FromTo K/E: A Korean-English Machine Translation System based on Idiom Recognition and Fail Softening

Byong-Rae Ryu, Youngkil Kim, Sanghwa Yuh & Sangkyu Park

Knowledge Information Department
Computer • Software Technology Laboratory
Electronics and Telecommunications Research Institute (ETRI)
161 Kajong-Dong, Yusong-Gu, Taejon, 305-350, South Korea

{ryu,skpark}@computer.etri.re.kr
{kimyk,shyuh}@etri.re.kr

Abstract

In this paper we describe and experimentally evaluate **FromTo K/E**, a rule-based Korean-English machine translation system adapting transfer methodology. In accordance with the view that a successful Korean-English machine translation system presumes a highly efficient robust Korean parser, we develop a parser reinforced with "Fail Softening", i.e. the long sentence segmentation and the recovery of failed parse trees. To overcome the language-typological differences between Korean and English, we adopt a powerful module for processing Korean multi-word lexemes and Korean idiomatic expressions. Prior to parsing Korean sentences, furthermore, we try to resolve the ambiguity of words with unknown grammatical functions on the basis of the collocation and subcategorization information. The results of the experimental evaluation show that the degree of understandability for sample 2000 sentences amounts to 2.67, indicating that the meaning of the translated English sentences is almost clear to users, but the sentences still include minor grammatical or stylistic errors up to max. 30% of the whole words.

1 Introduction

We present in this paper **FromTo K/E**, a rule-based Korean-English machine translation system adapting transfer methodology, which has been developed by Electronics and Telecommunications Research Institute (ETRI) [10]. The project for developing the Korean-English machine translation system, **FromTo K/E**, was carried out 1996-1998 by Computer • Software Technology Laboratory of ETRI, former Systems Engineering Research Institute (SERI), partly in collaboration with Seoul National University (SNU) and Korea Research & Development Information Center (KORDIC) [11]. The project was completed in 1998. The initial purpose of this system is to provide Korean information to the foreigners who do not understand the Korean language [5].

FromTo K/E is a rule-based machine translation system adapting the transfer methodology, into which recent illuminating approaches to the various tricky ambiguity problems in MT have been integrated. They particularly aimed for minimizing the parsing complexity for Korean parser. Three major refinements can be pointed out: The first one is HMM-based Korean tagger, which resolves the morphological ambiguity generated by a morphological analyzer. The second one is Korean dependency parser with idiomatic expression recognizer to reduce the syntactic ambiguities and to produce natural English translation. The last one is the resolver of the correct case roles and grammatical functions of Korean sentences based on the collocation. Collocation-based Korean-to-English transfer selects the best English target words for ambiguous Korean words. Those collocations are extracted from the 400,000 word Korean corpus.

The results of the experimental evaluation show that the degree of understandability for sample 2000 sentences amounts to 2.67, indicating that the meaning of the translated English sentences is by and large clear to users, but the sentences still include minor grammatical or stylistic errors up to max. 30% of the whole words. This system is originally designed for a stand-alone system, but the functionality of this systern is now being extended to a web translation server.

2 System Overview

Figure 1 shows the configuration of the Korean to English machine translation system that is based on the idiom recognition and robust parsing with fail-softening process. The system largely consists of 4 modules: morphological analyzer, syntactic parser, Korean to English transfer, and English generator.

