THE TRANSLATOR'S ASSISTANT: A MULTILINGUAL NATURAL LANGUAGE GENERATOR BASED ON LINGUISTIC UNIVERSALS, TYPOLOGIES, AND PRIMITIVES

by

TOD JAY ALLMAN

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON
December 2010

ACKNOWLEDGEMENTS

First I want to thank my wife, JungAe, for her extraordinary help with this dissertation. Many doctoral students acknowledge that their dissertations would not have been possible without the support of their spouses. While that is certainly true in this case, JungAe’s help went far beyond that of the typical supportive spouse. The entire Korean grammar for this project was developed by JungAe; I simply entered the rules that she suggested. The results of the Korean experiments exceeded my most optimistic hopes, and JungAe deserves full credit for that work. She developed the Korean lexicon and grammar, helped with the experiments to test for increased productivity, and developed the questionnaires for evaluating the texts. This project most certainly would not have been possible without her help.

Second I would like to thank Dr. Jerold Edmondson who patiently worked through this document numerous times. The final product is much better than the original draft, and credit for that goes largely to him. He taught me that even in a lengthy document, every word must be carefully selected with the reader’s perspective in mind.

I also want to thank Dr. ShinJa Hwang for reading this dissertation very carefully. Her attention to detail is remarkable; several times she pointed out mistakes in my analyses, thereby causing me to reanalyze particular issues.

I also want to thank Drs. Laurel Stvan and Colleen Fitzgerald for encouraging me to keep this dissertation focused and concise.

Finally, I want to thank God for enabling and allowing me to serve Him in this way. It has been a pleasure.

November 19, 2010

ABSTRACT

THE TRANSLATOR'S ASSISTANT: A MULTILINGUAL NATURAL LANGUAGE GENERATOR BASED ON LINGUISTIC UNIVERSALS, TYPOLOGIES, AND PRIMITIVES

Tod Allman, PhD.

The University of Texas at Arlington, 2010

Supervising Professor: Jerold Edmondson

The Translator's Assistant (TTA) is a multilingual natural language generator (NLG) designed to produce initial drafts of translations of texts in a wide variety of target languages. The four primary components of every NLG system of this type are 1) the ontology, 2) the semantic representations, 3) the transfer grammar, and 4) the synthesizing grammar. This system's ontology was developed using the foundational principles of Natural Semantic Metalanguage theory. TTA's semantic representations are comprised of a controlled, English influenced metalanguage augmented by a feature system which was designed to accommodate a very wide variety of target languages. TTA's transfer grammar incorporates many concepts from Functional-Typological grammar as well as Role and Reference grammar. The synthesizing grammar is intentionally very generic, but it most closely resembles the transformational-generative model. The meaning-based theory of translation underlies the TTA system.

The fundamental question that this research proposes to answer is as follows: if the semantic representations contain sufficient information, and if the grammar possesses sufficient capabilities, then is TTA able to generate drafts of sufficient quality that they improve the productivity of experienced mother-tongue translators? To answer this question, software was developed that allows a linguist to build a lexicon and grammar for a particular target language. Then semantic representations were developed for one hundred and five chapters of text. Two unrelated languages were chosen to test the system, and a partial lexicon and grammar were developed for each test language: English and Korean. Fifty chapters of text were generated in Korean, and all one hundred and five chapters were generated in English. Then extensive experiments were performed to determine the degree to which the computer generated drafts improve the productivity of experienced Korean mother-tongue translators. Those experiments indicate that when experienced mother-tongue translators use the rough drafts generated by TTA, their productivity is typically quadrupled without any loss of quality.

TABLE OF CONTENTS

APPENDIX

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION TO *THE TRANSLATOR'S ASSISTANT*

1.1 Introduction

This dissertation will describe a natural language generator (NLG) called *The Translator's Assistant* (TTA). A natural language generator is defined as any computer program that takes a representation of information, applies a language's lexicon and grammar to that representation, and then generates text in that language communicating part or all of the original information. Figure 1-1 below shows a typical natural language generation system.

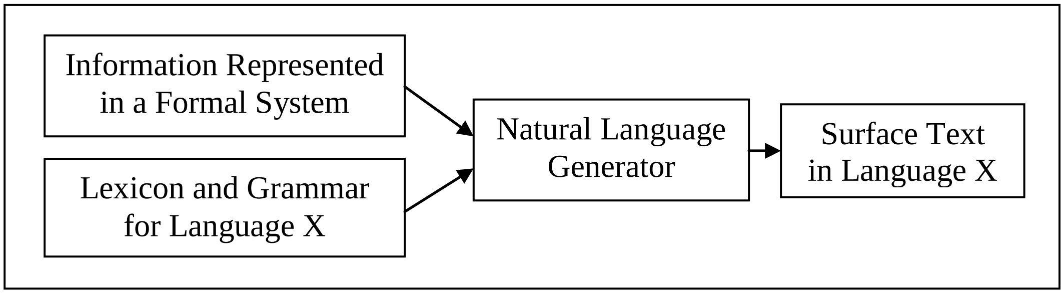

Information Represented in a Formal System
Lexicon and Grammar for Language X
Natural Language Generator
Surface Text in Language X

Figure 1-1 A Typical Natural Language Generation System

As seen in the figure above, the system begins with information of some type represented in a formal system. The generation system also has access to a lexicon and grammar for the target language. The natural language generator then takes those two sources of information and generates text in the target language that communicates all or part of the information in the original representation. Natural language generators are often used to generate texts in several languages, so a multilingual situation is shown below in figure 1-2.

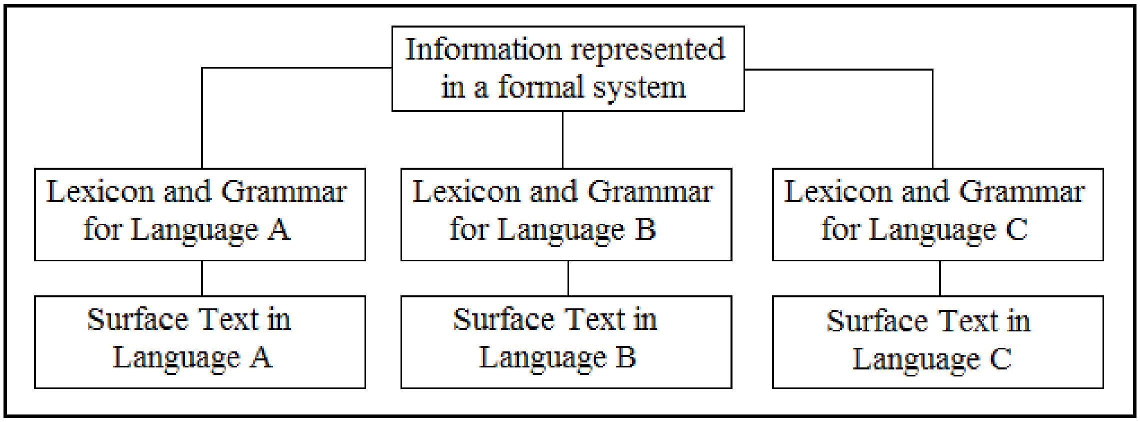

Information represented in a formal system
Lexicon and Grammar for Language A
Lexicon and Grammar for Language B
Lexicon and Grammar for Language C
Surface Text in Language A
Surface Text in Language B
Surface Text in Language C

Figure 1-2. A Typical Application of a Natural Language Generator

As seen in the figure above, the natural language generator starts with information that is represented in a prescribed formal system. The NLG then takes that information, applies the lexicon and grammar for a particular language to that information, and then generates surface text in that language that communicates part or all of the original information. The NLG then applies the lexicon and grammar of another language to the same representation of information, and generates surface text in that language that communicates part or all of the original information, etc. The information that is represented in the formal system may be numerical such as weather or medical data, or it may be an abstract representation of a proposition or text.

The research field of natural language generation first began to develop in the 1970s with two pioneering doctoral dissertations. The first was Goldman's (1975) *Computer Generation of Natural Language from a Deep Conceptual Base*, and the second was Davey's (1979) *Discourse Production: a computer model of some aspects of a speaker*. During this period it became clear that "natural language generation is the subfield of artificial intelligence and computational linguistics that focuses on computer systems that can produce understandable texts in English or other human languages" (Reiter and Dale 2000:1). The field began to mature in the 1980s with two distinct perspectives emerging: some systems were developed by artificial intelligence researchers, while other projects were founded on linguistics and computational linguistics (Reiter and Dale 2000:19). Notable contributions from that era include McKeown's work (1985) on schemas entitled *Text Generation*, and Appelt's research (1985) on reference entitled *Planning English Referring Expressions*.

Numerous NLG systems have been developed during the past two decades, and they may be divided into two broad categories: those that use numerical data as their source, and those that use abstract semantic representations as their source. The NLG system developed for this dissertation uses abstract semantic representations as its source, so that subdomain will be the focus of this dissertation. Within that subdomain, there are presently two large scale NLG systems that have been developed and are comparable to TTA: the KPML system which is being developed at the University of Bremen in Germany (Bateman 2010b), and the KANT system which was developed at Carnegie Mellon University (Nyberg 2004). Both of these systems were developed from the perspective of computational linguists as will be demonstrated in the next chapter. To date there has not been a large scale NLG system developed solely from the perspective of a linguist, and this dissertation fills that gap.

The Translator's Assistant is a practical, linguistically based engineering solution to the unsolvable conceptual problem of universal perfect translation. This project draws from a number of linguistic theories, and makes principled compromises in order to reach an operational solution. Many issues were encountered during the development of this project. Some of the issues were resolved either completely or partially, some were circumvented, and others remain unsolved. This dissertation will describe the system in its current form. The final chapter of this dissertation will discuss areas that require additional research, and the issues that still need to be resolved.

There are two fundamental differences between TTA and other NLGs: 1) TTA was designed and developed using theoretical frameworks that are familiar to linguists, and 2) TTA is intended to generate texts in a very wide variety of target languages. Other NLG systems will be described in chapter 2, but they are generally designed using models and terms that are unfamiliar to most linguists, and they are generally intended to generate texts in a small number of related languages. The concepts and structures underlying TTA's design take advantage of recent typological research, thereby enabling linguists to quickly develop their lexicons and accurately model their grammars for languages found around the world. TTA then uses those lexicons and grammars to generate texts in those languages. Additionally, because TTA is intended to generate texts in many languages, the semantic representational system and the grammar that were developed for TTA are very different from those designed for other NLGs. TTA's semantic representational system contains more information than is included in other semantic representational systems, and TTA's grammar includes more capabilities than do the grammars of other NLGs. Several of the more notable differences between TTA and other NLGs will be illustrated below in section 1.3.

1.2 The Fundamental Question Addressed by this Research

The fundamental question that this research answers is as follows: If TTA's semantic representations contain sufficient information, and if TTA's grammar possesses sufficient capabilities, then is TTA able to generate surface drafts of sufficient quality that they improve the productivity of experienced mother-tongue translators? The research that was done for this dissertation indicates that the answer is clearly yes. Extensive experiments in Korean indicate that the drafts generated by TTA typically quadruple the productivity of experienced mother-tongue translators (cf. Chapter 6).

1.3 Features Distinguishing TTA from other NLG Systems

As was mentioned above, the two primary characteristics which distinguish TTA from other NLGs are 1) its semantic representational system is much richer than the systems employed by other NLGs, and 2) its grammar possesses more capabilities than do the grammars of other NLGs. Specific examples of some of TTA's distinguishing features and capabilities will be briefly listed below, but they will be discussed more thoroughly at the end of chapter 2 after other very successful NLG systems have been described. Then the remainder of this dissertation will describe in detail TTA’s semantic representational system and its generating grammar.

1.3.1 Distinguishing Features of TTA’s Semantic Representational System

Because TTA is intended to generate texts in a wide variety of target languages, its semantic representational system must contain more information than is required by any one particular language or language family. Several specific examples illustrating how TTA’s semantic representational system contains much more information than do the semantic representational systems of other NLGs are listed below.

- TTA’s semantic representational system includes a feature system that is comprised of primitives that have been gathered from a large array of languages. For example, every noun in the semantic representations is marked for number, and the possible values are ‘Singular’, ‘Dual’, ‘Trial’, ‘Quadrial’, ‘Paucal’, and ‘Plural’. All of these values are necessary because some languages morphologically encode each of them. The feature system also includes discourse information (e.g., Discourse Genre, Notional Structure Schema, and Salience Band) (Longacre 1996:10, 28, 36), speaker and listener information (e.g., Speaker, Listener, Speaker’s Attitude, Speaker’s Age, and Speaker to Listener’s Age), Participant Tracking values (Longacre 1995:702; Prince 1981:230), Participant Status values (Longacre 1995:701; Bartsch 1995:47), etc. This feature system will be described in detail in section 3.3.2.
- The concepts in TTA’s semantic representations come from TTA’s ontology, which was developed using the foundational principles of Natural Semantic Metalanguage (NSM) theory. NSM theorists such as Anna Wierzbicka (Wierzbicka 1992; 1996; Goddard 1998; 2008) have proposed that every word in every language may be explicated using a small set of innate and indefinable semantic primitives (Wierzbicka 1996:13). These theorists are in the process of empirically identifying the universal semantic primitives, and developing a universal grammar that describes how the semantic primitives may be

combined. The details of NSM, the NSM foundational principles, and the rationale for adopting this approach will be presented in section 3.2.6.

- The concepts in TTA's ontology are very narrowly defined and used consistently throughout all the semantic representations. For example, TTA's ontology includes twenty-five distinct senses of BE. A few of these senses include BE-D, which is used in all attributive constructions (e.g., *John is tall.*), BE-E, which is used in all existential constructions (e.g., *There are lions in Africa.*), BE-F, which is used in all locative constructions (e.g., *John is in Africa.*), BE-M, which is used in all social role constructions (e.g., *John is a teacher.*), etc. Many of the concepts in TTA's ontology have multiple senses, and each sense is very narrowly defined and used in virtually identical constructions throughout all the semantic representations. The details of this approach will be provided in section 3.3.1.
- TTA's ontology includes *semantically complex* concepts which are inserted into the semantic representations automatically if a particular target language has a lexical equivalent. For example, the English concept *veterinarian* is semantically complex, and few languages will have a lexical equivalent for it. Therefore, whenever *veterinarian* appears in a source document that is to be translated by TTA, it is explicated as "doctor that treats sick animals." Each of the concepts in "doctor that treats sick animals" is considered semantically simple, and therefore other languages are more likely to have lexical equivalents for each of them. Semantically simple and complex concepts will be discussed in more detail in section 3.3.1.1, and the insertion of semantically complex concepts into the semantic representations will be discussed in sections 3.3.1 and 4.3.1.
- TTA's semantic analysis is considerably richer and more detailed than the analysis systems used by other NLGs because it attempts to identify the reasons why surface structures in the source documents have the forms that they do. The representational

systems used in other NLG systems identify the surface forms that occur in the source documents, but they don't identify the underlying reasons for those forms. For example, when an English source document is being analyzed in order to develop a semantic representation of it, if a pluperfect verb form occurs, there are two possible reasons: 1) the author is using *flashback* (i.e., describing an event that occurred in the past but has current relevance) (e.g., *John had seen Mary earlier that day.*) (Longacre 1996:28), or 2) the author is using a *counterfactual* construction (e.g., *If John had passed that test, …*). Many languages don't have grammatical constructions that correspond to a pluperfect, but every language has its own mechanisms for signaling flashback and counterfactuals. Therefore, when a pluperfect occurs in a source document, the reason for its use is determined and encoded in the semantic representation. TTA's semantic analysis also includes markers to indicate the beginning of each episode, the beginning of each scene, a variety of nominal-nominal relationships, etc. The analysis used to develop TTA's semantic representations will be discussed further in section 3.3.3.9.

1.3.2 Distinguishing Features of TTA's Grammar

Because TTA is intended to generate texts in many languages, its generating grammar must possess sufficient capabilities to produce surface text in those languages. Specific examples illustrating several of the capabilities that are unique to TTA's grammar are listed below.

- TTA's grammar is intentionally theory-neutral, but it has been designed with sufficient flexibility to permit linguists to develop their grammars using a variety of contemporary theoretical models. TTA's grammar will be described in chapter 4.
- The transfer component of TTA's grammar includes rules which are able to generate honorifics. The transfer grammar in TTA will be described in section 4.3, but generally it is responsible for transforming the semantic representations into new underlying

representations that are appropriate for each target language. It appears that none of the other NLGs that were examined for this dissertation attempt to deal with honorifics, yet encoding the proper honorific forms is crucial in many languages. The rules which enable TTA to encode honorifics will be described in section 4.3.3.

- The transfer component of TTA's grammar includes the capability to map a single source concept to multiple target words based on the context. It is well documented that every word in every language has its own collocation range and restrictions (Sinclair 1991:112). It is also well known that certain languages will have a single word for a range of concepts, but other languages will have multiple words for those concepts. For example, English has the word *to carry*, and it is used whether the item being carried is in one's hand, on one's head, in one's pocket, in a bag, etc. A language such as Tzeltal has specific verbs for each of these situations (Larson 1984:89), and it doesn't have a generic verb meaning *to carry.* In order to handle situations such as these, collocation correction rules were added to TTA's transfer grammar, and they will be presented in section 4.3.6. Those rules permit linguists to map the concept *carry* to many different target equivalents based on the other concepts in the environment.
- The synthesizing component in TTA's grammar was designed to resemble as closely as possible the descriptive grammars that field linguists routinely write. The synthesizing grammar is responsible for generating surface structure forms from an abstract underlying representation, and it will be presented in detail in section 4.4. That section shows that TTA's synthesizing grammar includes 1) feature copying rules, 2) spellout rules, 3) clitic rules, 4) movement rules, 5) phrase structure rules, 6) anaphora identification and spellout rules, and 7) word morphophonemic rules. All of these rules are very familiar to field linguists and appear often in their descriptive grammars.

- TTA's grammar takes advantage of recent typological research. Extensive typological research has been done in the areas of tense (Comrie 1985; Dahl 1985), aspect (Comrie 1976; Dahl 1985), mood, illocutionary force, discourse genres (Longacre 1996:10), salience bands (Longacre 1996:28), and the construction of relative clauses (Comrie 1989:13-163) and object complement clauses (Givón 1990:515-561). Various aspects of this research have been incorporated into TTA's feature system and grammar. For example, the typological research that has been done with respect to relative clauses guided the design of the Relativization Structures, Strategies, and Hierarchy dialog which will be presented in section 4.3.5.

The factors listed above make TTA unique, and each of these factors will be discussed in detail at the end of chapter 2 after other NLG systems have been examined.

1.4 Overview of this Dissertation

The second chapter of this dissertation will briefly discuss the translation process, and then describe why machine translation projects have generally failed. Because fully automatic, high quality machine translation has proven elusive, computational linguists have recently begun developing natural language generators. NLGs avoid many of the difficulties associated with machine translation, and have therefore been reasonably successful. That chapter will describe the two main categories of NLGs, and present several of the most successful NLG systems in each of those two categories. Then that chapter will briefly introduce the NLG that was developed for this dissertation, and describe in more detail why TTA is distinct from the other NLGs.

Chapter 3 will thoroughly describe the semantic representational system that was developed specifically for TTA. That chapter will begin with a brief survey of the semantic systems that were potential candidates for this project (Montague 2002; Jackendoff 1990; Lakeoff 1987; Langacker 1986; Talmy 1980; Nirenburg 2004; Wierzbicka 1996; Goddard 1998), but those systems were ultimately rejected in favor of the semantic system that was developed for TTA. Then chapter 3 will describe in detail TTA's ontology, its feature system, and the syntactic structures that are permitted in the semantic representations.

Chapter 4 will describe the organization and the capabilities of TTA's grammar. The grammar in TTA consists of two distinct components: a *transfer grammar* and a *synthesizing grammar*. Detailed models for these grammars will be presented, and the various types of rules and their capabilities will be illustrated. There is also a small grammar component in TTA's target lexicon, and it will be described.

Chapter 5 will discuss the experiments that were performed with the two test languages: Korean and English. Korean is an Altaic language spoken by approximately 67 million people living on the Korean peninsula, and an additional 6 million Korean expatriates (Cho et al. 2000:1). English is a West Germanic language that initially arose in the Anglo-Saxon Kingdoms of England, and has become the world's leading language of international discourse. For both test languages, graphs will be presented which demonstrate that as a linguist builds his target lexicon and grammar, TTA systematically acquires the lexical and grammatical knowledge required to generate texts in that language. Therefore each subsequent chapter of text generated by TTA requires less input from the linguist.

Chapter 6 will discuss the experiments that were performed with the Korean texts in order to answer the following two questions:

- Are the texts generated by TTA of sufficient quality that they significantly improve the productivity of experienced mother-tongue translators?
- After mother-tongue speakers edit TTA's drafts, are the edited drafts of the same quality as manually translated texts?

Numerous machine translation projects have been cancelled because the developers found that editing the computer generated texts was more time consuming and costly than manually translating the same texts (Whitelock 1995:83). Therefore experiments must be done for every computer assisted translation project to determine whether the computer's rough drafts are of sufficient quality that they actually reduce the amount of manual labor involved during the translation process. As was mentioned above, the experiments performed for this project indicate that TTA's rough drafts typically quadruple the productivity of experienced mother-tongue translators. Then extensive human evaluations were performed to confirm that the edited computer drafts are equivalent in quality to manually translated texts.

Chapter 7 will present the final conclusions and discuss areas requiring additional research. This project is by no means complete; as more experiments are done with additional languages, the author is certain that additional information will be required in the semantic representations, and more capabilities will be required of the generating grammar. This dissertation will report the progress made to date.

1.5 TTA's Contributions to the Field of Linguistics

The Translator's Assistant is a tool which may be used by both theoretical linguists and applied linguists. TTA may be used by:

- theoretical grammarians who wish to test their hypotheses regarding particular grammatical issues,
- field linguists who do translation work in relatively unstudied languages,
- semanticists who are testing the hypotheses of NSM,
- lexicographers who develop bilingual dictionaries, and
- language preservationists who hope to document a language and simultaneously promote its use by providing the speakers with empowering texts.

As was mentioned above, there are two components in this tool which make TTA distinct from other NLG systems that have been developed to date: 1) the semantic representational system, and 2) the grammatical apparatus. The semantic representational system, composed of concepts, features, and structures, is a new method of representing meaning. No other semantic representational system has classified its concepts according to their varying degrees of semantic complexity, nor has any other semantic representational system gathered the pertinent features and feature values from such a large array of languages. The grammatical apparatus developed for TTA is a linguistically informed system which permits theoretical linguists to validate and correct their insights to particular language phenomena. When there are multiple solutions to a particular grammatical problem, each may be modeled within TTA's apparatus and evaluated objectively by generating texts in the target language. Thus TTA's grammatical apparatus increases objectivity, elucidates unnoticed problems, and hones grammatical solutions.

TTA also has several very practical applications for linguists. For example, after a linguist has developed a lexicon and grammar for a target language, TTA enables him to easily produce a bilingual dictionary, bilingual machine-tractable corpora, and a topically organized grammar sketch. Similarly because TTA is intended to generate texts in a wide variety of target languages, it may be used by linguists doing research in a variety of fields. For example, TTA may be used by semanticists who are either confirming or disproving the NSM claim that the semantic primitives are present in every language (Wierzbicka 1996:13). TTA may be used by grammarians who are testing the validity of the accessibility hierarchy (Comrie 1989:155), the complementation scale of event integration (Givón 1990:537), methods of encoding discourse peak (Longacre 1996:38), etc. Thus TTA is a tool which may be used by theoretical grammarians, translators, semanticists, lexicographers, and language preservationists.

1.6 Methodology

In order to gather the data necessary to develop the Korean lexicon and grammar, the language informant[1] was shown individual propositions from the semantic representations. The informant was then shown the concepts in the proposition, the various semantic roles of each referent, the proposition's illocutionary force and salience band, and the other pertinent features. After the informant had seen all of the relevant information for a particular proposition,

[1] All Korean data is courtesy of JungAe Lee, a Korean Ph.D. student studying linguistics at the University of Texas at Arlington. All mistakes are my responsibility. The data was gathered under IRB number 07.275s.

the informant was asked how to express that proposition in Korean. After the proposition's Korean equivalent had been established, each constituent in the semantic representation was reexamined in order to determine its contribution to the Korean surface structure. After establishing relationships between each constituent in the semantic representation and each constituent in the surface structure, the necessary lexical items were entered into the lexicon, and the necessary rules were entered into the grammar. Then TTA generated its form of the Korean text, and that form was compared with the informant's original version. If the two forms were sufficiently close, the next proposition in the semantic representations was examined. If the two forms were not sufficiently similar, then the differences were identified and discussed. Then the grammatical rules were modified in order to make TTA's generated version become as close as possible to the language informant's version.

1.7 Conclusions

This dissertation will show that The Translator's Assistant is able to generate drafts of texts in a very wide variety of target languages. TTA is primarily a tool that drastically reduces the amount of work required by linguists to produce drafts of translations of documents in other languages. This tool works equally well for languages that are thoroughly studied, languages that have only slightly been studied, and languages that are endangered[2]. Similarly, this tool works equally well for languages that are typologically diverse with respect to their morphological and syntactic features; it works for languages that are coranking or clause chaining, for languages that are nominative-accusative or ergative-absolutive, for languages that are highly isolating or highly polysynthetic, for languages that are fusional or highly agglutinative, etc. This tool enables linguists to document a language and simultaneously generate texts for the speakers of that language. It is hoped that this tool will empower

[2] Stephen Beale, who is a research professor in the Department of Computer Science and Electrical Engineering at the University of Maryland in Baltimore, is in the process of applying for a NSF grant through the Documenting Endangered Languages program. He plans to use TTA in order to document two endangered languages in Vanuatu, and then generate drafts of several texts in those languages.

speakers of minority languages around the world by providing them with translations of vital information, which will not only enable them to live longer, healthier, and more productive lives, but it will also enable them to participate in the larger world. The remainder of this dissertation will describe this tool, present examples of how it has been used in particular languages, and discuss the results of numerous experiments which demonstrate that drafts generated by TTA significantly improve the productivity of experienced mother-tongue translators.

CHAPTER 2

INTRODUCTION TO NATURAL LANGUAGE GENERATION

2.1 Introduction

This chapter will introduce natural language generation. The field of natural language generation has developed during the past several decades primarily because fully automatic, high quality machine translation has not been achieved. Section 2.2 provides an overview of the translation process, and then describes why machine translation projects have generally failed. That section also describes how natural language generators (NLGs) avoid the difficulties associated with machine translation. Section 2.3 provides an overview of existing NLG systems. Section 2.3.1 will describe the two broad categories of NLG systems, and section 2.3.2 describes the techniques that have been developed by computational linguists to build these systems. Then sections 2.3.3 and 2.3.4 will present several of the most successful NLG systems that have been developed. After the other NLG systems have been described, section 2.4 will describe *The Translator's Assistant*, the NLG system that was developed for this dissertation. That section will conclude by elaborating on some of the features which make TTA distinct from other NLGs.

2.2 The Translation Process

Translating a document from one language to another is a very complex, labor intensive, highly skilled task. Producing a natural translation in a language that is unrelated to the source language requires a thorough knowledge of the source and receptor languages, cultures, and audiences. During the translation process, a myriad of linguistic and sociolinguistic factors must be taken into consideration. Although translation is a very complex process, it is usually divided into three fundamental steps:

1) analyze the source text to determine its meaning,

2) reconstruct that meaning[3] using the target language's structures, lexemes, and world view, and

3) synthesize the final surface forms.

These three steps are often summarized as *analysis*, *transfer*, and *synthesis*.

The information revolution of the past several decades has created a demand for translation far beyond what human translators are capable of fulfilling. Therefore people have looked to computers for assistance. However, after nearly half a century of computational linguistic research, it has become clear that fully automatic, high quality machine translation is not possible given our current state of technology and linguistic knowledge. The most developed machine translation system to date is Google-Translate, and it works reasonably well when translating a document from one language to a related language. However, even with Google's vast resources, when translating a document to an unrelated language, the results are generally unusable. An example of this is provided in appendix A. A short story was randomly selected from a sixth grade Korean textbook, and several paragraphs of that story were entered into Google-Translate. The original text, the results of Google's translation into English, and a human's translation of those paragraphs are shown at the end of appendix A. As seen in that example, the text produced by Google-Translate is incomprehensible. Many machine translation projects have been cancelled because they produced texts that were of unusable quality. For example, the TAUM AVIATION project was intended to translate aircraft maintenance manuals from English to French. However, after several years of development, the project was cancelled because the manual editing of the computer generated texts cost twice as much as manual translation (Whitelock 1995:83).

