# A Test Suite for Evaluation of English-to-Korean Machine Translation Systems

**Sungryong Koh, Jinee Maeng, Ji-Young Lee, Young-Sook Chae, Key-Sun Choi**

Korea Terminology Research Center for Language and Knowledge Engineering (KORTERM)
Korea Advanced Institute of Science and Technology (KAIST)
Kusong-dong Yusong-gu Taejon 305-701 Korea
{koh,aphroditejin,jinny206}@world.kaist.ac.kr, pinochae@chollian.net, kschoi@cs.kaist.ac.kr

**Abstract**

This paper describes KORTERM's test suite and their practicability. The test-sets have been being constructed on the basis of fine-grained classification of linguistic phenomena to evaluate the technical status of English-to-Korean MT systems systematically. They consist of about 5000 test-sets and are growing. Each test-set contains an English sentence, a model Korean translation, a linguistic phenomenon category, and a yes/no question about the linguistic phenomenon. Two commercial systems were evaluated with a yes/no test of prepared questions. Total accuracy rates of the two systems were different (50% vs. 66%). In addition, a comprehension test was carried out. We found that one system was more comprehensible than the other system. These results seem to show that our test suite is practicable.

**Keywords**

Evaluation, English-to-Korean, Yes-No Question, Comprehension Test, Linguistic Phenomena

## 1. Introduction

It has been emphasized that we have to evaluate the quality of translation from the specific purpose of an evaluation. The purpose of an evaluation is generally related to who a user is and what a task is. For example, a manager may want to read a letter or an email from foreign employees using an MT system. A developer of an MT system wants to test the performance of specific processes. The explicit description of the purpose helps to identify what characteristics of translation of a MT system should be measured. Since a manager or an end-user is usually interested in a general performance of a system, the degree of comprehensibility or fidelity of translation could be measured using a questionnaire. On the other hand, errors of a system about a variety of linguistic phenomena could be measured, since a system developer is interested in whether specific processes of translation have a problem.

JEIDA (1992) showed a neat method that relates users' needs to MT systems. In order to help the user evaluation of economic factors, several questions about the conditions of translation work and the user's needs are prepared. Their answers are analyzed into 14 parameters such as present translation needs, type of document and so on. The analysis result is represented as a radar chart and is compared with radar charts that characterize seven groups of MT systems. This comparison makes it possible to identify a system close to the user's needs.

Recently the importance of the usability of a product has been recognized and stressed. EAGLES Evaluation working group (1999) proposed a general framework for evaluation following ISO quality model. Their report emphasized the importance of quality in use as well as quality of a product. They defined quality in use as the user's view of the quality of a system containing software, claiming that it is measured in terms of the results of the use of software like an MT system, that is in terms of effectiveness, productivity, and satisfaction of users.

Considering that we didn't have almost any practicable evaluation methods for English-to-Korean MT systems, our urgent problem must be to provide a systematic and objective evaluation of the technical status of several commercial English-to-Korean MT systems. Systematic evaluation primarily concerns the analysis of a source language that should be handled by an MT system. Problems from analysis of a source language can be classified into two types. One type of problems comes from lack of knowledge. For example, if a system doesn't have the information that 'Bush' is a proper noun in its dictionary, it cannot translate 'Bush' as a name. Another type of problems comes from the inappropriate use of knowledge. When a string, 'Bush' is encountered, it cannot immediately be translated as a proper name, since 'Bush' can be used as a common noun. A system presumably tries to resolve the ambiguity using other kinds of information. If the string, 'Bush' is encountered in the middle of a sentence, it is easily disambiguated into a proper name because a proper name starts with upper case in English. However, if it appears as the first word of a sentence, the rule above is not useful since another rule that every sentence starts with upper case justifies its use of a common noun. In this case, a system should check other kinds of information such as animate information (*e.g.*, a verb in this fragment 'Bush said' describes a human behavior). The example shows that a systematic evaluation needs to check these two types of problems.

We believed that one way to systematically detect the lack of knowledge and the inappropriate use of knowledge was to specify the use of lexical and structural component in detail if possible and to collect many examples of one linguistic phenomenon in order to see the interaction with other phenomena in various local and global contexts, especially to see the conditions of interaction (In the above 'Bush' case, the position of a sentence). Thus, we have built large-scale test-sets that could support several types of evaluation such as internal evaluation, or comparison evaluation, and to test several commercial English-to-Korean MT systems.

