENCODING SCHEME

It is not easy to find a consensus concerning the expression of pronoun substitutes and address terms. The expression heavily depends on pragmatic circumstances or the way of how people act to each other. In the study of vocative in English, it can be a *noun phrase* denoting to the one or more persons to whom it is addressed [9]. English does not make use of the vocative case inflectionally, but expresses the notion by using an optional noun phrase with a distinctive *intonation* [10]. The vocatives are said to express attitude, politeness, formality, status, intimacy, or role relationship, and most of them mark the speaker *characterizing* him or her to the addressee [11].

Regarding the position of English vocatives, there is no absolute construction but can occur freely in the following possible positions [12].

1. Initial position, where the vocative precedes the clause of the utterance in the closest position, for example, "*Ladies and gentlemen*, and let me ask you something."
2. Medial position, where the vocative occurs in the clause of the utterance, for example, "How are we doing, *folks*, and with the scales"
3. Final position, where the vocative follows the clause of the utterance in the closest position, for example, "That's a heavy load, *girl*."

In the study of some Asian languages, many efforts have been made to understand the language specific features which are not in the scope of antecedent standards for English and European languages. A common standard for Asian language resources which is compatible with the international standard (MILE) is proposed in a framework of lexical entries and upper layer ontology. Inflection, classifier, orthographic variants, reduplication, and affixation in noun phrase construction which are the language specific features observed in the Asian languages, are discussed in the study [13]. It is also reported that the terms of pronoun substitutes and address terms are derived not only from nouns but also demonstrative and locative pronouns [2].

To collect an initial set of pronoun substitutes and address terms for Asian language study, a scheme to encode the necessary information has been proposed by grouping into the following set of information. The part of morphological

information covers the main phrase structure. The part of syntactic information is to denote the grammatical role types. The part of semantic information is to provide a linkage to WordNet for connecting the terms in the target languages. The part of pragmatic information is to characterize the nature of relationships between the speaker and the addressee. Followings are the groups of information to be a guideline for data collection.

- Morphological information
 - Expression
 - Phrase structure
- Syntactic information
 - Formal features
 - Title
 - Pronoun substitute (speaker)
 - Pronoun substitute (addressee)
 - Address term
- Semantic information
 - WordNet ID
- Pragmatic information
 - Social prestige

1. “ขอบคุณพ่อसानแล้วก็น้องปูนมากนะคะ” ขอบคุณ (pronoun_substitute_addressee (title พ่อ) (proper_noun สาน)) แล้วก็น้อง (pronoun_substitute_addressee (title น้อง) (proper_noun ปูน)) มากนะคะ (Thai PBS “บริษัทพ่อบ้าน” ตอนที่ 1 สาเหตุแห่งหนึ่ง) [lit. “Thank you very much, father and brother Pun, very much.” Thank you very much (pronoun_substitute_addressee (title father) (proper_noun San)) and (pronoun_substitute_addressee (title brother) (proper_noun Pun)) very much.]
2. “ศิษย์น้องสองคน! ตรวจสอบดู” (address_term (common_noun ศิษย์) (common_noun น้อง) (number สอง) (classifier คน)) ! ตรวจสอบดู (Thai PBS “บริษัทพ่อบ้าน” ตอนที่ 3 หนี้ที่ไม่ใช่แค่เรื่องเงิน) [lit. “Two younger disciples! Check.” (address_term (common_noun disciple) (common_noun younger) (number two) (person classifier)) !_check]
3. “ขอโทษ หมอ เสียใจ” ขอโทษ (pronoun_substitute_speaker (common_noun หมอ)) เสียใจ (หมยันต์ 2015. (พิมพ์รวมเล่มครั้งที่สิบเอ็ด) “คูกรม ๒” pp.39 ณ บ้านวรรณกรรม) [lit. “Sorry, doctor has to say sorry.” Sorry_(pronoun_substitute_speaker (common_noun doctor))_has to say sorry]
4. “มากันครบหรือยัง เด็กๆ” มากันครบหรือยัง (address_term (common_noun เด็กๆ)) (ดร. สรรตน์ จิรวรรณวิสุทธิ์ บทละครโทรทัศน์ “นาคี” ตอนที่ 1) [lit. “Have you all come, children?” Have you all come_(address_term (common_noun children))]

Fig. 1. Example of labelled original Thai text with English literal translation

Example sentences are provided to show the position of pronoun substitutes and address terms in an utterance as shown in Fig. 1.

