ON TRANSLATING 2nd PERSON PRONOUN (ENGLISH AND INDONESIAN) : A CASE STUDY ON BPPT PARALLEL CORPUS

Prihantoro,
Diponegoro University, Middle Java, Indonesia

prihantoro2001@yahoo.com

Abstract

Many linguists have posited a number of distinct types of translation method, but almost all agree on the existence of word-for-word translation, which to some extent must be avoided to generate natural translation in the target language. Some other methods are proposed to obtain natural translation but in this paper, we refer to communicative translation (Newmark, 1988). However, it does not suggest that some other methods such as faithful or word-for-word translation are useless. It is necessary that these methods apply to the whole text, as some items require faithful or word-for-word translation as well. However, these methods must be avoided when translating culture based item, e.g an entity that does not exist in the target language. On one hand, almost all languages make use of pronouns. On the other hand, the social dynamics of pronoun itself requires translators to shift from word-for-word translation method to another method as the culture polarity shift from source to target language. The data in this research is obtained from English-Indonesian Parallel Corpus by Indonesian Agency for the Assessment and Application of Technology (2008), which consists of written and spoken data. The result of this research suggests that some speaker-hearer relations cannot be fully expressed by English pronouns when translated to Indonesian. There are cases when culture based polarity requires pronoun to shift to proper name. Even when pronoun-to-pronoun translation is preserved, the paradigm changes: such as inclusive/exclusiveness which functions as in/out group identity marker.

Keywords: culture specific pronouns, speaker-hearer relation, social dynamics.

1. INTRODUCTION

In 2008, *Badan Pengkajian dan Penerapan Teknologi Indonesia* (BPPT) or Agency for the Assessment and Application of Technology of Indonesia had completed a design for Machine Translation Framework. This framework is expected to automate the translation of English to Indonesian text and vice versa. Even though the method is statistical, it is important to highlight that the data in the training parallel corpus, are manually supported by human translators (which is considered gold standard). Rather than the computational aspect, this paper studies the corpus from the translation aspect, the increasingly important aspect in the study of language comparison. The key aspect of this study is the study of pronoun translation: English and Indonesian with a focus on 2nd person pronoun. The data is obtained from the BPPT parallel corpus, considering that the translation in this stage (see section 3 of this paper) is performed by human translators instead of statistical machine translation.

This paper is aimed at examining the following aspects of translation in the corpus data: shifts in the pronoun translation, the social dynamics reflected by the choice of pronouns of both Indonesian and English.

The first part of this paper introduces the background and the aims. I briefly review the literatures concerning the grammar, semantic and pragmatic dimensions of the use of pronouns. In order to link the literature review to the finding and the discussion, I present the research method in section 3. Section 3 briefly outlined the methodology, concerning the corpus data and procedure of collecting data. Finding and discussion are integrated in section 4 of this paper. The result is synthesized compactly in section 5 that also concludes the finding and discussion in this paper.

2. LITERATURE REVIEW

Culture specific items differ from one place to another: hence, it requires a specific method of translation. When the item does not exist in the target language, maintaining its original signifier is preferable with reference to footnote. Another method is by paraphrasing or describing the items by using longer lexical chains in the target language. The culture specific items can refer not only concrete concept to the extent of abstract concept. One of

the items is pronoun. This problem was discussed by Baker (2011: 21-42). That is to say, equivalence above word level is sometimes necessary.

Mostly considered from the grammatical aspects (plural markers, case etc), the use of pronouns also reflects the social dynamics as well. For instance, in English, third person pronouns distinguish male and female. In Korean, 1st person pronoun is often dropped. In French and some other European languages such as Spanish and German, pronouns can be used as standard or honorific expressions, which are mostly known as T-V distinction. In some other languages, pronouns used in oral and written communication differ. Brown & Gilman (1960) described the use of pronouns across languages, which concerns the use of pronouns as social practice, beyond grammar. The significant aspects of the social practice of pronouns are the power and solidarity dimension. Let us put it simply by the following examples from Indonesian pronouns.