[3] This project is founded on the meaning-based theory of translation as portrayed by Mildred Larson (Larson 1984). In this context, 'meaning' includes both information and reference of language expressions.

Researchers have found that the vast majority of the difficulties associated with machine translation are encountered during the source analysis stage (Arnold 1994:93-116). In particular, fully automatic part of speech disambiguation, word sense disambiguation, and structural disambiguation have proven elusive. Therefore, in an attempt to circumvent the many difficulties associated with automatic source analysis, researchers have begun developing natural language generators.

Natural language generation is defined as "the process of mapping internal computer representations of information into human language" (Reiter and Dale 2000:3). As was stated in the previous chapter, a natural language generator (NLG) is defined as any computer program that takes a representation of information, applies a language's lexicon and grammar to that representation, and then generates text that communicates part or all of the original information. The representation of information used by an NLG system may be a table or sequence of numbers, or it may be an abstract representation of a sentence or text. Because NLGs do not use natural language text as their input, they avoid the difficulties associated with automatic source analysis.

Manually developed abstract syntactic representations of texts are generally called semantic representations. During the manual development of a semantic representation, each word's part of speech is identified, each word's lexical sense is specified, and the structure of each phrase, clause and sentence is made explicit. Because the manual development of semantic representations eliminates the need for automatic source analysis, NLGs are only required to perform the last two steps of the translation process, namely *transfer* and *synthesis*. Transfer and synthesis are much more mechanical in nature than is analysis and therefore more suitable for computational techniques.

The fundamental tasks that must be performed by every NLG system include: sentence construction, lexicalization, referring expression generation, and linguistic realization (Belz 2007:3). A wide variety of techniques have been developed to accomplish these tasks, and many NLG systems that serve vastly different purposes have been developed. The subsequent sections of this chapter will present several of the most successful NLG systems, and then introduce the NLG that was developed for this dissertation.

2.3 Notable Natural Language Generators

This section begins with a brief description of the two general categories of NLGs, and then provides a high level overview of the four design techniques that computational linguists have developed for building NLG systems. Then this section briefly describes five of the most significant and successful NLGs.

2.3.1 Two Categories of Natural Language Generators

Many notable NLGs have recently been developed using a variety of techniques and for a myriad of purposes. These NLGs may be divided into two broad categories:

- those that use numerical data as their input, and
- those that use semantic representations as their input.

The systems that use numerical data as their source generally begin by summarizing the data, and then generating short, coherent texts in one or more languages that present the most significant facts within that data. The vast majority of these NLG systems use either medical or weather data as their source. Other NLG systems use manually analyzed texts as their input, and then generate drafts of translations of those texts in multiple languages.

2.3.2 Four Design Techniques of Natural Language Generators

NLG systems generally use one of four design techniques:

- *Template based* systems have predefined sentence templates, and words or numbers are substituted into the slots in the templates. The SumTime system described in section 2.3.3.2.2 below is a template based system.

- *Linguistically based* systems generally have lexicons and grammars which are modeled after current linguistic theories. This is the most common approach and most of the systems described in the following paragraphs are linguistically based systems.[4]
- *Corpus based* systems use extensive bilingual corpora. These systems are generally used for translation purposes, and they search source language corpora for sentences that are very similar to the sentences in the text being translated. After finding a similar sentence in a source language corpus, these systems modify and output the corresponding sentence from the target corpus. None of the systems presented below are corpus based because that technique is inappropriate for the system being developed for this project. This project has been designed to deal with a very wide variety of target languages, many of which will have very little written literature and no bilingual corpora.
- *Stochastic based* systems use statistics derived from training corpora in order to select the most likely target realization. A very interesting example of a stochastic based NLG system is called pCRU (Probabilistic Context-free Representationally Underspecified) and was developed from 2006 to 2009 by Anja Belz at the University of Brighton, UK. (Belz 2007)

A helpful website listing many different NLG systems that use these various approaches is entitled NLG Systems Wiki (Bateman 2010a).

NLGs that use numerical data as their input are by far the most common type of NLG being developed today. Therefore the next section will present three of the most successful systems that use numerical data. However, those descriptions will be quite brief because those systems are very different from the system being developed for this dissertation project. In

[4] Although these systems are linguistically based, the depth of their semantic analysis is shallow, and the grammars in these systems are generally able to accommodate only one grammatical model. Because the semantic analysis is shallow, the target languages must be closely related to the source language. Examples of the semantic representational systems and grammars in the KPML and KANT projects will be presented in sections 2.3.4.1 and 2.3.4.2.

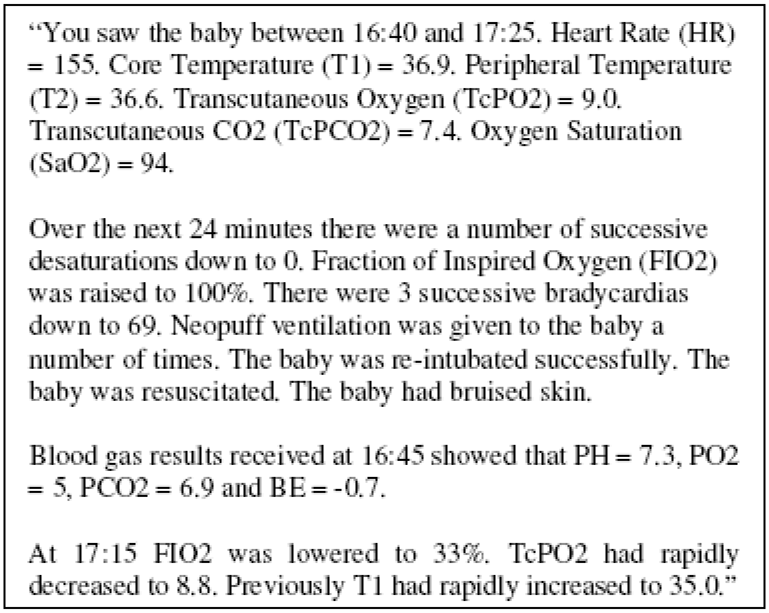
those systems, content determination and text structuring are a large portion of the task, but for systems that use semantic representations, the semantic representations determine the content and structure of the generated texts. The subsequent section will then present two systems that use semantic representations as their input; those discussions will be considerably more thorough because those systems are directly comparable to the NLG developed for this project.

2.3.3 NLG Systems using Numerical Data as the Source

NLG systems that use numerical data as their input scan through the data in order to determine the significant events, and then produce texts in one or more languages summarizing those events. NLG systems of this type have been developed for a very wide variety of domains, but the two most common domains are *medicine* and *weather*. This section will present very high level overviews of one medical NLG system and two weather NLG systems. These systems are very different from TTA, but they have been included here because they have been very successful.

2.3.3.1 Medical Data NLG Systems

The most common domain of numerical data NLG systems is medicine. Since the late '90s numerous systems have been developed to summarize various types of medical information. One system of this type is called BabyTalk (Hunter et al. n.d.a), and it is being developed at the Department of Computing Science at the University of Aberdeen, UK. This project was begun in 2007, and has the following three goals: 1) Interpret physiological data such as heart rate, blood pressure, temperature, O_2 and CO_2 saturations, etc., and also interpret data related to individual events such as laboratory results, probe removal, drug administration, and other procedures performed by the attending nurses or doctors. 2) Create written summaries of the data. 3) Tailor the summaries for three particular target audiences: the doctor, the attending nurses, and family members. A sample text generated for a doctor is shown below in Figure 2-1. The text in that figure illustrates both the content and quality of the texts generated by the BabyTalk project.



"You saw the baby between 16:40 and 17:25. Heart Rate (HR) = 155. Core Temperature (T1) = 36.9. Peripheral Temperature (T2) = 36.6. Transcutaneous Oxygen (TcPO2) = 9.0. Transcutaneous CO2 (TcPCO2) = 7.4. Oxygen Saturation (SaO2) = 94.

Over the next 24 minutes there were a number of successive desaturations down to 0. Fraction of Inspired Oxygen (FIO2) was raised to 100%. There were 3 successive bradycardias down to 69. Neopuff ventilation was given to the baby a number of times. The baby was re-intubated successfully. The baby was resuscitated. The baby had bruised skin.

Blood gas results received at 16:45 showed that PH = 7.3, PO2 = 5, PCO2 = 6.9 and BE = -0.7.

At 17:15 FIO2 was lowered to 33%. TcPO2 had rapidly decreased to 8.8. Previously T1 had rapidly increased to 35.0."

Figure 2-1. Sample of Text Generated by BabyTalk (Hunter et al. 2008:3)

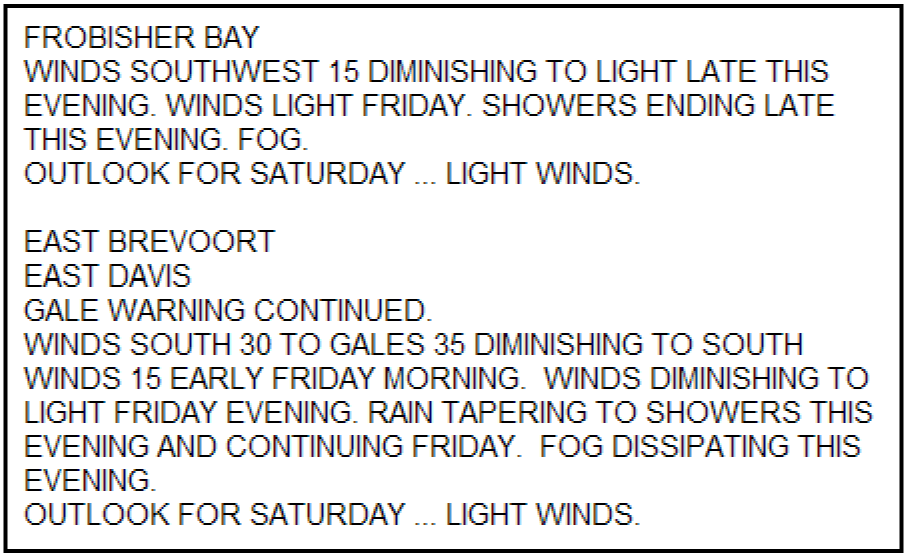
The developers of BabyTalk performed two sets of experiments in order to ascertain the utility of the generated texts. The purpose of the first set of experiments was to determine the quality of the generated summaries, and the purpose of the second set of experiments was to determine the quality of the decisions made by the medical staff after looking at the computer generated summaries. The first set of experiments indicated that the quality of the generated summaries was sufficient because "the decisions made by medical and nursing staff after reading the summaries were as good as those made after viewing the currently available graphical presentations with the same information content" (Hunter et al. 2008:1). The second set of experiments compared the appropriateness of decisions made by the medical staff after looking at 1) the graphical data, 2) summaries of the graphical data written by humans, and 3) summaries generated by BabyTalk. The average score after looking at the graphical data was .33, the average score after looking at the human written summaries was .39, and the average score after looking at the computer generated summaries was .34 (Hunter et al. 2008:4). The developers concluded that they can improve their score by increasing the amount of discourse content included in the computer generated summaries.

2.3.3.2 Weather Data NLG Systems

The second most common domain of numerical data NLG systems is weather. Weather conditions change rapidly, so forecasts must be updated several times each day. These forecasts are used by the general public, airlines, commercial fishermen, farmers, the military, and many other personnel. In regions like Europe and Canada where multiple languages are spoken and people travel extensively from one region to another, people need to know the very latest weather forecasts. Translating all of these forecasts into multiple languages several times each day is prohibitive. Therefore many systems have been developed in Europe and Canada which use tables of weather data as their input, and they generate summaries of that data and forecasts in one or more languages. Two of these systems include Fog and SumTime.

2.3.3.2.1 FoG (Bateman 2009)

FoG (Forecast Generator) is an NLG system developed to produce weather forecasts in English and French for the Canadian Weather Agency. It was developed from 1989 to 2000 by Eli Goldberg, Norbert Driedger, and Alain Polguère who work for CoGenText (Reiter and Dale 2000:9), and Richard Kittredge at the University of Montreal. This system produces textual weather forecasts from numerical weather data that has been annotated by a human forecaster. Because FoG is multilingual, it first produces a language-independent abstract representation of the forecast text, and that representation is then mapped to each output language using the appropriate lexical and grammatical resources. The system uses Meaning ↔ Text Theory during the generation of the surface texts (Sripada et al. 2004:762). A sample forecast generated by FoG in English is shown below in Figure 2-2. The text in this figure illustrates the content and style of the weather forecasts generated by FoG.



```
FROBISHER BAY
WINDS SOUTHWEST 15 DIMINISHING TO LIGHT LATE THIS
EVENING. WINDS LIGHT FRIDAY. SHOWERS ENDING LATE
THIS EVENING. FOG.
OUTLOOK FOR SATURDAY ... LIGHT WINDS.

EAST BREVOORT
EAST DAVIS
GALE WARNING CONTINUED.
WINDS SOUTH 30 TO GALES 35 DIMINISHING TO SOUTH
WINDS 15 EARLY FRIDAY MORNING.  WINDS DIMINISHING TO
LIGHT FRIDAY EVENING. RAIN TAPERING TO SHOWERS THIS
EVENING AND CONTINUING FRIDAY.  FOG DISSIPATING THIS
EVENING.
OUTLOOK FOR SATURDAY ... LIGHT WINDS.
```

Figure 2-2. Sample Forecast Generated by FoG (Reiter and Dale 2000:11)

As seen above in figure 2-2, the generated text intentionally has a “telegraphic” style, meaning that the verbs are generally not marked for tense, articles are often omitted, and other stylistic words are not included in order to mimic the style used by human forecasters (Goldberg et al. 1994:5). No experiments have been performed to evaluate the quality of the texts generated by FoG.

2.3.3.2.2 SumTime (Hunter et al. n.d.b)

The SumTime weather system is being developed in the Department of Computer Science at the University of Aberdeen, Scotland, in cooperation with Aerospace and Marine International. The project was started in 2001 and is still under development. The goal of the project is to develop a computer program that will accurately summarize time-series weather data in English. SumTime uses numerical weather data as its input, and generates short texts consisting of a few sentences which summarize that data. The system is currently being used to produce drafts of marine forecasts for the offshore oilrigs near Aberdeen. The system produces 150 draft forecasts each day for Weathernews Ltd UK, and those drafts are then edited by forecasters and released to oil company staff who support the offshore oilrig operations in the North Sea (Sripada et al. 2004:760).

In order to compare the quality of the texts generated by SumTime with similar weather forecasts written by meteorologists, the developers selected five weather forecasts that had been written by five different meteorologists in September of 2000 (Reiter et al. 2005:2). They then used the same wind data that the meteorologists had used for their forecasts, and generated summaries of that data. Using both the computer generated summaries and the meteorologists' summaries, they produced a hybrid summary by editing the texts written by the meteorologists so that they used the same style, words, and punctuation as the software. They then asked seventy-two people who had significant experience reading weather forecasts to read, interpret and compare the human-written forecasts, the hybrid forecasts, and the computer generated forecasts. In particular these people were asked to judge the three forecasts to determine which was easiest to read, and which was most appropriate given the wind data. The results of the first experiment are as follows: 51% said the hybrid forecasts were easiest to read, 33% said the hybrid and manually written forecasts were equally easy to read, and 17% said the manually written forecasts were easiest to read. In the second experiment where people were asked to judge which of the three forecasts was most appropriate, 43% of the respondents said the computer generated forecasts were most appropriate, 30% said the computer generated and manually written forecasts were equally appropriate, and 27% said the manually written forecasts were most appropriate. The differences in the second experiment were too small to be statistically significant.

2.3.4 NLG Systems that Use Meaning Specifications as the Source

All of the NLG systems presented in the previous section use either numerical medical data or numerical weather data as their input. Those systems scan through the data, extract the significant events, and then generate texts describing or summarizing those events. Those systems are much more common and have been much more successful than the systems that will be described in this section. This section will describe two of the most significant NLG systems that use meaning specifications as their input. A meaning specification is any abstract, annotated representation of a sentence or text. The format and content of these meaning specifications vary widely from one project to another, as do the domains that they cover. Several domains covered by this type of system include (Bateman 1997:3): the generation of letters responding to customer queries (Springer et al. 1991, Coch et al. 1995), the automatic production of technical documentation (Reiter et al. 1995, Rosner and Stede 1994), of instructional texts (Paris et al. 1995), of patent claims (Sheremetyeva et al. 1996), and in natural language interfaces to databases and information systems (Bateman and Teich 1995). This section will present two systems in this domain. The first system, called KPML, generates drafts of large texts in multiple languages, and the second system, KANT, generates repair manuals in multiple languages.

2.3.4.1 KPML (Bateman 2010b)

The Komet-Penman Multilingual (KPML) system is the most widely used NLG system that uses semantic representations for its input. It is a large scale grammar development environment based on Systemic Functional Grammar (SFG). KPML has a very large English grammar, moderately sized grammars of German and Dutch, and small illustrative grammars of Spanish, French, Japanese, Czech, Russian, Bulgarian, Greek, and Chinese. Grammars of other languages are currently being developed.

2.3.4.1.1 KPML's Generation Methodology

KPML accepts several knowledge representation systems as input, and also includes a module so that users can transform a new knowledge representation system into a format that KPML recognizes. The most common type of input used in KPML is called the Sentence Planning Language (Reiter and Dale 2000:171). An example of the Sentence Planning Language is shown below in Figure 2-3. This example illustrates the type and depth of analysis that is done manually to produce KPML's source texts.

```
(S1 / generalized-possession
 :tense past
 :domain (N1 / time-interval
            :lex march
            :determiner zero)
 :range (N2 / time-interval
            :number plural
            :lex day
            :determiner some
            :property-ascription
            (A1 / quality :lex rainy)))
```

Figure 2-3: Sentence Planning Language Representation of *March had some rainy days.*

In the figure seen above, note that both nouns are marked with a field called ‘determiner’, and the value for *March* is ‘zero’ while the value for *days* is ‘some’. The field ‘tense’ has a value of ‘past’, and the proposition’s event is ‘generalized-possession’. The analysis of the proposition is very shallow in comparison to the semantic analysis required in TTA’s semantic representations as will be demonstrated in the next chapter.

In order to generate target text from these meaning representations, KPML uses SFG which categorizes the resources of a language according to their functions. John Bateman, the principal architect of the KPML system, claims that SFG is more appropriate than a generative grammar during the generation process (Bateman 1997:9):

> Rather than adopting the structural/generative paradigm made dominant by American linguistics, NLG looks equally, if not more, to the functional linguistic paradigm including text linguistics, functional linguistics, and pragmatics in its broadest sense. Approaches to linguistics from this perspective consider the relation of language to social and psychological factors as determinative of its organization. It is usual within this paradigm to investigate the conditions of use of particular expressions found in natural languages and this is precisely what is necessary for a sentence generation program that is to generate sentences appropriate to their contexts of use. The functional linguistic paradigm includes a very broad and active range of work—often not formal, and usually with no thoughts of computational application. For computational sentence generation, however, it is necessary not only to find functional descriptions but also to find such descriptions that may be made sufficiently precise as to be embodied in computer programs.

Therefore the resources in the KPML sentence planning language are represented in a systemic network, and each node in the network represents a set of choices, each choice serving a particular function. So a systemic grammar is a system of choices, and at each node the function of the utterance determines the choice that is made. The KPML authors provide an illustration of the mood system for English as shown in Figure 2-4.

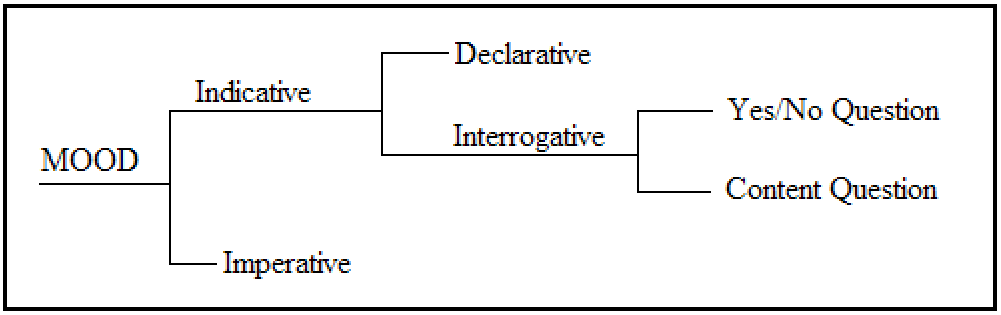


Figure 2-4. Systemic Grammar Representation of English Mood (Reiter and Dale 2000:176)

After a systemic network for a subsection of a language has been developed, the KPML grammar execution module walks through the network making decisions at each node based on the information in the meaning representation. The system starts with the highest ranking unit in the meaning representation, typically a sentence. The system then walks through the systemic network making decisions relevant to that unit. At each node there is a default choice so that if the meaning representation does not include a specification for that particular node, the grammar executor chooses the default. After walking through the network, the process is then repeated with the next highest ranking unit, typically noun phrases. Eventually the grammar executor has walked through the network for each unit in the meaning representation and made the relevant decisions. The module that makes the decisions at each node is called Inquiry Semantics (Reiter and Dale 2000:177). That module provides a bridge between the semantics in the meaning representation and the generation of surface forms by gathering the pertinent information and examining the options that are available. After examining the network to determine the options that are available, and after gathering the pertinent information from the relevant sections of the meaning representation, the Inquiry Semantics calls a module known as Choosers (Reiter and Dale 2000:177).

Choosers are small sections of software specifically developed for one or more nodes in the network, and they decide which path in the network is appropriate. A Chooser is shown below in Figure 2-5.

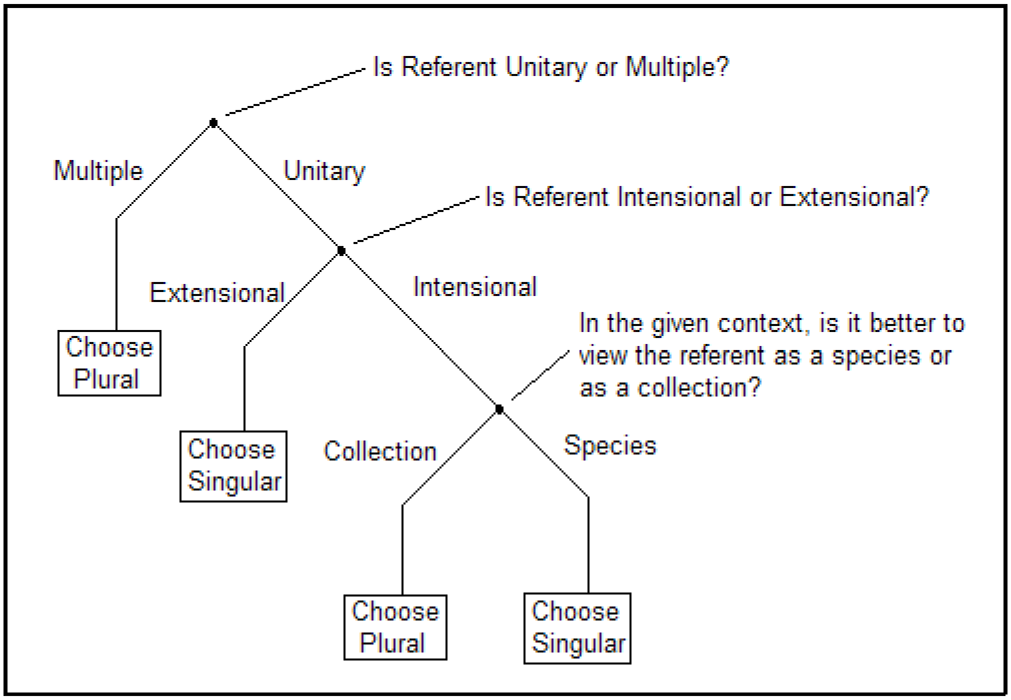

Is Referent Unitary or Multiple?
Multiple
Unitary
Is Referent Intensional or Extensional?
Extensional
Intensional
Choose Plural
In the given context, is it better to view the referent as a species or as a collection?
Choose Singular
Collection
Species
Choose Plural
Choose Singular

Figure 2-5. A Chooser in KPML (Reiter and Dale 2000:177)

The Chooser shown above is responsible for deciding whether the singular or plural form of a particular noun should be used. The term ‘chooser’ and the other terms in the figure above are generally foreign to linguists because this system was developed from a computational linguist’s perspective.

After a Chooser has made a decision, the grammar executor calls a Realization Statement which contributes to the building of the surface representation. A Realization Statement for English mood is shown below in Figure 2-6.

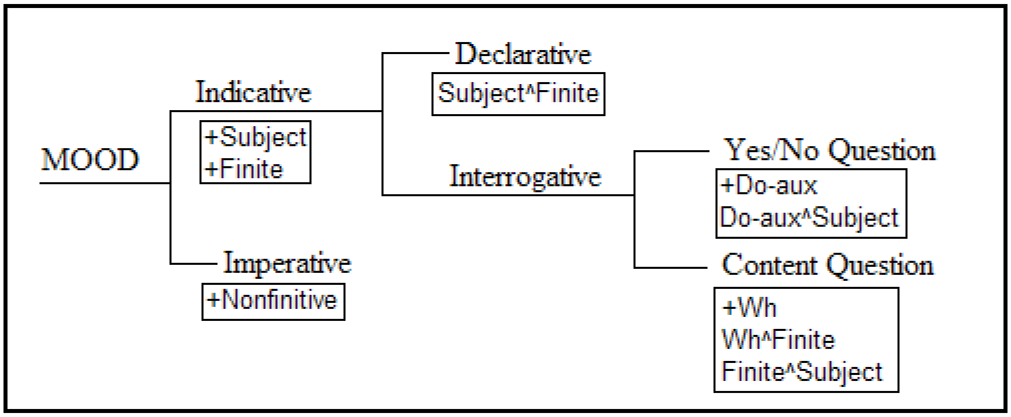


Figure 2-6. A KPML Realization Statement for English Mood (Reiter and Dale 2000:178)

The mood terms shown in the figure above are certainly familiar to linguists, but the other terms in the figure and the style of notation were developed by computational linguists in order to facilitate computational algorithms.

Some sample texts in English generated by KMPL were sent by John Bateman and are shown below (letter to the author, September 2008). These texts illustrate the quality and complexity of the English texts generated by KPML. Unfortunately the source texts for these sample texts were unavailable.

> Flu is primarily a seasonal disease, occurring almost exclusively in Autumn and Winter. It is a very contagious viral infection that affects the respiratory system. The disease is transmitted by direct contact with contaminated particles emitted during coughing, sneezing, or in normal conversation. (Health info, 1993)
>
> Behrens's principal activities were architecture and industrial design. He made electrical appliances and prototype flasks. He built the high tension plant and the turbine factory for AEG in 1908 - 1910. He built a housing area for the workers of AEG in Henningsdorf. He created a number of monumental buildings, such as the administration building of Mannesmann in Duesseldorf and the German embassy in St. Petersburg. (Text planning, 1994)
>
> At the next two junctions go straight on, and then, turn left at a t-intersection. Turn left after a chemist. At the next junction turn right, and then, at the next junction turn right again. Turn left at a t-intersection. Turn right after a supermarket. (Route planning, 2005)

2.3.4.1.2 Evaluating the Texts Generated by KPML

When the author of this dissertation asked Professor Bateman for a text that had been generated in two or more languages, he wrote that the grammars for the various languages cover very different areas, and he did not have a significant text that was generated in multiple languages (letter to the author, September 2008). Therefore it was not possible to compare the quality of a text generated in one language with the quality of the same text in another language. No discussion of experiments which evaluate KPML's generated texts was found in the literature.

However, another project called Automatic Generation of Instructions in Languages of Eastern Europe (AGILE)[5] used the KPML generator, and those researchers performed experiments to ascertain the quality of the texts generated by their system. The purpose of AGILE was to generate drafts of CAD-CAM (computer assisted design – computer assisted manufacturing) instruction manuals in Bulgarian, Czech, Russian. The texts generated by AGILE were evaluated according to two parameters: acceptability and grammaticality. In order to evaluate the texts for acceptability, six speakers of each target language were selected from groups of people who had experience in either writing or translating software manuals. These evaluators were told that the AGILE texts had been produced by human translators, and they were asked to compare the texts with other instructions which had been manually translated. The results of the Czech experiment indicate that the AGILE texts were considered of the same quality as the manually translated texts in 35% of the cases, the manually translated texts were considered of better quality in 24% of the cases, and the AGILE texts were considered better in 41% of the cases. The results of the experiments in the other two languages were not reported. In order to evaluate the texts for grammaticality, two linguists from each target language were asked to examine the generated texts. None of the evaluators found any grammatical errors,

[5] AGILE was developed at the University of Brighton, UK from 1999 to 2001. The project's web site is at www.itri.brighton.ac.uk/projects/agile accessed December 11, 2010.

but in a few cases word order issues were mentioned. However, those issues were considered to be stylistic rather than actual grammatical errors (Hana 2001:65).

