Interpersonal meaning annotation for Asian language corpora: The case of TUFS Asian Language Parallel Corpus (TALPCo)

Hiroki Nomoto Kenji Okano Sunisa Wittayapanyanon Junta Nomura
Tokyo University of Foreign Studies
{nomoto, okanok, sunisa, nomura.junta.q0}@tufs.ac.jp

Abstract

Natural conversations in many Asian languages require far more interpersonal meaning information than similar conversations in European languages. This paper demonstrates this point based on examples from the TUFS Asian Language Parallel Corpus and describes how we annotated the corpus with interpersonal meaning information.

1 Introduction

Interpersonal meanings are concerned with the relationship of the speaker with his/her interlocutor and the person(s) s/he refers to in a sentence. It is conventionalized and hence semantic in nature rather than pragmatic, and it belongs to the meaning dimension called expressive/use-conditional meaning, which is distinct from the descriptive/truth-conditional meaning.[1] Interpersonal meanings are manifest in pronouns, pronoun substitutes and address terms. For example, in (1), the speaker chooses the pronoun *anata* 'you' to refer to the addressee. The number in brackets after the free translation indicates the sentence's ID in the TUFS Asian Language Parallel Corpus (TALPCo) (Nomoto et al. 2018).[2]

(1) Japanese (pronoun)

その コップは あなたのです。
sono koppu-wa anata-no-desu.
that cup-TOP you-GEN-COP.POL
'That cup is yours.' [3289]

The interpersonal meaning eoncoded by *anata* constrains the possible situations in which its use is appropriate. Thus, (1) is appropriate when used, say, by a customer with his/her salesperson, but it sounds impolite when used in the opposite direction, even when it conveys true information. In the latter situation, role names are used as a pronoun substitute as in (2).[3]

(2) Japanese (pronoun substitute)

その コップは お客様のです。
sono koppu-wa okyakusama-no-desu.
that cup-TOP Mr./Ms.customer-GEN-COP.POL
'That cup is yours.'

In European languages, including English, interpersonal meanings concerning the speaker and addressee are almost entirely expressed by address terms (e.g. ***Miss**, this is my mother.* [2744]), with only second person pronouns having a polite-familiar distinction (e.g. *vous* vs. *tu* in French), if any, and lacking pronoun substitutes altogether. Many Asian languages have far more complex systems involving multiple pronouns and pronoun substitutes.[4] First and second person references are not a simple matter of 'I' and 'you'. Natural conversations in these languages require far more interpersonal meaning information than similar conversations in European languages.

Given the importance and complexity of interpersonal meanings in Asian langauges, it is useful to annotate them in corpora of Asian languages. We thus decided to annotate TALPCo with interpersonal meaning information. The present paper describes our annotation system.

This paper is organized as follows. Section 2 gives a brief overview of TALPCo and its recent developments. Section 3 demonstrates why interpersonal meanings should be annotated in the corpus. Section 4 describes the annotation scheme and the features we used for annotation. Lastly, section 5 discusses some implications of this study for practical applications.

2 Recent developments of TALPCo

The development of the TUFS Asian Language Parallel Corpus or TALPCo started in 2018 as a parallel corpus consisting of five languages: Japanese, Burmese (Myanmar), Malay, Indonesian and English. TALPCo is modelled after the Asian Language Treebank (ALT) Parallel Corpus (Riza et al. 2016), which is the first openly available paralell corpus involving multiple Asian languages.[5]

[1] See Kroeger (2018:Ch. 11) for a textbook introduction to use-conditional meanings.

[2] https://github.com/matbahasa/TALPCo

[3] Non-standard abbreviations not available in the Leipzig Glossing Rules: PART: particle; POL: polite.

[4] Chinese and Phillipine-type langauges are major exceptions.

[5] http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/index.html

Table 1: TALPCo and ALT

	TALPCo	ALT
Source language	Japanese	English
Translations	Burmese, Malay, Indonesian, Thai, Vietnamese, English	Burmese, Malay, Indonesian, Thai, Vietnamese, Japanese, Bengali, Filipino, Khmer, Lao
Register	formal, conversational; (quasi-)spoken	formal, journalistic (*Wikinews*); written
Size	1,372 sentences	20,106 sentences
Expert check	yes	no

TALPCo supplements ALT in several respects. Table 1 summarizes major differences between the two.