Consider 2nd person pronoun used by the employer to its employee. It is relatively simple in English as *you* is used regardless of power and solidarity. However, consider another language like Indonesian, where *you* has several equivalents: *kamu, Anda, kau*. The employer can use *kamu* (T) as s/he reserves the right due to his/her power. As the power differs, the use of 2nd person pronoun is not reciprocal. Considering the power dimension, the employee will avoid the use of *kamu* and shift to *Anda* (V). However, the pragmatic constraint restricts the use of *Anda*. Most likely the employee will prefer kinship terms *Ibu* or *Bapak* (literally translated as *mother* or *father*, but frequently used as honorific adresses).

Some newspapers have separate columns, which aim on different segment: for instance, a column for teens. This background affects the use of pronoun. In the counseling section, the counselor preserves the use of T pronoun, *kamu*. In this case, it is influenced not only by power dimension (bearing in mind that the counselor is much older than the reader, or more specifically the one who consult), but also the solidarity dimension. The use of *kamu* is expected to increase solidarity level, as this is a commonly used pronoun among teens. According to Brown & Levinson (1987) who formulated politeness strategies, this strategy is called as 'using in group identity marker'. The address is used as a marker to indicate that speaker and hearer belong to the same social group: hence increasing solidarity level.

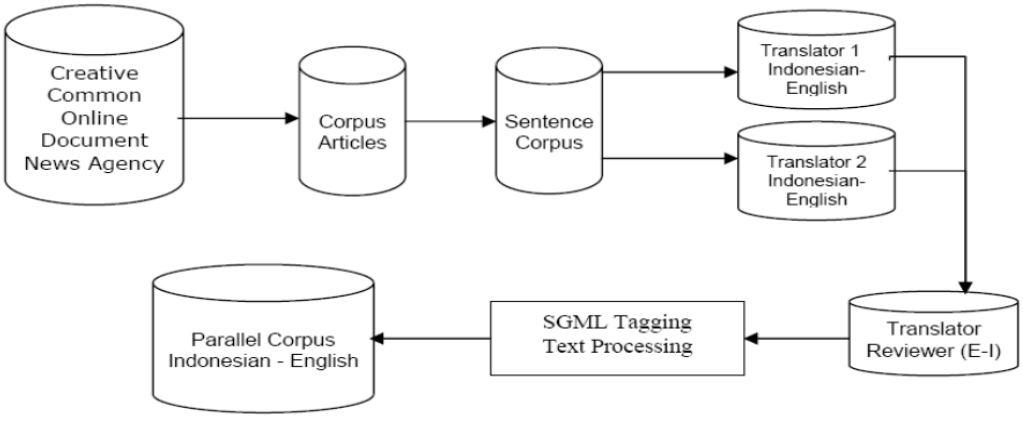
3. METHODOLOGY

3.1 Collecting Corpus Data

Corpus data in this research is downloaded from http://www.panl10n.net/indonesia/. The linguistic resource, software and resource listed in this page are the result of collaborative project under the name of PAN Localization project. This project is aimed at developing computing environment in Asia, which includes Indonesia. In Indonesia, the collaboration agency is BPPT. As for this project, it aims on creating a design for automatic translation (English & Indonesian).

The initial stage of the design of the machine translation framework is crucial to highlight in this paper. Before statistical computing is performed, the design requires corpus data. There are two types of corpora: monolingual and parallel corpus. However, the later is the most crucial aspect in creating the design of the parallel corpus presented by PAN Localization Research Report phase 1.1. Consider figure 1:

