Automatic English-to-Korean Text Translation of Telegraphic Messages in a Limited Domain[1]

Clifford Weinstein
Lincoln Laboratory, MIT
Lexington, MA 02173
USA
cjw@sst.ll.mit.edu

Dinesh Tummala
Lincoln Laboratory, MIT
Lexington, MA 02173
USA
tummala@sst.ll.mit.edu

Young-Suk Lee
Lincoln Laboratory, MIT
Lexington, MA 02173
USA
ysl@sst.ll.mit.edu

Stephanie Seneff
MIT LCS,SLS
Cambridge, MA 02139
USA
seneff@lcs.mit.edu

Abstract

This paper describes our work-in-progress in automatic English-to-Korean text translation. This work is an initial step toward the ultimate goal of text and speech translation for enhanced multilingual and multinational operations. For this purpose, we have adopted an *interlingua* approach with natural language understanding (TINA) and generation (GENESIS) modules at the core. We tackle the ambiguity problem by incorporating syntactic and semantic categories in the analysis grammar. Our system is capable of producing accurate translation of complex sentences (38 words) and sentence fragments as well as average length (12 words) grammatical sentences. Two types of system evaluation have been carried out: one for grammar coverage and the other for overall performance. For system robustness, integration of two subsystems is under way: (i) a rule-based part-of-speech tagger to handle unknown words/constructions, and (ii) a word-for-word translator to handle other system failures.

1 Introduction

The overall goal of our translation work is automatic text and speech translation for limited-domain multilingual applications. The primary target application is enhanced communication among military forces in a multilingual coalition environment where translation utilizes a Common Coalition Language as a military interlingua. Our development effort was initiated with a speech-to-speech translation system, called CCLINC (Tummala et al., 1995), which consists of a modular, multilingual structure including speech recognition, language understanding, language generation, and speech synthesis in each language. The system architecture of CCLINC is given in Figure 1. Note that the system design provides for verification of the system's understanding of each utterance to the originator, in a paraphrase in the originator's language, before transmission on the coalition network.

This paper describes our current work in automatic English-to-Korean text translation of telegraphic military messages,[2] which is an initial step toward the ultimate goal of producing high quality text/speech translation output.[3] The core of our text translation system consists of an analysis module and a generation module. The analysis module produces a semantic frame, which is an *interlingua* representation of the input sentence. The intractable ambiguities of natural language are overcome by restricting the domain and the grammar rules which specify the semantic co-occurrence restrictions of head categories. The structural difference between the source (English) and the target (Korean) language is easily captured by the flexible *interlingua* representation and the strictly modularized target language grammar template, external to the core generation system. The simplicity of the system enables us to detect problems and provide solutions easily. Currently the system has a vocabulary of 1427 words. The system runs on a SPARC 10 workstation. The Korean translation outputs are displayed on a *hangul* window running on UNIX. In addition, we are in the process of porting the system to a Pentium laptop running on Linux.

This paper is organized as follows: In section 2 we describe our system architecture, along with the grammar rules which drive the core system. In section 3 we summarize the characteristics of our source language text comprised of naval operational messages. In section 4 we give our system evaluation. In section 5 we discuss the integration of two subsystems for system robustness: rule-based part-of-speech tagger to handle unknown words/constructions, and a word-for-word translator to produce partial translations in the event of system failure. Finally we summarize the paper in Section 6.

2 System Description

The core of our translation system consists of two modules: the understanding/analysis module, TINA, and the generation module, GENESIS.[4] These modules are driven by a set of files which specify the source and target language grammars. The process flow of our text translation system is given in Figure 2.

2.1 Language Understanding

The language understanding system, TINA, described at length in (Seneff, 1992), integrates key ideas from context free grammar, augmented transition network and unification concepts. With the context free grammar rules of English as input, the system produces the parse tree of an input sentence. The parse tree is then mapped onto a semantic frame, which plays the role of an interlingua. The parse tree and the semantic frame of the input sentence "0819 z uss sterett

[1]This work was sponsored by the Defense Advanced Research Projects Agency. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Air Force.