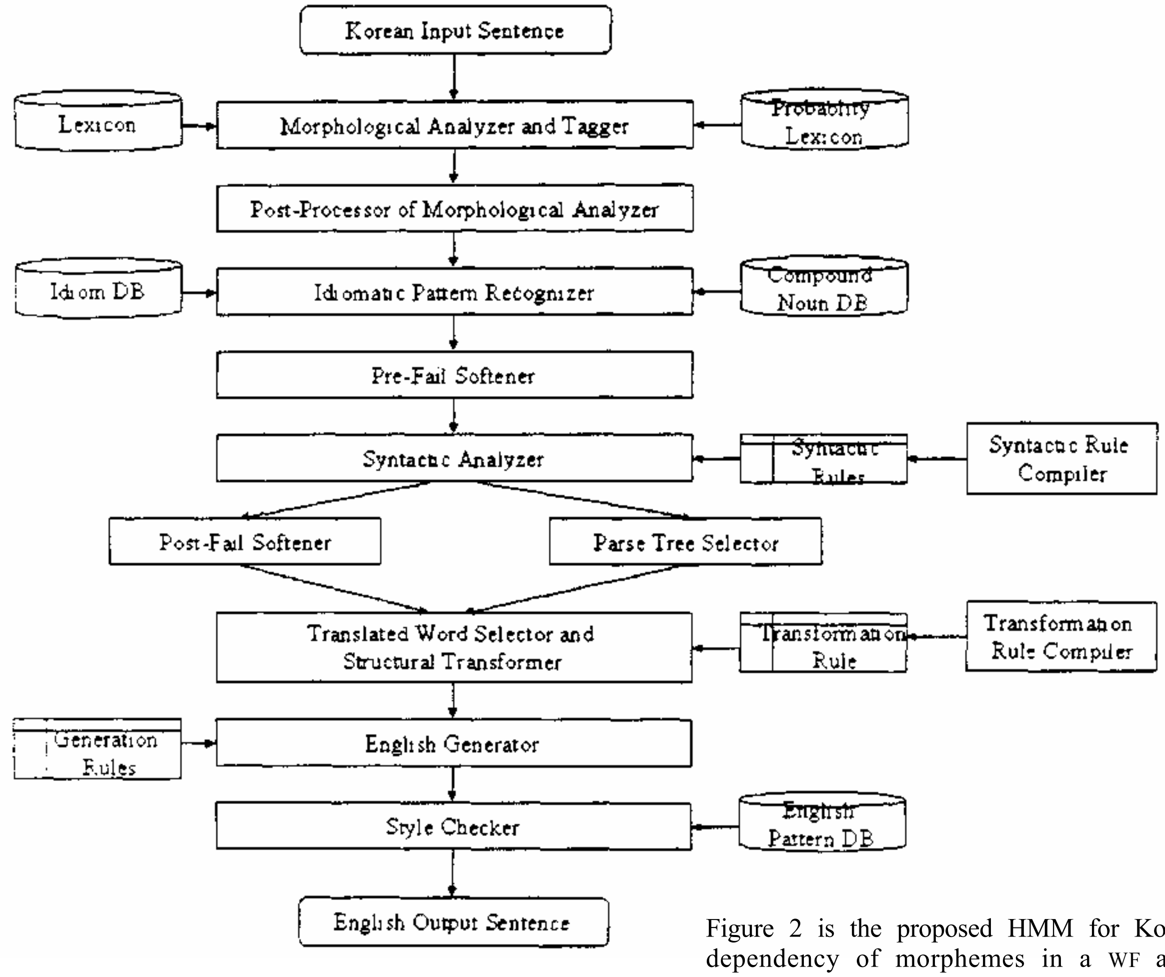


Figure 1 System Configuration of FromTo K/E

We use a dependency parser, based on the idiom recognition and fail softening. The pre-fail softener cuts an input sentence into two or three segments that the syntactic analyzer can parse. In case of parsing failure, the post-fail softener uses fragmental trees to recover the parse tree. And the style checker checks the stylistic errors of output English sentences, using error patterns of generated output sentences.

2.1 Morphological Analysis and Tagging

A Korean word form is a unit of Korean sentences. A WF consists of one or more morphemes so that Part-of-Speech (POS) ambiguity and lexical segmentation ambiguity duplicate the difficulty of Korean morphological analysis. The morphemes in a WF have a tight constraint of their POS sequence.

We design the model for Korean morphological analysis based on FST (finite state transducer) and use HMM (hidden Markov model) for POS tagging which resolves the morphological ambiguity generated by the morphological analyzer [8]. The tagger reflects the model of Korean word form to raise the tagging accuracy.