2.3.4.2 KANT (Nyberg 2004)

The KANT (Knowledge-based, Accurate Natural-language Translation) system was developed at the Center for Machine Translation at Carnegie Mellon University. The project began in 1989, but was terminated in the early 2000s because the department shifted its resources away from knowledge based techniques toward statistical techniques. The KANT system is a suite of commercially available software tools that enables users to do the following:

- develop source documents using the KANT document authoring tools,
- automatically and interactively analyze the source documents in order to produce a suitable interlingua representation,
- build grammars and lexicons for target languages so that the generator can produce drafts of translations of texts in these languages, and
- generate target texts from the interlingua representations in multiple languages.

This system was used primarily to translate technical texts in very specific subdomains such as maintenance manuals for power plants and heavy industrial equipment. The system's primary customer was Caterpillar tractors, and at the project's peak, this system was translating hundreds of thousands of pages per year from English to French for Caterpillar. The system's domain sublanguage contains approximately 65,000 words and technical phrases (Nyberg et al. 1997:3), and extensive grammars have been developed for French, Spanish, and German (Nyberg and Mitamura 2000:2). Prototype grammars have also been developed for Chinese (Li et al. 1996:1) and Japanese (Nyberg and Mitamura 1992:5). In the late 1990s the system went through a major redesign in order to take advantage of new computational techniques. The system was then called KANTOO, where the "OO" stands for "object oriented", and the system's new architecture is shown below in figure 2-7.

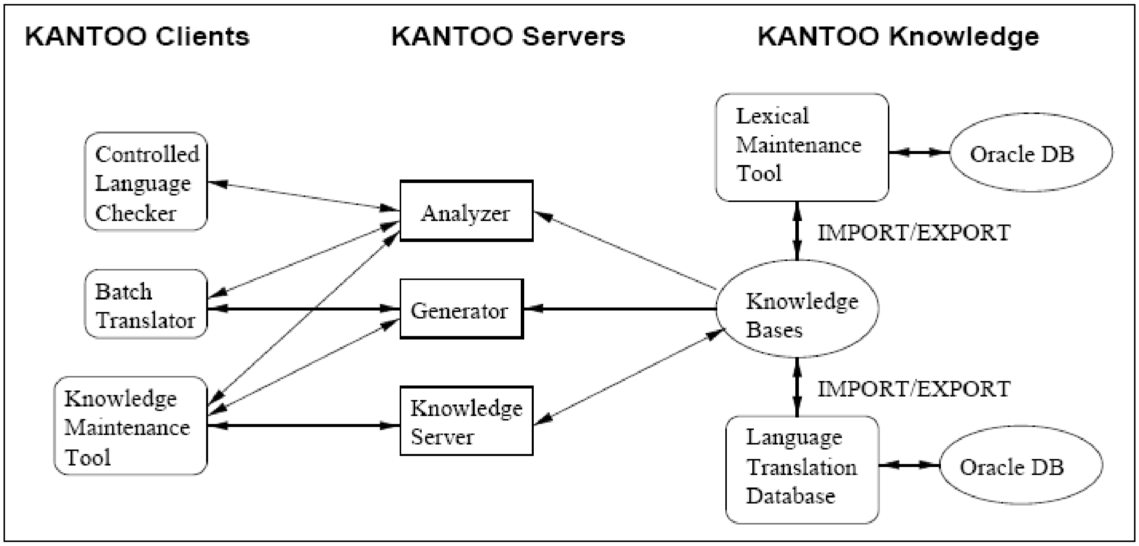
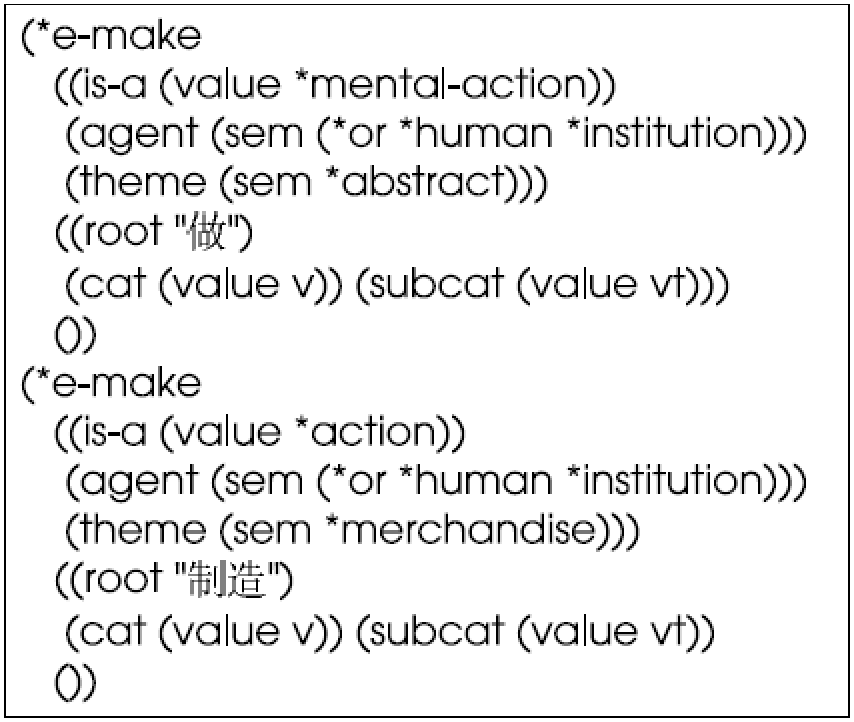


Figure 2-7. KANTOO System Architecture (Nyberg and Mitamura 2000:2)

The controlled language checker seen in the upper left corner of figure 2-7 is used while writing a document to verify that the author is using words, phrases, and sentence structures that are permitted in the interlingual representations (Mitamura and Nyberg 2001:2). After a document has been written using the controlled language checker, the analyzer module, which is also shown in figure 2-7 above, produces an interlingual representation of that document. An example of the interlingual representations will be provided below. The lexical maintenance tool, which is shown in the upper right corner of figure 2-7 above, is used to create, modify, and navigate through the source concepts that are permitted in the interlingual representations. The language translation database, which is shown in the lower right corner of figure 2-7 above, stores all the target words that are used in the translation process. The knowledge maintenance tool, which is shown in the lower left corner of the figure, stores the grammars for the source and target languages, and allows the users to modify the various grammatical rules (Mitamura et al. 2001:1). Several examples of the target language rules will be provided below. This system is primarily a knowledge based system, but it also uses stochastic techniques when appropriate (Carbonell et al. 1992:4). The following discussion will focus on the interlingua and the generation component because those are the sub-systems most directly comparable to other NLG systems.

2.3.4.2.1 KANT’s Generation Methodology

KANT’s interlingua is based on the notion of *concept frames* (Li et al. 1996:3). Concept frames are generally composed of objects, events, and properties, which typically represent nouns, verbs, and adjectives. Each concept has its frame specified in the ontology. For example, the concept frames for two senses of *to make* are shown below in figure 2-8.

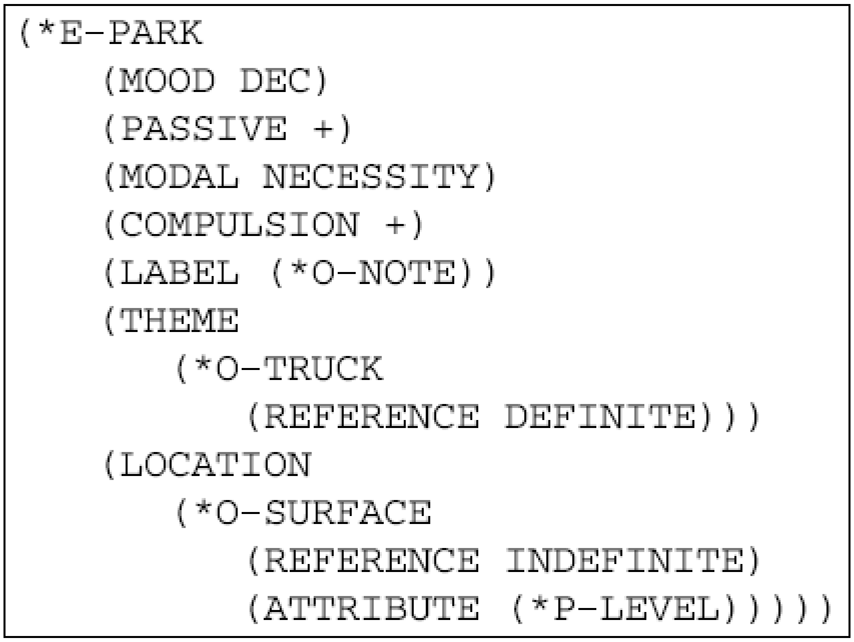


```
(*e-make
  ((is-a (value *mental-action))
   (agent (sem (*or *human *institution)))
   (theme (sem *abstract)))
  ((root "做")
   (cat (value v)) (subcat (value vt)))
  ())
(*e-make
  ((is-a (value *action))
   (agent (sem (*or *human *institution)))
   (theme (sem *merchandise)))
  ((root "制造")
   (cat (value v)) (subcat (value vt))
  ())
```

Figure 2-8. A Section of KANT’s Ontological Entry for MAKE (Li et al. 1996:5)

As seen in the figure above, the first sense of MAKE deals with mental actions, it takes a human or an institution as an agent, and its theme is an abstract object. The second sense of MAKE deals with generic actions, it takes a human or an institution as an agent, and its theme is some type of merchandise. The first sense of MAKE occurs in a sentence such as *The man made alterations to the device.* The second sense of MAKE occurs in a sentence such as *The antifreeze was made by this company.* Also seen in the figure above is the target language equivalent for each sense of MAKE; in this particular figure, the Chinese equivalent is shown. During the generation process, the various senses of each word in the interlingual representation are examined along with their associated frames, and the sense that has the argument structure which most closely fits the sentence being translated is selected. Then the target equivalent of that sense is inserted into the generation process. If a transformation is required by the target language equivalent, then the necessary transformation rules are also stored in the concept's entry. Those rules will be executed after the sense has been selected and before the target equivalent is inserted into the text.

In the KANT system, a semantic representation of a source sentence or text is a semantic network composed of instances of language independent concepts from the ontology. A concept frame consists of a concept name and an arbitrary number of slot-value pairs. So a sentence typically consists of an event frame which has a specified predicate followed by the relevant arguments. Additional information such as modality, speech act, focus, etc., is also encoded in the concept frame. An example of an interlingual representation of a sentence is shown below in figure 2-9. This example illustrates the type of analysis that is used in the KANT interlingua.



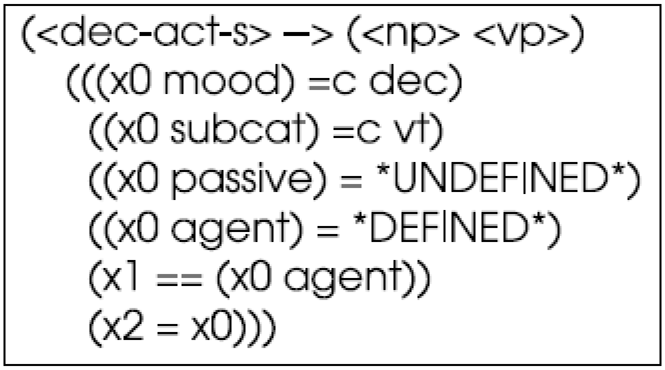
```
(*E-PARK
    (MOOD DEC)
    (PASSIVE +)
    (MODAL NECESSITY)
    (COMPULSION +)
    (LABEL (*O-NOTE))
    (THEME
       (*O-TRUCK
          (REFERENCE DEFINITE)))
    (LOCATION
       (*O-SURFACE
          (REFERENCE INDEFINITE)
          (ATTRIBUTE (*P-LEVEL)))))
```

Figure 2-9. KANT's Interlingual Representation of *The truck must be parked on a level surface.* (Li et al. 1996:4)

As seen in figure 2-9 above, the interlingual representation of a sentence begins with the verb, and each verb has particular semantic roles that are specified in the lexical maintenance tool. The mood, tense, and punctuation are specified, and the NPs that fill the various semantic roles are also specified. The analysis shown above is linguistically shallow, and the representation is foreign to linguists.

In order to generate target surface text from an interlingual representation, three steps are performed: 1) lexical selection, 2) f-structure creation, and 3) syntactic generation (Li et al. 1996:4). In the lexical selection step, the most appropriate target lexeme or phrase is selected for each frame in the interlingual representation. Then the interlingual representation and the target lexemes are examined to construct a syntactic functional structure (f-structure) for the target proposition. Then the syntactic generator uses the target language's rules to inflect and properly order the target constituents.

The generator uses a grammar formalism called *Pseudo Unification Grammar* (Li et al. 1996:6). Each rule consists of a context-free phrase structure description and a cluster of pseudo equations. The equations are used to check attribute values and to either construct or disassemble the sentence’s f-structure. A sample rule is shown below in figure 2-10.

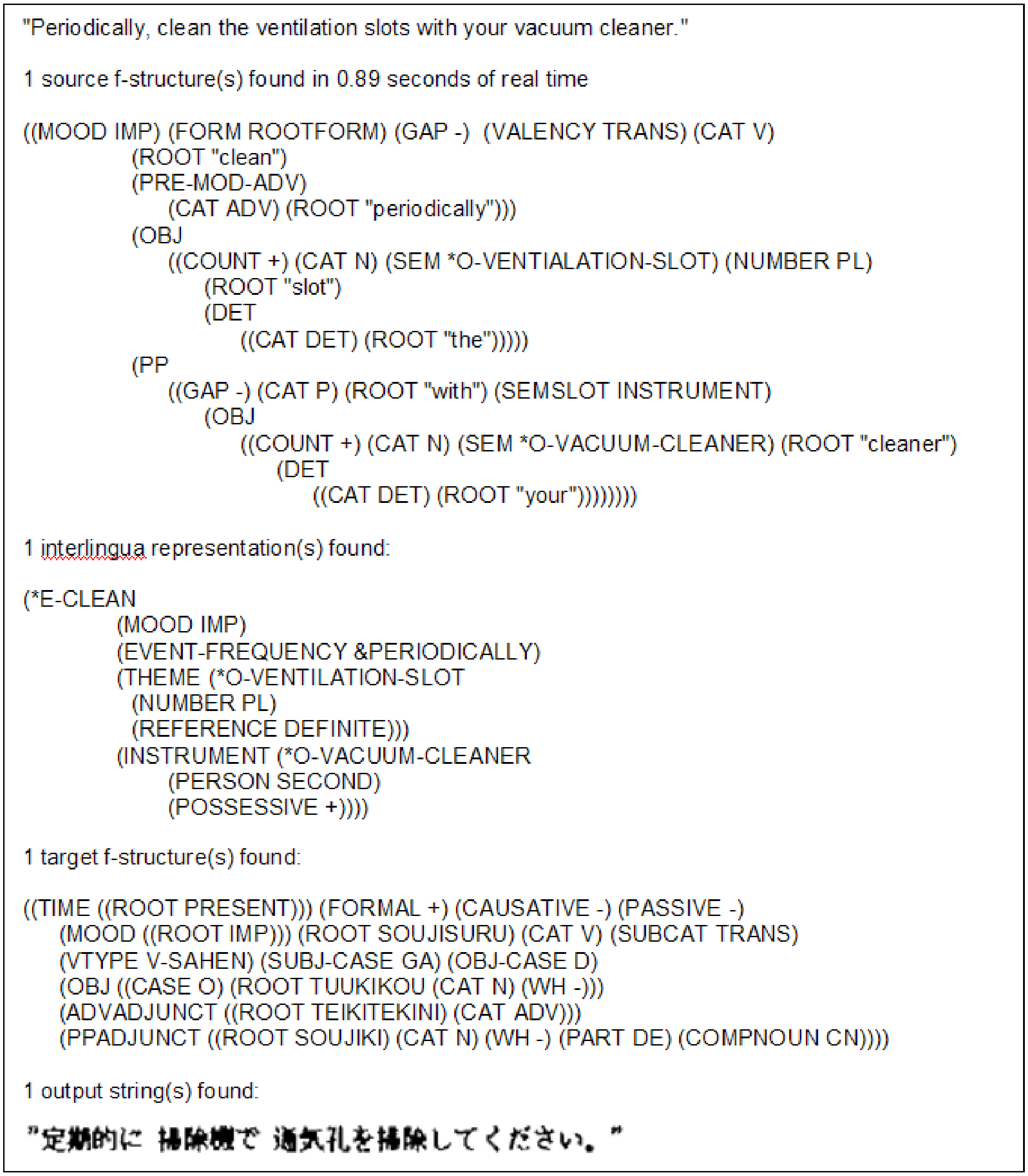


```
(<dec-act-s> —> (<np> <vp>)
   (((x0 mood) =c dec)
    ((x0 subcat) =c vt)
    ((x0 passive) = *UNDEFINED*)
    ((x0 agent) = *DEFINED*)
    (x1 == (x0 agent))
    (x2 = x0)))
```

Figure 2-10. Example of a KANT Generation Grammar Rule

The rule shown above in figure 2-10 looks to see if a particular sentence has declarative mood, a transitive verb, and is active. When that situation is found, the rule will move the agent NP into X1, so it will be realized as an NP, and the rule will put the rest of the sentence into X2, so it will be realized as a VP. Although the rule shown above is performing tasks that linguists routinely discuss, the style of representation is unfamiliar to linguists.

The complete KANT analysis and generation process is illustrated below in figure 2-11 for Japanese.



```
"Periodically, clean the ventilation slots with your vacuum cleaner."

1 source f-structure(s) found in 0.89 seconds of real time

((MOOD IMP) (FORM ROOTFORM) (GAP -)  (VALENCY TRANS) (CAT V)
            (ROOT "clean")
            (PRE-MOD-ADV)
                (CAT ADV) (ROOT "periodically")))
            (OBJ
                ((COUNT +) (CAT N) (SEM *O-VENTIALATION-SLOT) (NUMBER PL)
                    (ROOT "slot")
                    (DET
                        ((CAT DET) (ROOT "the")))))
            (PP
                ((GAP -) (CAT P) (ROOT "with") (SEMSLOT INSTRUMENT)
                    (OBJ
                        ((COUNT +) (CAT N) (SEM *O-VACUUM-CLEANER) (ROOT "cleaner")
                            (DET
                                ((CAT DET) (ROOT "your"))))))))

1 interlingua representation(s) found:

(*E-CLEAN
          (MOOD IMP)
          (EVENT-FREQUENCY &PERIODICALLY)
          (THEME (*O-VENTILATION-SLOT
            (NUMBER PL)
            (REFERENCE DEFINITE)))
          (INSTRUMENT (*O-VACUUM-CLEANER
                (PERSON SECOND)
                (POSSESSIVE +))))

1 target f-structure(s) found:

((TIME ((ROOT PRESENT))) (FORMAL +) (CAUSATIVE -) (PASSIVE -)
    (MOOD ((ROOT IMP))) (ROOT SOUJISURU) (CAT V) (SUBCAT TRANS)
    (VTYPE V-SAHEN) (SUBJ-CASE GA) (OBJ-CASE D)
    (OBJ ((CASE O) (ROOT TUUKIKOU (CAT N) (WH -)))
    (ADVADJUNCT ((ROOT TEIKITEKINI) (CAT ADV)))
    (PPADJUNCT ((ROOT SOUJIKI) (CAT N) (WH -) (PART DE) (COMPNOUN CN))))

1 output string(s) found:

"定期的に 掃除機で 通気孔を掃除してください。"
```

Figure 2-11. Generation of Japanese Equivalent for *Periodically, clean the ventilation slots with your vacuum cleaner.* (Nyberg and Mitamura 1992:5)

As seen at the top of figure 2-11, the source sentence being translated is *Periodically, clean the ventilation slots with your vacuum cleaner.* The analyzer that was shown in figure 2-7 above constructs a source f-structure for the source sentence, and that f-structure is displayed in the top portion of this figure. Then the interlingual representation for that sentence is produced, and it is shown in the middle region of figure 2-11. From the interlingual representation, the generator that was seen in figure 2-7 above constructs a target f-structure, and it is shown in the bottom portion of the figure. Finally the system produces the Japanese translation of the sentence, and it is shown at the bottom of figure 2-11.

2.3.4.2.2 Evaluating the Texts Generated by KANT

Experiments were performed to compare how much time was required to manually translate a particular document with how much time was required to post-edit KANT's generated draft of the same document. In order to edit KANT's texts quickly, a native French speaker who was an experienced translator, but who did not have any expertise in this particular subdomain, was trained to use the KANT post-editing tools, and he was instructed to do the minimal amount of editing required to make the translations understandable and accurate (Nyberg et al. 1997:4). The experiments were performed with documents which typically take approximately an hour to manually translate. The trained French editor was able to edit KANT's drafts of these documents in approximately 10 to 15 minutes, thereby suggesting an increase of productivity by a factor of 4 to 5 (Nyberg et al. 1997:4). The minimally post-edited output was then reviewed by the customer who confirmed that the text was of sufficient quality that it would be "highly useful." However, the customer then performed his own set of experiments using his translators who had a broad range of experience. The customer's experiments indicated that using KANT's texts approximately doubled the productivity of his translators (Nyberg et al. 1997:4). The researchers developing KANT concluded that the customer's translators were doing more than the minimal editing required while post-editing KANT's texts.

2.4 The Translator's Assistant

This section will provide a brief overview of the natural language generator called The Translator's Assistant. TTA is distinct from other NLGs in two very specific ways that were presented in the previous chapter: 1) TTA was designed and developed using contemporary typological research and theoretical frameworks that are familiar to linguists, and 2) TTA is intended to generate texts in a very wide variety of target languages. Now that other NLGs have been described, several of the notable differences between TTA and the other systems will be discussed in more detail.

2.4.1 Distinguishing Features of TTA's Semantic Representational System

- TTA's semantic representational system includes a feature system that is comprised of primitives that have been gathered from a large array of languages. The feature systems used in the other NLGs and portrayed in the figures above are generally adequate for Indo-European languages. For example, figure 2-9 shows an example of the Sentence Planning Language for KPML. That figure show a 'Tense' value of 'Past', and the nominals are marked with a 'Determiner' value of 'zero', 'some', or 'the', and a 'Number' value of 'Plural'. Values such as those are perfectly adequate when generating texts in Indo-European languages, but many languages don't have anything equivalent to the English articles, they have many degrees of past and future tenses, and their nominals may be morphologically marked for singular, dual, trial, quadrial, and plural. Therefore, the feature system developed for TTA's semantic representations contains much more information than do the feature systems developed for the other NLGs.
- The concepts in TTA's semantic representations come from TTA's ontology which was developed using the foundational principles of Natural Semantic Metalanguage (NSM) theory. Because TTA is intended to be used in a wide variety of target languages, its semantic representations are comprised primarily of semantic primitives and other compositionally simple concepts. By using semantically simple concepts, the probability of other languages having lexical equivalents is increased. The rationale for this approach and the NSM foundational principles will be presented in section 3.2.6. The systems portrayed above generally do not limit the concepts in their semantic

representational systems to semantically simple concepts, primarily because they are dealing with very sophisticated concepts (e.g., “oxygen saturation,” “power grids,” “vacuum cleaners,” etc.), and it is assumed that their target languages generally have either lexical or analytic equivalents for these concepts.

- The concepts in TTA’s ontology are very narrowly defined and used consistently throughout all the semantic representations. Similar to TTA’s approach, other NLGs use ontologies that include concepts which have multiple senses. For example, figure 2-8 shown above illustrates that the concept *make* in KANT’s ontology has at least two senses. The ontology in TTA uses the same approach, but the meaning of each concept is more narrowly defined than in the other systems. As was mentioned in chapter 1, TTA’s ontology includes twenty-five senses of BE. Many of the concepts in TTA’s ontology have multiple senses, but each sense is very narrowly defined and used in virtually identical constructions throughout all the semantic representations.
- TTA’s ontology includes semantically complex concepts which are inserted into the semantic representations automatically if a particular target language has a lexical equivalent. If the texts generated by TTA were to consist solely of semantically simple concepts, the texts would seem unnecessarily long and drawn out. For example, English speakers don’t want to read texts that include constructions such as “doctors who treat sick animals”; English speakers would rather read about “veterinarians.” Sections 3.3.1 and 4.3.1 will describe how semantically complex concepts are automatically inserted into the semantic representations if the target language has a lexical equivalent.
- TTA’s semantic analysis is considerably richer and more detailed than the analysis systems used by other NLGs because it attempts to identify the reasons why surface structures in the source documents have the forms that they do. As was mentioned above, figure 2-3 shows nominals marked with a ‘Determiner’ value. TTA’s nominals

are marked with a feature called ‘Participant Tracking’ rather than ‘Determiner’ because it is a nominal’s participant tracking value that is the underlying universal which triggers particular determiner values in Indo-European languages. None of the NLGs described above refer to a text’s discourse genre or a proposition’s salience band, but Longacre points out that these two parameters are very significant when determining the proper surface form of a proposition’s verb (Longacre 1996:27). For example, a salience band value of ‘Backgrounded Action’ requires the participial verb form in English, and a value of ‘Flashback’ triggers the past perfect form. Section 3.3.3.9 will discuss this analysis in more detail.

2.4.2 Distinguishing Features of TTA’s Grammar

- TTA’s grammar is intentionally theory neutral, and it has been designed with sufficient flexibility to permit linguists to develop their grammars using a variety of theoretical models. TTA’s grammar is able to accommodate Transformational-Generative models, Lexical-Functional models, Minimalist models, Role and Reference models, and Functional-Typological models. Rules in TTA’s transfer grammar are able to insert into the underlying representations an INFL node, CP and CP-Spec nodes, or any other constructs that the linguist desires. Then the synthesizing grammar is able to manipulate these nodes according to the linguist’s specifications. TTA’s lexicon and grammar will be described in detail in chapter 4.
- The transfer component of TTA’s grammar includes rules which are able to generate honorifics when people speak to one another. None of the NLGs presented above discuss the ability to generate appropriate honorific forms when people speak to one another, but in a language such as Korean, generating the appropriate honorifics is essential. If a source document is in English, there is no method of encoding honorifics in the surface text, and it appears that no other NLGs have a method of generating

honorifics. A detailed example of how TTA generates honorifics for Korean will be presented in section 5.2.

- The transfer component of TTA's grammar includes the capability to map a single source concept to multiple target words based on the context. As was mentioned in chapter 1, some languages have a single word for a range of concepts, but other languages will have multiple words for those same concepts. Collocation correction rules were added to TTA's transfer grammar to deal with this specific problem, but no other NLG appears to have dealt with this issue. If an NLG generates texts only in Indo-European languages, then this problem is perhaps sufficiently small that it may be ignored. However, if a NLG generates texts in a very wide variety of target languages, this problem becomes significant and must be resolved. Therefore collocation correction rules were added to TTA's transfer grammar, and they will be discussed in section 4.3.6.
- The synthesizing component of TTA's grammar was designed to resemble as closely as possible the descriptive grammars that field linguists routinely write. As was mentioned above in section 2.3.4.1.1, the KPML project uses Systemic Functional Grammar for the reasons cited, but SFG is generally unfamiliar to field linguists. That grammar includes "choosers" and "realization statements", neither of which ever appear in descriptive grammars written by field linguists. In the KANT project, surface structure is generated from the target f-structure by using Pseudo Unification Grammar. The KANT project's phrase structure rule shown in figure 2-10 above includes symbols and notations that computational linguists are generally familiar with (e.g., "==", X0, X1, X2, etc.), but these rules generally appear foreign to linguists, and their symbols and abbreviations rarely or never appear in linguistic literature. Opposed to this is TTA's synthesizing grammar which includes 1) feature copying rules, 2) spellout rules, 3) clitic rules, 4) movement rules, 5) phrase structure rules, 6) anaphora identification and

spellout rules, 7) word morphophonemic rules, and 8) find/replace rules. All of these rules except the find/replace rules are very familiar to field linguists and appear often in their descriptive grammars. TTA's synthesizing grammar will be described in detail in section 4.4.