In the following we present KORTERM's test suite based on systematic classification of linguistic phenomena and two evaluation tests, a yes/no question test and a comprehension test to see their practicability.

## 2. Construction of Test-sets

### 2.1 Characteristics of Test-sets

Two important problems concerning the construction of the test-sets were coverage and objectivity. In order to collect examples that cover a variety of linguistic phenomena, we initially classified linguistic phenomena, which will be described later. And we attempted to collect a variety of examples that can be assigned to specific linguistic phenomena.

In order to perform an objective evaluation method, we prepared one yes/no question for one example sentence. As pointed out in Isahara (1995), this yes/no question about a linguistic phenomenon enabled us to evaluate MT systems objectively. This objectivity can be a basis of a fair comparison between MT systems. One example of our test-sets is presented in figure 1.

[Serial Id]100
[Grammar Id] 10102080000
[English] August 15 is an unforgettable day to us Koreans.
[Korean] *8wol 15il-eun uri hangugin-egenun ijeul su eobs-neun nal-ida*. (August 15-TOPIC we Korean DATIVE-TOPIC forgettable-NEG-MOD day-FIN.
[Question] Are two nouns in "us Korean" translated into an appositive?
[Source] English High School Textbook- ii-a-1

Figure 1. A Sample of KORTERM's Test-sets for English-to-Korean MT Systems

As shown in figure 1, each test-set consists of an ID number, a number for grammatical category, an English sentence, a model Korean translation, a yes/no question, and a source.

### 2.2 Collection of Example Sentences

From the late 1999 to September 2000, English sentences were collected from several high school textbooks and other grammar books related to them, because we believed that English sentences included in textbooks are compact enough to show one linguistic phenomenon well. We collected about 5000 example sentences each of which consists of less than 15 words, and used 3431 sentences for evaluating two commercial systems.

Since October 2000, we have extended sampling domains to Web news and have been collecting about 3500 examples from a business news site, which will be tested. Their length was less than 30 words.

### 2.3 Classification of Linguistic Phenomena

When an MT system translates a sentence, it usually identifies lexical and structural components like a noun and a verb, and constructs syntactic and semantic relations for translation, although specific processes depend on the linguistic resources and algorithms of a specific MT system. From this processing perspective, we divided grammatical phenomena into the structural part and the selectional part as like JEIDA's test-sets. The structural part contains parts of speech, partial structure, and sentence structure. The selectional part contains lexical, syntactic and semantic ambiguities that can occur due to a choice between possible candidates. The distribution of 3431 examples that were used to test two commercial MT systems is in figure 2 partially (See Appendix for complete set). We subcategorized the structural and selectional part as shown in Appendix, referring to grammar books such as Hornby and JEIDA's categorization. Figure 2 shows that many English sentences were assigned to verb (12%) adverb (9%), and infinitive in partial structure (13%). After the sub-categorization, we analyzed them further depending on forms or their uses if the analysis is allowed, since we believed that this fine-grained classification helps to pin down the problems of an MT system. Article category is specified into three (indefinite article, definite article, and ellipsis and repetition of article), noun category into ten, and so on. In sum, the final classification contains 371 linguistic phenomena.

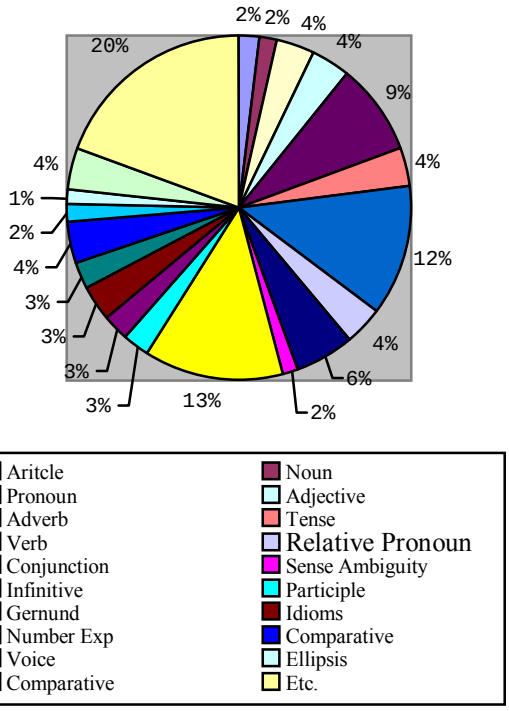


Figure 2.Items to Test Linguistic Phenomena

## 3. Evaluation

We carried out two 'pilot' evaluation tests to see the practicability of the test-sets. One test was a yes/no question test that was designed with our test-sets. The other was a comprehension test to see whether the test-sets can be used for other types of evaluations.

