Interlingua-Based Broad-Coverage Korean-to-English Translation in CCLINC

Young-Suk Lee
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420
U.S.A
1-781-981-2703
YSL@LL.MIT.EDU

Wu Sok Yi
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420
U.S.A
1-781-981-4609
WUYI@LL.MIT.EDU

Stephanie Seneff
MIT/LCS
77 Mass Avenue
Cambridge, MA 02673
U.S.A
1-617-254-0456
SENEFF@LCS.MIT.EDU

Clifford J. Weinstein
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420
U.S.A
1-781-981-7621
CJW@LL.MIT.EDU

ABSTRACT

At MIT Lincoln Laboratory, we have been developing a Korean-to-English machine translation system CCLINC (Common Coalition Language System at Lincoln Laboratory). The CCLINC Korean-to-English translation system consists of two core modules, language understanding and generation modules mediated by a language neutral meaning representation called a semantic frame. The key features of the system include: (i) Robust efficient parsing of Korean (a verb final language with overt case markers, relatively free word order, and frequent omissions of arguments). (ii) High quality translation via word sense disambiguation and accurate word order generation of the target language. (iii) Rapid system development and porting to new domains via knowledge-based automated acquisition of grammars. Having been trained on Korean newspaper articles on "missiles" and "chemical biological warfare," the system produces the translation output sufficient for content understanding of the original document.

1. SYSTEM OVERVIEW

The CCLINC The CCLINC Korean-to-English translation system is a component of the CCLINC Translingual Information System, the focus languages of which are English and Korean, [11,17]. Translingual Information System Structure is given in Figure 1.

Given the input text or speech, the language understanding system parses the input, and transforms the parsing output into a language neutral meaning representation called a *semantic frame*, [16,17]. The semantic frame — the key properties of which will be discussed in Section 2.3 — becomes the input to the generation system. The generation system produces the target to the generation system, the semantic frame can be utilized for other applications such as translingual information extraction and language translation output after word order arrangement, vocabulary replacement, and the appropriate surface form realization in the target language, [6]. Besides serving as the input question-answering, [12].* In this paper, we focus on the Korean-to-English text translation component of CCLINC.[1]

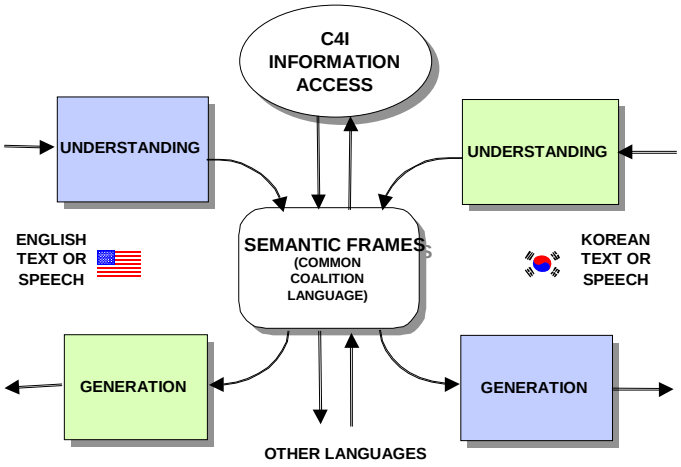

Figure 1. CCLINC Translingual Information System Structure

2. ROBUST PARSING, MEANING REPRESENTATION, AND AUTOMATED GRAMMAR ACQUISITION

* This work was sponsored by the Defense Advanced Research Project Agency under the contract number F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Air Force.

[1] For other approaches to Korean-to-English translation, the readers are referred to *Korean-to-English translation* by Egedi, Palmer, Park and Joshi 1994, a transfer-based approach using synchronous tree adjoining grammar, [5], and Dorr 1997, a small-scale interlingua-based approach, using Jackendoff's lexical conceptual structure as the interlingua, [4].

1.1 Robust Parsing

The CCLINC parsing module, TINA [16], implements the top-down chart parsing and the best-first search techniques, driven by context free grammars rules compiled into a recursive transition network augmented by features, [8]. The following properties of Korean induce a great degree of ambiguity in the grammar: (i) relatively free word order for arguments --- given a sentence with three arguments, subject, object, indirect object, all 6 logical word order permutations are possible in reality, (ii) frequent omissions of subjects and objects, and (iii) the strict verb finality, [10]. Due to the free word order and argument omissions, the first word of an input sentence can be many way ambiguous --- it can be a part of a subject, an object, and any other post-positional phrases.[2] The ambiguity introduced by the first input word grows rapidly as the parser processes subsequent input words. Verbs, which usually play a crucial role in reducing the ambiguity in English by the subcategorization frame information, are not available until the end, [1,3,11].