- TTA's grammar takes advantage of recent typological linguistic research. As was stated in chapter 1, linguists have done extensive typological research in many areas, and this research guided the development of the feature system in TTA's semantic representational system, and also the generating grammar. Examples of how TTA incorporates current typological research will be presented throughout chapters 3 and 4.

Because TTA was designed to perform transfer and synthesis for a very wide variety of target languages, extraordinary demands are placed on its semantic representational system and its grammar. Therefore this dissertation will describe in detail the content of the semantic representations and the capabilities required of TTA's grammar. Thus the goals of this dissertation are to: 1) determine the information that must be included in the semantic representations, and 2) determine what capabilities the grammar must possess in order to generate texts in many different languages. The underlying hypothesis for this project is as follows: if the semantic representations contain sufficient information, and if the grammar possesses sufficient capabilities, then TTA will be able to generate texts of sufficient quality that they improve the productivity of experienced mother-tongue translators in a very wide variety of target languages. This system is not intended to produce high quality literature for highly educated people, nor is it intended to generate texts that use all the typologically rare features of each particular target language. Instead this system generates very simple sentences that are easily understandable, grammatically correct, and both semantically and referentially equivalent to the source representations. The texts that are generated by TTA are generally at a sixth grade reading level[6].

Stochastic techniques were not an option for TTA because the intended target languages generally have very few if any written texts. Therefore training corpora for developing the necessary statistical information are unavailable. *Template techniques* were considered inappropriate for this project due to the variety of texts that are to be generated. Templates work well for form based documents, but the texts that are to be generated by this project are not form based.

TTA's grammar is intentionally very generic in order to accommodate a wide range of language models, but it most closely resembles the transformational-generative tradition because the intended users of TTA are linguists who are familiar with the generative approach to grammar. Functional typological grammar was very influential in determining the features that must be included in the semantic representations, but the grammar in TTA is based on the generative approach. A high level model of TTA is shown below in Figure 2-12.

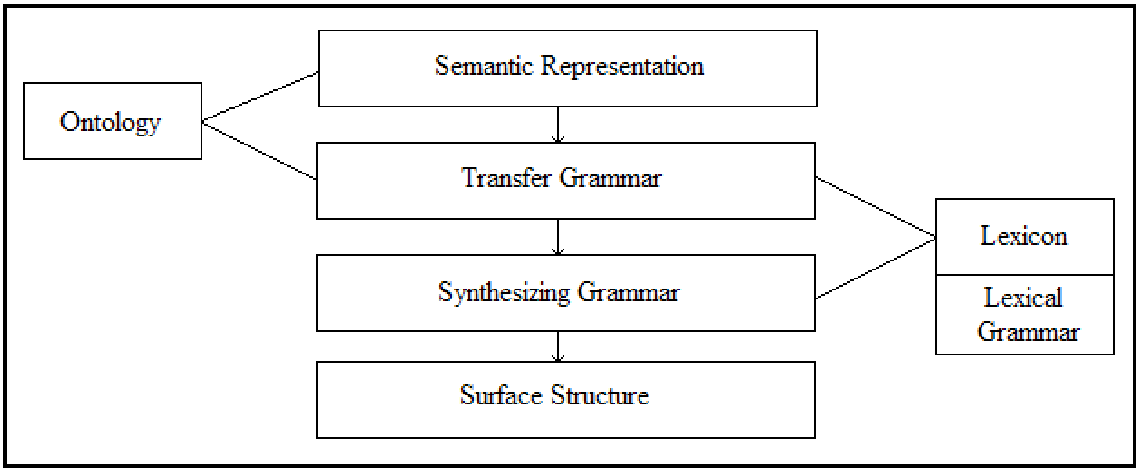


Figure 2-12. Underlying Model of The Translator's Assistant

[6] This was a strategic decision because there is a direct correlation between the complexity of the semantic representations and the complexity of the generating grammars. In order to keep the generating grammars reasonably simple, the complexity of the semantic representations has been constrained. The 'sixth grade reading level' was determined by applying the tool in Microsoft Word to the texts that were generated in English.

The model shown above indicates that TTA begins with a semantic representation of a proposition or text, and then executes all the rules in the target language's transfer grammar. The output of the transfer grammar is a deep structure representation of that proposition or text, and it consists of the target language's lexemes and structures. That deep structure representation is then supplied to the synthesizing grammar where all the target language's synthesizing rules are executed. The output of the synthesizing grammar is target language surface structure text. The pertinent concepts in the ontology are included in the semantic representation, and they are also available to the rules in the transfer grammar. Similarly the words in the target lexicon are available to both the transfer and synthesizing grammars.

2.5 Conclusions

This chapter has provided a brief introduction to natural language generation. It described the two main categories of NLG systems that have been developed, and it presented examples of the most successful systems from each of those two categories. This chapter also briefly introduced The Translator's Assistant, which is the NLG developed for this dissertation. TTA is distinct from the other NLGs for many reasons, and several of those reasons were listed and discussed.

The next chapter of this dissertation will describe the semantic system that was developed specifically for TTA. Every NLG is based on some type of semantic system, and many semantic systems have been developed. The next chapter will begin with a survey of six semantic systems that were potential candidates for TTA, but each of those systems was ultimately rejected as being either unsuitable or impractical. Then a new system of semantic representation that was developed specifically for this project will be presented. The concepts, features, and structures of this new semantic system will be thoroughly described in chapter 3.

CHAPTER 3

THE REPRESENTATION OF MEANING

3.1 Introduction

This chapter will introduce the semantic representational system that was developed specifically for The Translator's Assistant. Every natural language generator that uses semantic representations as its input requires a thoroughly specified, formal semantic system for its source texts. Many different semantic systems have been developed for NLGs, and many semantic systems were considered potential candidates for TTA. However, after investigating the existing semantic systems, each was found either unsuitable or impractical[7]. Therefore it was decided that a new semantic representational system must be developed. This chapter will present the new semantic system that was developed specifically for TTA.

This chapter begins with an overview of the perplexing, philosophical issues associated with the representation of meaning. Then six different semantic systems that were potential candidates for TTA will be described, and the reasons for their rejection in this particular project will be enumerated. These six semantic systems are the following:

- Formal Semantics developed by Richard Montague (Montague 2002)
- Conceptual Semantics developed by Ray Jackendoff (Jackendoff 1990; 2007)
- Cognitive Semantics developed by George Lakoff, Ronald Langacker, and Leonard Talmy (Lakoff 1987; Langacker 1986; Talmy 1985)

[7] It is important to note that this project is not arguing for or against the validity of these particular semantic systems. The focus here is not to develop a conceptual solution to the representation of meaning, but rather a practical solution that is both linguistically based and machine tractable. The semantic system developed for TTA and presented in section 3.3 below is a working hypothesis subject to refutation and counterevidence.

- Generative Semantics developed by McCawley, Postal, Ross, Katz, and Fodor (Fodor 1977)
- Ontological Semantics developed by Sergei Nirenburg and Victor Raskin (Nirenburg and Raskin 2004)
- Natural Semantic Metalanguage theory (NSM) developed by Anna Wierzbicka and Cliff Goddard (Wierzbicka 1992, 1996; Goddard 1998, 2008)

These six semantic systems will be described, and the features that make them unsuitable for this particular project will be delineated. Then this chapter will present the semantic system that was developed for TTA. This semantic representational system is comprised of 1) an *ontology*, 2) *features*, and 3) *structures*. TTA's ontology and its associated concepts will be described in section 3.3.1, the features associated with this semantic system will be described in section 3.3.2, and the structures permitted in this system will be described in section 3.3.3. This chapter will conclude with an example illustrating how the semantic system developed for TTA was applied to the opening paragraph of a short text that describes how to prevent eye infections.

The representation of meaning is an issue that has perplexed linguists, philosophers, lexicographers, logicians and cognitive scientists for decades. Fodor writes, "… what is meaning? This question has been repeatedly asked, and variously answered, throughout the history of philosophy and related disciplines. Along with problems about free will, the nature of time, and so on, it has seemed one of the ultimate metaphysical puzzles" (Fodor 1977:9). Because language is a tool for expressing meaning, meaning must be at least partially independent of language. Therefore it initially seems plausible that meaning should be representable in a method independent of language. But in order to discuss and describe meaning, language must be used. Tarksi (1956) showed that to define meaning, one must use a metalanguage in order to avoid paradoxes. Therefore meaning cannot possibly be described or represented in a completely language neutral way (Hutchins 1992:75). Due to this unavoidable fact, language must be used in the representation of meaning, and the language chosen will invariably influence what meaning is represented.

Existing NLG systems have used many different formats for their semantic representations. The fundamental question underlying the development of each semantic representation system is: “What type of semantics will be used?” Some linguistic theories and NLG systems in the past based their semantic representations on formal semantics such as Montague Grammar. Other NLG systems have tried using other types of linguistic semantics such as Jackendoff’s conceptual semantics (Jackendoff 1990), cognitive semantics developed by Lakoff (Lakoff 1987) and Langacker (Langacker 1986), generative semantics developed by Ross, Postal, and McCawley, or ontological semantics developed by Nirenburg and Raskin (Nirenburg 2004). Some NLGs have attempted to use a hybrid of these semantic systems in their semantic representations. Each of these semantic systems will be described below in sections 3.2.1 through 3.2.6, but they were ultimately found either unsuitable or impractical for TTA’s purposes. Therefore they were rejected, and the reasons will be presented below. Rather than using an existing semantic system, a new system of meaning representation was developed based on the foundational principles of Natural Semantic Metalanguage theory (NSM). NSM will be described in section 3.2.6, and then the new system that was developed specifically for TTA will be described in section 3.3.

3.2 Semantic Systems that Represent Meaning

This section will present six semantic systems that were candidates for TTA’s semantic representational system. After each semantic system has been described, the reasons for not using it will be presented.

3.2.1 Formal Semantics

The first semantic system that was considered for TTA is called *Formal Semantics*. The most widely acclaimed theory within formal semantics during the 1970s and 1980s is Montague Semantics (Montague 2002). Developed in the late 1960s and early 1970s by

Richard Montague, this theory proposes that there is no basic difference between natural languages and the artificial languages of logicians. Montague's theory claims that natural languages may be described and analyzed just as rigorously and precisely as the formal languages of logicians (Montague 1974:2). He described his system in a compact account of the syntax and semantics of ordinary English called PTQ (The Proper Treatment of Quantification in Ordinary English). His system is comprehensive, having both a semantic component and a syntactic component which are capable of analyzing and interpreting the syntax and semantics of natural languages. His theory radically influenced the study of linguistic semantics for the next several decades. An underlying premise of his theory is that when interpreting a sentence, the construction of meaning is rule governed just as is the construction of a syntactically well formed sentence. Therefore the meaning of an expression is determined by the meaning of its parts and how they have been put together, as is stated in Frege's Principle[8].

3.2.1.1 Montague's Apparatus

Using model theory, type theory, intensional logic, logical relations, and operators, Montague developed an elaborate system that was thought to be capable of describing a subset of many natural languages (Partee 2001). Several of his logical relations were conjunction '&', inclusive disjunction 'v', implication '→', equivalence '↔', and identity '='. A few of his operators were negation '~', possibility '<>', and necessity '□'. He observed that predicates operate on noun phrases and these operations may be represented by Q(x), P(x,y) or R(x,y,z) where Q represents all intransitive verbs, P represents all transitive verbs, R represents all ditransitive verbs, and x, y and z represent noun phrases. He then used a context-free phrase structure grammar to describe the resulting structures. He incorporated the lambda operator 'λ', originally developed by Alonzo Church in 1940, to handle constructions

[8] Gottlob Frege is widely accredited for the first modern formulation of this principle. However, the idea appears in Plato's *Theaetetus*.

such as the passive. Two quantifier operators, universal '∀' and existential '∃', were added to his grammatical apparatus to represent natural language quantifiers. He introduced the tense operators which are applied to verbs in order to produce tensed verbs. The present tense operator is denoted by 'Pres', past tense by 'Past', and future by 'Fut'.

3.2.1.2 Reasons for Rejecting Formal Semantics

Montague's elaborate system has been used by scholars in many disciplines to partially describe the semantics of many natural languages. It has also been very insightful into the structure and interpretation of human languages as well as cognitive processing. It has been especially insightful in natural language phenomena such as scope and entailment, and it also provides a formal way of representing some meaning.

However, there are notable problems with Montague's PTQ (Rosner and Johnson 1992:xv). Essentially he attempted to replace words with monosemic logical symbols which possess rigid boundaries. But his symbols could not do justice to the shades of meaning encapsulated by words or to polysemy. Neither linguists nor philosophers find Montague's approach entirely satisfying. Linguists claim that the system cannot be extended to include other syntactic or semantic phenomena exhibited in natural languages, and philosophers do not believe that higher order intensional logic is the proper tool for exploring issues such as intensionality or inference.

Montague's grammar was rejected for this project because 1) it is not capable of covering the many shades and nuances of meaning that every natural language is capable of expressing, and 2) a strictly logical representation may not always adequately constrain the surface text that is to be generated. For example, Bateman points out that the logical representation

$$\mathrm{on}(x,y) \land \mathrm{book}(x) \land \mathrm{red}(x) \land \mathrm{table}(y)$$

may be intended to represent *The red book is on the table*, or *It is the red book that is on the table,* or *The red book on the table …* or *The book on the table is red*, or *There is a book on the* *table. It is red* (Bateman 1997:2). For practical purposes, the semantic representations for an NLG system must contain sufficient information and be sufficiently constrained so that only one target form will be generated.

3.2.2 Conceptual Semantics

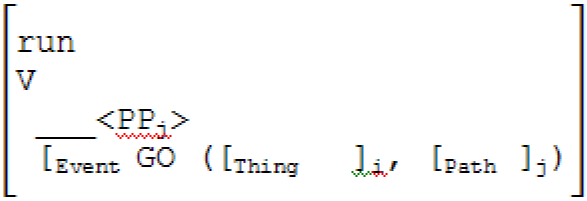
The second semantic theory that was considered for TTA is called *Conceptual Semantics*. Conceptual semantics is a theory of meaning that is organized in a psychological framework and attempts to integrate theories of linguistics, perception, cognition and conscious experience. This theory was proposed by Ray Jackendoff (Jackendoff 1990), and its purpose is to characterize the mental resources that make it possible for humans to think, know, and experience the real world. The goal of the theory is to determine the form of the internal mental representations that constitute conceptual structure (Jackendoff 1990:10). This theory of meaning is based on principles which parallel the foundational principles of generative syntax and phonology.

3.2.2.1 Jackendoff's Apparatus

Jackendoff proposes a set of conceptual primitives augmented by features and expanded with tiers in order to represent meaning in a structured and precise way. Jackendoff argues that all humans are endowed with an innate Universal Grammar that narrowly restricts the limitless variations of grammars that might exist. Similarly he believes that a person's stock of lexical concepts must be developed from an innate collection of possible concepts. One of his goals is to determine how much of the human ability to comprehend sentential concepts is innate, and how much is learned. Therefore the central issue of this theory is to determine the innate units and principles of organization that make human lexical and sentential concepts both possible and learnable. He believes that meaning is represented mentally and people should be able to articulate these representations clearly. Syntax gives us a glimpse into these representations because syntax evolved as a means of expressing conceptual structure.

Therefore there must be a great deal of correspondence between syntax and conceptual structures; each major syntactic constituent must correspond to a major semantic constituent.

Jackendoff postulated eight fundamental ontological categories: Thing, Event, State, Action, Place, Path, Property and Amount (Jackendoff 1990:22). Every lexical item is assigned to one of these eight categories. Additionally, every lexical item has a conceptual structure which takes zero or more arguments. These conceptual structures may be elaborated according to his proposed schema. For example, in the simple sentence *John ran into the room*, *run* is a movement verb, so it takes GO as an operator with an optional path; optionality being indicated by underlining. Therefore the conceptual structure for RUN may be represented as shown in Figure 3-1 below (Jackendoff 1990:45):



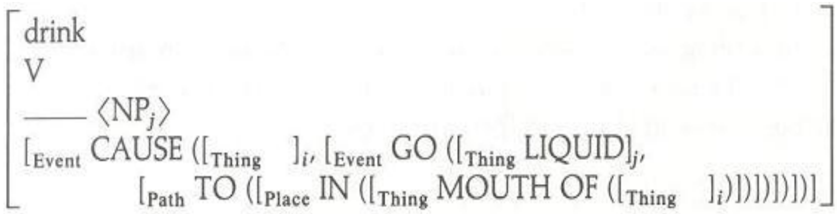
$$\begin{bmatrix} \text{run} \\ \text{V} \\ \underline{\quad\quad}\langle PP_j\rangle \\ [_{\text{Event}}\ \text{GO}\ ([_{\text{Thing}}\quad]_i,\ [_{\text{Path}}\ ]_j) \end{bmatrix}$$

Figure 3-1. Jackendoff's Conceptual Representation of RUN

In Jackendoff's representations, the thing phrase subscripted with 'i' always indicates the agent, actor, or experiencer. Jackendoff introduced features to modify these operators. These features include ±Contact, so a verb like *touch* is $\text{GO}_{\text{+Contact}}$ (Jackendoff 1990:107). Jackendoff's final addition to his apparatus is the *action and thematic tiers* which parallel the tier theory in phonology. He claims that conceptually, there is a thematic tier which identifies each NP in a sentence with a thematic role. However, there is also an action tier which identifies each NP as either an Actor, a Patient, or nothing.

3.2.2.2 Reasons for Rejecting Conceptual Semantics

This semantic theory was rejected for this project for four reasons: 1) The representations are more complex than the concepts that they represent. A simple concept such as DRINK[9] is represented as shown below in Figure 3-2 (Jackendoff 1990:53):



$$\begin{bmatrix} \text{drink} \\ \text{V} \\ \underline{\qquad} \langle \text{NP}_j \rangle \\ [_{\text{Event}}\ \text{CAUSE}\ ([_{\text{Thing}}\ \ ]_i,\ [_{\text{Event}}\ \text{GO}\ ([_{\text{Thing}}\ \text{LIQUID}]_j, \\ \qquad [_{\text{Path}}\ \text{TO}\ ([_{\text{Place}}\ \text{IN}\ ([_{\text{Thing}}\ \text{MOUTH OF}\ ([_{\text{Thing}}\ \ ]_i)])])])])] \end{bmatrix}$$

Figure 3-2. Jackendoff's Conceptual Representation of DRINK

This formula shown above is to be read "drinking is an event where something causes liquid to go into its own mouth." When developing semantic representations for large texts, it is much more convenient to simply use DRINK rather than the composed formula shown above. 2) The notation of the representational system is obscure and cumbersome. Deciphering the semantic structures developed by Jackendoff requires a thorough knowledge of the apparatus. When linguists use TTA, it is much simpler to ask them for their target language equivalent[10] of DRINK rather than asking them for the equivalents of the representation shown above. 3) There seems to be no limit to the number of primitives, features, operators, and rules that may be added to this system. Jackendoff himself says, "… two general questions constantly arise. The first is to what extent new conceptual functions ought to be added as simple primitives in their own right, and to what extent they should be added by elaborating old primitives in terms of a feature system" (Jackendoff 1990:87). The correspondence rules which deal with mismatch between conceptual arguments and syntactic positions are extremely complex and also appear

[9] It's important to note that the symbols for these concepts don't represent the meaning that is in a speaker's mind, nor do they include the knowledge that is required to use a concept in a discourse. These symbols simply represent the concepts that are in a speaker's mind. Only the symbols for prime concepts represent the meaning of those concepts.

[10] The target language equivalents aren't necessarily lexemes that correspond directly to the source concepts. As will be discussed in chapter 4, the target language equivalent may be expressed using very different lexemes and structures than those used in the source language. Examples of this will be provided showing how Korean expresses the equivalence of WEIGH-A and WEIGH-B

unbounded (Jackendoff 1990:155-243). 4) Jackendoff's apparatus is a metalanguage, and all metalanguages are degenerate forms of a natural language (Allan 1986:268, Goddard 1998:66). In order to understand a representation, one must first translate it back into ordinary English. But repeatedly coding and decoding these representations is inefficient and wasteful.

3.2.3 Cognitive Semantics

The third theory that was considered a potential candidate for TTA's semantic representational system is called *Cognitive Semantics*. Developed primarily by George Lakoff (Lakeoff 1987), Ronald Langacker (Langacker 1986), and Leonard Talmy (Talmy 1985), this theory uses image schemas rather than language specific words in order to represent meaning. Similar to conceptual semantics, this theory is concerned with the mental representation of the world and its relation to language. But these theorists reject propositional representation, and instead claim that meaning is derived from "the pre-conceptual structuring of bodily experiences which is impossible to represent in verbal terms" (Goddard 1998:79). Lakoff says that "the idea that all internal structure is of a building-block sort, with primitives and principles of combination, does not seem to work at the basic level of human experience. … basic level concepts cannot be considered elementary atomic building blocks within a building-block approach to conceptual structure" (Lakoff 1987:270).

3.2.3.1 Kinesthetic Image Schemas

Rather than using primitive concepts, these theorists propose kinesthetic image schemas. For example, the CONTAINER schema defines the basic distinction between IN and OUT (Lakoff 1987:271). Of the CONTAINER schema Lakoff says, "On our account, the CONTAINER schema is inherently meaningful to people by virtue of their bodily experience. The schema has a meaningful configuration, from which the basic logic follows. … Thus, schemas are not understood in terms of meaning postulates and their interpretations. Rather, meaning postulates themselves only make sense given schemas that are inherently meaningful because they structure our direct experience" (Lakoff 1987:273). These theorists propose additional schemas such as:

- the PART-WHOLE schema in which "we experience our bodies as wholes with parts (Lakoff 1987:273),"
- the LINK schema which begins with one's umbilical cord but also includes strings, wires, tape and anything else that links one thing to another,
- the CENTER-PERIPHERY schema which says that our bodies and other things have certain constituents at the center and other constituents at the periphery,
- the SOURCE-PATH-GOAL schema which says that all movement consists of a starting point, an ending point, and a contiguous set of points joining the starting point to the ending point, etc.

These theorists reject the idea of conceptual primitives, but hold to the idea of semantic compositionality. They claim that there are basic level concepts, image schemas, and rules of semantic composition that build complex concepts from less complex concepts, but nothing that satisfies the definition of a primitive.

A sample of their symbolic representation is shown below in Figure 3-3. The simple proposition *The cat is out of the bag* is represented by a series of diagrams (Langacker 1987:95).

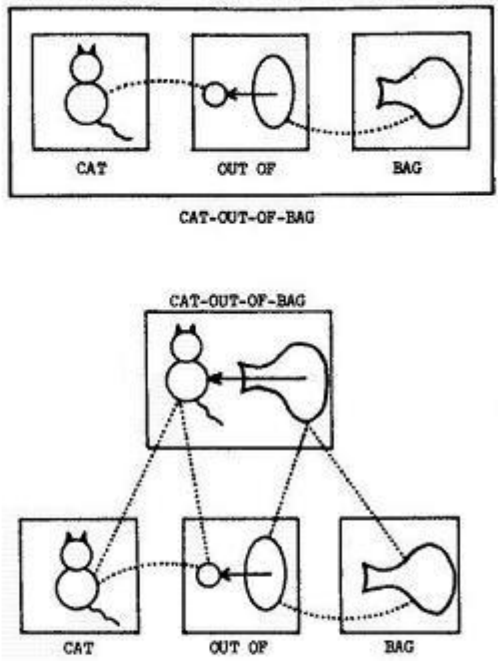


Figure 3-3. Langacker's Symbolic Representation for *The cat is out of the bag*.

3.2.3.2 Reasons for Rejecting Cognitive Semantics

Goddard points out several issues with this theory (Goddard 1998:81). He says that it is not clear how the various representational devices interact with one another. Similarly it is difficult to see how the diagrams and kinaesthetic image schemas can interface with the propositional aspects of meaning.

This theory was rejected for this project because it is impractical, perhaps impossible, to put the diagrams and kinaesthetic image schemas into machine tractable forms. Developing symbols for each of the necessary nominals is entirely impractical; the symbols would need to have labels such as CAT and BAG, and it is much simpler to use the labels rather than the images. Developing machine tractable forms for the schemas is an unnecessary complication that would neither enhance this project nor improve the quality of the generated text.

3.2.4 Generative Semantics

A semantic theory which initially showed great promise for TTA is called *Generative Semantics.* This theory was developed primarily by McCawley, Postal, Ross, Katz, and Fodor in the mid 1960s. Although this approach has been rejected by contemporary linguists, it has features that are quite attractive to developers of NLGs. Therefore this approach was considered for this project.

The goal of generative semantics was to state the characterization of correlations between surface structure and meanings (Fodor 1977:5). The developers argue that their theory must

> make available some format for the precise representation of meanings, both of lexical items and of phrases and sentences. It must specify the nature of the rules that will relate the meaning representations of phrases and sentences to representations of the meanings of the lexical items they contain and of the syntactic configurations in which they appear. And we can also expect it to provide formal definitions of meaning-dependent properties of expressions and meaning-dependent relations between expressions (Fodor 1977:6).

Katz and Fodor claimed that the grammar has a self-contained syntactic component which specifies the syntactic structure and lexical content of each sentence of the language. Therefore their semantic theory must perform two tasks: 1) provide specifications of the meanings of lexical items, and 2) identify the recursive rules that operate over syntactic structures for building up meaning specifications for phrases and sentences out of the meaning specifications for lexical items (Fodor 1977:64). These theorists claimed that there is a deep structure level which is identical to the semantic level, and it has no interpretive semantic rules. Then the projection rules begin with the lexical items and amalgamate larger and larger constituents until eventually they have accounted for the entire proposition (Fodor 1977:68).

3.2.4.1 The Apparatus of Generative Semantics

Proponents of this theory claimed that every language is a system that maps meaning to expression (Frantz 1974:1). They proposed universal semantic structure which they represented with phrase structure trees. They claimed that semantic structure must be grouped and put into hierarchies using phrase structure trees in order to account for scope (Frantz 1974:2) and logical inference (Frantz 1974:4). In addition to the universal deep structure representations, they proposed universal derivations or transformations that map deep structure to surface structure. They also proposed universal lexemes which are semantically simple and may be combined to form the lexical items of languages. Examples of universal lexical decomposition using their universal lexemes include (Frantz 1974:27):

- *look for* = [TRY [FIND]$_{P}$]$_{P}$
- *persuade* = [CAUSE [BECOME [INTEND]$_{P}$]$_{P}$]$_{P}$
- *convince* = [CAUSE [BECOME [BELIEVE]$_{P}$]$_{P}$]$_{P}$
- *deny* = [SAY [NEG]$_{P}$]$_{P}$
- *kill* = [CAUSE [BECOME [NEG [ALIVE]$_{P}$]$_{P}$]$_{P}$]$_{P}$

The developers of this theory explicitly described the semantic content of the deep structure representation. As implied by the name, generative semanticists placed more emphasis on semantic and pragmatic issues than on syntactic issues. They used language neutral predicate logic rather than language specific lexical items to represent meaning. Because their deep structure representations were logical representations which contain different categories and structures than do the English surface representations, their theory required powerful transformations to render the surface equivalents. To precisely identify and describe the semantic content of the deep structure, generative semanticists adopted a method called lexical decomposition. Rather than using semantically complex lexemes from a particular language, they used abstract nonlexical forms which were semantically simple, but were then realized through a series of transformations as language specific words. Their efforts were eventually supplanted by and incorporated into models such as Montague's which used explicit semantics.

3.2.4.2 Reasons for Rejecting Generative Semantics

This approach would have been very appealing for this project. The claim that many or all languages could have the same or similar deep structures, and that surface structure is realized through a series of transformations matches the purposes of this project very closely. However, this approach to semantic research is subject to the same criticisms as Montague grammar, and was abandoned in the 1980s and supplanted by cognitive semantics. Therefore it was not used for this project.

3.2.5 Ontological Semantics

A new semantic theory that was developed specifically for computational natural language systems is called *Ontological Semantics.* This theory was recently developed by Sergei Nirenburg and Victor Raskin (2004). Ontological semantics deals specifically with the extraction, representation, manipulation and generation of meaning within natural language texts by computers. The information acquired from these tasks may be used in a variety of natural language processing projects such as machine translation, text summarization, question and answering, advice giving, etc. Ontological semantics differs from the other types of semantics discussed above in the following four ways (Nirenburg and Raskin 2004:103): 1) It introduces an ontology with a rich set of language independent primitives related to one another in an *inheritance hierarchy*. 2) It is a comprehensive theory, composed of a multitude of microtheories, that integrates lexical semantics with compositional semantics and also includes pragmatics. 3) It is designed to adjust semantic description depth according to the needs of a particular application. 4) It emphasizes full coverage of all the phenomena in a text at a predefined level of detail.