Figure 1. Corpus Collection Process



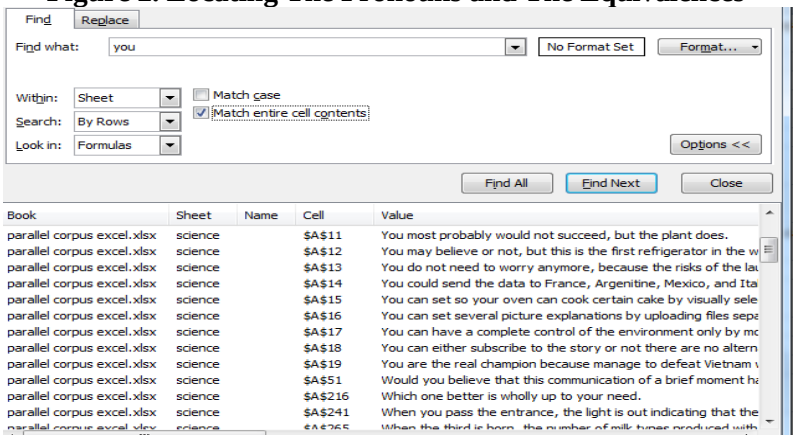
The figure can briefly be outlined as follow. The corpus contains spoken and written data from news agencies. The texts are then aligned to sentences. Human translators then translate the sentences to English. The results are assessed by human reviewer. It results on parallel corpus, which requires computing for text processing. For this moment, we do not take the SGML tagging and Text Processing into account as this is more relevant for

computational linguistics rather than the translation aspect itself. The data is obtained from ANTARA news Agency. However, the corpus also makes use of some data from internets, which are already bilingual pairs in two languages. The quality of the translation is considered valid as it is performed and reviewed by human translators.

3.2 Data Alignment and Indexing

The early forms of parallel corpus that I research are two files: one is Indonesian, and another is English. The format is .txt extension or notepad file. The content of these two files are assigned to two different columns (English and Indonesian) in spreadsheet. The alignment process takes places in the spreadsheet to confirm that the English text on the left column and the Indonesian text in the right column are equivalent. As for the samples to analyze, pronouns must be located in the first place. For pronouns recognition, I automatically indexed them by using find and replace function in the spreadsheet. The indexing can be performed, either in Indonesian or English column. And, when the pronouns on one column is located, its equivalences on the other column are noted.

Figure 2. Locating The Pronouns and The Equivalences



4. FINDING AND DISCUSSION

This section presents findings from the research, and discusses the findings by referring to the literature review and some other sources as well. It begins by outlining the finding for 2nd person pronoun and its diverse equivalents in Bahasa Indonesia. Bearing in mind that English distinguishes the use *you* as subject and possessive pronoun, the equivalence table is divided into separate rows. Consider table 1:

Table 1. Equivalents for 2nd Person Pronouns

English	Indonesian
You (sub/obj)	*Kamu*
	Anda
	ϕ
	Saudara-saudara
	Kalian
	Tamu
	Voice changing
Your (poss)	*{-mu}*
	Milik Anda
	milik saudara

4.1 Near-Equivalent Pronouns

The most frequently used equivalence for *you* and *your* is *Anda*, followed by *kamu*. In written data, the writer does interact directly with the listener. As for this, a much safer strategy is by using *honorific* pronoun *Anda* regardless of the readers' background.

Note that the use of *kamu* in the corpora is frequently used in spoken data. One occurrence of *kamu* in written data is identified only on the international news corpus, and contextually it is a direct quotation from a father to

his son. The role of a father authorizes him to use such address to his son in Indonesian, but the opposite case is not acceptable in Indonesian. However, pronoun of English *you* does not show this power gap, which to some extent we can say, more democratic. Consider example 1:

(1) My son said, 'Am **I** going to have limits like this my whole life?', and I said, 'No, when you move away **you** can set your own screen limits', Gates recounted, to audience laughter.

*Anakku berkata, Haruskah **saya** menghadapi pembatasan ini sepanjang hidup saya?', dan saya berkata, 'Tidak, ketika **kamu** pindah, kamu dapat menetapkan batas waktumu sendiri di depan layar komputer', kata Gates, sementara pendengarnya tertawa.*

From example 1, the Indonesian translation suggests that the social dynamics restricts the use of standard pronouns from one of higher position. The supporting evidence for this is the use of 1st person pronoun *saya* (V) when it is addressed from a son to his father. The son avoids the use of *aku* (T) as it suggests an equally powerful dimension to his father. Therefore, he prefers on using *saya* instead of *aku* (T).