[2]We are also working on Korean-to-English text translation on the same domain, which we do not include in this paper.

[3]Refer to (Kim, 1994) for other ongoing efforts in English/Korean text translation including (Choi, 1994). See (Lee, 1995) for speech translation work with Korean as the source language.

[4]Both modules are developed under ARPA sponsorship by the Spoken Language Systems Group at the MIT Laboratory for Computer Science.

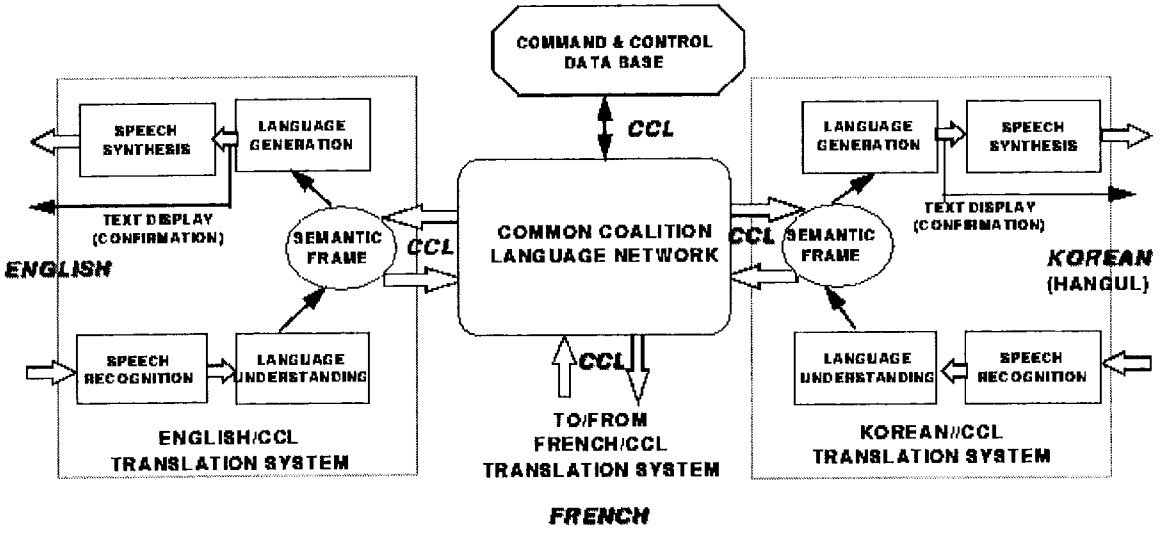

Figure 1: System structure for multilingual speech-to-speech translation

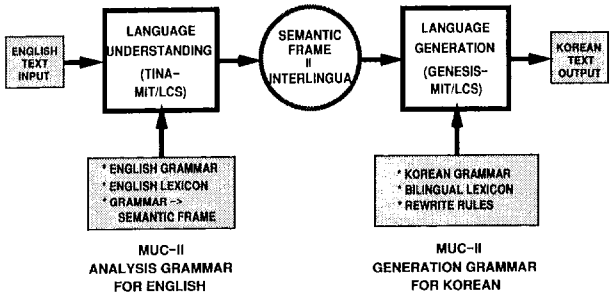

Figure 2: Process Flow of Text Translation

taken under fire by kirov with ssn-12's" are given in Figure 3 and Figure 4, respectively.