3.2.5.1 The Apparatus of Ontological Semantics

A variety of resources have been developed to perform ontological semantics. These resources include 1) an ontology which is a constructed hierarchical network of real world unambiguous concepts within a particular domain, 2) a fact database that includes instances of concepts as they have occurred in texts, as well as information about the concepts, 3) a lexicon that maps the words of a natural language to the concepts in the ontology, and 4) an onomasticon which is a collection of proper names from a particular natural language. The ontology and fact database deal with concepts rather than natural language words; these concepts are intended to be language neutral, unambiguous, and together form a metalanguage which is capable of representing meaning. The lexicon and onomasticon are language specific; each natural language requires the development of its own lexicon and onomasticon (Nirenburg and Raskin 2004:10).

Nirenburg and Raskin subscribe to the “weak” artificial intelligence thesis - that computers can be made to perform the same tasks that humans perform and they can achieve the same results. However, this will not be accomplished by making the computers model human problem solving techniques; instead it will be achieved by using other approaches which will enable the computers to achieve functional equivalence (Nirenburg and Raskin 2004:14). They also believe that it is impossible to develop a single comprehensive theory that is capable of accounting for all of the phenomena encountered in natural language processing projects. Therefore they have resorted to the development of a multitude of microtheories, each one being responsible for one particular domain such as aspect, negation, relative clauses, etc. (Nirenburg and Raskin 2004:30). These microtheories incorporate target language devices for realizing the underlying semantics. Nirenburg and Raskin have also recognized that it is a practical impossibility to develop a comprehensive set of features that will adequately describe all of the phenomena in natural languages, so they are focusing on language universals and incorporating as many of the universals as possible into their microtheories.

As implied by the name, the ontology is the key component in ontological semantics. The ontology contains knowledge about objects, processes and properties in the real world, but it uses a metalanguage to represent this information. They assert that it is impossible to use elements of the real world to represent meaning, so they have developed a set of meaning elements, or primitive symbols, which are used as substitutes for the objects and processes in the real world (Nirenburg and Raskin 2004:82). But their ontology is more than just a collection of primitive meanings; it is also a model of the real world because it includes hierarchical relationships between the primitives. Thus it is a “language independent compendium of information about the concepts underlying elements of natural language” (Nirenburg and Raskin

2004:77). "The function of the ontology is to supply world knowledge to lexical, syntactic and semantic processes" (Nirenburg and Raskin 2004:114). Figure 3-4 below shows the highest level of their ontology.

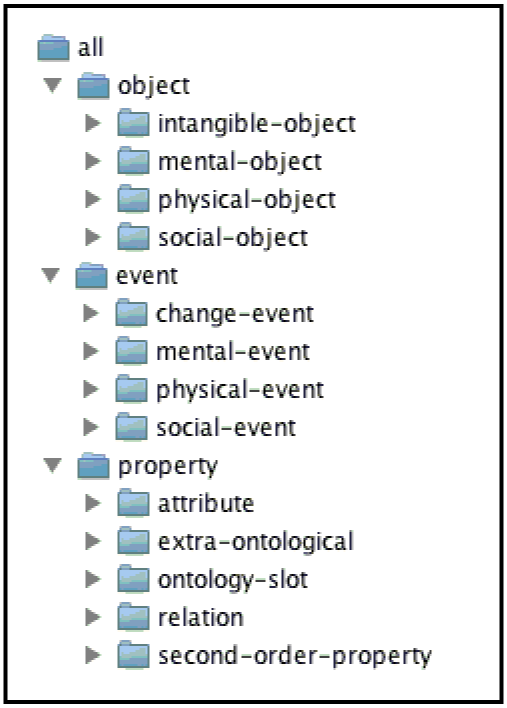


Figure 3-4. The Ontology Developed for Ontological Semantics

The figure above shows that there are three fundamental categories of concepts according to the ontological semantics perspective: objects, events, and properties. Each of these categories is then subdivided, and each subdivision contains additional subdivisions or concepts.

3.2.5.2 Reasons why Ontological Semantics wasn't used in TTA

Stephen Beale, who is a co-developer of Ontological Semantics with Sergei Nirenburg, contributed significantly to the development of this project's ontology and semantic representations. Because Beale and Nirenburg's work is intended for general use in a variety of applications, their approach differs significantly from the approach developed for this project, but the techniques do overlap to some degree. In particular, the ontologies for the two projects are fairly similar. However, Nirenburg and his team claim that their ontology is comprised of language independent concepts (Nirenburg and Raskin 2004:77). No such claim is made for this project's ontology. The ontology for this project will be described thoroughly below, but it consists of semantically simple concepts that have been lexicalized by English speakers. This author ascribes to the view that only a few dozen concepts can be described as truly "language independent"; therefore the ontology for this project cannot possibly be described as consisting of language independent concepts.

3.2.6 Natural Semantic Metalanguage Theory

The final semantic theory that was considered a potential candidate for TTA's semantic representational system is called *Natural Semantic Metalanguage* theory (NSM). Although this theory could not be adopted for TTA for reasons which will be presented later, the foundational principles of this theory significantly influenced the development of TTA's ontology and semantic representational system. This theory has been developed primarily by Anna Wierzbicka (Wierzbicka 1996) and Cliff Goddard (Goddard 1998). The primary focus of NSM is to identify a set of semantic primitives which Wierzbicka and others claim are innate. NSM theorists have proposed a table of fundamental concepts which they believe are both innate and indefinable, and they have also proposed a grammar that describes how these primitive concepts may be combined into propositions. These concepts and grammar together form a universal mini-language which they believe is present in the mind of every child and thus constitutes a language neutral universal deep structure representation. NSM theorists believe that the semantic primitives, being indefinable themselves, may be used to define every word in every language. The goal of NSM is to determine which concepts are the semantic primitives, and to describe the syntax governing the combinations of these primitives. After they have completed that task, they believe they will have identified a language neutral method for representing deep structure meaning (Wierzbicka 1996:22).

3.2.6.1 The Apparatus of Natural Semantic Metalanguage Theory

Wierzbicka refers to the small set of innate concepts as the "alphabet of human thoughts" (Wierzbicka 1992:210). She writes:

> For example, *this* is an English word, and *hic*, a Latin one, but both can realize the same "atom" of human thought. We could say, therefore, that the set of indefinables is universal, although every language has its own, language-specific "names" for them. Consequently, the number of indefinables is probably the same in all languages, and the individual indefinables can be matched cross-linguistically. Of course the indefinables of different languages cannot be expected to be equivalent in all respects; they can, nonetheless, be regarded as SEMANTICALLY equivalent (Wierzbicka 1992:210).

The following list of indefinables is provided in (Goddard 2008:58):

Substantives: I, YOU, SOMEONE, PEOPLE, SOMETHING/THING, BODY
Relational Substantives: KIND OF, PART OF
Determiners: THIS, THE SAME, OTHER/ELSE
Quantifiers: ONE, TWO, SOME, ALL, MANY/MUCH
Evaluators: GOOD, BAD
Descriptors: BIG, SMALL
Mental Predicates: THINK, KNOW, WANT, FEEL, SEE, HEAR
Speech: SAY, WORD, TRUE
Actions, events, movement, and contact: DO, HAPPEN, MOVE, TOUCH
Existence, location: THERE IS, BE (SOMEWHERE)
Possession, specification: HAVE, BE(SOMEONE/THING)
Life and death: LIVE, DIE
Time: WHEN/TIME, NOW, BEFORE, AFTER, A LONG TIME, A SHORT TIME, FOR SOME TIME, MOMENT
Space: WHERE/PLACE, HERE, ABOVE, BELOW, FAR, NEAR, SIDE, INSIDE
Logical concepts: BECAUSE, IF, MAYBE, CAN, NOT
Intensifier, Augmentor: VERY, MORE
Similarity: LIKE

NSM theorists claim that these concepts are *lexical universals* (Goddard 1998:59). Additionally, the strong form of the NSM hypothesis claims that every language has a morpheme or word corresponding to each of these concepts. These concepts are indefinable, but by using these concepts, every word in every language can be defined or explicated. An example of their explications for the English word *promise* follows:

promise as in X *promised* Y (to do A) = (Goddard 1998:147)
X said to Y:
I want you to know I will do A
when X said it, it was as if X was saying at the same time:
I know you want me to do this
I know you think that maybe I will not do it
I don't want you to think this
I know if I don't do it after saying this, people will think something bad about me

Sometimes it is necessary to include concepts that are not primitives in an explication. NSM theorists have proposed that non-primitive words which frequently occur in explications be called 'semantic molecules' (Goddard 1998:254). Examples of semantic molecules include the following (Goddard 1998:255):

Body parts: head, mouth, teeth, lip, nose, hand, finger, foot, claw, tail, ear, leg, arm, hair, neck, back, fur, skin, etc.
Actions and activities: make, drink, eat, hold, pick up, put down, chase, catch, fight, kill, climb, jump, touch, run, fly, bite, crush, dig, kick, jump, suck, pour, roll, keep in, get out, carry, cut, hunt, smell, look after
Postures: sit, lie, stand
Shape and dimensions: long, short, flat, round, curved, pointy, thick, thin, stick out, wide, narrow, straight, bent, open, closed
Parts of shapes: top, middle, bottom, side, front, back, end, edge
Physical properties: hard, soft, smooth, rough, sharp, rigid, flexible, light, heavy
Secondary qualities: hot, cold, loud, soft
Manners: quickly, slowly
Environment: ground, air, sky, sun, water, grass, trees
Colors: black, yellow, green, brown, red, gray
Sex: male, female

By explicating words using *semantic molecules* rather than only the primitives, the explications become much easier to understand.

As indicated by the name, NSM theorists are proposing a metalanguage that has the same expressive power as a full natural language (Goddard 1998:60). The lexicon of this language consists of the semantic primitives, and the syntax prescribes the allowable combinations of these primitives. Just as all languages share certain basic concepts, these theorists claim that there are universally shared grammatical patterns across all languages (Goddard 1998:329). The basic unit in this grammar is analogous to a clause which consists of substantive phrases and a predicate. Substantive phrases may consist of one of the substantives listed above, as well as optional modifiers such as *these two people, good things,* *after a short time,* etc. Certain combinations of modifiers are not allowed, such as **the same some people*, but more research is required to determine which combinations are universal, which are allowed but not universal, and which are not allowed. English examples illustrating universal sentences include:

- *These two people said many good things.*
- *Maybe something bad happened.*
- *That place is far from here.*
- *I want to do this.*

These "simple" sentences are purportedly translatable into every language without distortion or loss of content. But regarding "simple" sentences Wierzbicka writes:

> The semantic structure of an ordinary human sentence is about as simple and 'shallow' as the structure of a galaxy or the structure of an atom. Looking into the meaning of a single word, let alone a single sentence, can give one the same feeling of dizziness that can come from thinking about the distance between galaxies or about the impenetrable empty spaces hidden in a single atom (Wierzbicka 1996:233).

3.2.6.2 Reasons for Rejecting Strict Adherence to the NSM Approach

The fundamental claims made by the NSM theorists are fairly similar to the claims made by the conceptual semanticists, the cognitive semanticists, and the generative semanticists: words consist of semantically simple components which are innate and arranged in some type of structure. Clearly the system of representation adopted by NSM is much easier to understand than the other systems because it closely reflects natural language. Even people who are unfamiliar with NSM can easily understand the explications and simple sentences. Unfortunately, using just the semantic primitives and the proposed molecules to represent the meaning of a text with relatively complex semantic content is unwieldy. Therefore strict adherence to the NSM approach had to be rejected for this project.

3.2.6.3 Fundamental Principles of NSM Adopted by TTA

NSM has revealed two principles which have proven very helpful for this project: 1) *When preparing a document that is to be translated, using semantically simple lexemes rather than semantically complex lexemes will reduce the complexity of the translation task*. This follows from the fact that texts comprised solely of the NSM primitives are translatable into every language without distortion. 2) *Semantically simple lexemes can be identified in a principled manner using the same procedure that NSM theorists used to identify their primitives.* A fundamental principle of NSM states that every word of every language may be defined using just the innate primitives. From this principle a guideline may be derived for identifying semantically simple lexemes: lexemes which are frequently used in the definitions of other lexemes may be considered semantically simple. Because words should always be defined using simpler words, words that appear frequently in definitions may be categorized as semantically simple; words that appear infrequently in definitions should be considered complex. For example, the English word *walk* is used in the definitions of many other words such as *saunter* "to walk in a slow and relaxed way, especially so that you look confident or proud" (Longman Dictionary 2003:1458), *waddle* "to walk with short steps, with your body moving from one side to another – used especially about animals or birds with fat bodies and short legs" (Longman Dictionary 2003:1848), and *wade* "to walk through water that is not deep" (Longman Dictionary 2003:1848). Because *walk* appears in numerous definitions, it must be semantically simpler than words like *saunter*, *waddle* or *wade* which appear in very few if any other definitions[11]. From this guideline one may conclude that the more frequently a word appears in definitions of other words, the semantically simpler the word. By using semantically simple words in a document that is to be translated, the complexities of the translation task may be reduced because the target languages are more likely to have lexicalized the same bundle of

[11] The definitions in Longman's *Dictionary of Contemporary English* and the Longman Defining Vocabulary will be described in section 3.3.1 below.

primitives. Using anything other than the innate primitives in a source text will certainly result in lexical mismatch when translating the text into another language, but by using semantically simple concepts in the source text, the problem of lexical mismatch may be reduced.

3.3 The Semantic System Developed for The Translator's Assistant

This section will describe the semantic representational system developed for TTA. This semantic representational system consists of the following three components:

- an ontology with its various classes of concepts, concept senses, and concept environments,
- features and feature values that are associated with each semantic category, and
- the small set of structures that are permitted in TTA's semantic representational system.

This section will conclude with an example illustrating how this semantic representational system was applied to the opening paragraph of a text that describes how to prevent eye infections.

The purpose of The Translator's Assistant is to generate drafts of texts that have significant semantic content in a very wide range of target languages. Therefore, as was stated in section 2.4.2, one of the goals of this project was to determine what information must be included in the semantic representations so that TTA could generate texts in many different languages. The semantic representations must include concepts that are semantically more complex than the NSM primitives, and the feature system must be much richer than those developed for the systems described in the previous chapter. For example, marking a noun as *singular* or *plural*, *definite* or *indefinite*, may be adequate for Indo-European languages, but it is completely inadequate for most other languages. Similarly, marking a verb as past, present or future may be sufficient for all Indo-European languages, but it is completely inadequate for languages in many other families. Therefore a very elaborate feature system was developed for this project, each feature being an exhaustive list of the values that are pertinent to the world's languages. Certainly more features will have to be added in the future in order to accommodate languages that have particular needs. Nevertheless the system developed to date has worked well for the two primary test languages, as well as the other languages on which small experiments have been performed.

As stated at the beginning of this chapter, it is impossible to represent meaning in a language neutral way. All meaning must be conveyed through language, so a natural language had to be chosen for this project's semantic representations. It was decided that a controlled subset of English would be used for the representational system because the author and most users of TTA speak English as their first language. Therefore the semantic representations consist of semantically simple English lexemes in simple English structures, and the English world view is used (e.g., The sun rises in the morning and sets in the evening, people catch colds, etc.). These English lexemes, structures, and the feature system will now be described in detail.

3.3.1 The Ontology

The ontology is the most problematic issue in this project. Constructing an ontology that will work well for every language is a daunting task. The speakers of every language have lexicalized the concepts that are significant to them and in accord with their world view. Every language has very particular lexemes which require phrases, clauses, or sometimes entire paragraphs in order to convey their meaning in another language. And even when two languages have lexicalized a similar concept, the range of meanings for those lexemes and the distribution of their use will probably not overlap entirely. When translating a document from one language to another, lexical mismatch is the norm and must be dealt with. Many examples of lexical mismatch and how this project deals with it will be provided in chapters 4 and 5.

During the past several decades, numerous ontological models have been developed. Ontological semantics, discussed in section 3.2.5 above, is a good example of an ontological system. A variety of definitions for *ontology* have been proposed. Farrar defines *ontology* as "a repository of linguistic knowledge that attempts to ground linguistic constructs in concepts of time, space, causality and human interaction" (Farrar et al 2002:6). Noy defines *ontology* as "some formal description of a domain of discourse, intended for sharing among different applications, and expressed in a language that can be used for reasoning" (Noy and Hafner 1997:53). Some ontologies are generic while others are specific to a particular domain such as medicine (UMLS – unified medical language system) or genetics (GENSIM – genetic simulation system). The best known, most thoroughly developed, and most commonly used ontology is WORDNET, developed primarily by George Miller at Princeton. It contains nouns, verbs, adjectives, and adverbs, and it groups synonyms together into sets called synsets. WORDNET is a taxonomy because it does not have structured concepts or axioms (Noy and Hafner 1997:60). John Bateman, the principal architect of the KPML system described in chapter 2, proposed a Generalized Upper Model (GUM) for ontologies. The current version, called GUM 2.0, is "a general task and domain independent linguistically motivated ontology that supports sophisticated natural language processing while significantly simplifying the interface between domain-specific knowledge and general linguistic resources" (Bateman et al. 2005:1). Another popular and powerful ontology is called Suggested Upper Merged Ontology (SUMO). This ontology is comprised of eleven sections (Pease et al. 2002:2): 1) the Structured Ontology contains the definitions for the relations that serve as the framework for defining the ontology proper, 2) the Base Ontology consists of very fundamental ontological notions such as abstract entity and the distinction between objects and processes, 3) the Set/Class Theory section consists of basic set theoretic content, 4) the numeric section provides definitions of basic arithmetic functions, 5) the Temporal section contains temporal relations, 6) the Mereotopology section contains a basic axiomatization of part/whole relations, 7) the Graph Theory section provides general graph theoretic notions, 8) the Unit of Measure section provides definitions of the unit systems, and the remaining sections of the ontology provide subhierarchies and axioms relating to process types, object types, and attribute types. Another project called General

Ontology for Linguistic Description (GOLD) is attempting to apply reasoning to the Semantic Web. GOLD is built on SUMO, and “attempts to give an account of the most basic categories and relations used in the scientific description of human language” (Farrar 2007:176). GOLD organizes linguistically related concepts into four major domains: expressions, grammar, data constructs, and metaconcepts.

The ontology developed for TTA is considerably simpler than the systems described above because TTA is strictly a language generator. TTA’s ontology will not be used for natural language processing, nor will it be used for reasoning. Instead, TTA’s ontology is a simple taxonomy. Like other ontologies, TTA’s ontology must specify very precisely the meaning of each concept, as well as the environment in which that concept may be used. In order to maximize the probability that other languages will have lexical equivalents for the concepts in the semantic representations developed for this project, the ontology consists primarily of semantically simple English lexemes (Goddard 1998:57, 61).

3.3.1.1 Four Semantic Complexity Levels

TTA’s ontology contains concepts that have been categorized into four semantic complexity levels: 1) NSM primitives, 2) semantic molecules, 3) complex concepts that have been explicated, and 4) inexplicable concepts that are not NSM primitives. Every concept in every language has a semantic complexity level according to Wierzbicka, who states that “the complexity of a concept can be viewed as the distance separating it from the level of indefinables” (Wierzbicka 1996:212). Based upon this principle, each concept in TTA’s ontology has been assigned a semantic complexity level ranging from 1 to 4, and these four categories will be described below.

3.3.1.1.1 Semantic Complexity Level 1: The NSM Primitives

The first semantic complexity level contains the NSM primitives which were listed above in section 3.2.6. According to NSM theorists, these primitives have lexical or morphological equivalents in every language, and they cannot be defined.

3.3.1.1.2 Semantic Complexity Level 2: The Semantic Molecules

As was mentioned in section 3.2.6, when NSM theorists explicate semantically complex concepts, they frequently use ‘semantic molecules’. They define semantic molecules as ‘non-primitive words which frequently occur in explications’ (Goddard 1998:254). They have developed a list of semantic molecules, but their list was insufficient for this project. In order to identify additional semantically simple English lexemes, one of the foundational principles of Natural Semantic Metalanguage theory was used: words which are used frequently in the definitions of other words must be semantically simpler than words which are used less frequently in the definitions of other words. Based upon this principle, semantically simple English lexemes may be identified by observing which words occur frequently in the definitions of other words. This is precisely the approach that was used during the development of Longman’s *Dictionary of Contemporary English*. The developers of the Longman dictionary state:

> The Longman Defining Vocabulary of around 2000 common words has been used to write all the definitions in this dictionary. The words in the Defining Vocabulary have been carefully chosen to ensure that the definitions are clear and easy to understand, and that the words used in explanations are easier than the words being defined (Longman 2003:1943).

Therefore for this project, words that are in Longman’s defining vocabulary are designated as semantic molecules.

3.3.1.1.3 Semantic Complexity Level 3: Explicated Semantically Complex Concepts

Unfortunately a problem arises when a text consists solely of the NSM primitives and Longman’s defining vocabulary. As was mentioned earlier, every language has words that are semantically complex, and if those words are not used when they are appropriate, the text seems long, drawn out, and unnecessarily wordy. For example, English has the complex concept *veterinarian*. A *veterinarian* is defined in Longman’s dictionary as “someone who is trained to give medical care and treatment to sick animals” (Longman 2003:1835). One of the texts that was developed for this project is a short story published by the Indonesian branch of the Summer Institute of Linguistics, and it describes how to prevent the spread of Avian Influenza. That source text repeatedly mentions veterinarians, but *veterinarian* is not in Longman's defining vocabulary. Therefore *veterinarian* was explicated for this project as "a doctor that treats sick animals."[12] One of the propositions in this source text says, "Veterinarians who know about this disease go to the market each day." Therefore the semantic representation for that sentence is, "Doctors that treat sick animals and that know about this disease go to the market each day." However, when generating this text in English, English speakers do not want to read sentences that contain *doctors that treat sick animals* because English has the word *veterinarian.* Therefore another category of concepts was allowed in TTA's ontology: complex concepts that have been explicated and have an associated Complex Concept Insertion rule. In this particular case, if a language has a lexeme corresponding to VETERINARIAN, then the user can activate the complex concept insertion rule for VETERINARIAN. Then all occurrences of 'doctor that treats sick animals' in the semantic representations will automatically be replaced with the complex concept VETERINARIAN. The Complex Concept Insertion rules will be described thoroughly in the next chapter.

3.3.1.1.4 Semantic Complexity Level 4: Inexplicable Semantically Complex Concepts

The final category of concepts in TTA's ontology includes concepts that are inexplicable. Inexplicable concepts such as proper names, cardinal and ordinal numbers, a variety of relationship markers, and several particles are included in the ontology.

3.3.1.2 Seven Semantic Categories

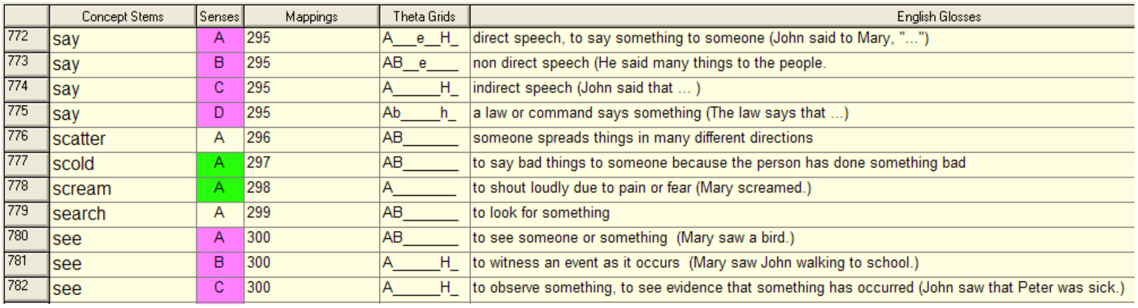
The concepts in TTA's ontology have been organized into seven semantic categories: 1) Objects, 2) Events, 3) Object Attributes, 4) Event Attributes, 5) Relations, 6) Conjunctions, and 7) Particles. These categories were chosen because the concepts in TTA's ontology have been significantly influenced by English, and all English words belong to the corresponding

[12] Note that 'doctor,' 'treat,' 'sick,' and 'animal' are all in Longman's defining vocabulary.

seven syntactic categories: 1) Nouns, 2) Verbs, 3) Adjectives, 4) Adverbs, 5) Prepositions, 6) Conjunctions, and 7) Particles. At the present time the concepts in each category are in simple lists; they are not subcategorized or in any type of hierarchical relationship with one another. However, one of the improvements planned for this project is to restructure the ontology into a hierarchical network. This improvement will be discussed in chapter 7.

3.3.1.3 A Sample of TTA's Ontology

A sample of TTA's ontology is shown below in figure 3-5. In the ontology, concepts which are NSM primitives are indicated with purple cells in the Senses column, the semantic molecules are marked with yellow cells in the Senses column, and the complex concepts that are inserted only if the user activates the associated rule are indicated by green cells in the Senses column. Inexplicable concepts are indicated with blue cells in the senses column, but none of the events shown in figure 3-3 are in that semantic level.



	Concept Stems	Senses	Mappings	Theta Grids	English Glosses
772	say	A	295	A___e__H_	direct speech, to say something to someone (John said to Mary, "...")
773	say	B	295	AB__e____	non direct speech (He said many things to the people.
774	say	C	295	A______H_	indirect speech (John said that ...)
775	say	D	295	Ab_____h_	a law or command says something (The law says that ...)
776	scatter	A	296	AB_______	someone spreads things in many different directions
777	scold	A	297	AB_______	to say bad things to someone because the person has done something bad
778	scream	A	298	A________	to shout loudly due to pain or fear (Mary screamed.)
779	search	A	299	AB_______	to look for something
780	see	A	300	AB_______	to see someone or something (Mary saw a bird.)
781	see	B	300	A______H_	to witness an event as it occurs (Mary saw John walking to school.)
782	see	C	300	A______H_	to observe something, to see evidence that something has occurred (John saw that Peter was sick.)

Figure 3-5. A Section of the Events Category in TTA's Ontology

As seen in the figure, many of the concepts have multiple senses, each sense having a very specific meaning and occurring in a very specific environment. For example, as seen in the first four rows of figure 3-5, SAY has four senses:

- SAY-A always occurs with direct quotes, the first proposition of the direct quote being in a patient proposition. That event also has an optional destination phrase meaning that sometimes the event occurs with a destination object phrase (e.g., *John said to Mary,* "…"), and sometimes there is not a destination object phrase (e.g., *John said,* "…").

- SAY-B always occurs when there is a patient object phrase rather than a patient proposition (e.g., *John said many things to the people.*). Again the destination object phrase is optional.
- SAY-C is always used for indirect speech (e.g., *John said that Mary read that book.*).
- SAY-D is always used in the semantic representations when a law, command, or message says something. For example, English speakers can say *The law says that we must pay taxes.* In Korean that structure must be revised to say *According to the law, we must pay taxes.*

By precisely defining each concept and using them in consistent environments, users of TTA are able to write grammar rules that will restructure the propositions according to the target language's requirements.

The events SCOLD-A and SCREAM-A are semantically complex as indicated by the green cells in figure 3-5, so they do not occur in the semantic representations. Longman's dictionary defines *scream* as "to make a loud high noise with your voice because you are hurt, frightened, excited, etc." Whenever the concept *scream* occurs in a source text, it is explicated in the semantic representation in one of two ways: 1) "X shout loudly because X be afraid", or 2) "X shout loudly because body part of X hurt". It is certainly possible that a language may have two different words for these two explications, but at this time, TTA's ontology has only one sense of SCREAM. If a target language has a good lexical match for *scream* that fits both of the explications above, the user can activate the complex concept insertion rule associated with SCREAM. Then all occurrences of "X shout loudly because X be afraid" and "X shout loudly because body part of X hurt" in the semantic representations will be replaced with X SCREAM-A.

The concept *scold* is even more complex. *Longman's* dictionary defines *scold* as "to angrily criticize someone, especially a child, about something they have done." Whenever *scold* occurs in a source text, it is replaced in the semantic representation with "X angrily criticize Y".