Our solution to the ambiguity problem lies in a novel grammar writing technique, which reduces the ambiguity of the first input word. We hypothesize that (i) the initial symbol in the grammar (i.e. **S**entence) always starts with the single category ***generic_np***, the grammatical function (subject, object) of which is undetermined. This ensures that the ambiguity of the first input word is reduced to the number of different ways the category *generic_np* can be rewritten. (ii) The grammatical function of the *generic_np* is determined after the parser processes the following case marker via a trace mechanism.[3]

Figure 2 illustrates a set of sample context free grammar rules, and Figure 3 (**on the next page**) is a sample parse tree for the input sentence "*URi Ga EoRyeoUn MunJe Reul PulEox Da* (We solved a difficult problem)."[4]

(i) sentence → generic_np clause sentence_marker
(ii) clause → subject generic_np object verbs
(iii) subject → subj_marker np_trace

Figure 2. Sample context free grammar rules for Korean

The *generic_np* dominated by the initial symbol *sentence* in (i) of Figure 2 is parsed as an element moved from the position occupied by *np_trace* in (iii), and therefore corresponds to the category *np_trace* dominated by *subject* in Figure 3 (**placed on the next page for space reasons**). All of the subsequent *generic_np*'s, which are a part of a direct object, an indirect object, a post-positional phrase, etc. are unitarily handled by the same trace mechanism. By hypothesizing that all sentences start with *generic_np*, the system can parse Korean robustly and efficiently. The trace mechanism determines the grammatical function of *generic_np* by repositioning it after the appropriate case marker.

Utilization of overt case markers to improve the parsing efficiency precisely captures the commonly shared intuition for parsing relatively free word order languages with overt case markers such as Korean and Japanese, compared with parsing relatively strict word order languages with no overt case markers such as English: In languages like English, the verb of a sentence plays the crucial role in reducing the ambiguity via the verb subcategorization frame information on the co-occuring noun phrases, [1,3,11]. In languages like Korean, however, it is typically the case marker which identifies the grammatical function of the co-occuring noun phrase, assuming the role similar to that of verbs in English. The current proposal is the first explicit implementation of this intuition, instantiated by the novel idea that all noun phrases are moved out of the case marked phrases immediately following them.

2.2 Meaning Representation and Generation

The CCLINC Korean-to-English translation system achieves high quality translation by (i) robust mapping of the parsing output into the semantic frame, and (ii) word sense disambiguation on the basis of the selection preference between two grammatical relations (verb-object, subject-verb, head-modifier) easily identifiable from the semantic frame, [13]. The former facilitates the accurate word order generation of various target language sentences, and the latter, the accurate choice of the target language word given multiple translation candidates for the same source language word. Given the parsing output in Figure 3, the system produces the semantic frame in Figure 4:[5]

[2] Post-positional phrases in Korean correspond to pre-positional phrases in English. We use the term post-positional phrase to indicate that the function words at issue are located after the head noun.

[3] The hypothesis that all sentences start with a single category *generic_np* is clearly over simplified. We can easily find a sentence starting with other elements such as coordination markers which do not fall under *generic_np*. For the sentences which do not start with the category *generic_np*, we discard these elements for parsing purposes. And this method has proven to be quite effective in the overall design of the translation system, especially due to the fact that most of *non generic_np* sentence initial elements (e.g. coordination markers, adverbs, etc.) do not contribute to the core meaning of the input sentence.

[4] Throughout this paper, "subj_marker" stands for "subject marker", and "obj_marker", "object marker".

[5] Strictly speaking, the meaning representation in Figure 4 is not truly language neutral in that the terminal vocabularies are represented in Korean rather than in interlingua vocabulary. It is fairly straightforward to adapt our system to produce the meaning representation with the terminal vocabularies specified by an interlingua. However, we have made a deliberate decision to leave the Korean vocabularies in the representation largely (1) to retain the system efficiency for mapping parsing output into meaning representation, and (2) for unified execution of automation algorithms for both Korean-to-English and English-to-Korean translation. And we would like to point out that this minor compromise in meaning representation still ensures the major benefit of interlingua approach to machine translation, namely, *2 x N* sets of grammar rules for N language pairs, as opposed to 2^N.