As is reflected in the parse tree, both syntactic and semantic categories are utilized in our grammar specification. Top level categories such as 'sentence,' 'subject,' etc. are syntax-based, whereas lower-level categories such as 'ship_name,' 'time_expression,' etc. are semantics-based. The main advantage of adopting semantic categories is that we can easily specify the co-occurrence restrictions of head categories (e.g., the parse tree specifies that the category *ships* occurs with a small subset of nominal modifiers including *uss* which we call 'ship_mod'), and therefore reduces the ambiguity of the input sentence. In addition, it provides for easy access to the meaning of domain-specific expressions. The parse tree directly encodes the knowledge that *sterett* and *kirov* are ship names, *ssn-12* a submarine name, and *z* stands for Greenwich Mean Time.

As for mapping from parse tree to semantic frame, we reduce all the major parse tree constituents into one of the three syntactic roles, i.e. clause, topic, and predicate. All clause-level categories including statements, infinitives, etc., are mapped onto *clause*. All noun phrases are mapped onto *topic*, and all modifiers as well as verb phrases are mapped onto *predicate*. However, there is no limit to the number of semantic frame categories, and we can easily create new categories for a more elaborate representation. In Figure 4, we have additional categories like 'time_expression.' Whether or not we add more categories to the semantic frame depends on how elaborate a translation output is desired. If elaborate translations are required, we increase the number of semantic frame categories. The flexibility of the semantic frame representation makes the TINA system an ideal tool for machine translation of various (i) purposes (i.e., whether a detailed or rough translation is required), and (ii) languages (e.g., some languages require a more elaborate tense representation or honorification than others, and the appropriate categories can be easily added.)

Table 1: Sample English/Korean Bilingual Lexicon

be	V	*i* ROOT *i* PRES *i* PAST *iess*
V2	V	" " ING *goiss* PP *ess* PRES *n* PAST *ess*
cause_en	V2	*cholaytoy*
visually	AV	*sikakulo*
cap_aircraft	N	*centhwu cengchalki*

2.2 Language Generation

The language generation system, GENESIS (Glass, Polifroni and Seneff, 1994), produces target language output on the basis of the semantic frame representation. It is driven by three submodules: a lexicon, a set of message templates, and a set of rewrite rules. These modules are language-specific and external to the core generation system. Consequently, porting the generation system to a new language is confined to developing these submodules.[5]

2.2.1 Lexicon

Since the semantic frame uses English as its specification language, and is the basis for constructing the target language grammar and lexicon, entries in the lexicon contain words and concepts found in the semantic frame, expressed in English, with corresponding surface realization forms in Korean. A sample fragment of a bilingual lexicon is given in Table 1.

2.2.2 Message Templates

Message templates are target language grammar rules corresponding to the input sentence expressions represented in the semantic frame. For instance, the word order constraint of the target language is specified in this module. A set of message templates used to produce the Korean translation from the semantic frame in Figure 4 is given in Table 2.

[5]A pilot study of applying GENESIS to Korean language generation can be found in (Yang, 1995).