4.2 Anaphoric Reference and Address

Often the direct reference of the pronoun is preferred to come into the text. This strategy is related to anaphora. Anaphora refers to the interpretation of a linguistic unit that derives from the previously expressed unit. In language testing, like TOEFL, test takers often encounter this kind of test item. They are usually asked to which oh the choices that refers to the pronoun in the text. Here test takers are requested to resolute the anaphoric reference of pronoun. Consider (2):

(2) **If <u>you</u> detain people, <u>you</u>** must have good enough reason for detaining them and have a chance for there being a successful prosecution, he told ITV news, after anti-terror raids detained nine people in Birmingham last week.

***Jika <u>Ø</u> menahan orang, <u>Anda</u>** harus memunyai alasan agar berkesempatan berhasil dalam pengadila, katanya kepada televisi berita ITV sesudah penggerebekan anti-teror menahan sembilan orang di Birmingham pekan lalu.*

To some extent, the case in (2) suggests the anaphoric reference of pronoun *you*. In Indonesian version, the subject pronoun is dropped. Here, any Indonesian speaker will understand this sentence, and it is a natural sentence in Indonesian. However, in English, the subject pronoun is not omitted, as both clauses (dependent and independent) require a subject each.

When the reference is clear, besides writing the pronoun-pronoun equivalence, It is also possible to mention the reference directly such as example (3):

(3) I am calling on **you** at Bank of Indonesia and leaders of other banks to make available your credits to finance economic growth.

*Saya mengajak **saudara-saudara** di Bank Indonesia dan pimpinan dunia perbankan untuk mengalirkan kredit saudara membiayai semua yang menunjang 'growth' tadi.*

The Indonesian version of the object pronoun is *saudara-saudara* (honorific). Literally, *saudara* refers to any relatives, but the meaning shifts (wide) into a honorific plural address in Indonesian. In English, however, equal address does not exist. Therefore, the translator believes that pronoun *you* is the nearest equivalence as it can also refer to plural objects as well. The anaphoric resolution for *saudara-saudara* is a group of executives in Indonesian Central Bank (Bank Indonesia) and other banks as well. Consider the anaphoric resolution tracking on the sentences preceding *saudara-saudara*.

President Susilo Bambang Yudhoyono has asked **national banks** to boost economic growth with their equity by providing credits to productive sectors and micro, small and medium businesses UMKM. All financial and capital resources like bank credits should take advantage of the construction of power plants, transportation, and infrastructure. Without these three elements, the economy will not run well, and reduce competitiveness and

investment, he said here on Wednesday. In addition, **the national banks** should also support the development of investment, boosting exports, industries, agriculture and services as well as the real sector for a substantial contribution to economic growth. I am calling on **you** at **Bank of Indonesia** and **leaders of other banks** to make available your credits to finance economic growth.

The sentences are well-interconnected one and each other. The text is considered coherent and cohesive. Therefore, readers will find anaphoric resolution relatively easy. The four word forms *national banks, bank of indonesia,* is replaced by pronoun *you.* As pronoun *you* is plural, another reference comes in to the text, which is *leaders of other banks* (not only bank of Indonesia). Tracking anaphoric resolution on a text that is not cohesive or coherent appear to be relatively complex. Consider example (4):

(4) When **you** pass the entrance, the light is out indicating that the electricity has been produce.

*Ketika **tamu** melewati pintu, lampunya menyala, menandakan listrik sedang diproduksi.*

Consider pronoun **you** in (4), where in the Indonesian translation no pronoun can be found. Instead, we find the direct reference of the pronoun. We understand that the function of a pronoun is to replace the reference, avoid the repetition of the reference in the text. When the word *tamu* (guest) is preferred as the equivalence for *you*, it is assumed that the text is addressed to guests. However, tracking back to few previously existing sentences, I have failed to identify that in English, the word you can actually refer to guest. The text does not seem to be cohesive and coherent. Consider the excerpt of the corpus in (5):

(5) **Whereas** addresses start with uncommon letter has spam quantity up to one-fifth of their inboxes. **Whereas** 9% acknowledge that they have not use the network sites, but have planned to try. **Where** the bleeding is, a blood clot forms that hardens and heals the wound in due time. **Where** justice could be enforced? He/she wrote in English. **Where** as, sound intencity posibly reached of even over the dangerous level for our ear. **Whenever** a new bee looks in on the same flower, it smells the scent and understands that the flower is of no use and so goes on directly towards another flower. **When you** pass the entrance, the light is out indicating that the electricity has been produce.

