A Comparison of Two Variant Corpora: The Same Content with Different Sources

Kyonghee Paik[1], **Kiyonori Ohtake**[1], **Kazuhide Yamamoto**[1,2]

[1] ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City", Kyoto 619-0288 JAPAN
{kyonghee.paik, kiyonori.ohtake}@atr.jp

[2] Nagaoka University of Technology
1603-1 Kamitomioka, Nagaoka City, Niigata 940-2188 JAPAN
yamamoto@fw.ipsj.or.jp

Abstract

In order to investigate the effect of source language on translations, we investigate two variants of a Korean translation corpus. The first variant consists of Korean translations of 162,308 Japanese sentences from the ATR BTEC (Basic Expression Text Corpus). The second variant was made by translating the English translations of the Japanese sentences into Korean. We show that the source language text has a large influence on the target text. Even after normalizing orthographic differences, fewer than 8.3% of the sentences in the two variants were identical. We describe in general which phenomena differ and then discuss how our analysis can be used in natural language processing.

1. Introduction

We compare two corpora in order to investigate the effect of source language effect on translation. In particular, we present an analysis of paraphrased sentences extracted from bilingual travel corpora. The corpora used for this research consist of 324,616 Korean sentences. Half of the Korean sentences (162,308 sentences) were translated from Japanese, and the other half were translated from English sentences that match the original Japanese.

Although the two Korean corpora should be equivalent in meaning, they have different characteristics, since they were translated originally from such different languages as English and Japanese. English has a relatively fixed word order (SVO), and complements such as subject and object are obligatory. On the other hand, Japanese has a relatively free word order, although it is strongly verb-final, and complements can be freely omitted when their referents are clear from the context. The grammar itself and the set of available grammatical constructions are quite different from each other. It is also obvious that the ways of perceiving and conceptualizing facts, activities and emotional events reflected in basic vocabulary vary with the language. In this respect, Korean is closer to Japanese rather than to English.

The following examples show the differences in linguistic structure that are extracted from our two corpora. For example, (1) comes from the Korean corpus translated from Japanese [1] and (2) comes from the Korean corpus translated from English[2]

(1) 이 케이블카를 타면 호텔에 갈 수 있습니다.
this cable-car-acc take-if hotel-loc go can-decl

(2) 케이블카가 호텔에 데려다 줄 겁니다.
cable-car-nom hotel-loc take give future

The Korean translation in (2) does not sound natural, but the one in (1) does. This translation difference comes from the influence of the source language. That is, the two translations reflect their original source language. Even though the original translation pairs are matched between English and Japanese, the results of the translation are very different. The differences in translations due to the source text have not been extensively studied. In this paper, we examine whether a clear difference exists throughout the corpus and investigate further whether we can use this difference to improve the quality of machine translation. Furthermore, we confirm that the differences can be applied to paraphrasing the source or target languages, which may lead to improving the quality of machine translation.

2. Comparing Two Corpora

We have a basic travel expression corpus (BTEC) that is a collection of Japanese sentences and their English translations for Japanese travelers (an early version is described in Takezawa *et al.* (2001)). This corpus covers such topics related to travel as "shopping", "hotel/restaurant reservation", "airport", "lost and found" and so on. So far, we have developed four language corpora — Japanese, English, Korean, and Chinese, but we exclude Chinese data since they are not relevant to the present paper. Henceforth, we call them K_E (the corpus translated from English) and K_J (the corpus translated from Japanese). The size of the corpora and amount of redundancy is shown in Table 1.

	Japanese	English	K_J	K_E
All	162,320	162,320	162,320	162,308
Unique	102,247	97,326	103,051	92,816
Redundancy	37.0%	40.0%	36.5%	42.8%

Table 1: Summary of BTEC (K_J & K_E)

[1]The original Japanese sentence is "このケーブルカーに乗れば、ホテルに行くことができます。"(lit. trans. "If you take this cable car, to go to the hotel is possible."

[2]The original English sentence is "This cable car will take you to the hotel."