The target languages in this study are Burmese, Indonesian, Japanese, Javanese, Korean, Malaysian, Thai, and Vietnamese. The expression of pronoun substitutes and address terms are collaboratively collected online according to the designed encoding scheme as shown in Fig. 2.

English and Japanese translations of the expression and sample sentences are preferred for better access across the

languages since finally the analysis of pronoun substitutes is to be conducted to understand its mutual expression.

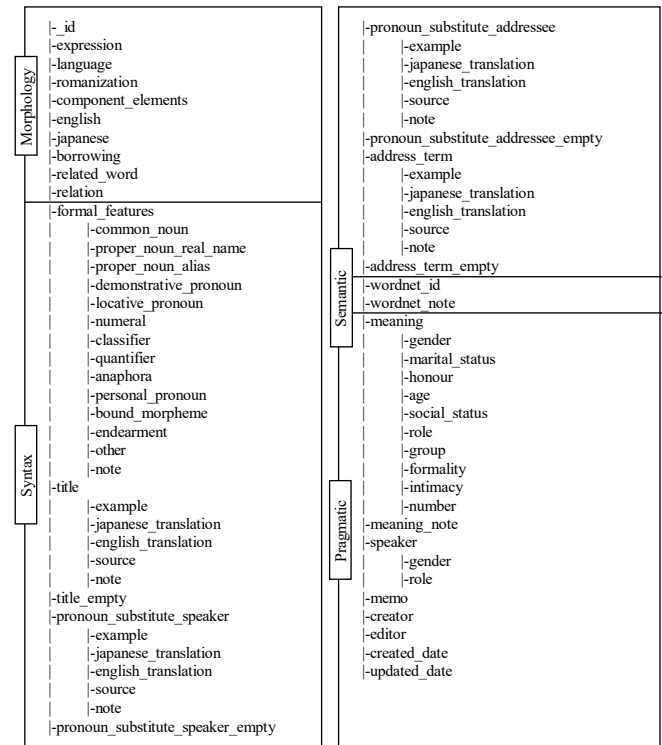


Fig. 2. Pronoun substitutes encoding scheme

IV. WORDLIST COMPLETION

WordList is a workbench designed for online text data input for the restricted contribution member only. It is developed as a web application for online collaboration in adding and editing the word list. The collaborator is registered by an authentication system to avoid the undesired operations and in the same time, to allow the traceability of the editing.

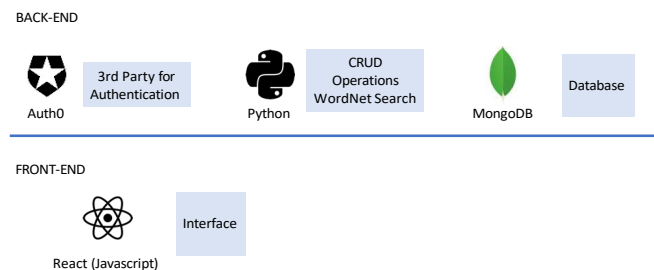


Fig. 3. WordList web application architecture

WordList is implemented on the opensource solution composed of MongoDB as the main database system, FastAPI for API establishment to external resources, and React as the web frontend framework, as shown in Fig. 3.

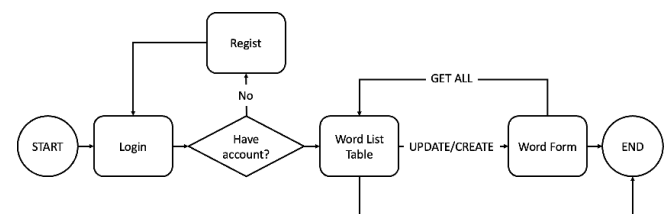
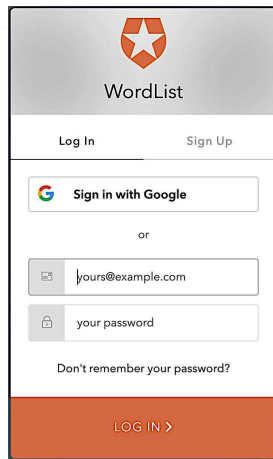


Fig. 4. Process flow in WordList

The total flow of the process is shown in Fig. 4. The collaborator gets started by acquiring an account through the registration process and can login after obtaining an account. The top page of summary is shown and allow the collaborators to edit or create a new entry with a pre-defined form. The entry is updated and stored only after receiving the submission command.