### 3.1 A Yes/No Question Test

A yes/no test was designed with the construction of the test-sets to see the technical status of MT systems from the developer's point of view. A tester judges whether the answer of a question is yes or no, after reading a translation of MT system. As expected, the judgment was obvious in most cases. If the answer is yes, one 'yes' button is checked in the evaluation tool made for this test. Otherwise, a 'no' button is checked.

Initially we started to test five MT systems. One system was excluded from analysis because of too poor translation and two are being analyzed. The test results of two MT systems are reported here as in table 1. The percent of each cell represents the number of success over total number of examples assigned to that linguistic phenomenon.

| | | System A | System B |
|---|---|---|---|
| Structural Part | | | |
| | Part of speech | 82 | 70 |
| | Partial structure | 40 | 60 |
| | Sentence structure | 49 | 67 |
| Selectional Part | | 38 | 55 |

Table 1. Accuracy (%) of Two MT Systems

The two MT systems showed better performance in the structural part than in the selectional part as shown in Table 2. Within the structural part, both systems showed better performance in parts of speech than in the other two parts. Considering the relative performance between the two, total accuracy rate of system B were higher than that of system A. (66% *vs*. 50%). System A showed better performance in parts of speech than system B, but showed not better performance in the other two of the structural part and the selectional part than system B.

These results indicate that the two systems handle the structural identification of components. However, the two systems, especially system A cannot identify a unit well, as the unit becomes larger. Furthermore, both systems, especially system A were poor in disambiguation at the lexical, syntactic, and semantic level. In sum, the yes/no question evaluation method using our test-sets showed how well one system handles a variety of linguistic phenomena and how different two MT systems are.

### 3.2 A Comprehension Test

Using our test-sets, we carried out a sentence comprehension test to look at the general performance of MT system to get information for an end-user. Reading an English sentence, a model Korean translation and a translation of MT system, one evaluator judges whether the translation of the system is comprehensible or not. The evaluator tries to judge as soon as possible. If she judges such that the translation conveys the message perfectly, it is assigned as 'good' (10 points). If the translation conveys the message partially, it is assigned as 'not good and not bad' (5 points). If the translation doesn't convey any message, it is assigned as 'bad' (0 points). The results are presented in table 2.

| | System A | System B |
|---|---|---|
| Good (10) | 592 | 970 |
| Not good and not bad (5) | 732 | 996 |
| Bad (0) | 2107 | 1465 |

Table 2. Comprehension Scores of Two MT Systems

The score of each cell refers to the total number of example assigned to the evaluation class. As we expected from a yes/no question test, the average of system B was higher than that of system A (4.27 *vs*. 2.97). This result seems to indicate that our test sets are practicable.

## 4. Discussion

The results of the two tests seem to support that the test-sets are quite practicable. However, there are several things to be improved. One limitation of the results of the two 'pilot' tests is that the results were based on one evaluator. In the case of a yes/no question test, it may not be a serious problem, since the answer was quite obvious and we don't expect much different judgments depending on evaluators. On the other hand, since a comprehension test is mainly based on subjective judgment, the number of evaluators can be a crucial problem. Many evaluators seem to be needed to obtain reliable results.

Another limitation concerns the frequency of each linguistic phenomenon in our classification. The frequency of each linguistic phenomenon seems to be influenced by genre and other factors. Without considering this point, the test-sets have been evaluated.

One issue concerns whether our classification of linguistic phenomena is too fine-grained. We pursued the fine-grained classification compared with the classification of JEIDA's test-sets. The classification of JEIDA may help us to detect the problems from lack of knowledge. However, it may not be easy to see the problems from the inappropriate use of knowledge in local and global contexts. When one sentence is analyzed, different types of linguistic knowledge are used and they presumably interact during analysis. The interaction influences the quality of translation of MT systems a lot. If an analysis is fine-grained, we might see this interaction well. Consider (1).

(1) He said in an interview that he was disappointed that the public project was sending its manuscripts.