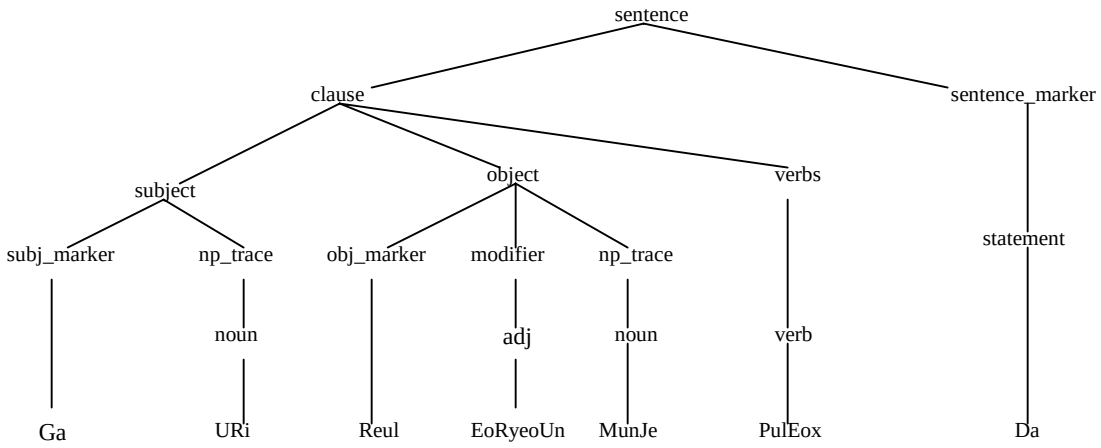

Figure 1. Parse Tree for the Sentence *URi Ga EoRyeoUn MunJe Reul PulEox*

{c statement
:topic {q pronoun
:name "**URi**" }
:pred {p **pul_v**
:topic {q problem
:name "**MunJe**"
:pred {p EoRyeoUn } } }

Figure 4. Semantic Frame for the input sentence *"URi Ga EoRyeoUn MunJe Reul PulEox Da."*

The semantic frame captures the core predicate-argument structure of the input sentence in a hierarchical manner, [9,10] (i.e. the internal argument, typically object, is embedded under the verb, and the external argument, typically subject, is at the same hierarchy as the main predicate, i.e. verb phrase in syntactic terms). The predicate and the arguments along with their representation categories are bold-faced in Figure 4. With the semantic frame as input, the generation system generates the English translation using the grammar rules in (1), and the Korean paraphrase using the grammar rules in (2).

The semantic frame captures the core predicate-argument structure of the input sentence in a hierarchical manner, [9,10] (i.e. the internal argument, typically object, is embedded under the verb, and the external argument, typically subject, is at the same hierarchy as the main predicate, i.e. verb phrase in syntactic terms). The predicate and the arguments along with their representation categories are bold-faced in Figure 4. With the semantic frame as input, the generation system generates the English translation using the grammar rules in (1), and the Korean paraphrase using the grammar rules in (2).

(1) a. statement :topic :predicate
b. pul_v **:predicate :topic**
(2) a. statement :topic :predicate
b. pul_v **:topic :predicate**

(1b) and (2b) state that the topic category for the object follows the verb predicate in English, whereas it precedes the verb predicate in Korean.

The predicate-argument structure also provides a means for word sense disambiguation, [13,15]. The verb *pul_v* is at least two-way ambiguous between *solve* and *untie*. Word sense disambiguation is performed by applying the rules, as in (3).

(3) a .pul_v b .pul_v
problem pul+solve_v thread pul+untie_v

(3a) states that if the verb *pul_v* occurs with an object of type **problem**, it is disambiguated as *pul+solve_v*. (3b) states that the verb occurring with an object of type **thread** is disambiguated as *pul+untie_v*. The disambiguated verbs are translated into *solve* and *untie*, respectively, in the Korean-to-English translation lexicon.