Figure 2 is the proposed HMM for Korean WFs. Both dependency of morphemes in a WF and the dependency between WFs are reflected in a Markov network.

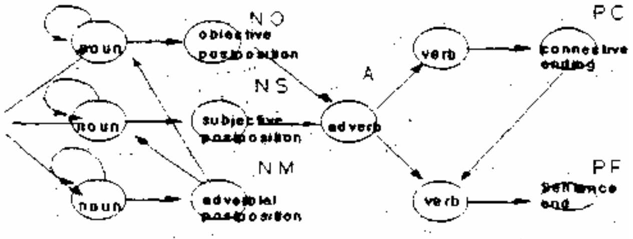


Figure 2 Markov model for Korean word forms

2.2 Pre-Fail Softening

We use the pre-fail softener, prior to parsing Korean sentences, to avoid the failure of parsing too long sentences that the syntactic grammar rule can't cope with. The system segments sentences that are too long to parse into partial fragments that the parser can process. We collect Korean sentences that have 10 or more, and analyze the cutting points. We notice that the connective endings, specific part-of-speech, and some lexical forms can be used as crucial cutting points. We use the workbench to build pre-fail softening rules.

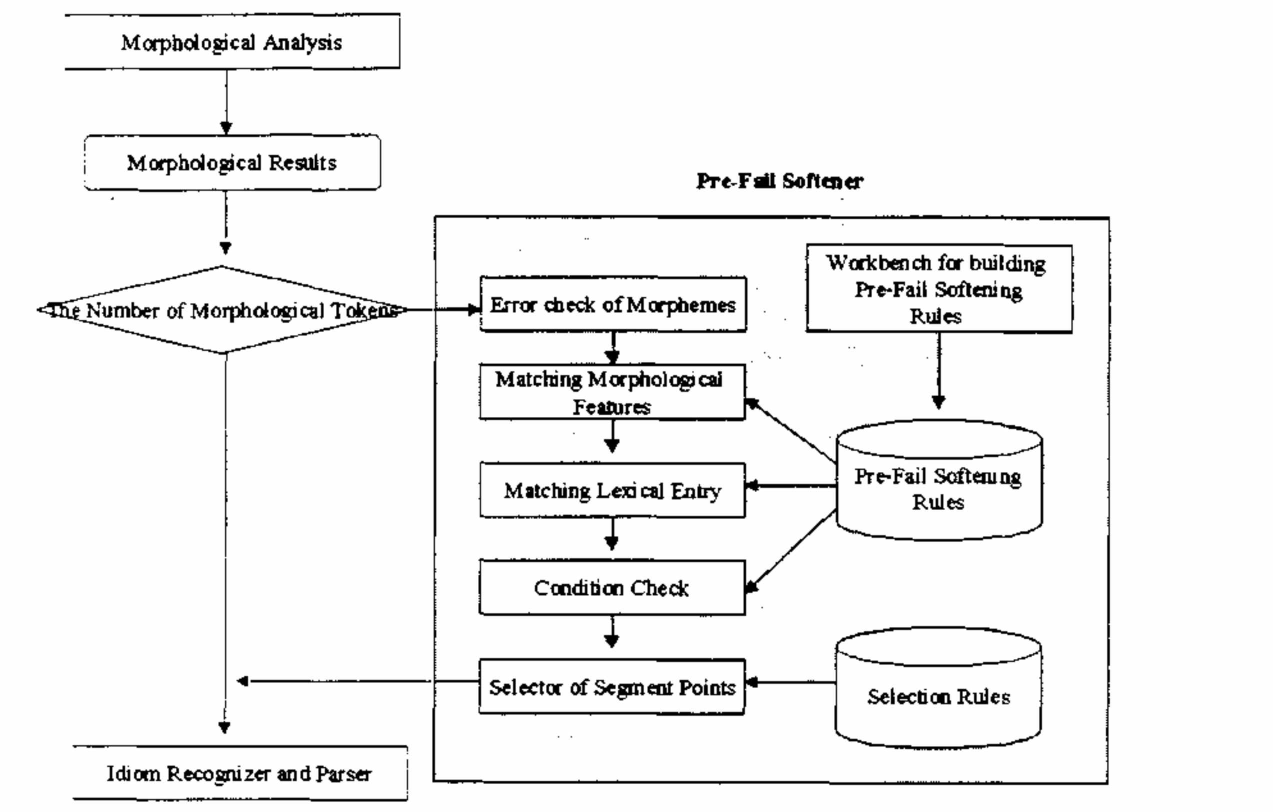

Morphological Analysis
Morphological Results
The Number of Morphological Tokens
Pre-Fail Softener
Error check of Morphemes
Matching Morphological Features
Matching Lexical Entry
Condition Check
Selector of Segment Points
Workbench for building Pre-Fail Softening Rules
Pre-Fail Softening Rules
Selection Rules
Idiom Recognizer and Parser

Figure 3 System Configuration of Pre-fail Softener of FromTo K/E

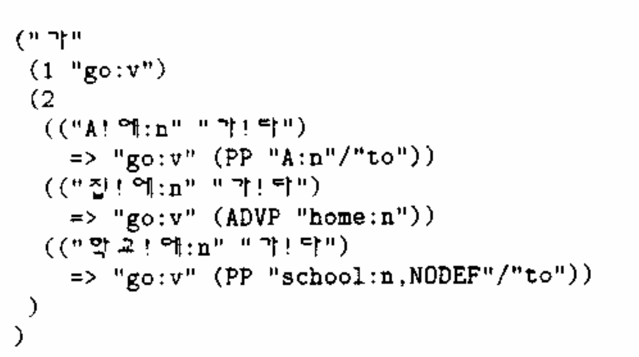
Figure 3 presents the configuration of pre-fail softener of **FromTo K/E**. This module receives the output tokens from the morphological analyzer and tagger. The pre-fail softening is performed prior to idiom recognition and parsing. First, the system checks the number of morphological tokens and decides whether the parser can succeed in parsing the input sentence. If an input sentence is too long to parse, the pre-fail softener tries to segment it into feasible fragments. The softener refers to the pre-fail softening rules to detect cutting points. When the softener detects several cutting points, it uses statistical data, provided by the selection rules, to selects the best one. Finally, the parser can parse the feasible fragments, and combine each parse tree fragments to complete the parse tree of the original input sentence.

2.3 Idiom-Based Syntactic Analysis