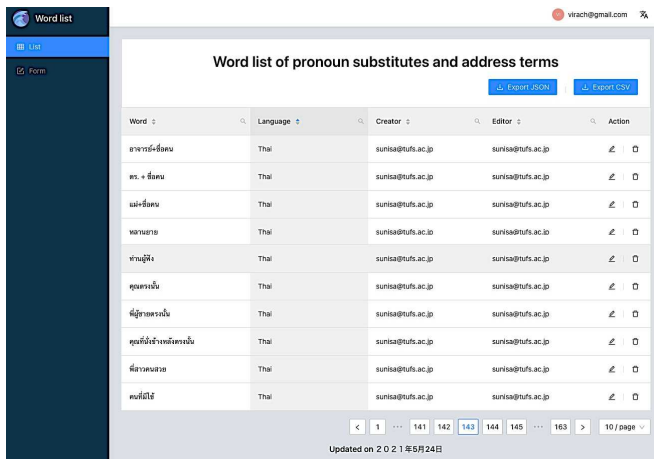


The login page for WordList features a header with the WordList logo. Below the logo are two tabs: 'Log In' and 'Sign Up'. A 'Sign in with Google' button is present, followed by an 'or' separator. Below this is a form with a text input field for an email address (placeholder: 'yours@example.com') and a password input field (placeholder: 'your password'). A link 'Don't remember your password?' is located below the password field. At the bottom is a large orange button labeled 'LOG IN >'.

Fig. 5. Login page for WordList

New collaborator needs to register in 'Sign Up' tab to get an account. The registered collaborator gets started via the authentication process in 'Log In' tab as shown in Fig. 5.

The input templates are prepared according to the encoding scheme. The top page, shown in Fig. 6, displays the summary of the input entry list which can be retrieved and sorted by entry, language, creator, and editor. Once an entry has been created it can be edited and deleted by the collaborators. The menu is currently interchangeable between English and Japanese. JSON and CSV exporters are prepared to download the total dataset.



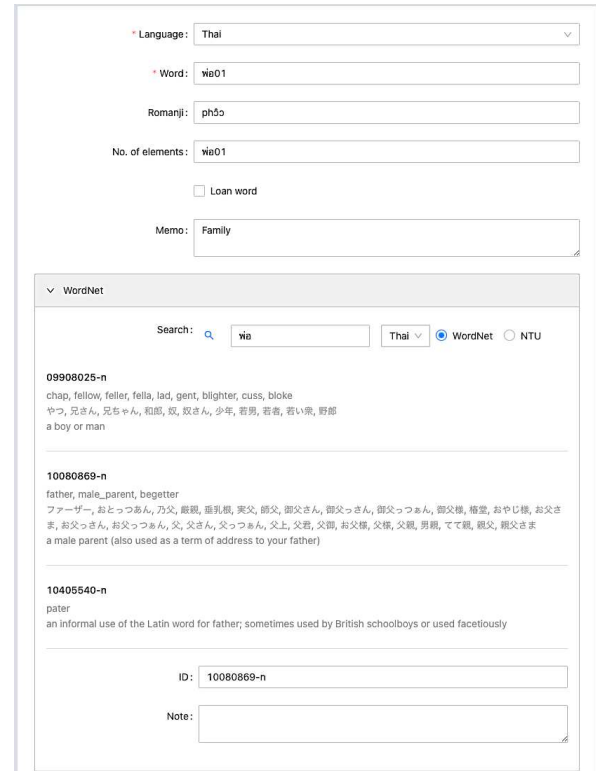
The screenshot shows the 'Word list' interface. It has a sidebar with 'List' and 'Form' tabs. The main area is titled 'Word list of pronoun substitutes and address terms'. It includes buttons for 'Export JSON' and 'Export CSV'. Below is a table with columns: Word, Language, Creator, Editor, and Action. The table lists several entries in Thai. At the bottom, there are pagination controls showing '1' to '163' and '10 / page'. The page is updated on 2021年5月24日.