	Similarity Score										
	0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0	Total
All sentences	100	1,910	11,006	23,126	33,755	34,888	28,083	17,400	7,693	4,347	162,308
Unique sentences	58	1,243	7,876	19,351	29,053	30,149	24,382	14,946	6,434	3,037	136,529

Table 2: Distribution of the sentences on similarity score

		Similarity Score								
Category	**Phenomenon**	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Sentential Type	Identical	0	0	0	0	0	0	3	0	18
	Free translation	6	26	36	13	8	4	4	0	0
	Mistranslation	3	8	6	7	1	2	2	6	0
Lexical Choice	Noun	0	30	98	110	98	70	23	15	1
	Verb	0	32	112	193	186	88	37	11	2
	Interrogative	0	2	11	9	10	2	4	1	0
	Other	1	17	68	115	89	40	16	1	1
Syntax	Classifier	0	2	5	11	6	2	0	0	0
	Other	1	20	71	109	156	96	34	15	5
Orthography	Alphabet	1	0	1	7	7	0	0	0	0
	Number	0	6	14	18	20	25	10	0	0
Number of samples		12	78	192	290	300	243	149	64	30

Table 3: Paraphrasing categories

2.1. Analysis of Our Corpora

We compared the sentences using the perl module `String::Similarity` (Lehman 2000). It returns a similarity score based on the edit distance (the number of characters that need to be deleted, added or substituted to change one string into another), normalized to give a score between 0 and 1 (Myers, 1986). Two completely different strings have a score of zero, while two identical strings have a score of one.

There were 136,529 sentences. Their distribution is given in Table 2. Less than 2% of the sentences were the same in both corpora. Most sentences were reasonably similar (0.4–0.8). Very few sentences were identical (less than 4%), and even fewer were totally dissimilar (less than 0.1%). There were some minor orthographic differences (described in Section 2.2.4.), but even when they were resolved only 8.3% of the sentences were identical.

2.2. How similar they are: K_E and K_J

Here we present an analysis of how similar the two corpora are. First, we divided linguistic phenomena into four categories: sentential type (how the sentence was translated), lexical choice (difference due to lexical choice), syntax (difference due to syntax) and orthography (difference due to orthographic variation). The distribution is shown in Table 3. A random sample was chosen for each of 9 similarity bands from 0.1–0.2 to 0.9–1.0. The sample size was proportional to the number of sentences in each band.

2.2.1. Sentential Type

The more identical sentences are found, then of course the higher similarity score is. On the other hand, in the low similarity score we found more free translations. For example, 받아요 (trans.“Here you are”) in K_E is given as 이것이면 됩니까? (trans. “Will it be okay with this?”) in K_J. At a glance, this is not a good translation, but when you consider a certain situation, these two translations can be compatible. Analyzing the sentences of similarity score 0.0-0.4, we find many sentences that are expressed in a different way. Take the following examples. Both of the sentences mean “I do not have an appetite.”

(3) 밥 생각이 없어요.[3] : K_E

(4) 식욕이 없습니다.[4] : K_J (from similarity 0.3-0.4)

If you want to translate the Korean sentence (3) into other languages, there is a high chance of translating it wrongly. It does not give you the real meaning when you translate it word to word, or phrase to phrase. However, if we can paraphrase (3) with a simpler expression like (4), then the paraphrased sentence will be easily and correctly translated. We can extract these kinds of paraphrases at low similarity scores from 0.0 to 0.4. As Table 3 shows, the extracted number of free translations is relatively high in these similarity bands.

Finally, there are a few cases where the two corpora do not match each other because the Japanese and English bilingual corpus contain a translation mismatch. However, overall, the number of mistranslations is negligible.

2.2.2. Lexical Choice

The main source of variation is differences in lexical choice. We divided the phenomena into four classes. Then we compare the two sentences and check whether any noun, verb or other lexical items such as adjective, adverb and related phrase has been replaced. For example, “욕실 딸린 싱글룸 예약하고 싶습니다” in K_E is compared with “욕조 딸린 싱글 룸을 부탁합니다” in K_J, meaning “A single room with a bath, please.” Comparing the two sentences, we found that 욕실 *yoksil* “bath” is used instead of 욕조 *yokco* “bath”, which is noun, and 예약하고 싶습니다 “would like

[3] Lit. trans. “There is no thought of rice.” It is translated from the English sentence “I do not have an appetite.”

[4] The sentence was translated from the Japanese sentence “食欲がない.”, meaning “There is no appetite.”

to reserve" is used instead of 부탁합니다 "do me a favor" or "please", which is considered verb alternation. However, the use of different nouns and verbs accounts for about 20% of the variation throughout the corpora.

2.2.3. Syntax

The next most common category is differences in syntax. This includes the deletion or addition of case markers, scrambling, and changes in voice (active/passive). As we can see in Table 3, these occur throughout the entire range of similarities. The higher the similarity is, the more syntactic changes are found. At the same time, we examine how often numeral classifiers are used in our corpora since it is necessary for both Korean and Japanese. We found that about 1% of the numeral classifiers are floated. Among them, some classifiers are used in one corpus but omitted in the other. For example, in (5) coffee is counted (a cup of coffee), but tea isn't (some tea). Both choices are valid.