Korean is a relatively free word order language. It is frequently pointed out in literature that dependency grammar is quite adequate for parsing the free word order languages. We developed Korean dependency parser including idiomatic pattern recognizer, fail-softeners and parse tree selector [7]. The parser is implemented according to CYK algorithm and the time complexity is $O(n^3)$. The output of the parser can be more than one parse tree. 37 heuristic rules for the best parse selection is applied to select the best one.

One of the most difficult problems in machine translation is how to deal with those expressions that cannot be translated compositionally. And the expressions can be thought of as idiom [16]. In machine translation, the idiom can be defined as the information that can be used for parsing syntactic structure that is not analyzed by the regular grammar [14]. In **FromTo K/E** Korean-English machine translation system, idioms are detected and processed before syntactic analysis. The structural or semantic ambiguity is resolved in terms of the recognized idiomatic information.

Let's take an example: 'go home'(' 집(house)에(to) 가다 (go)') and 'go to school' ('학교(school)에(to) 가다'(go)').



```
("가"
 (1 "go:v")
 (2
  (("A!에:n" "가!다")
    => "go:v" (PP "A:n"/"to"))
  (("집!에:n" "가!다")
    => "go:v" (ADVP "home:n"))
  (("학교!에:n" "가!다")
    => "go:v" (PP "school:n,NODEF"/"to"))
 )
)
```

The expression of an idiom follows the bracket notation using the dependency structure. In idiom base. the symbol "!" distinguishes a grammatical morpheme from a lexical morpheme. And the lexical morphemes are represented as two types, fixed or variable component. For example, the word "학교" (school) in the phrase "학교!에" (to school) is a fixed component and the symbol 'A' in ""A!에:n""가!다"" (to go to A) is an variable component. The left-hand side of "→," represents a transformational information that is used in the transfer. The structural information makes it possible to resolve the semantic ambiguity in the transfer process. We use pre-defined table of the grammatical morpheme representatives for effectively matching the idiomatic expressions.