However, the semantic representations do not indicate whether or not a particular referent is a child. If a man angrily criticizes his boss, English speakers do not want the concept SCOLD inserted into the semantic representation. In order to clearly indicate where “X angrily criticize Y” may or may not be replaced with SCOLD, the semantic representation contains “X angrily criticize/scold Y.” By combining the semantic molecule “criticize” with the complex concept “scold” as in “criticize/scold”, the semantic representations are able to clearly indicate where this particular complex concept may or may not be inserted. If the user activates the complex concept insertion rule for SCOLD, then these constructions will automatically be changed from “X ANGRILY CRITICIZE/SCOLD Y” to “X SCOLD Y.” If the user does not activate the rule for SCOLD, then the semantic representation will be changed to “X ANGRILY CRITICIZE Y.” This situation will be discussed and illustrated more thoroughly in the next chapter under the Complex Concept Insertion rules.

3.3.2 The Features

As was mentioned above, an elaborate feature system has been developed for this project. Each feature is a comprehensive list of values that are pertinent to the world’s languages. Certainly more features and feature values will have to be added to this project in the future, but the set compiled to date is working well. The complete set of features and their values are listed in appendix E. The discussion here will be limited to the most salient features.

3.3.2.1 Object Features

Every object in the semantic representations is marked with six features which are listed below in Tables 3-1 through 3-6. Additionally every object has an Object List Index value so that the grammar rules are able to determine if two nominals refer to the same referent or different referents. For example, a proposition such as *One man said to another man, “…”* is represented in the semantic representations as [NP-Agent MAN$_1$] [VP SAY] [NP-Destination MAN$_2$] [“…”], where the ‘1’ and ‘2’ are indices which distinguish the two men from one another. If the proposition were *One man said to himself, “…”*, the representation would be [NP-Agent

MAN_1] [VP SAY-Reflexive] [NP-Destination MAN_1] ["…"], where both occurrences of MAN have the same Object Index value. These Object Index values are also used to identify the relativized referent in relative clauses, and in the grammar rules which identify referents that may be realized with pronouns. These latter two uses will be discussed more thoroughly in the next chapter.

3.3.2.1.1 Object Number

Table 3-1. Object Number

Number	Singular, Dual, Trial, Quadrial, Plural, Paucal

The values for object number are listed above in Table 3-1. All of these values are necessary because some languages morphologically distinguish each of these values. However, most languages only distinguish Singular and Plural, so users of TTA are able to write a feature collapsing rule that will merge Dual, Trial, and Quadrial with Plural. The feature collapsing rules will be described in the next chapter under the Transfer Grammar.

3.3.2.1.2 Object Participant Tracking

Table 3-2. Object Participant Tracking

Participant Tracking	First Mention, Integration, Routine, Exiting, Offstage, Restaging, Generic, Interrogative, Frame Inferable

The values of Participant Tracking are listed above in Table 3-2. This list was developed by Longacre (Longacre 1995:702), but the values Generic (e.g., *There are <u>lions</u> in Africa.*), Interrogative (e.g., *Which <u>book</u> did John read?*), and Frame Inferable (e.g., *The steering wheel on my new car is broken.*) (Prince 1981:230) were added for this project. Longacre includes three additional values: Confrontation and/or role change, marking of locally contrastive/thematic status, and an intrusive narrator evaluation. These values were not included in this list because they have not yet been needed; if a need for these values arises, they will be added.

3.3.2.1.3 Object Polarity

Table 3-3. Object Polarity

Polarity	Affirmative, Negative

Most of the objects in the semantic representations have their Polarity value set to Affirmative. In a sentence such as *No man has climbed that mountain*, the Polarity of MAN is set to Negative.

3.3.2.1.4 Object Proximity

Table 3-4. Object Proximity

Proximity	Near Speaker and Listener, Near Speaker, Near Listener, Remote within sight, Remote out of sight, Temporally Near, Temporally Remote, Contextually Near with Focus, Contextually Near, Not Applicable

The values for object proximity are listed above in Table 3-4. Similar to the values for object number, very few languages will distinguish each of these values. English distinguishes only two values for proximity: 'near' encoded with *this/these* and 'far' encoded with *that/those.* Korean has three values: 'near speaker' encoded with 이 [i][13], 'near listener' encoded with 그 [geu], and 'that over there away from speaker and listener' marked with 저 [jeo] (Cho et al. 2000:320). Tuscan Italian also has a three way distinction in proximity: *qui* 'here by me,' *costí* 'there by you,' and *la* 'there away from both of us' (Comrie 1985:14). For languages that do not distinguish all of these values, users of TTA are able to write a feature collapsing rule which will transform this list into the values that are pertinent to their particular target language. The first five values in this list (e.g., Near Speaker and Listener, Near Speaker, Near Listener, Remote within sight, and Remote out of sight) are used only in direct quotes. The values Temporally Near and Temporally Remote occur both in direct quotes and in narrative discourse. Examples of the last four values are shown below:

- Temporally Near: *This year we went on a vacation.*

[13] Throughout this dissertation all romanization of the Korean texts was done at the following web site: http://www.kawa.net/works/ajax/romanize/hangul-e.html accessed December 11, 2010. This site uses the revised romanization of Korean which is currently considered the official romanization system in South Korea. This system was developed by the South Korean government and released in July of 2000.

- Temporally Remote: *That year we didn't go on a vacation.*
- Contextually Near with Focus: *A certain man was living in California. This man …*
- Contextually Near: *A certain man was living in California. That man …*

3.3.2.1.5 Object Person

Table 3-5. Object Person

Person	First, Second, Third, First & Second, First & Third, Second & Third, First & Second & Third

All of the logically possible values for Person are listed above in Table 3-5. The First & Second value is equivalent to First Person Inclusive which is common in many languages, and First Person Plural is equivalent to First Person Exclusive. The values First & Third, Second & Third, First & Second & Third have not been used in the semantic representations, but they have been included here for completeness.

3.3.2.1.6 Object Participant Status

Table 3-6. Object Participant Status

Participant Status	Protagonist, Antagonist, Major Participant, Minor Participant, Major Prop, Minor Prop, Significant Location, Insignificant Location, Significant Time, Not Applicable

The values for this feature come from Longacre (Longacre 1995:701) and Bartsch (Bartsch 1995:47). Bartsch states that some languages overtly distinguish major and minor participants, props, locations, and events (Bartsch 1995:47). Therefore this feature is used in order to differentiate the significant referents from the incidental referents, and languages may use this information in the following ways:

- Method of introduction – major characters may be formally introduced while minor characters may simply appear.
- Fronting – significant characters, places, or times may be moved to sentence initial position as in *On my birthday I went to beach.*
- Morphological marking – a special morpheme may be used to indicate the most significant characters as in Algonquian languages which use a "proximate" suffix to

indicate the main character, and an “obviative” suffix to mark the lesser characters (Bartsch 1995:48).

- Referential expressions – major characters may be referred to by their names or titles.
- Pronouns – A language may use one set of pronouns when referring to significant characters, and another set of pronouns when referring to background characters, or a language may not use pronouns at all when referring to highly honored characters such as kings, queens, mothers, fathers, etc.
- Speech – Major participants may have their speech in direct quotes while minor participants have their speech marked in indirect quotes (Bartsch 1995:50). So if a minor participant has a direct quote in the semantic representations, a rule could change the direct quote to an indirect quote.
- Imperatives – Major participants may give unmitigated commands while minor participants always have their commands mitigated.
- Events – A language may have a small set of verb pairs, one member being used with honorable agents, the other member being used with non-honorable agents such as the Korean verb pairs 자다 [ja da] ‘to sleep’ (non-honorable agent) and 주무시다 [ju mu si da] ‘to sleep’ (honorable agent), 먹다 [meok da] ‘to eat’ (non-honorable agent) and 들다 [deul da] ‘to eat’ (honorable agent), 죽다 [juk da] ‘to die’ (non-honorable agent) and 돌아가다 [dor a ga da] ‘to die’ (honorable agent), etc.

3.3.2.2 Event Features

Every event in the semantic representations is marked with four features which are listed below in Tables 3-7 and 3-10 through 3-12. Two of these features are particularly problematic: Time and Aspect. When dealing with Time, the question is whether absolute time or relative time should be used. When dealing with Aspect, it is well known that many languages use their aspectual systems to distinguish the various salience bands. Therefore when developing the semantic representations and marking each event's aspect, one must determine whether the surface aspect in the source text is due to the salience band or the author's portrayal of the event. Both Time and Aspect will be discussed below.

3.3.2.2.1 Event Time

Table 3-7. Event Time

Time	Discourse, Present, Immediate Past, Earlier Today, Yesterday, 2 to 3 days ago, 4 to 6 days ago, 1 to 4 weeks ago, 1 to 5 months ago, 6 to 12 months ago, 1 to 9 years ago, 10 to 20 years ago, During Speaker's lifetime, Historic Past, Eternity Past, Unknown Past, Immediate Future, Later Today, Tomorrow, 2 to 3 days from now, 4 to 6 days from now, 1 to 4 weeks from now, 1 to 5 months from now, 6 to 12 months from now, 1 to 9 years from now, 10 to 20 years from now, during speaker's lifetime, Historic future, Eternity future, Unknown Future, Timeless

All languages have a concept of time (Comrie 1985:3), and tense is a grammaticalized expression of location in time (Comrie 1985:9). Absolute tense refers to tenses that take the present moment as their deictic center (Comrie 1985:36), while relative tense refers to tenses that take their deictic center from the context (Comrie 1985:56). It is well documented that many languages, particularly those in West Africa and Australia, have various degrees of past and future tenses. In order to compile the list of Time values shown in Table 3-8 above, data from a wide variety of languages was examined. Samples of this data are listed below (Comrie 1985:88-99; Dahl 1985:121).

- Languages with two degrees of past tense:
  Luganda: 'earlier today' and 'before today'
  Ancash Quechua: 'earlier today' and 'before today.'

- Languages with three degrees of past tense:
  Haya: 'earlier today,' 'yesterday,' and 'before yesterday'
  Hixkaryana: 'earlier today,' 'before today and going back several months,' and 'more than several months ago'
  Burera: 'earlier today,' 'within the past few days,' and 'before that'
  Kamba: 'earlier today,' 'yesterday to a week ago,' and 'a week or more ago'
  Mabuiag dialect of Kalaw Lagaw Ya: 'earlier today,' 'the past few days,' and 'more than a few days ago'
  Saibai dialect of Kalaw Lagaw Ya: 'earlier today,' 'yesterday,' and 'more remote'

- Languages with four degrees of past tense:
  Kamba: 'earlier today,' 'yesterday to a week ago,' 'more than a week ago but within the past few months,' and 'more than a few months ago'

Mabuiag: ‘earlier today,’ ‘last night[14],’ ‘yesterday,’ and ‘more remote’
Bamileke-Ngyemboon: ‘earlier today,’ ‘yesterday,’ ‘within the past few days,’ and ‘a long time ago (year or more)’

- Languages with five degrees of past tense:
  Bamileke-Dschang: ‘immediate past,’ ‘earlier today,’ ‘yesterday,’ ‘two days to several days ago,’ and ‘a year or more ago’
  Yandruwandha: ‘very recent,’ ‘within the last couple of days,’ ‘within the last few days,’ ‘weeks or months ago,’ and ‘distant past’
  Araona: ‘earlier today,’ ‘yesterday to several weeks ago,’ ‘several weeks to several years ago,’ ‘distant past,’ and ‘remote past’
  Yagua: ‘earlier today,’ ‘yesterday,’ ‘within a few weeks,’ ‘within a few months,’ and ‘distant or legendary past’

- Languages with seven degrees of past tense:
  Kiksht: ‘just now,’ ‘earlier today,’ ‘yesterday to a few days ago,’ ‘last week,’ ‘from a week ago to a year ago,’ ‘from one to ten years ago,’ and ‘ten or more years ago’

Table 3-8 shown below summarizes the data for the various degrees of past tense.

Table 3-8. Summary of the Various Degrees of Past Tense (P=Past)

	Unknown Past	Eternity Past	Historic Past	Speaker’s Lifetime	10 to 20 years ago	1 to 9 years ago	6 to 12 months ago	1 to 5 months ago	1 to 4 weeks ago	4 to 6 days ago	2 to 3 days ago	Yesterday	Earlier Today	Immediate Past	Now
Luganda	◄	—	—	—	—	—	—	—	—	—	—	P2	◄	P1	
Ancash Quechua	◄	—	—	—	—	—	—	—	—	—	—	P2	◄	P1	
Haya	◄	—	—	—	—	—	—	—	—	—	P3	P2	◄	P1	
Hixkaryana	◄	—	—	—	—	P3	◄	—	—	—	—	P2	◄	P1	
Burera	◄	—	—	—	—	—	—	—	—	P3	◄	P2	◄	P1	
Kamba	◄	—	—	—	—	—	—	—	P3	◄	—	P2	◄	P1	
Mabuiag dialect of Kalaw Lagaw Ya	◄	—	—	—	—	—	—	—	—	P3	◄	P2	◄	P1	
Saibai dialect of Kalaw Lagaw Ya	◄	—	—	—	—	—	—	—	—	—	P3	P2	◄	P1	
Kamba	◄	—	—	—	—	—	P4	◄	P3	◄	—	P2	◄	P1	
Mabuiag	◄	—	—	—	—	—	—	—	—	—	P3	P2	◄	P1	
Bamileke-Ngyemboon	◄	—	—	—	—	—	—	—	—	P4	P3	P2	◄	P1	
Bamileke-Dschang	◄	—	—	—	—	—	—	—	—	P5	P4	P3	P2	P1	
Yandruwandha	◄	—	—	—	—	P5	◄	—	P4	P3	◄	P2	◄	P1	
Araona	◄	—	P5	◄	P4	◄	—	P3	◄	—	—	P2	◄	P1	
Yagua	◄	—	—	—	—	—	P5	P4	◄	—	P3	P2	◄	P1	
Kiksht	◄	—	—	—	P7	P6	◄	—	P5	P4	◄	P3	P2	P1	

[14] The value ‘last night’ has to be collapsed with ‘yesterday’ in this system because there isn’t a time slot for ‘last night.’

As was stated above for the Object Number and Object Proximity features, there are far more values in this list than any one language will use. Therefore linguists who use TTA may write a feature collapsing rule in order to change the values in the semantic representations to the values that are pertinent to their particular target languages.

Many languages also have multiple degrees of future tense; a few examples are listed below (Comrie 1985:88-99, Dahl 1985:121):

- Languages with two degrees of future tense:
  Haya: 'tomorrow' and 'after tomorrow'

- Languages with three degrees of future tense:
  Kamba: 'today,' 'tomorrow to a month from now,' 'after a month from now'

- Languages with four degrees of future tense:
  Bamileke-Ngyemboon: 'later today,' 'tomorrow,' 'within the next few days,' and 'later in the future'

- Languages with five degrees of future tense:
  Bamileke-Dschang: 'immediate future,' 'later today,' 'tomorrow,' 'two days to several days from now,' and 'several days or more hence'

Table 3-9 shown below summarizes the data for the various degrees of future tense.

Table 3-9. Summary of the Various Degrees of Future Tense (F=Future)

	Now	Immediate Future	Later Today	Tomorrow	2 to 3 days from now	4 to 6 days from now	1 to 4 weeks from now	1 to 5 months from now	6 to 12 months	1 to 9 years from now	10 to 20 years	Speaker's Lifetime	Historic Future	Eternity Future	Unknown Future
Haya		F1 ——►			F2 ——►										
Kamba		F1 —►		F2 ——►				F3 ——►							
Bamileke-Ngyemboon		F1 —►		F2	F3	F4 ——►									
Bamileke-Dschang		F1	F2	F3	F4 —►		F5 ——►								

In this system a day begins when the sun rises and people get up, so typically around 6 or 7 am rather than midnight, and this is in agreement with most languages (Comrie 1985:89). For example, in English the phrase *last night* spans from about 10pm to 6am. The vast majority of the mainline events in the semantic representations are marked with a Time value of Discourse. Each language has its own tense system for each particular type of discourse, so when an event is marked with a Time value of Discourse, language specific rules in the grammar will mark the verb with the appropriate tense. In the semantic representations, the only mainline events that are marked with a Time value other than Discourse are those that are in direct quotes. For example, if someone is speaking and says *I went to the store*, the event GO will be marked with a value such as 'Yesterday', '1 week ago', etc., whichever is appropriate for the situation.

When marking the Time value for subordinate propositions, three options were considered: 1) marking them with absolute time, 2) marking them with relative time, or 3) marking them with both relative and absolute time. Some languages such as English adopt the first option, while other languages such as Korean and Imbabura Quechua adopt the second option (Comrie 1985:61). If the subordinate propositions in the semantic representations were marked with absolute time, then the languages which require relative time would have very complex rules to convert absolute time to relative time. However, if the subordinate propositions were marked with relative time, then the languages which require absolute time would have a very simple set of rules to convert relative time to absolute time. For this reason, the second option was adopted. Therefore the propositions that serve as agents or patients of an event (subject or object complements) and the propositions that modify objects (relative clauses) are marked with relative time, while mainline propositions and propositions that modify an event (adverbial clauses) are marked with absolute time. Examples illustrating these adopted guidelines follow:

- Patient Proposition: Infected Eye 1:7 *Alex knew* [*that Melissa's eyes were sick*]*.*

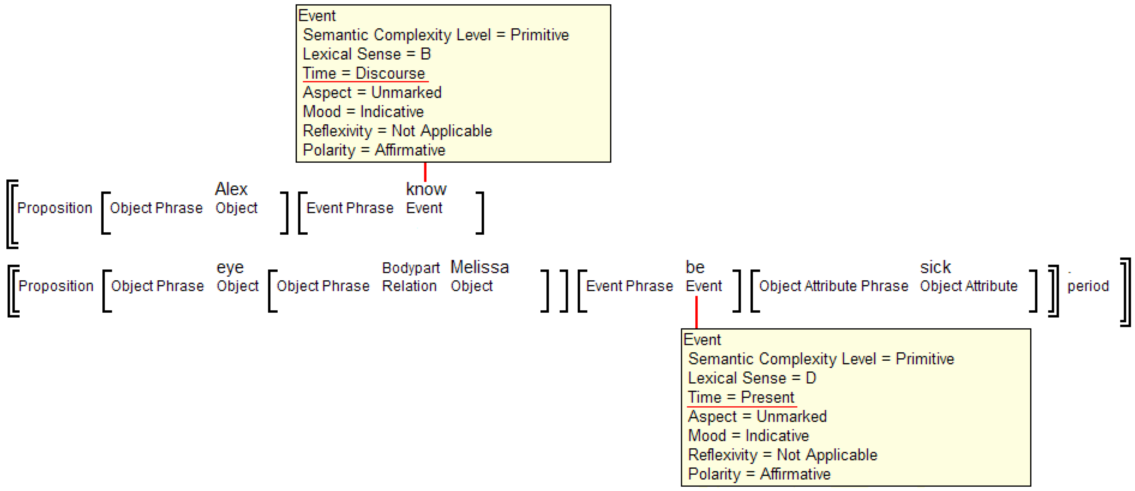


Figure 3-6 Semantic Representation of *Alex knew that Melissa's eyes were sick.*

In figure 3-6 above, the event KNOW in the matrix proposition has a Time value of Discourse as seen in the popup directly above KNOW, but the event BE in the patient proposition has a Time value of Present as shown in the popup below it. A Time value of Present in a subordinate proposition indicates that the event is simultaneous with the matrix event. The Korean equivalent of this proposition is shown below in (1A). In that sentence the verb 'be sick', which is in the object complement clause, is unmarked for tense, meaning that it is interpreted as present tense. However, in English that verb in the complement clause is marked with past tense as seen in the gloss. Because the semantic representations have relative time in patient propositions, Korean, which uses relative tense, does not need a rule to generate the proper time value in the patient proposition. English, which uses absolute time, requires a single rule that changes the Time value Present to Discourse when the matrix event is marked with a Time value of Discourse.

(1A) 알렉스-는 멜리사-의 눈-이 아프-다는 것을 알-았-다.
Alex-Topic Melissa-Gen eye-Topic sick-complementizer know-past-declarative
'Alex knew that Melissa's eyes were sick.' (Korean text generated by TTA)

- Object modifying proposition: Infected Eye 1:8 *A towel* [*that was hanging on a rope*] *was dirty.*

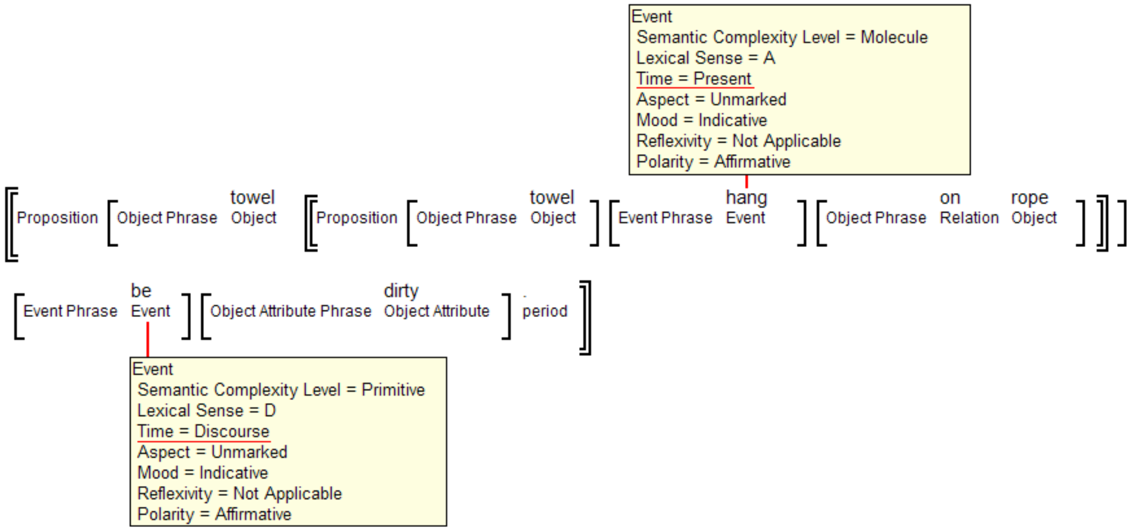


Figure 3-7 Semantic Representation of *A towel that was hanging on a rope was dirty.*

In figure 3-7 above, the event BE in the independent proposition has a Time value of Discourse, but the event HANG in the event modifying proposition has a Time value of Present. In the Korean equivalent of this proposition, the event in the object modifying proposition HANG is marked with the present tense relativizer −는 as seen in 1B below, but in English the verb in the relative clause is marked with past tense[15]. As mentioned in the previous example, the Korean grammar is able to use the Time value in the semantic representation, while English requires a rule to convert relative time to absolute time.

(1B) 밧줄 위에 걸려 있−는 수건−이 더러−웠−다.
rope on hang-Present.Relativizer towel-Topic be.dirty-past-declarative
'A towel that was hanging on a rope was dirty.' (Korean text generated by TTA)

- Event modifying proposition: Infected Eye 1:2 *But Melissa was not happy* [*because her eyes were very sore*].

[15] The Salience Band for the relative clause is set to Backgrounded Action which generates the past imperfective *was hanging* in English.

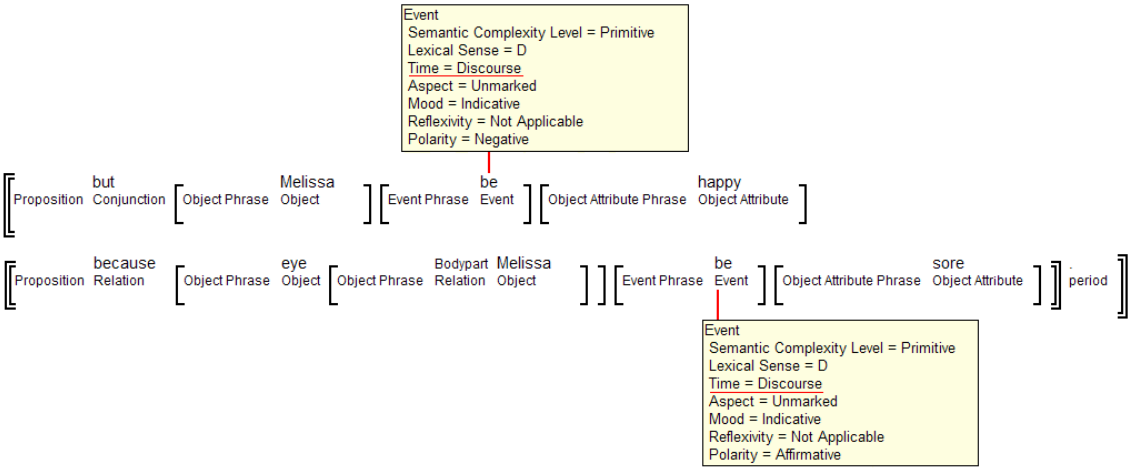


Figure 3-8 Semantic Representation of *But Melissa was not happy because her eyes were very sore.*

In figure 3-8 above, the event BE in the independent proposition has a Time value of Discourse, and the event BE in the event modifying proposition also has a Time value of Discourse. In both the Korean and English equivalents for this proposition, the events in the main proposition and in the event modifying proposition are marked with past tense as seen in (1C) below.

(1C) 멜리사−는 눈−이 매우 아팠−기 때문에 행복하−지 않−았−다.
Melissa-Topic eye-Topic very sick.Past-because be.happy-not-Past-Declarative
‘Melissa was not happy because her eyes were very sore.’ (Korean text generated by TTA)

In summary, all the events in mainline propositions and event modifying propositions are marked with a Time value of Discourse, except for the events in direct quotes. Events in mainline and event modifying direct quotes are marked with the appropriate absolute time value. Events in object modifying propositions are marked with relative time, and events in agent and patient propositions are also marked with relative time. This system may need adjustments after more languages have been examined, but it has worked well for the test languages to date.

3.3.2.2.2 Event Aspect

Table 3-10. Event Aspect

Aspect	Unmarked, Completive, Inceptive, Cessative, Continuative, Habitual, Gnomic, Imperfective

Aspect portrays the internal temporal constituency of an event (Comrie 1976:3), and is often portrayed as shown below in Figure 3-9.

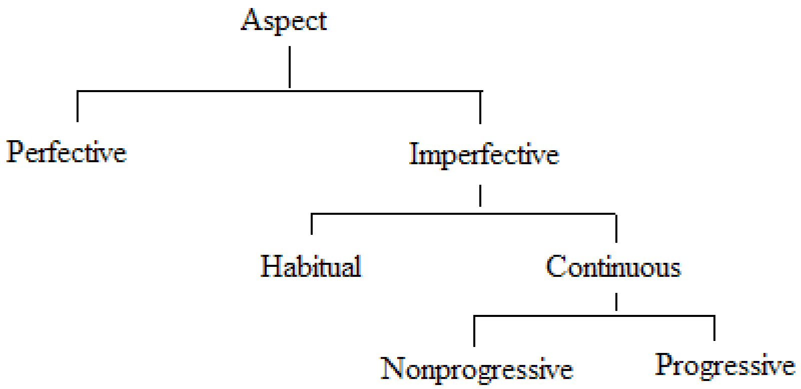


Figure 3-9. A Typical Representation of Aspect (Comrie 1976:24)

The values listed above in Table 3-10 do not correspond well with the values in figure 3-9 for a single reason: The values chosen for this system are those that emphasize a particular temporal component of the action, and are independent of the event's salience band and discourse genre. Often it is an event's salience band within a particular discourse genre that dictates the surface aspectual coding, and each language has its own rules for encoding aspect according to the salience band and discourse genre. For example, English uses imperfective aspect to encode backgrounded events in a narrative as in Infected Eye 1:2 *One day Melissa was sitting outside her house*. In this proposition the event SIT has an Aspect value of Unmarked, but its Salience Band is Backgrounded Action, and the proposition is in a Climactic Narrative Story. Therefore the rules of English generate *sitting* with imperfective aspect. Similarly in Kande's Story 1:7 *Kande's father had slept for many days*, the event SLEEP has an aspect value of unmarked, but its Salience Band is Flashback and the Discourse Genre is Climactic Narrative Story. Therefore the English grammar generates *had slept* with perfect aspect. The values chosen for this aspectual system listed in table 3-11 above are those that highlight a particular component of the event, regardless of the event's salience band or discourse genre. Examples of each of these aspects follow:

- Completive: *John finished reading a book.*

- Inceptive: *John started reading a book.*
- Cessative: *John stopped reading a book.*
- Continuative: *John continued reading a book.*
- Habitual: *John always/regularly/habitually reads this book.*
- Gnomic: *The sun rises in the east.*

The value Imperfective shown above in table 3-10 is only used in direct quotes such as in Infected Eye 1:6 *Melissa shouted, "Something is preventing me from opening my eyes."* In that proposition, the aspect of PREVENT is set to Imperfective because the speaker is portraying the action as ongoing. That particular proposition is not a backgrounded action, so the use of imperfective aspect in the source text is a reflection of the author's portrayal of the event.