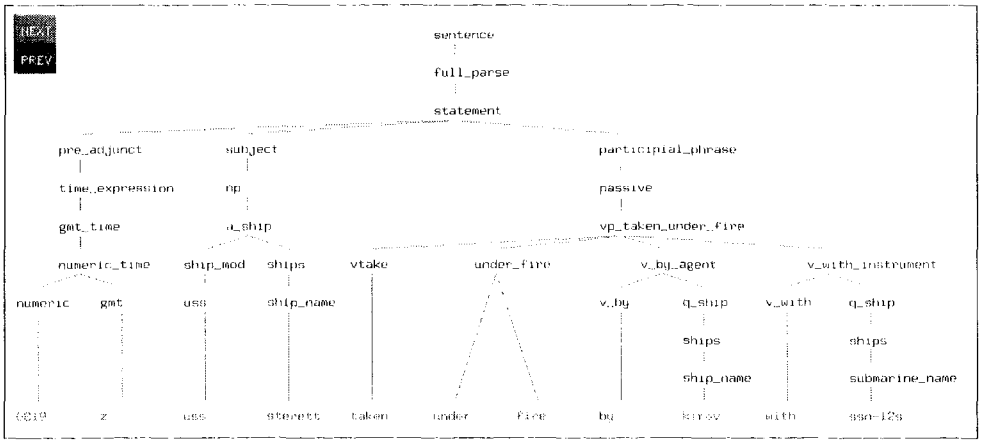

Figure 3: Parse Tree

```
{c statement
   :time_expression {p numeric_time
                       :topic {q gmt
                                 :name "z" }
                       :pred {p numeric
                                :topic 819 } }
   :topic {q ship_name
             :name "sterett"
             :pred {p ship_mod
                      :topic "uss" } }
   :subject I
   :pred {p taken_under_fire
            :mode "pp"
            :pred {p v_by
                     :topic {q ship_name
                               :name "kirov" }  }
            :pred {p v_with
                     :topic {q submarine_name
                               :name "ssn-12s" }  } } }
0819 z uss sterett taken under fire by kirov with ssn-12s
PARAPHRASE: 819 z USS Sterett taken under fire by Kirov with SSN-12s
TRANSLATION: 8시 19분 표준 시간  미군함 스테렛이 키로프에 의해 ssn 12s로  포격되었다
```

Figure 4: Semantic Frame and Translation

Table 2: Sample Message Templates

(a)	statement	:topic *i* :predicate *ta*
(b)	topic	:quantifier :noun_phrase
(c)	predicate	:topic :predicate
(d)	np-ship_mod	:topic :noun_phrase
(e)	ship_mod	:topic
(f)	np-missile_name	:topic :noun_phrase
(g)	missile_name	:topic
(h)	np-by	:topic *ey uyhay* :noun_phrase
(i)	np-with	:topic *lo* :noun_phrase

Template (a) says that a statement consists of a subject followed by a verb phrase. Note that all of the entries in a message template are optional so that a statement need not contain a subject or a verb phrase. Template (c) says that a verb phrase consists of an object followed by a verb. Other templates can be interpreted in a similar manner.

2.2.3 Rewrite Rules

The rewrite rules are intended to capture surface phonological constraints and contractions, in particular, the conditions under which a single morpheme has different phonological realizations. In English, the rewrite rules are used to generate the proper form of the indefinite article *a* or *an*. This module plays an important role in Korean due to numerous instances of phonologically conditioned allomorphs in the language. For instance, the so-called nominative case marker is realized as *i* when the preceding morpheme ends with a consonant, and as *ka* when the preceding morpheme ends with a vowel, as illustrated below. Similarly, the so-called accusative case marker is realized as *ul* after a consonant, and as *lul* after a vowel.

	Nominative Case	Accusative Case
Following a consonant	John-*i*	John-*ul*
Following a vowel	Mari-*ka*	John-*lul*

3 Data Summary

Our source language text is called MUC-II data, and consists of naval operational report messages.[6] There are 105 messages for system development and 40 messages set aside for system evaluation. These messages feature incidents involving different platforms such as aircraft, surface ships, sub-

[6]MUC-II stands for the Second Message Understanding Conference.

Table 3: Training Data (Data for System Development)

Data Set	Original Data	Modified Data	Total
A	101	105	206
A*		154	154
B	101	115	215
C	101	118	219
Total	303	337	794

marines and land targets. There are 12 words/sentence (average) and 3 sentences/message (average) (Sundheim, 1989). The original messages are highly telegraphic with many instances of sentence fragments, as illustrated in (1).

(1) At 1609 hostile forces launched *massive recon effort* from *captured airfield* against ctf 177 units transiting toward a neutral nation. Humint sources indicated 12/3 strike acft have launched (1935z) enroute *battle force. Have positive confirmation* that *battle force* is targeted (2035z). *Considered hostile act.*

For each original message, there is a corresponding modified message written in natural English. (2) is the modified counterpart of the original message (1). Note that the underlined parts in the modified message are missing in the original message.