The idiomatic expression recognizer reduces the input size of the parser by recognizing idioms, frozen expressions and compound nouns. It also helps the transfer process to produce natural English expressions for the Korean idiomatic ones. The recognizer finds all the types of idiomatic expressions in an input sentence on its search index using co-occurrence constraints of both functional and lexical morphemes, part of speech constraints, and syntactic constraints. Recognition reliability is evaluated as the dispersion of the idiomatic expression. As the matched words grow and span distance of the expression decrease, the dispersion is smaller.

2.4 Korean-English Transfer and English Generator

The Korean-English Transfer and the English Phrase Structure Generator integrated into **FromTo K/E** system have 5 subcomponents: idiom translation, basic word translation, translation of particles and endings. structure translation, and English morpheme generation.

In the transfer process, the most likely English target word of ambiguous Korean word is selected on the basis of collocation information, which is extracted from the 400.000 word Korean corpus. The basic structure of transfer dictionary is illustrated as follows:

```
("Keyword"
  (1 "default") ;; Default target word
  (2            ;; Idiomatic expressions
  ("Korean_idiom1" => "English_idiom1")
  ("Korean_idiom2" => "English_idiom2"))
  (3            ;; Collocation
    ("case"
     (("English_verb")
     ("Korean_noun" => "English_noun"))))
```

The English words are selected under consideration of the structural information specified in the structure descriptions of the Korean-English dictionary. The following illustrates how to use and store the structural information in the Korean-English dictionary.

```
"A!을" "B!라고" "부르!다"
 => "call:v" (OBJECT"A:n")(OCOMP"B:n")
```

The described English structure on the right-hand side of the entry above can be represented in a tree notation as follows:

```
HEAD ->  call
     OBJECT -> A
     OCOMP -> B
```

After the generation of English words, we transfer the Korean dependency structure with English words into English one. This process is necessary for the generation of the correct English structure, since Korean and English do not share the same structure. Thus, one should delete some nodes in the Korean dependency structures on the one hand, and add some nodes to English dependency nodes on the other.

For the generation of the final English structure, we put one further transfer module, where the English dependency structures are transformed into the English phrase structure trees. We use GWL (Grammar Writing Language) in writing the dependency structure transfer rules and the English phrase structure generation rules [6]. The English morpheme generation finally generates the correct English words, and regulate some grammatical awkwardness of the translated sentences.

2.5 Rules and Dictionaries

For the analysis of the syntactic structures of Korean, we lexico-syntactic properties develop a set of syntactic rules basically using Dependency Grammar. The reason for the choice of Dependency Grammar as grammatical framework is the belief that Dependency Grammar is adequate for the description of syntactic structures of such a language as Korean, which has a relatively free word order and rich morphologically distinguished surface features.

The parser often produces more than 1 parsed tree candidates. In order to choose just one correct parsed tree, we collect a set of parse tree selection rules, which heuristically choose the best single parse tree.

Table 1 Summary of rules

syntactic rules	206 rules
Korean-English dependency structure transfer rules	170 rules
English phrase structure generation rules	124 rules
sum	500 rules

Table 2 Summary of dictionaries

Korean dictionary	150.000
Korean-English dictionary	150.000
idiom dictionary	20.000
sum	320.000 entries

The number of the parse tree selection rules amounts to 38 in the **FromTo K/E** machine translation system. The parse tree selection rules are applied in order of the pre-defined priority hierarchy.

The computational linguistic rules developed for the **FromTo K/E** are summarized as follows: 206 syntactic rules, 170 Korean-English dependency structure transfer rules, and 124 English phrase structure generation rules.