1.2 Knowledge-Based Automated Acquisition of Grammars

To overcome the knowledge bottleneck for robust translation and efficient system porting in an interlingua-based system [7], we have developed a technique for automated acquisition of grammar rules which leads to a simultaneous acquisition of rules for (i) the parser, (ii) the mapper between the parser and the semantic frame, and (iii) the generator.

The technique ***utilizes*** a list of words and their corresponding parts-of-speech in the corpus as the knowledge source, ***presupposes*** a set of knowledge-based rules to be derived from a word and its part-of-speech pair, and gets ***executed*** according to the procedure given in Figure 5. The rationale behind the technique is that (i) given a word and its part-of-speech, most of the syntactic rules associated with the word can be automatically derived according to the **projection principle** (the syntactic

representation must observe the subcategorization properties of each lexical item) and the **X-bar schema** (major syntactic categories such as N, V, Adj, Adv project to the same syntactic structures) in linguistic theories, [2], and (ii) the mapping from the syntactic structure to the semantic frame representation is algorithmic. The specific rules to be acquired for a language largely depend on the grammar of the language for parsing. Some example rules acquired for the verb *BaiChiHa* (arrange) in Korean — consistent with the parsing technique discussed in Section 2.1 — are given in (4) through (7).

Initialization: Create the list of words and their parts-of-speech in the corpus.

Grammar Update: For each word and its associated part-of-speech, check to see whether or not the word and the rules associated with the corresponding part-of-speech occur in each lexicon and grammar.

If they already occur, do nothing.

If not:

(i) Create the appropriate rules and vocabulary items for each entry.

(ii) Insert the newly created rules and vocabulary items into the appropriate positions of the grammar/lexicon files for the parser, the grammar file for the mapper between the parser and the semantic frame, and the grammar/lexicon files for the generator .

Figure 5. Automated Gammar Acquistion Procedure

(4) Rules for the parser[6]
.verbs
[negation] vBaiChiHa [negation] [aspect] [tense] [auxiliary] [negation] [aspect] [tense] [and_verbs] [or_verbs]

.vBaiChiHa
#BaiChiHa

(5) Rules for the mapper from the parser to the semantic frame
.bachiha_v
vBaiChiHa

(6) Lexicon for the generation vocabulary
baichiha_v V2 “arrang”
V2 V “e” ING “ing” PP “ed” THIRD “es” ROOT “e” PAST “ed” PASSIVE “ed”

(7) Rules for the generation grammar
baichiha_v :predicate :conj :topic :sub_clause
np-baichiha_v :noun_phrase :predicate :conj :topic :sub_clause

The system presupposes the flat phrase structure for a sentence in Korean, as shown in Figure 3, and therefore the rules for the verbs do not require the verb subcategorization information, as in (4). The optional elements such as [negation], [tense], etc. are possible prefixes and suffixes to be attached to the verb stem, illustrating a fairly complex verb morphology in this language. The rules for the generation grammar in (7) are the subcategorization frames for the verb *arrange* in English, which is the translation of the Korean verb *baichiha_v*, as given in (6).

The current technique is quite effective in expanding the system’s capability when there is no large syntatically annotated corpus available from which we can derive and train the grammar rules, [14], and applicable across languages in so far as the notion of part-of-speech, the projection principle and the X-bar schema is language independent. With this technique, manual acquisition of the knowledge database for the overall translation system is reduced to the acquisition of (i) the bilingual lexicon, and (ii) the corpus specific top-level grammar rules which constitute less than 20% of the total grammar rules in our system. And this has enabled us to produce a fairly large-scale interlingua-based translation system within a short period of time. One apparent limitation of the technique, however, is that it still requires the manual acquisition of corpus-specific rules (i.e. the patterns which do not fall under the linguistic generalization). And we are currently developing a technique for automatically deriving grammar rules and obtaining the rule production probabilities from a syntactically annotated corpus.

3. EVALUATION AND RESEARCH ISSUES

We have trained the system with about 1,600 Korean newspaper articles on “*missiles*” and “*chemical biological warfare*”, as in Table 1.