(2) At 1609z hostile forces launched *a massive recon effort* from *a captured airfield* against ctf 177 units transiting toward a neutral nation. Humint sources indicated 12 strike aircraft have launched (1935z) enroute *to the battle force. CTF 177 has positive confirmation* that *the battle force* is targeted (2035z). *This is considered a hostile act.*

MUC-II data also have other typical features of raw text. There are quite a few instances of complex sentences, (3), coordination, (4) and (5), and multiple noun/verb modifiers, (5).

(3) Two uss-america based strike escort f-14's were engaged by unknown number of hostile su-7 aircraft near land9 bay (island target facility) while conducting strike against xxx guerilla camp.

(4) Kirov locked on with fire control radar and fired torpedo in spencer's direction.

(5) The deliberate harassment of uscgc spencer by hostile kirov endangers an already fragile political/military balance between hostile and friendly forces.

For our system training and development, we use both the original and the modified MUC-II data. We discuss the details of the data partition and training in the following section.

4 Training and Evaluation

4.1 Training

We first partitioned the MUC-II data corpus into three data sets. In addition, we also have a set of 154 MUC-II-like sentences that were created in an in-house experiment. We will discuss these sentences (which comprise data set A*) in Section 4.2.1. A summary of our training data is given in Table 3, where the numerals are the number of sentences.

We have trained the system on all the training data. The analysis rules are developed by hand, based on observed patterns in the data. These rules are then converted into a network structure. Probability assignments on all arcs in the network are obtained automatically by parsing each training sentence and updating appropriate counts (See (Seneff, 1992) for details). Table 4 gives the statistics of the current system in terms of the size of the analysis lexicon/grammar and generation lexicon/grammar.[7] The translation statistics for the training data are shown in Table 5.

Table 4: Current System Specifications

	Size of lexicon	Size of grammar (Number of Categories)
Analysis	1427	1297
Generation	997	763

Table 5: Translation Statistics for Training Data

# of Sentences	794
# of Parsed Sentences	743 (93.6%)
# of Correctly Translated Sentences	673 (84.8%)

4.2 Evaluation

We have carried out two types of system evaluations. The first evaluation specifically tests grammar coverage and the second evaluation tests overall system performance.

4.2.1 Evaluation of the Grammar Coverage

In order to evaluate grammar coverage, we created a MUC-II-like database (data set A*) which contains no unknown words in an in-house experiment.[8] In the experiment, we asked the subjects to study a list of data set A MUC-II sentences and then create about 10 MUC-II-like sentences on their own. Subjects were told to create sentences which illustrate the general style of the MUC-II sentences and which only use the vocabulary items occurring in the example sentences. We collected 154 MUC-II-like sentences in this experiment. We then evaluated the system's performance on these sentences. As Table 6 indicates, the system parsed over 50% of the A* sentences. Of the sentences that do parse, 91% of them are correctly translated.

4.2.2 Overall Evaluation

The results of system evaluation on the test data (consisting of 40 messages, 111 sentences) are shown in Table 7. The first point to note is that the system parsed 34.8% of the test sentences which have no unknown words. This figure is somewhat lower than the corresponding figure for data set A* (50.6%) because the test set is harder than data set A*. (The test sentences are completely new whereas A* sentences were fabricated by studying A sentences.) The second point to note is that system failure is due in large part to the presence of unknown words. In fact, as Table 7 indicates, about 41% of new MUC-II sentences fail solely because the system can not handle unknown words at this time. We discuss our ongoing efforts to tackle this problem in Section5.

4.3 Efficiency

Largely due to our effort to reduce the ambiguity of the input sentences, the system runs efficiently. For an average length sentence containing 12 words, it takes about 1.7 seconds to

[7] Table 4 gives the number of preterminal categories in the analysis grammar. The actual number of rules is much greater because TINA allows cross-pollination, the sharing of common elements on the right-hand side of rules. See (Seneff, 1992) for more details about cross-pollination.