3 Experiment and Evaluation

For the evaluation of the system described above, we use a test suite, which consists of 2000 sentences covering some language-specific linguistic phenomena. The basic properties of this test suite can statistically be summarized as in Table 3:

We empirically evaluate the system by using this test suite. As can be seen in Table 4. we classify the translated sentences of test suite into five classes on the basis of their quality, i.e. understandability of the translated sentences. We summarize the final result in Table 4.

The results of the experimental evaluation show that the degree of understandability for sample 2000 sentences amounts to 2.67, indicating that the meaning of the translated English sentences is almost clear to users, but the sentences still include minor grammatical or stylistic errors up to max. 30% of the whole words. The typical errors are found, among others, the treatment of definite and indefinite articles, the treat-

Table 3 Analysis of the test suite

criteria	result
number of sentences	2000 sentences
whole number of words	1.5643 words
average number of words per sentence	7.8 words
average number of phonemes per sentence	25.3 phonemes
average number of phonemes per words	3.2 phonemes

Table 4 Evaluation of the translated sentences

credit	evaluation criteria	results
4 (Perfect)	The meaning of the sentence is perfectly clear. No grammatical error of word translation exists.	310 (16%)
3 (Good)	The meaning of the sentence is almost clear. Minor grammatical error of word translation exists (20% of the whole words of the sentence).	930 (74%)
2 (OK)	The meaning of the sentence can only be understood after several repeated readings. Some grammatical error of word translation exists (20%-30% of the whole words of the sentence).	560 (5.6%)
1 (Poor)	The meaning of the sentence can be merely guessed only after a lot of readings.	190 (4.5%)
0 (Fail)	The meaning of the sentence cannot be guessed at all.	10 (0.5%)

Table 5 Experiment for Pre-Fail Softening

word count	sentences	segmented sentences	successful segmentation	failed segmentation
10	142	66	54	12
11	114	56	46	10
12	83	49	39	10
13	70	40	36	4
14	55	27	24	3
Sum	464	238	199	39

ment of agreement in person, tense, number in the English sentences generation. Nevertheless the users do not find any particular difficulties in understanding the translated English sentences.

The final understandability note for the system is 2.67: ((4 x 310) + (3 x 930) + (2 x 560) + (1 x 190 + 0) ÷ 2000.

Since the parser often fails to parse the Korean sentences having 10 or more words in Korean-English machine translation, it can be said that the pre- and the post-fail softening modules are interesting, and deserve to be evaluated. In this sense, we empirically evaluate the effectiveness of the Pre-fail Softner by using the long sentences with 10 or more words. Table 5 shows the rate of the successful and the failed segmentation.

If one accept the view that a successful Korean-English machine translation system presumes a highly efficient robust Korean parser. "Fail Softening"' technology, i.e. the long sentence segmentation and the recovery of failed parse trees, turns out to be very efficient modules in **FromTo K/E** .

4 Concluding Remarks

We describe in this paper the final report of **FromTo K/E**, a rule-based Korean-English machine translation system adapting transfer methodology, and experimentally evaluate it.

In sum, **FromTo K/E** is the first prototype of Korean-English machine translation system. This system integrates some recent technical innovations for tackling the long-lasting language-typological and ambiguity-related problems in the field of Korean-English machine translation. High accuracy HMM-based tagger reflecting the Korean WF structure, idiomatic expression recognizer, and collocation-based role selection of ambiguous functional morphemes effectively reduce the load of parser by minimizing the POS ambiguity and reducing the input units. Collocation information extracted from 400,000 Korean corpus has effectively supported the word sense disambiguation.

The results of the experimental evaluation show that the degree of understandability for sample 2000 sentences amounts to 2.67. indicating that the meaning of the translated English sentences is almost clear to users, but the sentences still include minor grammatical or stylistic errors up to max. 30% of the whole words.

We have much more things to do to make the system more practical. We recognize the following points as our future works.