Fig. 6. Screenshot of the WordList top page

WordNet linkage is an API call to consult WordNet for general terms, and Open Multilingual WordNet (OMW) for additional terms, especially pronouns since the original WordNet contains only noun, verb, adverb and adjective.

The current task of collecting the terms of pronoun substitutes needs a semantic identifier (ID) to define the common concept to share among the languages. Practically, Fig. 7 shows the options of WordNet and NTU for OMW to select the IDs. The ID is the key to link the terms semantically

among the languages. Collaborators can look up for the ID by inputting an entry, selecting the language and fixing the option to be WordNet or NTU. Then, the result list of synset ID and its synset in both English and Japanese appears for collaborators to click select the proper ID for the target word.



The interface shows a search form. It includes a 'Language' dropdown set to 'Thai', a 'Word' input field with 'wie01', a 'Romanji' input field with 'ph5o', and a 'No. of elements' input field with 'wie01'. There is a 'Loan word' checkbox and a 'Memo' text area with 'Family'. Below the form is a section titled 'WordNet' with a search bar containing 'wie'. It has radio buttons for 'Thai', 'WordNet' (selected), and 'NTU'. The results show synset IDs and their corresponding English and Japanese descriptions. For example, '09908025-n' is described as 'chop, fellow, feller, fella, lad, gent, blighter, cuss, bloke' in English and 'やつ, 兄さん, 兄ちゃん, 和郎, 奴, 奴さん, 少年, 若男, 若者, 若い衆, 野郎' in Japanese. Another example is '10080869-n' described as 'father, male_parent, begetter' in English and 'ファーザー, おとつあん, 乃父, 嚴親, 実父, 御父さん, 御父っさん, 御父っつあん, 御父様, おやじ様, お父ま, お父っさん, お父っつあん, 父, 父さん, 父っつあん, 父上, 父君, 父間, お父様, 父親, 父親, 男親, てて親, 親父, 親父さま' in Japanese. A third example is '10405540-n' described as 'pater' in English and 'an informal use of the Latin word for father; sometimes used by British schoolboys or used facetiously' in Japanese. At the bottom, there is an 'ID' input field with '10080869-n' and a 'Note' text area.

Fig. 7. WordNet and Open Multilingual WordNet API interfacing

Since December 2020 there are 16 contributors for the eight target languages. 1,538 entries have been created attaching to 168 unique WordNet synset ID. As of December 2021, the distribution of entries by languages is shown in Table I.

TABLE I. DISTRIBUTION OF ENTRY BY LANGUAGE

Language	Number of Entry	Number of unique WordNet ID
Burmese (Bur)	384	71
Japanese (Jpn)	354	13
Thai (Tha)	238	68
Malaysian (Mal)	232	57
Korean (Kor)	155	62
Vietnamese (Vie)	122	68
Indonesian (Ind)	85	2
Javanese (Jav)	58	1

The number of assigned IDs mainly vary by the availability of words in WordNet and OMW which reflects the readiness of WordNet development for the language.

TABLE II. DISTRIBUTION OF TYPE BY LANGUAGE

	Bur	Ind	Jpn	Jav	Kor	Mal	Tha	Vie	Total
Kinship	23	2	4	0	27	14	21	19	110
Title	35	0	7	0	15	30	39	31	157
Endearment	2	0	4	0	1	3	1	0	11
Connotation	22	0	2	0	21	8	5	18	76

Table II shows the possible types of pronoun substitutes and address terms collected in each language. Kinship and title show their high variation in using in the expressions.

The distribution of part of speech (POS) shown in Table III is also our interest in capturing the syntactic constituent of the expression. Common noun shows its widely used to express the pronoun substitutes and address terms.

TABLE III. DISTRIBUTION OF POS BY LANGUAGE

	Bur	Ind	Jpn	Jav	Kor	Mal	Tha	Vie	Total
Name	11	12	61	2	10	33	6	0	135
Personal Pron.	10	2	47	6	3	17	37	36	158
Common Noun	74	58	204	23	143	134	152	87	875
Demonstrative Pron.	19	4	19	0	0	17	12	5	76

It is noted that number of entry shown in Table I to III do not describe the frequency of the usage in each language. Instead, we conduct the survey to ensure the coverage of the collection.