[8] As we will discuss in the next section, it is difficult to use a MUC-II database to evaluate grammar coverage because many MUC-II sentences fail based strictly on the fact that they contain unknown words, i.e., words which are not in the system's lexicon.

Table 6: Data Set A* Evaluation Results

# of Sentences	154
# of Parsed Sentences	78/154 (50.6%)
# of Correctly Translated Sentences	71/78 (91.0%)

Table 7: Test Evaluation Results

# of Sentences	111
# of Sentences with No Unknown Words	66/111 (59.5%)
# of Parsed Sentences	23/66 (34.8%)
# of Correctly Translated Sentences	20/23 (87.0%)

translate. It takes an average of 2.28 seconds to translate a sentence containing 16 words. For a fairly complex sentence containing 38 words, it takes about 4 seconds to translate. Some examples of Korean translation output are given in Figure 5. The system runs on a SPARC 10 workstation, and the source code is written in C. Korean translation outputs are displayed on a *hangul* window running on UNIX.

5 Toward Robust Translation

At the moment, our system is not capable of dealing with a sentence containing (i) unknown words, cf. Section 4.2, and (ii) unknown constructions, cf. Section 4.1. In this section we discuss our on-going efforts to overcome these deficiencies: Integration of a part-of-speech tagger to handle unknown words/constructions, and a word-for-word translator to cope with other system failures, cf. (Frederking and Nirenburg, 1995).

5.1 Integration of Part-of-Speech Tagger

Regarding the unknown word problem, an obvious solution is to expand the lexicon. Concerning the problem involving unknown constructions, we could easily generalize the grammar to extend its coverage. However, both of these solutions are problematic. Handling the unknown word problem by increasing the size of the lexicon is not that straightforward given that most unknown words are open class items such as nouns, verbs, adjectives and adverbs. In addition, one can not generalize the grammar without side effects. Due to the highly telegraphic nature of the MUC-II data, generalizing the grammar will increase the ambiguity of an input sentence greatly, cf. (Grishman, 1989).[9] Hence, we need alternative solutions to deal with unknown words and unknown constructions. The most desirable solution is to (i) leave the current grammar intact since it efficiently parses even highly telegraphic messages, and (ii) tackle unknown words and unknown constructions by the same mechanism.

A potential solution to the unknown word problem is to: *Do part of speech tagging and replace unknown words with their parts-of-speech, and bootstrap the parts-of-speech (instead of the actual words) to the analysis grammar.* The unknown words would be replaced in the sentence string with their corresponding part-of-speech tag, and the semantic grammar would be augmented to handle generic adjectives, nouns, verbs, etc., intermixed in the rules at appropriate positions. The idea would be to include just enough semantic information to solve the ambiguity problem, effectively anchoring on words such as ship-name that have high semantic relevance within the domain.

This approach might also be effective as a backoff mechanism when the system fails to parse a sentence containing only *known* words. A set of semantically significant vocabulary items could be tagged as "immutable", and all the words in the sentence *except* these anchor words would be converted to part-of-speech prior to a second attempt to parse. The same grammar would be used in all cases.

[9]Recall that we resolve the ambiguity problem by constraining the grammar with semantic categories.

Table 8: Tagger Evaluation on TEST Data

Stage	Overall Accuracy	Unknown Word Accuracy
Before Training	1125/1287 (87.4%)	67/82 (82%)
After Training I	1249/1287 (97%)	70/82 (85%)
After Training II	1263/1287 (98%)	71/82 (87%)

For the solution sketched above, we have evaluated the Rule-Based Part-of-Speech Tagger (Brill, 1992) on the test data both before and after training on the MUC-II database. These results are given in Table 8. Tagging statistics 'before training' are based on the lexicon and rules acquired from the BROWN CORPUS and the WALL STREET JOURNAL CORPUS. Tagging statistics 'after training' are divided into two categories, both of which are based on the rules acquired from training on data sets A, B, and C of the MUC-II database. The only difference between the two is that in one case (After Training I) we use a lexicon acquired from the MUC-II database, and in the other case (After Training II) we use a lexicon acquired from a combination of the BROWN CORPUS, the WALL STREET JOURNAL CORPUS, and the MUC-II database. Since the tagging result is quite promising, despite the fact that the training data is of modest size, we are planning to integrate the tagger into the analysis module.