Currently, we are preparing to add two more languages to TALPCo, namely Thai and Vietnamese. At the time of writing this paper (January 2019), the third author of the present paper is translating the Japanese sentences into Thai. Her translation will be checked by an undergraduate Thai major student at Tokyo University of Foreign Studies (TUFS). The checked sentences will then be tokenized by a Thai tokenizers and checked by the same student.

As for Vietnamese, the translation was prepared by a native Vietnamese speaker who is a lecturer at the University of Da Nang and is currently conducting her graduate study at TUFS. The translation was checked by the fourth author of the present paper. The sentences were tokenized using the `word_tokenize` function of the Undersea - Vietnamese NLP Project[6] and then checked by the fourth author. *Từ Điển Tiếng Việt* (Hoàng 2003) was consulted when it was not immediately obvious whether a word sequence constituted a multiword expression.

3 The importance of interpersonal meaning information

We faced two major problems when translating Japanese into Thai and Vietnamese. Both problems arose due to the lack of necessary contextual information, in particular, information concerning interpersonal meanings in the Japanese sentences.

Two reasons exist for this lack of contextual information. First, each sentence in TALPCo is presented by itself rather than as part of a conversation, as in (3a). In formal conversations in Thai, sentences sound unnatural without a final particle indicating the speaker's gender: *khráp* for males and *khâ* for females. Using a wrong final particle will make the sentence infelicitous. Sentence (3a) alone provides no information about the gender of the speaker. To translate (3a) into the Thai sentence in (3b), which uses the particle *khráp*, one needs to know that the speaker is male.

(3) a. きのう わたしは 勉強しました。
kinoo watasi-wa benkyoosi-masi-ta
yesterday I-TOP study-POL-PST

b. เมื่อวาน **ผม** เรียน หนังสือ **ครับ**
mûawaan **phŏm** rian náŋsŭɯ **khráp**
yesterday I learn book PART
'I studied yesterday.' [1356]

Moreover, Thai has no gender-neutral first person pronoun that can be used in formal conversations like *watasi* 'I' in Japanese. Speakers choose an adequate pronoun or pronoun substitute (= kin terms, nicknames, roles) mainly based on their own gender, the addressee's age and the situation of conversation (Wittayapanyanon 2018). The pronoun *phŏm* in (3b) is used by males when talking to superiors or equals in a formal setting. Such detailed contextual information cannot be obtained from (3b) alone.

The second reason for the lack of sufficient contextual information is that for certain concepts, target languages make a finer distinction than Japanese. In other words, a Japanese word corresponds to more than one word in the target language. The most notable example involves the title *san*. While *san* in Japanese can be gender-neutral, all target languages in TALPCo but Thai lack a gender-neutral title corresponding to *san*. For example, Vietnamese has a number of equivalents of *san*, differing in the individual's gender and age difference relative to the speaker. The titles *chị* and *anh* in (4b) originate from kin terms meaning 'elder sister' and 'elder brother', respectively, and hence are used for someone in the same age group as the speaker's elder siblings. For those who are in different age groups, different kin terms are used as summarized in Table 2.

(4) a. 木村さんは 学生ですが、
Kimura-san-wa gakusei-desu-ga
Kimura-SAN-TOP student-COP.POL-but
田中さんは 会社員です。
Tanaka-san-wa kaishain-desu
Tanaka-SAN-TOP office.worker-COP.POL

b. ***Chị*** *Kimura là học sinh nhưng* ***anh***
Ms. Kimura COP student but Mr.
Tanaka là nhân viên công ty.
Tanaka COP staff company
'Ms. Kimura is a student, but Mr. Tanaka is an office worker.' [3110]

[6] `http://undertheseanlp.com/`

Table 2: *San* 'Mr./Ms.' in Vietnamese

Age group	'younger sibling'	'elder sibling'	'parent's elder sibling'	'grandparent'
Male	*em*	*anh*	*bác*	*ông*
Female		*chị*		*bà*

In fact, we have encountered the second case before, as Burmese, Malay, Indonesian and English all lack gender-neutral titles like *san*. We decided to follow the choice in the Malay translation, where three kinds of titles are distinguished: *Encik* 'Mr.', *Cik* 'Miss' and *Puan* 'Mrs.'. This strategy worked well but not perfectly. Inconsistencies occured among different languages. Furthermore, it cannot handle a new language that makes a finer distinction than Malay. Vietnamese is one such language (cf. Table 2). Therefore, it is necessary to explicitly specify relevant interpersonal meaning information in the corpus.