Table 1. Korean-to-English translation training data statistics

# of articles	# of sents/article	# of words/sent	# of distinct words
1,631	24	17	15,220

For quality evaluation, we have adopted a 5-point scale evaluation score, defined as follows. **Score 4**: Translation is both accurate and natural. **Score 3**: Translation is accurate with minor grammatical errors which do not affect the intended meaning of the input, e.g. morphological errors such as “*swam* vs. *swimmed*.” **Score 2**: Translation is partially accurate, and sufficient for content understanding. Most errors are due to inaccurate word choice, inaccurate word order, and partial translation. **Score 1**: Translation is word-for-word, and partial content understanding is

---

[6] The rules for the parser for the verb *tell* in English are given below, to illustrate the dependency of the rules acquired to the specific implementation of the grammar of the language for parsing:

.vp_tell
vtell [adverb_phrase] dir_object [v_pp]
vtell [adverb_phrase] indir_object dir_object
vtell [adverb_phrase] dir_object v_to_pp [v_pp]
vtell [adverb_phrase] dir_object that_clause
vtell [and_verb] [or_verb] [adverb_phrase] dir_object wh_clause

The contrast in complexity of verb rules in (4) for Korean, and (i) for English, reflects the relative importance of the role played by verbs for parsing in each language. That is, verbs play the minimal role in Korean, and the major role in English for ambiguity reduction and efficiency improvement.

possible. **Score 0**: There is no translation output, or no content understanding is possible.

We have performed the quality evaluation on 410 clauses from the training data, and 80 clauses from the test data. We have conducted the evaluation in 3 phases. **Eval 1**: Baseline evaluation after grammar and lexicon acquisition. **Eval 2**: Evaluation after augmenting word sense disambiguation rules. **Eval 3**: Evaluation after augmenting word sense disambiguation rules and accurate word order generation rules. The purpose of the 3-phase evaluation was to examine the contribution of parsing, word sense disambiguation and accurate word order generation to the overall translation quality. Once the score had been assigned to each clause, the translation score was obtained by the formula: (Sum of the scores for each clause * 25) / Number of clauses evaluated.

Evaluation results are shown in Table 2 and Table 3 in terms of parsing coverage **(P)** and the translation score **(T).**[7]

Table 2. Translation Quality Evaluation on Training Data

Eval 1		Eval 2		Eval 3	
P	T	P	T	P	T
92	**58**	94	**69**	94	**74**

Table 3. Translation Quality Evaluation on Test Data

Eval 1		Eval 2		Eval 3	
P	T	P	T	P	T
79	**55**	89	**63**	89	**65**

For both training and test data, the baseline translation quality score is over 50, sufficient for content understanding of the documents. Word sense disambiguation (Eval 1 vs. Eval 2) increases the translation score by about 10%, indicating that effective word sense disambiguation has a great potential for improving the translation quality.

We would like to point out that the evaluations reported in this paper are performed on clauses rather than sentences (which often consist of more than one clause). In a very recent evaluation, we have found out that evaluations on sentences decrease the overall translation score about by 15. Nevertheless, the translation quality is still good enough for content understanding with some effort. The primary cause for the lower translation scores when the evaluation unit is a sentence as opposed to a clause is due to either an incorrect clause boundary identification, or some information (e.g. missing arguments in embedded clauses) which cannot be easily recovered after a sentence is fragmented into clauses. This has led to the ability to handle complex sentences as the primary research issue, and we are working out the solution of utilizing syntactically annotated corpus for both grammar and probability acquisition, as discussed in Section 2.3.

[7] We would like to note that the evaluation reported here was a self-evaluation of the system by a system developer, primarily to identify the key research issues in system development. We will report evaluation results by non system developers who have no knowledge of Korean in the future. A system evaluation by a non-bilingual speaker will avoid the issue of implicitly utilizing the knowledge the evaluator has about the source language in the evaluation process.

4. SUMMARY AND ONGOING WORK

We have described the key features of the CCLINC interlingua-based Korean-to-English translation system which is capable of translating a large quantity of Korean newspaper articles on missiles and chemical biological warfare in real time. Translation quality evaluations on the training and test data indicate that the current system produces translation sufficient for content understanding of a document in the training domains. The key research issues identified from the evaluations include (i) parsing complex sentences, (ii) automated acquisition of word sense disambiguation rules from the training corpus, and (iii) development of discourse module to identify the referents of missing arguments. Our solution to the key technical challenges crucially draws upon the utilization of annotated corpora: For complex sentence parsing, we acquire both rules and rule production probabilities from syntactically annotated corpus. For automated word sense disambiguation, we utilize a sense-tagged corpus to identify various senses of a word, and obtain probabilities for word senses in various contexts. For discourse understanding, we are developing an algorithm for our 2-way speech translation work, [12], and plan to expand the module for document translations.