(5) 커피 한 잔 같이 안 하실래요? [5] : K_E

(6) 함께 차라도 어떻습니까? [6] : K_J

We found an interesting fact with respect to numeral classifiers. The kinds of numeral classifiers are far more in the corpus K_E, whereas the frequency of numeral classifiers are far more in the corpus K_J, as shown in Table 4.

Numeral classifier	K_E	K_J
Type	320	284
Token	7,878	8,928

Table 4: Comparison of numeral classifiers in K_J & K_E

2.2.4. Orthography

There are two kinds of orthographic difference. One arises due to differences in transliterating foreign words such as place names (e.g. 피카딜리 *pikadilli* "Piccadilly" vs 피카디리 *pikadili*), people's names and so on. The other is whether numbers are given using Hindu-Arabic numerals (1,2,3,...) or spelled out in Korean (일,이,삼....)[7]. There can also be some minor variation in punctuation (especially the use of question marks, which are optional in Japanese), although none appeared in our samples. After normalizing the differences in spelling out numbers and letters, the number of identical sentences more than doubled to 8.3% of the total. These orthographic differences are not important, but they are widespread. This highlights the importance of dealing with orthographic variation in any empirically based approach.

2.3. Linguistic phenomena between K_J and K_E

In this section we focus on four phenomena with quite different distributions in K_J and K_E: honorifics, zero pronoun, Kango (Chinese words), and loan words. The details are shown in Table 5 and discussed below.

[5]Lit. trans."How about joining me for a cup of coffee?"

[6]The sentence was translated from the Japanese sentence "いっしょにお茶でもどうですか。", meaning "How about having (a cup of) tea?"

[7]In theory, numbers could also be written in Hanja, but this did not occur in these two corpora.

2.3.1. Honorifics

Korean has more honorific speech levels than Japanese. Consider (7) and (8).

(7) 후추 있나요?
pepper exist-honorific/polite
"Do you have pepper?" (*polite*)

(8) 후추는 있습니까?
pepper-top exist-honorific/deferential
"Do you have pepper?" (*deferential*)

	deferential		polite		
	-nida.	-nikka?	-yo.	-yo?	Total
K_E	23,316	9,970	**33,617**	**25,481**	93,227
K_J	**33,351**	**34,872**	21,922	3,082	92,384

Table 6: Difference in honorifics between K_J and K_E

We found that K_J tends to use more deferential honorifics, whereas K_E tends to use more polite honorifics as shown in Table 6. This distinction is not made in either English or Japanese. Korean has six different speech level honorifics whereas Japanese has two. In the sense that Japanese has two lexicalised honorifics showing degree of politeness, we can say that the source language affects the degree or level of honorifics.

2.3.2. Zero Pronoun

We expect that there are some differences which comes from the source language before translation. In English subjects are obligatory. One of the main characteristics of translated Korean sentences from English is that many specific pronouns are used. This is fully reflected in the fact that English strongly express subject/agent of sentences, whereas Korean and Japanese tend to eliminate its counterpart. Furthermore, both Korean and Japanese lack obligatory determiners, whereas English generally requires some determiner. Compare (9) and (10).

(9) **제** 친구 집에 머물 예정입니다.
my-gen friend house-loc stay plan-beVERB
"I am planning to stay at my friend's house." (K_E)

(10) ϕ 친구 집에 묵습니다.
ϕ friend house-loc stay-VERB
"I stay at my friend's house." (K_J)

2.3.3. Kango

As Table 5 shows, K_J has a slightly stronger tendency to use words containing Chinese characters. Since the ratio of words of Chinese origin in both Korean and Japanese accounts for more than 50% (Paik and Bond:2001), we believe that if a shared Kango is used in K_J, then it will be easier to translate into the same word corresponding to Korean. 매장 (賣場) in (12) is a word of Chinese origin that is also used in Japanese. On the other hand, the phrase of the sentence from K_E is translated into native Korean words like (11).

(11) 화장품 **있는 데**는 어디죠?
cosmetics exist place-top where-be
"Where are the cosmetics?"

(12) 화장품 **매장**은 어디입니까?
cosmetics **selling place**-top where-be
"Where is the cosmetics section?"