4 Interpersonal meaning annotation

4.1 Annotation scheme

We annotated the target language data with the interpersonal meanings. Interpersonal meanings constrain the situations in which a linguistic form can be used appropriately (see section 1). Hence, the annotations restrict the contexts of use to those compatible with the annotations.

Our annotations consist of two levels, namely lexical and contextual. LEXICAL ANNOTATION targets tokens whose Japanese counterparts do not provide sufficient interpersonal meaning information for translation. CONTEXTUAL ANNOTATION is concerned with the speaker and the addressee, and hence it targets the whole sentence.

(5b) gives an example of two kinds of annotations for the Vietnamese sentence in (5a). In this example, the addressee information is revealed by the address term *thầy*, which is used for male teachers. The Japanese word aligned with it, namely *sensei*, is gender-neutral. The lexical annotation includes only the missing meaning, namely male. The pronoun substitute *em* can be used for younger addressees in general. However, given that the addressee is a teacher, it is limited to a student here.

(5) a. ***Thầy*** *ơi, đây là mẹ của* ***em***.
male. teacher VOC this COP mother POSS younger. sibling
'Sir, this is my mother.' [2744]

b. [J] 先生、こちらが 私の母です。
[V] **Thầy** ơi, đây là mẹ của **em** .
Lex. male student
Con. Spkr: student
Addr: male, teacher

Because Japanese is a radical *pro* drop language, it is often the case that the contextual information needs to be more specific in the target language, as in the Malay example in (6a). Notice that the second person pronoun *awak* is not aligned with any word in Japanese in (6b).

(6) a. ***Awak*** *nak beli kasut yang mana?*
you be.going.to buy shoes which
'Which shoes are you going to buy?' [1286]

b. [J] どの 靴 を 買い ます か。
[M] **Awak** nak beli kasut yang mana ?
Lex. equal_or_lower
Con. Spkr: ___
Addr: equal_or_lower

4.2 Feature set

Figure 1 summarizes the annotation features we employed. The features are hierarchically organized. Lower features add further specifications to higher features. Thus, sibling entails elder. In our annotation scheme, we will spell out the full specification as in sibling.elder. The nodes in normal font represent the categories to which the features in sanserif font belong. The number features do not pertain to interpersonal meanings, but were included in the annotation because they facilitate translation. When the specific features are irrelevant in relation to the corresponding Japanese expression (e.g. (5b)) or within the target language's lexical system, a category or subordinate feature is left unspecified. Some examples of annotations are given in (7). See also (5b) and (6b).

(7) a. さん 'Mr., Ms.'
[V] **anh** male, elder.sibling, hon; **bạn** friend
[B] **မစ္စတာ** male, hon, foreigner
[I] **Pak** male, mature, hon

b. わたし 'I'
[B] **ကျွန်တော်** male,
Spkr: male, Addr: equal_or_lower
[T] **ผม** male,
Spkr: male, Addr: equal_or_higher

c. ます (POL)
[B] **ခင်ဗျာ** Spkr: male; **ရှင်** Spkr: female
[T] **ครับ** Spkr: male; **ค่ะ** Spkr: female

d. 先生 'teacher'
[B] **ဆရာ** male; **ဆရာမ** female
[M] **cikgu** school; **pensyarah** university

It must be noted that the features given in Figure 1 are sufficient only for annotating TALPCo. More features are necessary to handle more diverse and larger data. Note also that although we present our feature system as a cross-linguistically applicable one, the precise semantics of some features may differ from language to language.