Table IV shows the top ten assigned IDs to describe the variation of the meaning of terms used in pronoun substitutes and address terms. Interestingly, except the 'driver.n.01' all are the kinship terms used in the sample sentences. Especially, the languages in Asia have a trend in using kinship terms to express their closer relationship in the conversation though it is a pseudo-kinship relation. The speaker and addressee play different functions around the kinship relation.

TABLE IV. FREQUENTLY USED WORDNET SYNSET ID IN THE SAMPLE SENTENCES

Frequency	ID	Synset
23	10736091-n	uncle.n.01
18	09823502-n	aunt.n.01
17	10142391-n	grandfather.n.01
17	10780632-n	wife.n.01
13	10048218-n	elder.n.01
13	10193967-n	husband.n.01
12	10142747-n	grandma.n.01
12	09918248-n	child.n.01
11	10624915-n	son-in-law.n.01
10	10034906-n	driver.n.01

V. CONCLUSION

For the early stage of pronoun substitute study, the encoding scheme has been proposed under the consideration of providing sufficient information for understanding the expression in the eight target Asian languages. In the current stage of collection, the WordList successfully achieves in text collection by the designed encoding scheme, and provides the semantic linkage by consulting the WordNet and OMW API for synset ID assignment. We found that the kinship terms for pronoun substitute expression are the topmost reference terms among the collaborators. In the future work, the collected entry list is to be an initial source for pronoun substitute and

address term labeling in general text corpus. The possible patterns of expression will be then extracted to create a model for understanding the speaker's intention. The synset ID from WordNet will be used to facilitate the cross language study in terms of their interchangeability of the expression of pronoun substitutes and address terms.

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.

REFERENCES

- [1] C. Fellbaum, ed. *WordNet: An Electronic Lexical Database*, MIT Press, 1998.
- [2] H. Nomoto, S. Wittayapanyanon (Saito), K. Okano, T. Hlaing, Y. Nam and S. B. 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]," in *Gogaku Kenkyuujo Ronshuu* 25, Tokyo University of Foreign Studies, 2021, pp. 63-78.
- [3] F. Bond and R. Foster, "Linking and extending an open multilingual wordnet." In *Proc. 51st Annu. Meeting of ACL*, Sofia, 2013, pp. 1352-1362.
- [4] T. Charoenporn, H. Isahara, and V. Sornlertlamvanich, "Construction of Dictionary Network for Asian Languages," In *Proc. 9th Annu. Meeting of The ANLP*, Mar. 2003, pp. 262-265.
- [5] V. Sornlertlamvanich, T. Charoenporn, and H. Isahara, "Language Resource Management System for Asian WordNet Collaboration and Its Web Service Application," In *Proc. 7th Int. Conf. on LREC*, Mediterranean Conference Center (MCC), Malta, May 17-23, 2010.
- [6] K. Robkop, S. Thoongsup, T. Charoenporn, V. Sornlertlamvanich and H. Isahara, "WNMS: Connecting the Distributed WordNet in the Case of Asian WordNet," In *Proc. 5th Int. Conf. of GWC*, Mumbai, India, Jan. 31-Feb. 4, 2010.
- [7] S. Gramley and K. Pätzold, "A Survey of Modern English," *London: Routledge*, 1992, pp. 289.
- [8] Z. M. Griffin, "Chapter 9 - Retrieving Personal Names, Referring Expressions, and Terms of Address," In *Psychology of Learning and Motivation*, B. H. Ross, eds., Academic Press, Vol. 53, 2010, pp. 345-387.
- [9] R. Quirk, et al., *A Comprehensive Grammar of the English Language*, USA: Longman Group Ltd., 1985, pp. 773.
- [10] D. Crystal, *A Dictionary of Linguistics and Phonetics*, 5th edition, Oxford: Blackwell Publishers Ltd., 2003.
- [11] A. M. Zwicky, "Hey, What's your name!," In *10th Regional Meeting of the Chicago Linguistic Society*, 1974, pp. 787-801.
- [12] D. Biber, et al., *Longman Grammar of Spoken and Written English*, London: Longman, 1999.
- [13] T. Takenobu, V. Sornlertlamvanich, T. Charoenporn, N. Calzolari, M. Monachini, C. Soria, C. R. Huang, Y. J. Xia, H. Yu, L. Prevot and K. Shirai, "Infrastructure for standardization of Asian language resources," In *Proc. of COLING/ACL*, Sydney, Australia, Jul. 15-23, 2006.