	Similarity Score								
Phenomenon	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Honorific (K_J)	1	10	16	82	101	76	32	11	2
Honorific (K_E)	0	8	15	20	32	31	9	2	0
Zero (K_J)	0	3	8	23	28	22	13	4	1
Zero (K_E)	0	2	5	7	15	8	9	1	1
Kango (K_J)	0	7	23	39	54	22	19	4	0
Kango (K_E)	0	2	10	25	35	19	12	0	0
Loan word(K_J)	0	5	16	14	15	7	6	0	0
Loan word(K_E)	0	0	7	11	9	11	3	0	0

Table 5: Linguistic phenomena causing differences between K_J and K_E

2.3.4. Loan word

By loan words we mean words in daily usage borrowed from languages other than Chinese. We expected that more loan words would be used in K_E because most loan words come from English. However, according to Table 5, the result is the opposite. After examining the source corpora of K_J, we found that many "Katakana words[8]," which include borrowed/foreign words, are used in Japanese sentences. Almost all of the Katakana words are translated into foreign words in Korean. In contrast, all of the English words are equally foreign, so for any single word, there is little pressure for it to be translated into a loan word rather than into native Korean.

2.4. Discussion

As we have seen, the source language has a large effect on the translation. Accordingly, it is necessary to reconsider the quality of parallel corpora because they are widely used in the NLP area. So far, we have taken it for granted that existing parallel corpus is used when necessary, if there is any. This is partly because that we do not have enough parallel corpora throughout the world and partly because many of the corpora have been created through English. We expect that many more parallel corpora will be built not only through English but also through many other languages from now on. Therefore, we need to put emphasis on the quality or characteristics of the corpora according to the application.

In particular, much research into machine translation uses parallel or comparable corpora without considering their characteristics. This will lead to different results. For example, as we observed in this paper, the result of a translation trained on the K_E corpus will be less polite compared to that of a translation trained on K_J, and it will also produce different results with respect to the linguistic phenomena such as zero pronouns, numeral classifers and so on. As for the numeral classifier, it is interesting that K_E uses more classifiers than K_J if we compare unique classifiers. In many cases, English does not use classifers, whereas Japanese has to use numeral classifiers for counting objects. At a glance, this is the opposite to what the existing corpora show. However, K_E and K_J closely reflect the effects of the source language. When the numeral classifier is implicitly expressed, it will be more freely translated. That is why K_E has more types of numeral classifiers.

[8]Japanese has two writing systems based on syllables: Hiragana and Katakana. Katakana is mainly used for writing loan words and names/places that can't be written in Kanji.

In addition, we can expect that the optimal translation strategy may be different between language pairs. In particular, a system translating between Japanese and Korean needs to put less effort into lexical and syntactic choice and more into the use of honorifics. On the other hand, a system going between English and Korean has a much harder task, since it must consider lexical and syntactic choice and zero pronoun resolution in addition to the use of honorifics for machine and human translation.

Furthermore, a pair of corpora such as these is a useful source of data on variation within a single language and can readily be exploited to learn paraphrasing rules. For example, as we have seen in 2.2.1., we can extract free translation relatively easily and with less cost.

3. Conclusions

We investigated two variants of a Korean translation corpus, one based on translations from Japanese and the other from English. We have shown that the source language text has a large influence on the target text in almost all areas — lexical choice, syntactic structure, use of zero-pronouns and honorifics, and even orthographic variation. One surprising result is how different the corpora were, even after normalizing orthographic differences: fewer than 8.3% of sentences in the two variants were identical. Also, we should emphasize that the characteristics and the effect of the source languages must be taken into consideration in constructing a pair of parallel corpora for the better machine translation and for the related application.

4. Acknowledgements

This research was supported in part by the Ministry of Public Management, Home Affairs, Posts and Telecommunications. We would like to thank Francis Bond of NTT for his useful advice and discussion.

5. References

Lehmann, Marc, 2000. String::Similarity. Perl Module (`cpan.org`). (v0.02).

Myers, Eugene, 1986. An O(ND) difference algorithm and its variations. *Algorithmica*, 1(2):251–266.

Paik, Kyonghee and Francis Bond, 2001. Multilingual generation of numeral classifiers using a common ontology. In *19th International Conference on Computer Processing of Oriental Languages: ICCPOL-2001*. Seoul. 141–147.

Takezawa, Toshiyuki, Satoshi Shirai, and Yoshifumi Ooyama, 2001. Characteristics of colloquial expressions in a bilingual travel conversation corpus. In *ICCPOL-2001*. Seoul. 384–389.