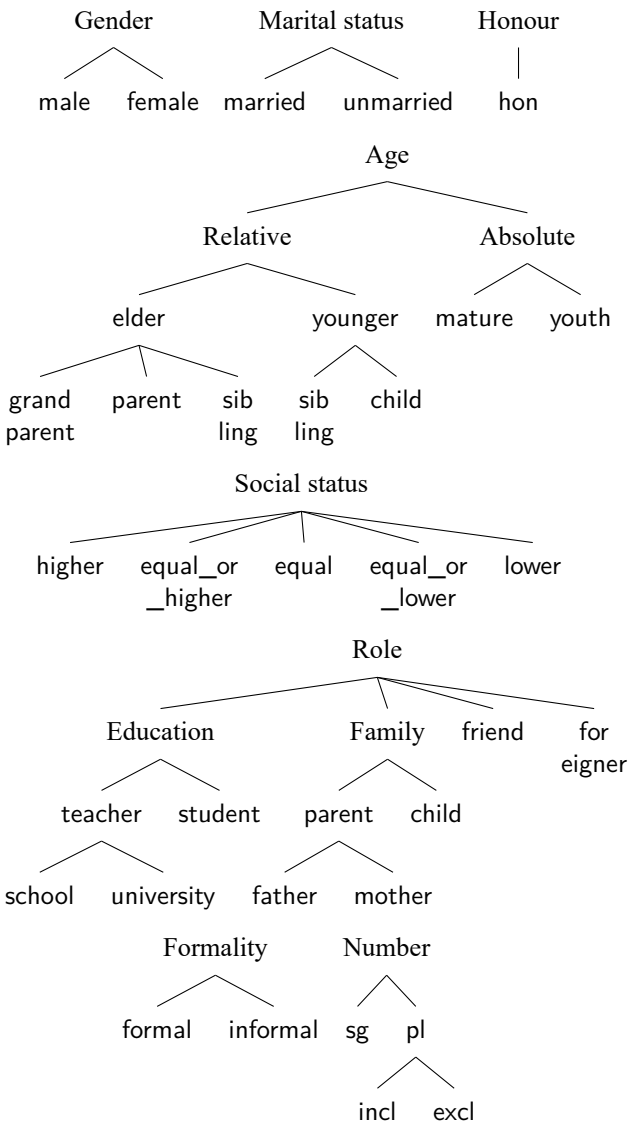

Figure 1: Annotation features

5 Conclusion

Interpersonal meaning information is essential for generating sentences that are not just accurate descriptively but also appropriate in the context of use. Thus, annotations like ours will help to develop better conversation applications in areas such as language education, machine translation (e.g. VoiceTra; Matsuda et al. 2013) and multilingual dialogue systems. Non-linguistic factors, such as people's voice and appearance, are also helpful in identifying interpersonal meanings and should be used alongside linguistic ones.

Conversely, understanding interpersonal meanings in language enables one to know more than the language itself. The choice from different linguistic forms reflects the speaker's official presentation of how s/he wants others to know what s/he thinks about how s/he sees himself/herself and others. The speaker's choice is not always predictable from objective facts; rather, it may involve the speaker's views and attitudes. For example, Vietnamese normally uses kin terms such as *bác* 'uncle, aunt' and *cháu* 'nephew, niece, grandchild' for first and second person reference. According to the actual age difference, one calls someone younger than himself/herself *cháu*. However, people sometimes use *bác*, which is supposed to be used for older addresses, for those who are obviously younger than them to show respect to the addressee (Shimizu 2011:135).

In the future, we would like to expand our corpus so that it can accommodate more varied interpersonal meanings and the expressions encoding them. Furthermore, an in-depth cross-linguistic study of interpersonal expressions is indispensable for improving the feature system in Figure 1.

Acknowledgements The research reported in this paper was conducted under the JSPS grant "Program for Fostering Globally Talented Researchers" offered to Tokyo University of Foreign Studies for a project entitled "Developing Human Resourses Taking the Lead in Research on Endangered and/or Minority Languages in an International Network."

References

Hoàng, Phê, ed. 2003. *Từ Điển Tiếng Việt*. Đà Nẵng: Nhà Xuất Bản Đà Nẵng.

Kroeger, Paul R. 2018. *Analyzing Meaning: An Introduction to Semantics and Pragmatics*. Berlin: Language Science Press.

Matsuda, Shigeki, Xinhui Hu, Yoshinori Shiga, Hideki Kashioka, Chiori Hori, Keiji Yasuda, Hideo Okuma, Masao Uchiyama, Eiichiro Sumita, Hisashi Kawai, and Satoshi Nakamura. 2013. Multilingual speech-to-speech translation system: VoiceTra. In *2013 IEEE 14th International Conference on Mobile Data Management*, volume 2, 229–233.

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

Riza, Hammam, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. 2016. Introduction of the Asian Language Treebank. In *Oriental COCOSDA*.

Shimizu, Masaaki. 2011. *Betonamugo [Vietnamese]*. Osaka: Osaka University Press.

Wittayapanyanon, Sunisa. 2018. Taigo deno ichininshouhyougen no shiyoujittai to taigokyouiku eno katsuyou [A utilization study of first person expressions in Thai and its application to Thai language education]. Paper presented at the 22nd meeting of the Japan Association of Foreign Language Education (JAFLE).