- Integrating a rule-based tagger with our HMM-based tagger as a post-processor of the HMM-based tagger to minimize the tagging error
- Ranking of too many ambiguities in Korean syntactic analysis
- Testing and tuning the dependency parser for much more 2,000 Korean sentences
- Minimizing the English syntactic error and morphological generation error
- Expanding the lexicon up to 50.000 words for practical MT system
- Robust processing for failed or ill-formed sentences

The Korean-English machine translation system **FromTo K/E** has been originally developed as stand-alone system, but extended its functionality to Korean-English machine translation web server.

References

[1] Brill. E. (1995). "Transformation-based Error-driven Learning and Natural Language Processing: a Case Study in Part-of-Speech Tagging". *Computational Linguistics.*

[2] Cho, J., Cho, Y., & Kim. G. (1996). "Automatic Acquisition of N(oun)-C(ase)-P(redicate) Information from POS-Tagged Corpus". *An International Journal of the Oriental Language Computer Society.*

[3] Jung. H., Kim, Y., Kim, T., & Park, D.I. (1998). "The Construction of English/Korean Bilingual Pattern from Morpheme-based Uninterrupted Pattern Extraction" . in *Proceedings of the 9th KIPS Spring Conference* (written in Korean).

[4] Katoh, N. & Aizawa. T. (1995). "Machine Translation of Sentences with Fixed Expression", in *Proceedings of the 4th Applied Natural Language Processing.*

[5] Kim, Y.-T., *et al.* (1997). *A Study on the Korean Syntax Analysis and Transfer for Korean-to-English Machine Translation.* Technical Report of Systems Engineering Research Institute (written in Korean).

[6] Kwon, C.-J., *et al.* (1990). "Grammar Writing language (GWL) in MATES/EK". in *Proceedings of Pacific Rim International Conference on Artificial Intelligence.* 263-267.

[7] Lee, H. (1994). *Recognition of Korean-English Bilingual Idioms using Idiom Dispersion Characteristics.* Ph.D. thesis. Seoul National University (written in Korean).

[8] Lee, S., Seo, J. & Oh, Y.-H. (1995). "A Korean Part-of-Speech Tagging System with Handling Unknown Words", in *Proceeding of the International Conference on Computer Processing of Oriental Languages.* 164-171.

[9] Nasukawa, T. (1996). "Robust Parsing Based on Discourse Information", in *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics,* 39-46.

[10] Park, D.-L. *et al.* (1997). *Development of Machine Translation System between Korean and English.* Technical Report of Korea Ministry of Information and Communication (written in Korean).

[11] Park, H.-R., *et al.* (1997). *Development of a Translation Unit Recognition System for Korean-to-English Machine Translation.* Technical Report of Systems Engineering Research Institute (written in Korean).

[12] Schenk, A. (1986). "Idiom in the Rosetta Machine Translation". in *Proceedings of International Conference on Computational Linguistics.*

[13] Yang, J.-H. & Kim, Y.-T. (1993). "Corpus-based Resolution of Functional Ambiguity in Parsing Korean", in *Proceedings of the InfoScience-93,* 517-524.

[14] Yoon, D.-H. (1993). *Korean Parsing using Idiomatic Information.* Ph.D. thesis, Seoul National University (written in Korean).

[15] Yoon, S.-H. (1992). "Idiomatical and Collocational Approach to Machine Translation", in *Proceeding of the 2nd Pacific Rim International Conference on Artificial Intelligence.*

[16] Yoon, S.-H. (1994). "Efficient Parser to Find Bilingual Idiomatic Expressions for English-Korean Machine Translation", in *Proceedings of International Conference on Computer Processing of Oriental Languages.*

[17] Yuh, S., Jung, H., Kim, Y., Choi, S.-K.. Kim, T., Park, D.-I., & Seo, J. (1998). Esoro K/E: A Korean-English Machine Translation System, in *Proceedings of the International Association of Science and Technology for Development* (IASTED) *International Conference Artificial Intelligence and Soft Computing.* 207-210.