Several traditional values of aspect are not included in the table 3-10 above, notably perfective, perfect, and pluperfect. Perfective aspect seems to be the default aspect for events that occurred in the past. Perfective aspect was not included in this feature because it does not emphasize a particular temporal component of the event. Perfect and pluperfect were not included in this system because they are methods of encoding flashback.

3.3.2.2.3 Event Mood

Table 3-11. Event Mood

Mood	Indicative, Definite Potential, Probable Potential, 'might' Potential, Unlikely Potential, Impossible Potential, 'must' Obligation, 'should' Obligation, 'should not' Obligation, Forbidden Obligation, 'may' (permissive)

The values of mood selected for this system are all very straightforward and will be illustrated below. Indicative mood is the default mood and will not be illustrated.

- Definite Potential: *John will definitely read this book.*
- Probable Potential: *John will probably read this book.*
- 'might' Potential: *John might read this book.*
- Unlikely Potential: *John might not read this book.*
- Impossible Potential: *John will definitely not read this book.*

- ‘must’ Obligation: *John must read this book.*
- ‘should’ Obligation: *John should read this book.*
- ‘should not’ Obligation: *John should not read this book.*
- Forbidden Obligation: *John must not read this book.*
- ‘may’ (permissive): *The teacher said to John, “You may read this book.”*

3.3.2.2.4 Event Polarity

Table 3-12. Event Polarity

Polarity	Affirmative, Negative, Emphatic Affirmative, Emphatic Negative

The vast majority of the events in the semantic representations have a Polarity value of Affirmative. The emphatic polarity values (e.g., Emphatic Affirmative and Emphatic Negative) are for situations such as: *John certainly read this book*, and *John certainly did not read this book.*

3.3.2.3 Object Attribute Features

Object attributes have a single feature called Degree, and the values are listed below in table 3-13.

Table 3-13. Object Attribute Features

Degree	Comparative, Superlative, Intensified, Extremely Intensified, ‘too’ or ‘overly’, ‘less’, ‘least’, Not Applicable

Examples illustrating each of these values follow:

- Comparative: *John is taller than Mary.*
- Superlative: *John is the tallest man.*
- Intensified: *John is very tall.*
- Extremely Intensified: *John is extremely tall.*
- ‘too’ or ‘overly’: *John is too tall.*
- ‘less’: *John is less important than Mary.*
- ‘least’: *John is the least important person.*

3.3.2.4 Object Phrase Features

Object phrases have two features: Sequence and Semantic Role. The Sequence feature specifies whether a particular phrase is the first in a sequence, the last in a sequence, in the middle of a sequence, or not in a sequence. All types of phrases have a Sequence feature, but it will not be discussed further here. The Semantic Role feature is considerably more controversial, and the values selected for this project are shown below in table 3-14.

Table 3-14. Object Phrase Features

Semantic Role	Most Agent-like, Most Patient-like, State, Source, Destination, Instrument, Beneficiary, Addressee, Not Applicable

Semantic roles were brought to a position of prominence in linguistic theory by Charles Fillmore (Fillmore 1968). Many linguists have suggested various sets of semantic roles. Some linguists have proposed large sets with finely differentiated values, while other linguists have proposed very small sets of generic semantic roles. A common set of semantic roles that is between these two extremes was proposed by Frantz (Frantz 1974:9-10). He proposed the following set of semantic roles with their definitions:

- Agent – Instigator of an action
- Experiencer – a psychologically affected patient; sentient being
- Object – a non-psychologically affected patient; an entity that is moved, changed, etc.
- Means – non-instigative cause of a predication/instrument
- Source – former state, location, or time
- Goal – later state, location, or time
- Referential – entity to which another is related by a predication; the point, line, or plane or reference
- Time

Givón points out that "one must further remember that in principle there are as many case-roles as there are verbs …" (Givón 1990:127). The view adopted for this project is similar to that of Givón's: each event has its own set of very specific semantic roles. However, since it is entirely impractical to have thousands of semantic roles, a small set of very generic semantic roles was adopted, as is shown in table 3-14 above. However, TTA allows a linguist to enter his own set of semantic roles that are pertinent to his language, and then each event in the ontology has its own theta grid adjustment rule which is capable of changing the generic semantic roles to whatever is appropriate for that particular target language. A brief discussion of each of the semantic roles used in this project follows.

- Most Agent-like: This is a generic role that includes agents, experiencers, forces, etc. It is the most salient nominal with respect to the event.
- Most Patient-like: This is another generic role that includes Themes, Undergoers, and all the nominals that are directly affected by the event.
- State: This is the second nominal in all BE propositions (e.g., *John is a man/doctor/my brother/with Mary/etc*.) and also the proper name in propositions such as *John named his son Steve*.
- Source: This is a generic role that indicates where an action originates (e.g., Infected Eye 1:20 *Mary took the garbage away from her house.*)
- Destination: This role marks a person or place to which the action is oriented (e.g, *John said to Mary …*).
- Instrument: The event is performed with a particular nominal (e.g., Infected Eye 1:9 *Melissa washed her eye with clean water*.)
- Beneficiary: This is the standard beneficiary role; the event is performed for the benefit of someone (e.g., *John washed the dishes for Mary.*).
- Addressee: This is the standard addressee role (e.g., *John, wash the dishes.*)
- Not Applicable: Nominals that are not directly related to the event are given a Semantic Role value of ‘Not Applicable’ (e.g., *John studied for the test in the library on Monday evening.*).

3.3.2.5 Proposition Features

Each proposition in the semantic representations is marked with fifteen features, but only nine of those features will be discussed here. A complete list of the features associated with propositions is in appendix E. Five of the features discussed here are used with direct quotes and are called Speaker, Listener, Speaker Attitude, Speaker's Age, and Speaker to Listener's Age. These direct quotation features have been included here because they are essential to the discussion of Korean honorifics in section 5.2. Each proposition is also marked with three discourse features called Discourse Genre, Notional Structure Schema, and Salience Band. The values for these three features all come from Longacre (Longacre 1996:10, 28, 36).

3.3.2.5.1 Proposition Type

Table 3-15. Proposition Type

Type	Independent, Restrictive Thing Modifier, Descriptive Thing Modifier, Event Modifier, Agent, Patient, Attributive Patient, Closing Quotation Frame

The Type feature indicates the proposition's type. All of the non-dependent propositions in the semantic representations are marked as Independent. The dependent propositions are each illustrated below:

- Restrictive Thing Modifier: *I saw a man* [*who was reading a book*].
- Descriptive Thing Modifier: *John* [*, whom I've known for ten years,*] *read a book.*
- Event Modifier: [*After John read a book,*] *he went for a walk.*
- Agent: *It is good* [*that John read a book*].
- Patient: *John thought* [*that Mary read a book*].
- Attributive Patient: *John is afraid* [*to read that book*].
- Closing Quotation Frame: *Melissa said to Janet, "Please look at my eyes. Is some sand on my eyes?"* [*Melissa said to Janet*].

All direct quotes that consist of multiple propositions have a subordinate proposition at the end of the direct quote, and that subordinate proposition's Type is set to Closing Quotation Frame. This subordinate proposition is necessary to accommodate languages which customarily repeat the speaker and listener in order to signal the end of the quotation. Languages such as Korean and English do not use these closing quotation frames, so the transfer grammar has an option to delete all of the closing quotation frame propositions from the semantic representations.

3.3.2.5.2 Proposition Illocutionary Force

Table 3-16. Proposition Illocutionary Force

Illocutionary Force	Declarative, Imperative, Content Interrogative, Yes-No Interrogative

This feature has the standard four values of illocutionary force.

3.3.2.5.3 Proposition Topic NP

Table 3-17. Proposition Topic Phrase

Topic NP	Most Agent-like, Most Patient-like

The vast majority of the propositions in the semantic representations have their Topic NP feature set to Most Agent-like. In English this corresponds to the active form of the sentence. Occasionally a proposition's Topic NP will be set to Most Patient-like for one of two possible reasons: 1) the patient is the referent in focus and its prominence in the discourse needs to be maintained, or 2) the agent of the action is either unknown or insignificant, so it needs to be made less prominent. In either of these cases, each target language will have its own mechanism for promoting the patient or demoting the agent. In English both of these situations correspond to the passive form.

Many languages permit a variety of semantic roles to be topicalized. For example, in Korean it is possible to topicalize almost any argument in a proposition. However, TTA's semantic representations only permit the Most-Agent like and Most-Patient like arguments to be topicalized.

3.3.2.5.4 Proposition Salience Band

Table 3-18. Proposition Salience Band

Salience Band	Pivotal Storyline, Primary Storyline, Secondary Storyline, Script Predictable Actions, Backgrounded Actions, Flashback, Setting, Irrealis, Evaluation, Cohesive Material, Not Applicable

Longacre developed this list of salience bands to distinguish mainline material from the various types of supportive material (Longacre 1996:28). This feature has proven extremely helpful because quite often a language employs its tense/aspect system to encode the various salience bands. Therefore each language has its own set of rules which look at the various salience bands, and then produce surface forms marked appropriately for tense, aspect, and perhaps also word order (Longacre 1996:23). For example, English uses imperfective aspect when the salience band is backgrounded action (e.g., Melissa's Eye 1:2 *One day a girl named Melissa was sitting outside her house.*), and it uses perfect aspect for flashback (e.g., Kande's Story 1:7 *Father had slept for many days.*).

3.3.2.5.5 Proposition Speaker

Table 3-19. Proposition Speaker

Speaker	Not Applicable, Adult Daughter, Adult Son, Angel, Animal, Boy, Brother, Crowd, Daughter, Demon, Disciple, Employee, Employer, Father, Girl, God, Government Leader, Government Official, Group of Friends, Holy Spirit, Husband, Jesus, King, Man, Military Leader, Mother, Prophet, Queen, Religious Leader, Satan, Servant, Sister, Slave, Slave Owner, Soldier, Son, Wife, Woman, Written Material to General Audience (letter, law, etc.)

Many languages have honorific systems, and these systems must be employed when people talk to each other. In these languages when children talk to their parents, they must use a certain type of speech, and when the parents talk to their children, they will use a different type of speech. Similarly when employers talk to employees, kings talk to servants, students talk to teachers, etc., each of these speech situations requires the proper use of honorifics. It is impossible to predict or determine all the situations in which a language may require the use of honorifics. Some cultures may honor warriors, other cultures may honor religious leaders, while other cultures may honor only the parents. Therefore, in order to accommodate all of the potential situations which may require honorifics, general categories were developed for the speaker and listener. Every proposition that is direct speech is tagged to indicate who is talking to whom. The categories of speakers are shown above in table 3-19.

3.3.2.5.6 Proposition Listener

Table 3-20. Proposition Listener

Listener	Not Applicable, Adult Daughter, Adult Son, Angel, Animal, Boy, Brother, Crowd, Daughter, Demon, Disciple, Employee, Employer, Father, Girl, God, Government Leader, Government Official, Group of Friends, Holy Spirit, Husband, Jesus, King, Man, Military Leader, Mother, Prophet, Queen, Religious Leader, Satan, Servant, Sister, Slave, Slave Owner, Soldier, Son, Wife, Woman

This set of values is identical to the set of values under Speaker, except the value Written Material to General Audience was not included. All direct quotes in the semantic representations are marked with both a Speaker and a Listener value.

3.3.2.5.7 Proposition Speaker's Attitude

Table 3-21. Proposition Speaker's Attitude

Speaker's Attitude	Not Applicable, Neutral, Familiar, Endearing, Honorable, Derogatory, Friendly, Antagonistic, Complimentary, Anger, Rebuke

When one person is talking to another, the speaker's attitude toward the listener at that moment may affect word choice, whether or not an imperative is mitigated, the degree of respect shown, etc. The value Neutral in the table above is used when people who do not know each other are speaking and there are no significant emotions involved. The value Familiar is used when the speaker knows the listener, but there are no significant emotions involved. If significant emotions are involved, the proposition is tagged accordingly.

3.3.2.5.8 Proposition Speaker's Age

Table 3-22. Proposition Speaker's Age

Speaker's Age	Not Applicable, Child (0-17), Young Adult (18-24), Adult (25-49), Elder (50+)

Children may use different words or speech styles than adults, and elders may speak differently than adults. Therefore this feature indicates the approximate age group of the speaker.

3.3.2.5.9 Proposition Speaker to Listener's Age

Table 3-23. Proposition Speaker to Listener's Age

Speaker to Listener's Age	Not Applicable, Older - Different Generation, Older - Same Generation, Essentially the Same Age, Younger - Different Generation, Younger - Same Generation

In many languages such as Korean, when someone speaks to a person that is significantly older, certain honorifics are required. Similarly when an adult speaks to a child, certain word choices or speech styles may be used. Therefore this feature indicates whether the speaker and listener are of the same generation or different generations.

3.3.3 The Structures

The introduction in chapter 2 described the three primary difficulties associated with fully automatic machine translation: 1) part of speech disambiguation, 2) word sense disambiguation, and 3) structural disambiguation. Natural language generators avoid these three difficulties by using manually developed semantic representations. The issues of word sense disambiguation and part of speech disambiguation and their resolution were described above in the ontology section. The difficulties associated with structural disambiguation are resolved in the structures of the semantic representations, and the structures that are permitted in this system of semantic representation will be described here.

As was stated above in section 3.3.1, the concepts in the ontology have been organized into seven semantic categories: 1) Objects, 2) Events, 3) Object Attributes, 4) Event Attributes, 5) Relations, 6) Conjunctions, and 7) Particles. The concepts in the semantic representations are each marked with the appropriate semantic category thereby eliminating all part of speech ambiguity. Then the concepts are put into phrases; there are four types of phrases in this system: 1) Object Phrases, 2) Event Phrases, 3) Object Attribute Phrases, and 4) Event Attribute Phrases. These phrases are then put into propositions. Noticeably absent from this system are relational (adpositional) phrases. There are no relational phrases in this system because they add considerable complexity without providing any benefits. In this system, relations occur in event modifying propositions (e.g., <u>*After*</u> *John read this book, …*) and in object phrases (e.g., <u>*under*</u> *the table*).

A sample from the semantic representations is shown below in figure 3-10. That figure shows the semantic representation for *John read a good book*. In this sample, all of the features associated with the concepts, phrases, and proposition have been hidden.

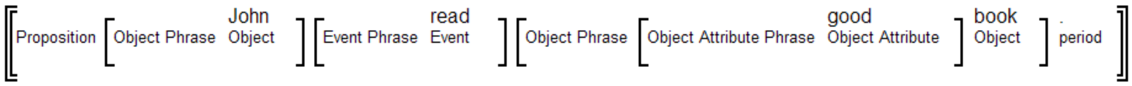
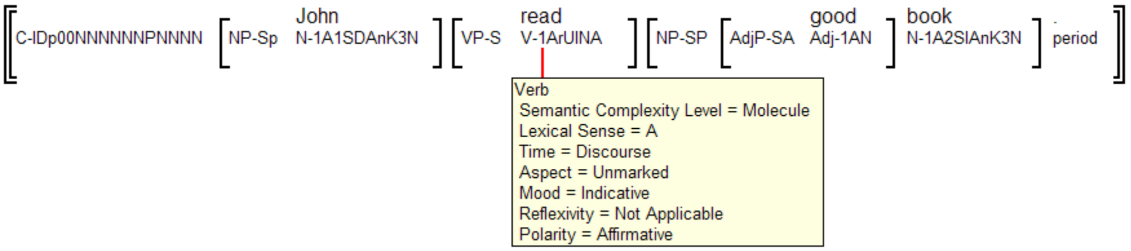


Figure 3-10. Semantic Representation without Features for *John read a good book.*

Figure 3-10 above indicates that objects are embedded in object phrases, events are embedded in event phrases, and object attributes are embedded in object attribute phrases. All of these phrases are embedded in a proposition. In figure 3-11 below is the same proposition, but the features are no longer hidden, and there is a popup explaining each of the feature values associated with the verb.



Verb
Semantic Complexity Level = Molecule
Lexical Sense = A
Time = Discourse
Aspect = Unmarked
Mood = Indicative
Reflexivity = Not Applicable
Polarity = Affirmative

Figure 3-11. Semantic Representation with Features for *John read a good book*.

The figure above shows how the features that were discussed in section 3.3.2 above are associated with specific objects, events, object phrases and propositions in the semantic representations.

3.3.3.1 Event Modifying Propositions

There are twenty-three types of event modifying propositions allowed in this system. This is a small subset of the event modifying propositions that Longacre discusses (Longacre 1996:51-100). The event modifying propositions that are permitted in the semantic representations are as follows:

- Comparison: *John likes Mary <u>more/less than Bill likes Susan</u>.* Longacre states that comparison is not a universal (Longacre 1996:60). For languages which do not permit this construction, a transfer rule can easily convert these propositions to two separate propositions which convey approximately the same information as the original: *John likes Mary a lot. But Bill likes Susan a little.* In these comparative constructions, the same event will always be used in both the matrix and dependent propositions, but the agent and/or the patient objects may be different. When the agents in both propositions are identical, some languages may reduce these to *John likes Mary more than Susan.* The semantic representation for *John likes this book more than Mary likes that book* is shown below in Figure 3-12.

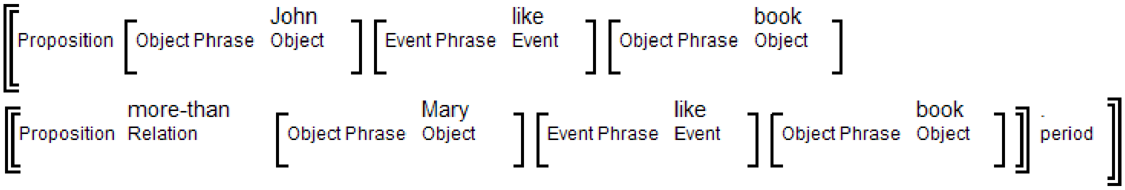


Figure 3-12. Semantic Representation of *John likes this book more than Mary likes that book.*

- Conditional: *If you go to the party, you'll see John.*
- Hypothetical: *If you were to go to the party, you'd see John.*
- Counterfactual: *If you had gone to the party, you would have seen John.* Counterfactual constructions are not universal (Longacre 1996:75) so the grammar library in TTA has a rule which converts these constructions to equivalent constructions with a reason event modifying proposition: *You did not see John because you did not go to the party.* That particular rule must look at the polarity of the events in both the matrix proposition and the event modifying proposition. Examples illustrating that rule are shown below in Table 3-24.

Table 3-24. Examples of Rule which Converts Counterfactuals to 'because' Propositions

Matrix Polarity	Event Modifier Polarity	Example
Affirmative	Affirmative	*If you had read the book, you would have passed the test. ->* *You didn't pass the test because you didn't read the book.*
Negative	Negative	*If you hadn't eaten the candy, you wouldn't be sick. ->* *You're sick because you ate the candy.*
Negative	Affirmative	*If you had eaten your lunch, you wouldn't be hungry. ->* *You're hungry because you didn't eat your lunch.*
Affirmative	Negative	*If you hadn't eaten your lunch, you'd be hungry. ->* *You're not hungry because you ate your lunch.*

- Concessive: *Although John loves Mary, he married Susan.*
- Concessive Conditional: *Even if it rains, we will go to the beach.*
- Manner: *John passed the test by studying every day.*
- Purpose: *John moved to Texas in order to study linguistics.*
- Reason: *John ate an apple because he was hungry.*
- Result-Enablement: *John went to the store so that Peter could sleep.*
- Result-Causative: *John went to the store so that Peter would sleep.*
- Simile: *John walks just like Mary walks.*
- Substitution (same agent): *John read a book instead of watching a movie.*
- Substitution (different agent): *John read this book instead of Mary reading that book.*
- Temporal (after): *After John walked, he read a book.*
- Temporal (before): *Before John walked, he read a book.*
- Temporal (since): *John has been reading since he was four years old.*
- Temporal (until): *John studied until Mary came to his house.* The concept UNTIL is probably not universal, so TTA's Grammar Library includes a rule which converts these constructions to an equivalent proposition using WHEN: *When Mary came to John's house, he stopped studying.*
- Temporal (when): *When Mary came to John's house, he started studying.*
- Temporal (whenever): *Whenever Mary comes to John's house, he stops studying.*
- Temporal (while): *John studied while Mary read a book.*
- 'unless': *John will study unless Mary comes to his house.* The concept UNLESS is probably not universal, so a rule in TTA's Grammar Library converts these constructions to an equivalent using the conditional IF: *If Mary does not come to John's house, John will study.* The semantic representations do not permit events with Negative polarity in these event modifying propositions, so the rule that converts these

propositions to the equivalent proposition using IF only has to look at the polarity of the matrix proposition. Examples illustrating this rule are shown in table 3-25 below.

Table 3-25. Examples of Rule which Converts UNLESS Propositions to 'if' Propositions

Matrix Polarity	Event Modifier Polarity	Example
Affirmative	Affirmative	*John will study unless Mary comes to his house. ->* *If Mary does not come to John's house, he will study.*
Negative	Affirmative	*John won't study unless it rains. -> If it doesn't rain, John won't study.*

The semantic representation of a proposition with an embedded event modifying proposition is shown below in Figure 3-13.

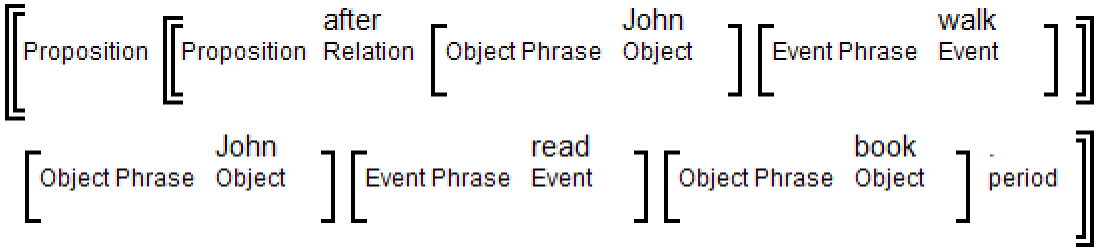
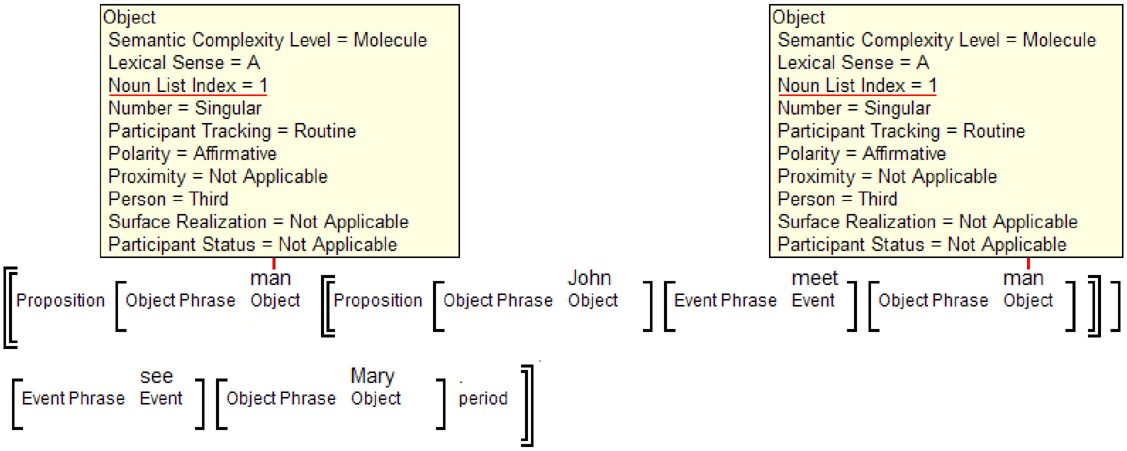


Figure 3-13. Semantic Representation of *After John walked, he read a book.*

3.3.3.2 Object Modifying Propositions

In the semantic representations, object modifying propositions are always embedded in the phrase that contains the modified object. The modified object and the object in the modifying proposition that is coreferential with the modified object both have the same Object Index value. Object modifying propositions are permitted in all object phrases except in genitives and objects of comparison. The reason for this restriction is that many languages do not permit relative clauses to be formed on these NPs. This will be discussed more thoroughly in the next chapter in the transfer grammar section. The semantic representation for *The man that John met saw Mary* is shown below in figure 3-14.


Object
Semantic Complexity Level = Molecule
Lexical Sense = A
Noun List Index = 1
Number = Singular
Participant Tracking = Routine
Polarity = Affirmative
Proximity = Not Applicable
Person = Third
Surface Realization = Not Applicable
Participant Status = Not Applicable

Object
Semantic Complexity Level = Molecule
Lexical Sense = A
Noun List Index = 1
Number = Singular
Participant Tracking = Routine
Polarity = Affirmative
Proximity = Not Applicable
Person = Third
Surface Realization = Not Applicable
Participant Status = Not Applicable

Figure 3-14. Semantic Representation of *The man that John met saw Mary.*

The two popups in figure 3-14 above show that both occurrences of MAN have the same Noun List Index value of 1.

3.3.3.3 Agent Propositions

Agent propositions function as the agent of an event. The semantic representation for *It pleased Mary that John read that book* is shown below in Figure 3-15. If a language requires an expletive pronoun and the postposing of the agent proposition as English does (Haegeman 1994:62), then the linguist must write the necessary rules to accomplish this.

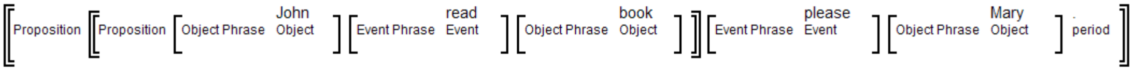


Figure 3-15. Semantic Representation of *It pleased Mary that John read that book.*

3.3.3.4 Patient Propositions

Patient propositions function as the patient of an event. The semantic representation for *John wants to read that book* is shown below in figure 3-16.

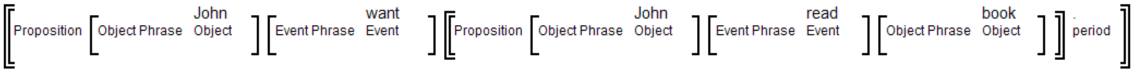


Figure 3-16. Semantic Representation of *John wants to read that book.*

3.3.3.5 Attributive Patient Propositions

Attributive patient propositions function as the patient of an object attribute and are embedded in the object attribute phrase. The semantic representation for *John is afraid to read that book* is shown below in figure 3-17.

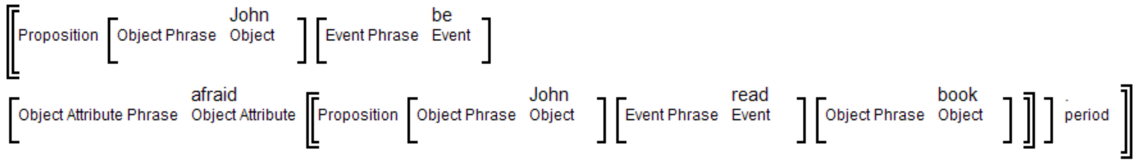


Figure 3-17. Semantic Representation of *John is afraid to read that book.*

3.3.3.6 Closing Quotation Frames

As was mentioned above in section 3.3.2.9.1, all direct quotes that consist of multiple propositions have an embedded subordinate proposition at the end of the quote, and its Type is set to Closing Quotation Frame. A sample of a closing quotation frame is shown below in figure 3-18.

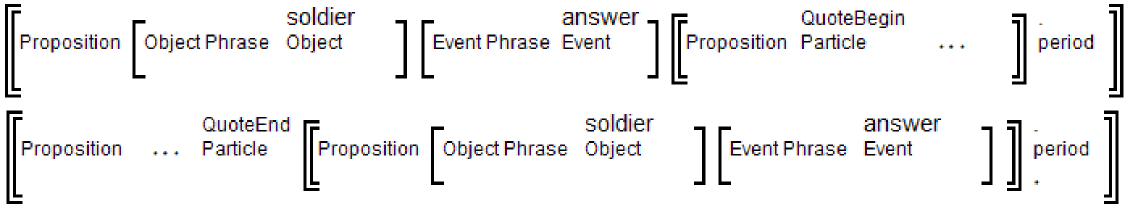


Figure 3-18. An Example of a Closing Quotation Frame Proposition

3.3.3.7 Object-Object Relationships

The proposal for this dissertation stated that this project will identify and categorize very precisely the many object-object relationships that are permitted in English. However, that task proved impossible because the object-object relationships permitted by English are virtually limitless[16]. Therefore, rather than very precisely identifying the relationships between various object-object combinations, a few very specific relationships were identified, and all other object-object relationships are marked in the semantic representations simply as "Generic

[16] For a detailed discussion of complex nominals, see The Syntax and Semantics of Complex Nominals by Judith Levi, 1978, New York: Academic Press.