In (1), one translation error of one MT system is that the 'that' clause modifies the noun 'an interview'. This error comes from a variety of uses of 'that'. The clause following 'that' can play a role of a complement clause or a relative clause and so on. We can imagine that two interpretations may compete in (1) and the length between the verb 'said' and 'that' may lead to the preference of a relative clause interpretation. If 'that' is treated as only one category of parts of speech in (1), it may be difficult to guess the cause of a translation error. This example above suggests that our test-sets based on fine-grained classification help us to diagnose translation problems of English-to-Korean MT systems accurately. Also to see the disambiguation in local and global contexts, it seems necessary to collect many examples of a certain linguistic phenomenon. The reason is our intuition that linguistic phenomena may compete under a certain condition. We can imagine that in the example above, an MT system might succeed in translation if the adverbial phrase 'in an interview' is located at the initial position of a sentence. This intuition suggests that locality condition is important in the use of knowledge. We will perform this kind of diagnosis on the basis of the results of the yes/no question test in the near future to see that a degree of processing complexity of a linguistic phenomenon could be described .as the conditions to reach a correct interpretation. This idea may lead us to group test sets depending on the degree of complexity of a specific linguistic phenomenon.

One realistic problem with the above discussion is in "cost". The construction of a large-scale and a test with it needs expertise and a lot of time. It is not easy to find an efficient way to save cost. One possible way may be to select small-size sample test sentences randomly from each linguistic phenomenon sentence pool of large-scale test-sets and run an evaluation test to sample and generalize the results to the whole test-sets. Another possibility may be to select small-size test sentences with the same degree or kind of complexity from a large pool concerning a specific linguistic phenomenon, and test them

## 5. Conclusion

We have constructed our test-sets, recognizing lack of an evaluation method in Korea. Using the test-sets we carried out a yes/no question test and a comprehension test with two commercial MT systems to see its practicability. The results of a yes/no question test were different in the two systems, showing how good each system handles various linguistic phenomena. In addition, the results of a comprehension test were similar to those of a yes/no question test. These results seem to show that the test-sets are practicable.


### Acknowledgements

This research has been being carried out by the technology service fund of Ministry of Science & Technology in Korea for 1999-2001 under the title "Large-scale Speech/Language/Image Database and Evaluation". We would like to appreciate Dr. Hyo-Sik Shin and Jong-Hoon Oh's comments.


### Bibliographical References


Bohan, N., Breidt, E., &Volk, M. (2000). Evaluating Translation Quality as Input to Product Development. Second International Conference on Language Resources and Evaluation. Proceedings, Vol. I.

EAGLES Evaluation Working Group (1999). Final Report.

JEIDA (1992). JEIDA Methodology and Criteria on Machine Translation Evaluation.

Isahara, H. (1995). JEIDA's Test-sets for Quality Evaluation of MT Systems – Technical Evaluation from the Developer's Point of View. Proceedings of MT Summit V.

White, J. (1995). Approaches to Black Box MT Evaluation. Proceedings of MT Summit V.

White, J. & O'Connell, T. (1996). Machine Translation Evaluation. AMTA Tutorial, Litton PRC, Montreal Canada.

## Appendix. Distribution of the Tested Items

1 Structural Part
- 1.1 Part of Speech 1711
  - 1.1.1 Article 64
  - 1.1.2 Noun 84
  - 1.1.3 Pronoun 127
  - 1.1.4 Adjective 129
  - 1.1.5 Adverb 300
  - 1.1.6 Preposition 180
  - 1.1.7 Verb 428
  - 1.1.8 Relative pronoun 128
  - 1.1.9 Conjunction 195
  - 1.1.10 Symbol 74
- 1.2 Partial Structure 452
  - 1.2.1 Infinitive 115
  - 1.2.2 Participle 90
  - 1.2.3 Gerund 39
  - 1.2.4 Idioms 119
  - 1.2.5 Number expressions 89
- 1.3 Sentence Structure 817
  - 1.3.1 Sentence Type 115
  - 1.3.2 Negation 73
  - 1.3.3 Special Constructions 57
  - 1.3.4 Comparative 133
  - 1.3.5 Subjunctive 54
  - 1.3.6 Voice 54
  - 1.3.7 Mood 13
  - 1.3.8 Insertion 50
  - 1.3.9 Ellipsis 52
  - 1.3.10 Inversion 16
  - 1.3.11 Parallel Structure 76
  - 1.3.12 Tense 124

2. Structural and Semantic Selectional Part 451
- 2.1 Main category ambiguity 95
- 2.2 Modification 79
- 2.3 Collocation 84
- 2.4 Word sense ambiguity 57
- 2.5 Style 130
- 2.6 Singular vs. plural difference 5
- 2.7 Collocation between a verb and a noun. 1