5.2 Integration of Word-for-Word Translator

Even though implementing the part-of-speech tagger and extending the analysis grammar to accept parts-of-speech as terminal strings will increase the grammar coverage, it is an almost impossible task to write a grammar which covers all freely occurring natural language texts, let alone have a robust parser to deal with this inadequacy.[10] Despite this difficulty in designing a complete translation system, an ideal translation system ought to be able to produce translations which are useful under any circumstances. Therefore, we are integrating a word-for-word translator[11], which provides tools to aid a human translator, as a fallback system.

Figure 6 shows the planned robust system architecture, with the part-of-speech tagger and the word-for-word translator integrated into the core understanding/generation system. Note that the system will provide an indication or flag to the user showing whether the translation is produced by TINA/GENESIS or by the word-for-word fallback system.

6 Summary

In this paper we have described our ongoing work in automatic English-to-Korean text translation of telegraphic messages. This is a part of our overall effort in text and speech translation for limited-domain multilingual applications. We have described the system architecture (Section 2), the source language text (Section 3), and the system evaluation results (Section 4). We have also discussed ideas on how to make the system robust, and proposed two specific solutions: integration of a part-of-speech tagger and a word-for-word translator (Section 5).

7 Acknowledgements

We would like to acknowledge the following people for their contributions to this project: Jack Lynch, Beth Carlson, Victor Zue, and SungSim Park. We also would like to thank Beth Sundheim of NRaD, Professor Ralph Grishman

[10]See (Sleator, 1991) for a design of a robust parser which handles unknown constructions.

[11]The word-for-word translator is being developed by GARJAK under a subcontract.