5. ACKNOWLEDGMENTS

We would like to acknowledge Dr. Jun-Tae Yoon, who provided us with a high-quality robust Korean morphological analyzer called *morany* during his stay at the Institute for Research in Cognitive Science, University of Pennsylvania as a postdoctoral fellow. *Morany* has served as a pre-processor of the understanding module in the CCLINC Korean-to-English translation system.

6. REFERENCES

[1] Srinivas Bangalore and Aravind Joshi. "Some Novel Applications of Explnation-Based Learning for Parsing Lexicalized Tree-Adjoining Grammars," *Proceedings of 33rd Association for Computational Linguistics*. pp. 268—275. 1995.

[2] Noam Chomsky. *Barriers*. Linguistic Inquiry Monograph 13. MIT Press, Cambridge, MA. 1986.

[3] Michael Collins. Three Generative, Lexicalized Models for Statistical Parsing. *Procceedings of the 35th Annual Meeting of ACL*. pp. 16—23. Madrid, Spain. July. 1997.

[4] Bonnie Dorr. "LCS-based Korean Parsing and Translation," Ms. Institute for Advanced Computer Studies and Department of Computer Science, University of Maryland. 1997.

[5] Diana Egedi, Martha Palmer, H-S. Park, Aravind Joshi. "Korean to English Translation Using Synchronous TAGs," *Proceedings of the First Conference of the Association for Machine Translation in the Americas*. pp. 48—55. Columbia, Maryland. October 1994.

[6] James Glass, Joe Polifroni and Stephanie Seneff. "Multilingual Language Generation across Multiple Domains,"

Proceedings of International Conference on Spoken Language Processing, pp. 983—986. Yokohama, Japan. September, 1994.

[7] W.J. Hutchins and H.L. Somers. *An Introduction to Machine Translation*. Academic Press. London. 1992.

[8] James Allen. *Natural Language Understanding*, 2nd Edition. Benjamin-Cummings Publisher. 1995

[9] Ken Hale. "Preliminary Remarks on Configurationality," *Proceedings of NELS 12*, pp. 86—96. 1982.

[10] Young-Suk Lee. *Scrambling as Case-Driven Obligatory Movement*. PhD Thesis (IRCS Report No.: 93-06). University of Pennsylvania. 1993.

[11] Young-Suk Lee, Clifford Weinstein, Stephanie Seneff, Dinesh Tummala, "Ambiguity Resolution for Machine Translation of Telegraphic Messages," *Proceedings of the 35th Annual Meeting of ACL*. pp. 120—127. Madrid, Spain. July 1997.

[12] Young-Suk Lee and Clifford Weinstein. "An Integrated Approach to English-Korean Translation and Translingual Information Access," *Proceedings of CSTAR Workshop*. Schwetzingen, Germany. September, 1999.

[13] Young-Suk Lee, Clifford Weinstein, Stephanie Seneff, Dinesh Tummala. "Word Sense Disambiguation for Machine Translation in Limited Domains," Manuscript. Information Systems Technology Group. MIT Lincoln Laboraotry. January 1999.

[14] Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. "Building a large annotated corpus of English: the Penn Treebank," *Computational Linguistics 19 (2)*. pp. 313—330. 1993.

[15] Philip Resnik. "Semantic Similarity in a Taxonomy: An Information-Based Measure and Its Application to Problems of Ambiguity in Natural Language," *Journal of Artificial Intelligence Research (JAIR) 11*. pp. 95—130. 1999.

[16] Stephanie Seneff. "TINA: A Natural Language System for Spoken Language Applications," *Computational Linguistics 18 (1)*. pp. 61—92. 1992.

[17] Clifford Weinstein, Young-Suk Lee, Stephanie Seneff, Dinesh Tummala, Beth Carlson, John T. Lynch, Jung-Taik Hwang, Linda Kukolich. "Automated English-Korean Translation for Enhanced Coalition Communications," *The Lincoln Laboratory Journal 10 (1)*. pp. 35—60. 1997.