The interconnection between one section and another seem to be lost. This might be caused by the alignment of the corpus is organized with reference to the alphabetical order of the sentence. Finding the reference is relatively easy when the organization is passage based as it preserves the coherence and cohesion of the text. It suggests that in order to perform anaphoric resolution, the text must be cohesive and coherent in the first place. The distance between the pronoun and its reference must not be too far, otherwise it will be relatively difficult to trace. It suggests that some of the corpora are organized in differently.

4.3 Voice Shift

Bearing in mind the previous points where the pronouns are translated to word-level equivalences (pronoun, address or reference), voice changing is also one of the strategies concerning the pronoun translation. Consider example 6:

(6) Whatever **you want to do**, **you do it**, I vowed.
*Jadi saya bersumpah, **apapun yang mau dilakukan**, **lakukanlah***

The second person pronoun in Indonesian is not overt. This happens, as the sentence voice is passive. Differently, in English, the sentence is active: therefore, the presence of *you* is required. Example (6) presents a case where the voice of the sentence differs in two languages (active in English and Passive in Indonesian). Note that passive voice is frequently used in Indonesian. The nature of passive voice in the both languages, however, is the same: agent is not required to occur. There are several possibilities for this such as the focus shift from agent (usually subject) to action (the verb). As the grammar allows the dropping of the agent, the agent of the action is unknown: or highly understandable from the context. The voice shift and pronoun (*you*) drop can also be considered as an effort to minimize face threatening act to the addressee (Brown & Levinson, 1987).

5. CONCLUSION

This paper has investigated the quality of the parallel corpus of BPPT, with focus on pronouns translation. It suggests that the complexity of Indonesian pronouns influences the strategies for pronoun translation. A pronoun in English might have several equivalences in Indonesian. By considering the anaphora reference, translators also

in some cases, prefers to translate a culture specific concept to pronoun. Although the expressive meaning differs, it might be the nearest equivalence considered. To some extent, voice changing is also preferred (passive-active or vice versa). Pronouns that are usually dropped in Indonesian, occurs in the English translation. These, in substance, are some efforts to preserve naturalness in the target language.

It is crucial to review the result of statistical machine translation. The title of the report (Adriani & Hamam, 2009) concerns the bidirectional translation (Indonesian–English and English-Indonesian). They claimed to achieve 92.1% translation quality for English to Indonesian direction. We need to understand the parameters of this claim and more importantly, this claim should ideally open for testing. It is true that the report provided sample translation. However, the testing update is always called for. Hence, the comparison, of pronouns translation for example, can be performed as it is necessary to evaluate how human and machine translation differs.

Up to the moment this paper is written, access to the machine translation website listed on the report (http://translator.iptek.net.id/PANL), is still restricted by id and password. When the access is completely open, it will invites more and more testing beyond project members. Other computer scientists and linguists can also take part in the evaluation and in consequence will improve translation quality of MT near human translation quality.

REFERENCES

[1] Adriani, M., & Hamam, R. (2008). *Research Report Phase 2.1: Final Design Report on Statistical Machine Translation Network.* Jakarta: Badan Pengkajian dan Penerapan Teknologi (BPPT).
[2] Adriani, M., & Hamam, R. (2009). *Research Report Phase 3.2: Final Report on Statistical Machine Translation for Bahasa Indonesia - English and English to Bahasa Indonesia.* Jakarta: BPPT.
[3] Baker, M. (2011). *In Other Words: A Coursebook on Translation 2nd Edition.* London: Routledge
[4] Brown, P., & Levinson, S.-C. (1987). *Politness: Some Universals in Language Use.* Cambridge: Cambridge University Press
[5] Brown, R., & Gilman, A. (1960). The Pronouns of Power and Solidarity. Dalam T.-A. Sebeok, *Style in Language* (hal. 253-276). Cambridge: MIT Press
[6] Catford, J.-C. (1965). *Translation, A Linguistic Theory of Translation.* London: Oxford University Press
[7] Newmark, P. (1988). *A Textbook of Translation.* London: Prentice Hall