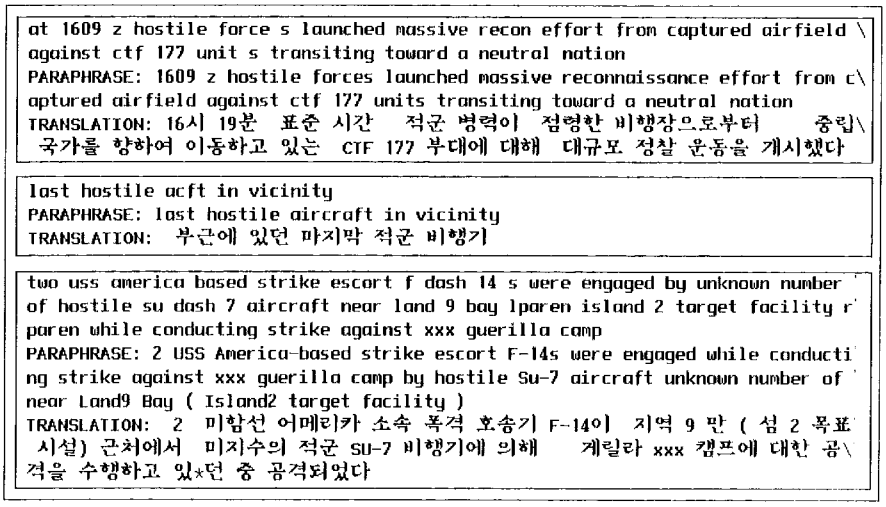

Figure 5: Sample Korean Translation Output

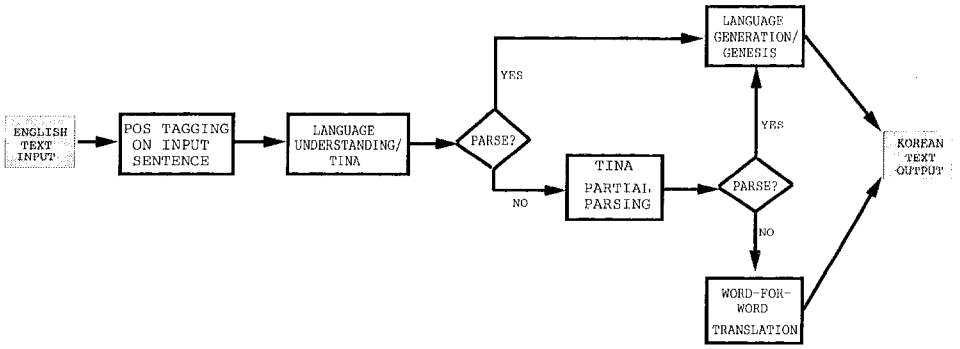

Figure 6: Robust Translation System

of NYU, Professor Key-Sun Choi of KAIST, Korea, and Willis Kim of MITRE. Beth Sundheim provided us with the MUC-II data as well as technical reports documenting the data. Dr. Grishman provided us with his grammar, dictionary, and semantic models for the MUC-II data. These materials helped us understand the linguistic properties of the MUC-II corpus. Dr. Choi provided us with documentation and software for KAIST'S MATES/EK English-to-Korean machine translation system. Kim provided us with electronic English/Korean dictionaries as well as a report on his work.

References

Eric Brill. 1992. A Simple Rule-Based Part of Speech Tagger. *Proceedings of the Third Conference on Applied Natural Language Processing.* Trento, Italy: ACL.

Key-Sun Choi, Seungmi Lee, Hiongun Kim, Deok-Bong Kim, Cheoljung Kweon, and Gilchang Kim. 1994. An English-to-Korean Machine Translator: MATES/EK. *Proceedings of the 15th International Conference on Computational Linguistics.* Kyoto, Japan.

Robert Frederking and Sergei Nirenburg. 1995. Three Heads are Better than One. C-STAR II Meeting, Pittsburgh.

James Glass, Joseph Polifroni and Stephanie Seneff. 1994. Multilingual Language Generation Across Multiple Domains. *Proceedings of the 1994 International Conference on Spoken Language Processing.* Yokohama, Japan.

Ralph Grishman. 1989. Analyzing Telegraphic Messages. *Proceedings of the 1989 DARPA Speech and Natural Language Workshop.* Cape Cod, Massachusetts.

W. Kim and W. Rhee. 1994. Machine Translation Evaluation. Seoul, Korea: MITRE.

Youngjik Lee, Young-Sum Kim, Jung-Chul Lee, Joon-Hyung Ryoo and Jae-Woo Yang. 1995. Korean-Japanese Speech Translation System for Hotel Reservation - Korean front desk side. *European Conference on Speech Communication and Technology.* Madrid.

Stephanie Seneff. 1992. TINA: A Natural Language System for Spoken Language Applications. *Computational Linguistics,* 18(1): 61-88.

Daniel D.K. Sleator and Davy Temperley. 1991. *Parsing English with a Link Grammar.* CMU-CS-91-196.

Beth M. Sundheim. 1989. *Navy Tactical Incident Reporting in a Highly Constrained Sublanguage: Examples and Analysis.* Technical Document 1477. Naval Ocean Systems Center, San Diego.

Dinesh Tummala, Stephanie Seneff, Douglas Paul, Clifford Weinstein, and Dennis Yang. 1995. CCLINC: System Architecture and Concept. Demonstration of Speech-to-Speech Translation for Limited-Domain Multilingual Applications. *Proceedings of the 1995 ARPA Spoken Language Technology Workshop.* Austin, Texas.

Dennis W. Yang. 1996. *Korean Language Generation in an Interlingua-based Speech Translation System.* Technical Report 1026. MIT Lincoln Laboratory.