Genitive." Then a special section was added to the transfer grammar to deal with the object-object relationships that are tagged as "Generic Genitive." Therefore there are now eleven possible ways that objects may be related to one another in the semantic representations; the first ten object-object relationships are very precise and easily identified, while the final relationship is completely generic and used for all object-object relationships that don't fit into one of the first ten categories. The hypothesis underlying this approach is that every target language will handle each occurrence of the first ten object-object relationships in the same way. For example, the first object-object relationship is Bodypart (e.g., *John's hand, John's eye,* etc.). Different languages certainly divide the human body and animal bodies into different parts, but the hypothesis adopted here is that whatever morphosyntactic method a language employs to encode *John's hand*, that language will employ the same method to encode *John's eye, John's foot,* etc.

- Bodypart: *Melissa's eye* is represented as shown in figure 3-19 below. All body parts are marked with the relation Bodypart.

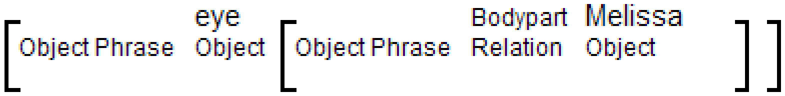


Figure 3-19 Semantic Representation of *Melissa's eye*

- Made of: *brick house* is represented as shown below in figure 3-20. Whenever one referent is made of another referent, the relation 'Made of' links the two referents.

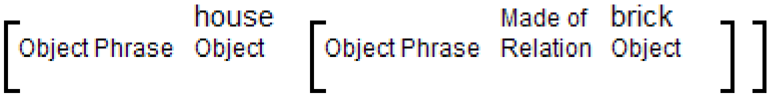


Figure 3-20 Semantic Representation of *brick house*

- Group: *herd of sheep* is represented as shown below in figure 3-21. English has different classifiers for different objects (e.g., *gaggle of geese, school of fish, range of mountains,* etc.). In the semantic representations, the generic term GROUP is used with the generic relation Group, and the target grammar must supply the appropriate terms for each target language.

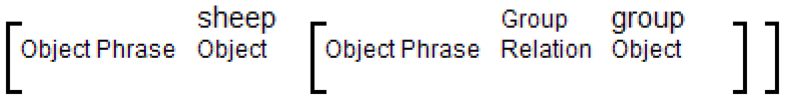


Figure 3-21. Semantic Representation of *a herd of sheep*

- Kinship: *Mary's mother* is represented as shown below in figure 3-22. All blood and legal kinships are marked with the relation Kinship.

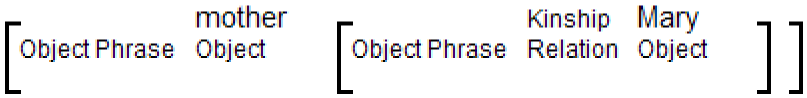


Figure 3-22 Semantic Representation of *Mary's mother*

- Name: *a man named John* is represented as shown below in figure 3-23. In English this is not an object-object relationship, but this structure has proven very productive and is identical to all of the other object-object relationships, so it has been included here.

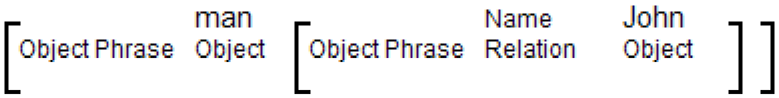


Figure 3-23 Semantic Representation of *a man named John*

- Owner: *John's book* is represented as shown below in figure 3-24.

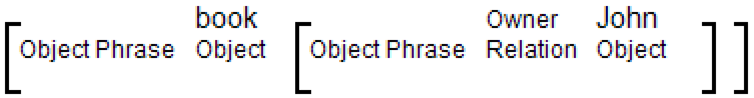


Figure 3-24 Semantic Representation of *John's book*

- Quantity: *two liters of oil* is represented as shown below in figure 3-25. The relation Quantity is used with all volume, weight, and distance measurements.

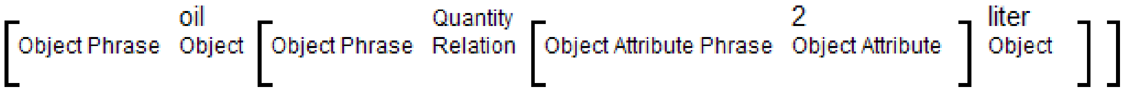


Figure 3-25 Semantic Representation of *two liters of oil*

- Region of Authority: *king of Babylon* is represented as shown below in figure 3-26.

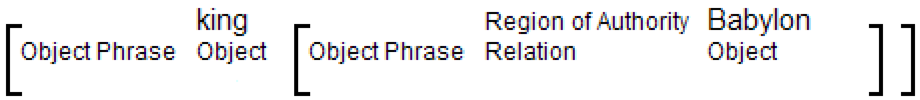


Figure 3-26 Semantic Representation of *king of Babylon*

- Part-Whole: *the door of the house* is represented as shown below in figure 3-27.

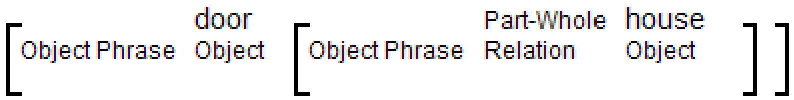


Figure 3-27 Semantic Representation of *the door of the house*

- Nationality: *a Hebrew slave* is represented as shown below in figure 3-28.

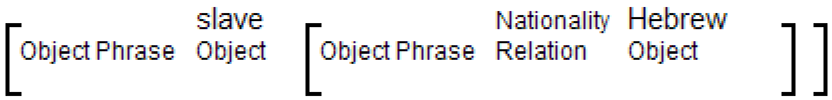


Figure 3-28 Semantic Representation of *Hebrew slave*

- Generic Genitive: All other object-object combinations in the semantic representations are marked simply with the relation 'Generic Genitive'. Each language will have its own methods of encoding each of these combinations, and certainly many of the noun-noun combinations that are permitted in English will not be permitted in other languages. Those combinations will have to be treated case by case in the transfer grammar. For example, in English it is perfectly acceptable to say something like *John's work was better than Bill's work*, but other languages may need to convert these particular object-object combinations to relative clauses such as *The work that John did was better than the work that Bill did*. A special section of the transfer grammar deals with these particular issues, and that section will be described in the next chapter. The semantic representation for *John's work* is shown below in figure 3-29.

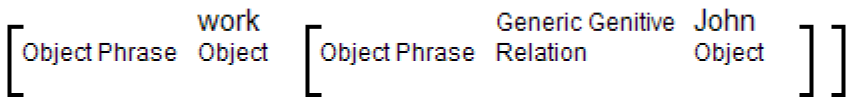
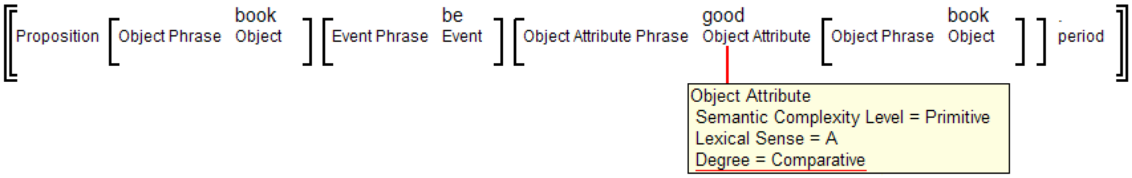


Figure 3-29 Semantic Representation of *John's work*

3.3.3.8 Comparisons

In addition to the comparative event modifying proposition described above in section 3.3.3.1, there are three other comparative constructions allowed in this system:

- Comparative Object Attribute: *This book is better than that book.* The semantic representation for this proposition is shown below in figure 3-30.



Object Attribute
Semantic Complexity Level = Primitive
Lexical Sense = A
Degree = Comparative

Figure 3-30. Semantic Representation of *This book is better than that book.*

The popup in figure 3-30 shows that the Degree of GOOD is Comparative. The standard of comparison is embedded in the object attribute phrase that is used predicatively. Language specific rules must insert a marker (e.g., *than*) if one is required in comparative constructions.

- Comparison with MORE or LESS in an object phrase: Luke 6:40 *A student does not know more things than his teacher.* The semantic representation for this proposition is shown below in figure 3-31.

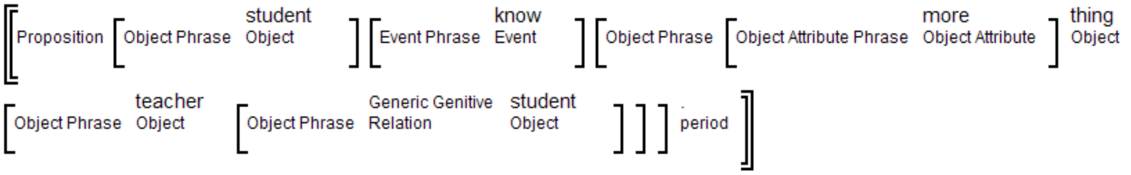
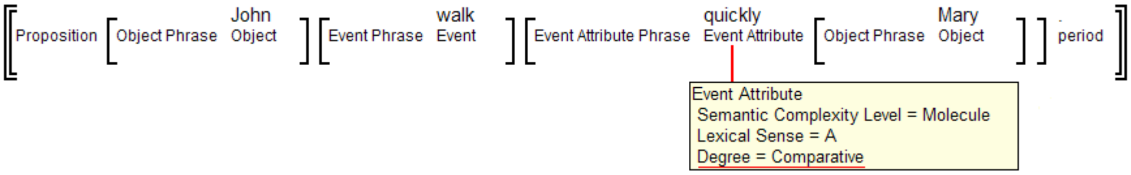


Figure 3-31. Semantic Representation of *A student does not know more things than his teacher.*

As seen above, the patient object is modified with the object attribute MORE, and the standard of comparison is in an object phrase which is embedded in the patient phrase.

- Comparative Event Attribute: *John walked more quickly than Mary.* The semantic representation for this proposition is shown below in figure 3-32.



Event Attribute
Semantic Complexity Level = Molecule
Lexical Sense = A
Degree = Comparative

Figure 3-32. Semantic Representation of *John walked more quickly than Mary.*

The popup in figure 3-32 shows that the Degree of QUICKLY is Comparative. The standard of comparison is embedded in the event attribute phrase.

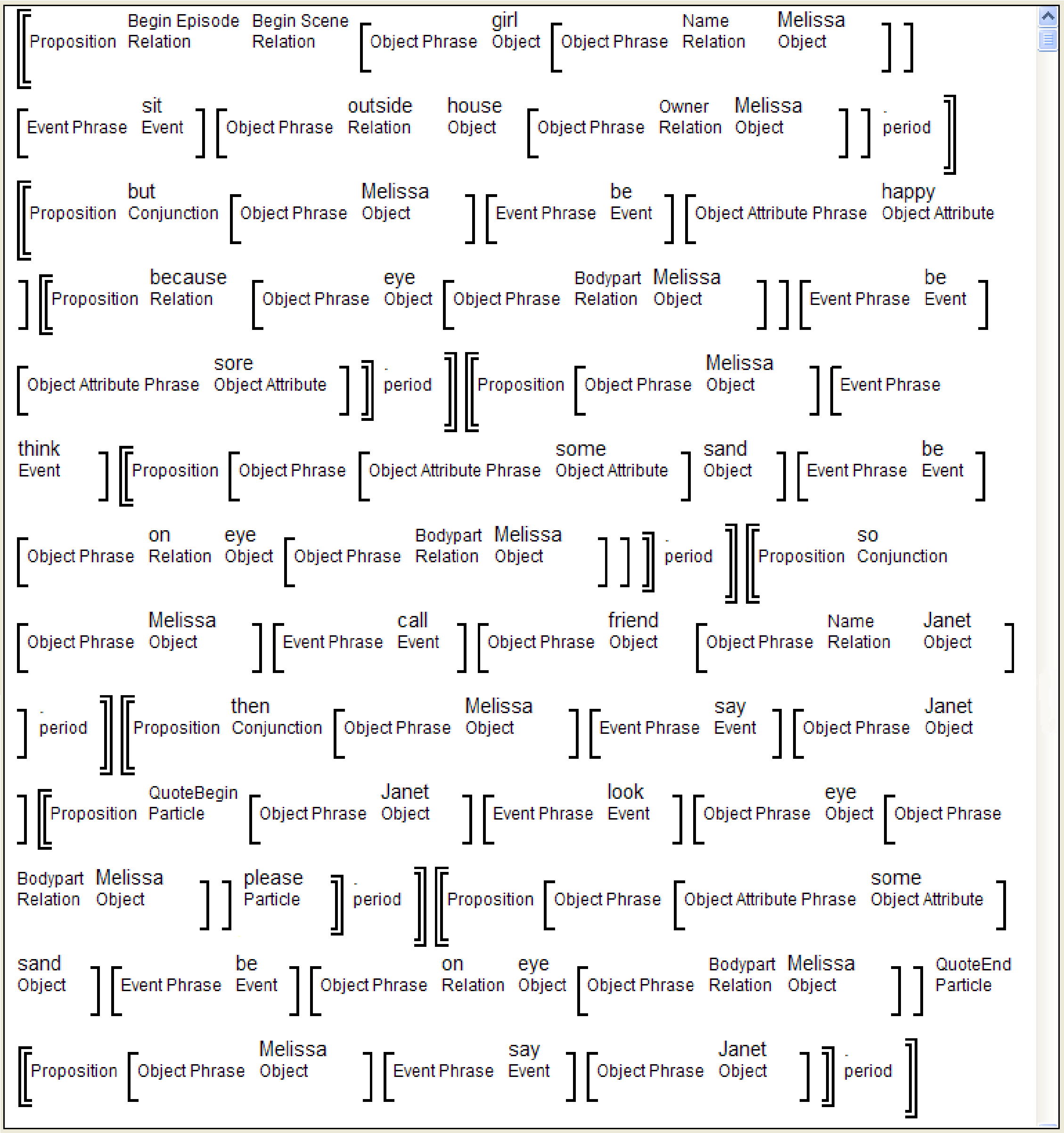
3.3.3.9 Semantic Representation of a Discourse

Using the concepts, features, and structures described above, semantic representations have been developed for many texts. A sample representation is shown on the following page in figure 3-33; this sample is the semantic representation for Infected Eye 1:2. A few comments describing this sample will be made here.

- Every episode begins with a semantic marker labeled “Begin Episode.” None of the test languages have linked this marker to a target equivalent. Every scene within an episode begins with the semantic maker “Begin Scene.” All of the test languages have mapped this particular marker to *one day*.
- The final subordinate proposition in figure 3-33 is a closing quotation proposition; those propositions were discussed in section 3.3.3.6 above. English and many other languages do not use those propositions, so they are automatically removed if the user activates the appropriate rule in the grammar.
- Prior to the direct quote in this verse, all the events have a Time value of Discourse except for the event BE in the patient proposition of THINK. The event BE in that patient proposition has a Time value of Present meaning that it is simultaneous with the matrix event. The English grammar generates a past tense form whenever a Time value of Discourse occurs, and another rule converts Present in the patient proposition to Discourse when the matrix event is Discourse. Those rules generate the past tense forms of the verbs in *Melissa thought that some sand was in her eyes.*

- The salience band of the first proposition is set to Backgrounded Action. Therefore the English grammar generates the imperfective form *was sitting*. The aspect of SIT is set to Unmarked.
- The Illocutionary Force of the last proposition is set to Yes/No Interrogative. Therefore the rules for English insert a CP node and move the verb *be* to that node. Then affix hopping rules convert *be* to *is.*
- There are no pronouns in the semantic representations; language specific rules must generate all the pronouns. The first occurrence of *Melissa* is realized in English with her proper name, but all subsequent references to Melissa in this verse are realized with pronouns which are generated by rules. Some of those rules look within a proposition (e.g., *One day Melissa was sitting outside <u>her</u> house.*), while other rules look to the preceding proposition (e.g., *But <u>she</u> was not happy …*). Before converting a noun to a pronoun, the rules must check to make sure that there is no intervening noun of the same number and gender.
- The propositions in the semantic representations are generally short and simple; each proposition contains only one event. When there are two or more propositions that do not contain subordinate propositions, language specific rules are able to combine them to form a longer proposition if that is appropriate. For example, the two propositions *So Melissa called a friend named Janet. Then Melissa said to Janet, …* are combined by a rule in the English grammar to produce *So she called a friend named Janet and said to her, …*
- The spatial relations in the semantic representations (e.g., IN, ON, ABOVE, BELOW, BESIDE, etc.) describe the spatial relationships between two objects. The sense of ON that is used in *some sand is on Melissa's eye* is ON-B which means that one object is on the surface of another object. However, English speakers do not talk about *sand* being *<u>on</u> an eye*, instead they talk about *sand* being *<u>in</u> an eye*. Therefore an English

collocation correction rule changes *sand on eye* to *sand in eye*. The collocation correction rules have proven very helpful and will be described thoroughly in the next chapter.



Begin Episode Begin Scene girl Name Melissa
[[Proposition Relation Relation [Object Phrase Object [Object Phrase Relation Object]]

sit outside house Owner Melissa .
[Event Phrase Event] [Object Phrase Relation Object [Object Phrase Relation Object]] period]]

but Melissa be happy
[[Proposition Conjunction [Object Phrase Object] [Event Phrase Event] [Object Attribute Phrase Object Attribute

because eye Bodypart Melissa be
] [[Proposition Relation [Object Phrase Object [Object Phrase Relation Object]] [Event Phrase Event]

sore . Melissa
[Object Attribute Phrase Object Attribute]] period]] [[Proposition [Object Phrase Object] [Event Phrase

think some sand be
Event] [[Proposition [Object Phrase [Object Attribute Phrase Object Attribute] Object] [Event Phrase Event]

on eye Bodypart Melissa . so
[Object Phrase Relation Object [Object Phrase Relation Object]]] period]] [[Proposition Conjunction

Melissa call friend Name Janet
[Object Phrase Object] [Event Phrase Event] [Object Phrase Object [Object Phrase Relation Object]

. then Melissa say Janet
] period]] [[Proposition Conjunction [Object Phrase Object] [Event Phrase Event] [Object Phrase Object

QuoteBegin Janet look eye
] [[Proposition Particle [Object Phrase Object] [Event Phrase Event] [Object Phrase Object [Object Phrase

Bodypart Melissa please . some
Relation Object]] Particle]] period]] [[Proposition [Object Phrase [Object Attribute Phrase Object Attribute]

sand be on eye Bodypart Melissa QuoteEnd
Object] [Event Phrase Event] [Object Phrase Relation Object [Object Phrase Relation Object]] Particle

Melissa say Janet .
[[Proposition [Object Phrase Object] [Event Phrase Event] [Object Phrase Object]] period]]

Figure 3-33. Semantic Representation from Infected Eye 1:2

Taking all of the above into consideration, the English grammar generates the following text for this verse: *One day a girl named Melissa was sitting outside her house. But Melissa was not* *happy because her eyes were very sore. She thought that some sand was in her eyes. So she called a friend named Janet and said to her, "Please look at my eyes. Is some sand in my eyes?"*

Using the apparatus described above, semantic representations have been developed for one hundred and five chapters of text[17]: Luke 1-15, Ruth 1-4, Esther 1-10, Daniel 1-12, Nahum 1-3, Genesis 1-50, Melissa'e Eye 1, Avian Influenza 1-5, and Kande's Story 1-5. Although this apparatus is unable to convey some of the more subtle nuances of meaning, it does successfully capture the vast majority of the meaning of a wide range of texts in a convenient machine tractable form. This method of building semantic representations is called the "rich interlingua" approach, and is similar to what Yorick Wilks calls 'Common Sense Semantics' (Rosner and Johnson 1992:262). Wilks' position has been summarized as follows: "there must be an understandable connection between the formal language of the representation and the world the language is used to describe, and that, ultimately, only natural language can serve the purpose of elucidating this connection" (Rosner and Johnson 1992:296).

3.4 Conclusions

As was stated in chapter 2, this dissertation has two goals: 1) determine the information that must be included in TTA's semantic representational system, and 2) determine the capabilities required of TTA's grammatical apparatus. This chapter described the information that is included in TTA's semantic representational system. This chapter began with a brief introduction to the fundamental philosophical issue associated with the representation of meaning. Then this chapter presented six semantic systems that were potential candidates to serve as TTA's semantic representational system. However, each of those six systems was

[17] Stephen Beale, who is a research associate professor in the Computational Linguistics Department at the University of Maryland in Baltimore, developed approximately one half of these semantic representations. Richard Denton, who is a research professor in the Physics Department at Dartmouth, has also contributed significantly to the development of these semantic representations.

found either unsuitable or impractical. Therefore it was decided that a new semantic representational system must be developed for TTA, and this chapter presented the apparatus that was developed specifically for TTA's semantic representational system. This chapter presented TTA's ontology, the feature system, and the structures that are permitted in TTA's semantic representations. Then this chapter concluded with an example showing how the semantic representational apparatus is applied to the opening paragraph of a short text that describes how to prevent eye infections.

The next chapter will describe the apparatus that was developed to generate target text from these semantic representations. That apparatus consists of a target lexicon, a transfer grammar, and a synthesizing grammar. The target lexicon and the two grammars are responsible for generating target language text that is easily understandable, grammatically correct, and semantically equivalent to the semantic representations.

CHAPTER 4

THE GENERATION OF SURFACE STRUCTURE:

THE LEXICON AND TWO GRAMMARS IN THE TRANSLATOR'S ASSISTANT

4.1 Introduction

This chapter will provide an overview of the three components within The Translator's Assistant that are responsible for generating target language text. These three components are 1) the target lexicon, 2) the transfer grammar, and 3) the synthesizing grammar. TTA's target lexicon will be described in section 4.2. The target lexicon enables linguists to enter their target stems, and define features and forms for those stems. Linguists are also able to write rules in the lexicon to generate the various lexical forms, and those rules will be presented below. TTA's transfer grammar will be described in section 4.3. The transfer grammar is responsible for performing the transfer step of the translation process. Therefore the transfer grammar consists of rules that adjust the semantic representations into new underlying representations that are appropriate for a particular target language[18]. The various types of rules in the transfer grammar will be listed and discussed, and a model of the transfer grammar will be presented. Finally the synthesizing grammar will be described in section 4.4. As indicated by the name, the synthesizing grammar is responsible for synthesizing the final surface forms of the target text. Each of the rule types in the synthesizing grammar will be discussed and illustrated. This chapter will conclude with an illustration of the complete generation process. The semantic

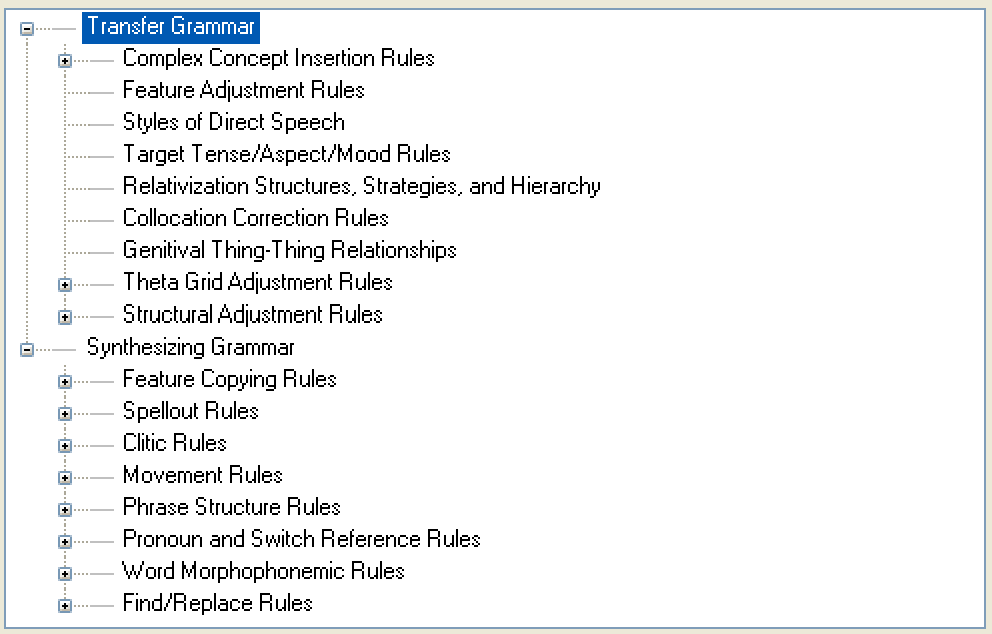
[18] It is important to note that the transfer grammar does not perform the types of transformations that were originally proposed by Chomsky. If a linguist chooses to develop a transformational-generative type grammar, the rules which execute the Chomskian transformations would probably be placed in the synthesizing grammar. For example, English questions are often viewed as being generated by transformations that are applied to declarative propositions. The rules which perform the necessary movement for English questions belong in the synthesizing grammar, and the process of generating English questions from declaratives will be discussed in the next chapter.

representation for Infected Eye 1:2, which was displayed at the end of chapter 3, will be shown again, but this time the Korean lexicon and grammars will be applied to that semantic representation. The generated target text for that passage will be shown for both English and Korean.

All multilingual natural language generators that use semantic representations for their source must perform two steps of the translation process: transfer and synthesis. In other words, the semantic representations that were described in the previous chapter are fed into the transfer grammar, and the transfer grammar makes the necessary adjustments so that the result is an appropriate underlying representation for a particular target language. Then the new underlying representations are fed into the synthesizing grammar, and the synthesizing grammar produces the target surface forms. In order to accomplish these two tasks, TTA has two grammars: a transfer grammar and synthesizing grammar. For the past several decades linguists have discussed in great detail the synthesizing grammars of languages[19]. However, they do not generally discuss transfer grammars because they are primarily interested in language description rather than translation. When a linguist writes a grammar for a language, that grammar generally describes the synthesizing process, and the linguist assumes that an appropriate underlying representation containing the language's structures, lexemes, features, and worldview is already present in the speaker's mind. For example, when a linguist describes a clause chaining language, it is assumed that the deep structure representation has clause chains. Similarly when a linguist describes a co-ranking language, it is assumed that the underlying representation has coordinated clauses. When a linguist describes a language that uses relative tense, it is assumed that a relative tense system is already present in the speaker's underlying representation, and when linguists describe a language that uses absolute tense, they assume an absolute tense system is present in the underlying representation.

[19] Thomas Payne provides a list of "successful" descriptive grammars in appendix 2 of *Describing Morphosyntax* (Payne 1997:372).

However, when an NLG generates texts in a target language, the process of converting the semantic representations to a new underlying representation appropriate for that target language must be described in complete detail. That is the task of the transfer grammar. Because the semantic representations use the same structures that are employed by English, the propositions in the semantic representations are *coordinated* rather than *chained*. In order to generate text in a clause chaining language, clause chains must be built, and that is one of the responsibilities of the transfer grammar. Similarly, the semantic representations use concepts that have been lexicalized by English. For example, English speakers can say things like, "*John weighed the box. The box weighed ten pounds.*" The first occurrence of *weigh* corresponds to WEIGH-B in TTA's ontology, and the second occurrence of *weigh* corresponds to WEIGH-A in TTA's ontology. Korean does not have a verb that corresponds to either of those senses of *weigh*. The Korean equivalent of those two sentences is "John measured the weight of the box. The weight of the box was ten pounds." Therefore the transfer grammar must restructure the semantic representations so that they contain the target language's structures and lexemes. Other common tasks performed by the transfer grammar include generating grammatical relations from semantic roles, performing theta grid adjustments, converting the semantic representation's TAM system to the target language's TAM system, combining propositions where appropriate, resolving collocational clash, inserting semantically complex concepts where appropriate, etc. After the transfer grammar has restructured the semantic representations into a new underlying representation that is appropriate for the target language, the synthesizing grammar is able to produce the final surface forms. TTA's synthesizing grammar has been designed to closely resemble the descriptive grammars that linguists routinely write. Therefore the synthesizing grammar includes phrase structure rules, spellout rules, morphophonemic rules, feature copying rules, etc. This chapter will discuss the target lexicon and these two grammars with each of their types of rules. TTA's grammar tree with all of its types of rules is shown below in figure 4-1.



- Transfer Grammar
  - Complex Concept Insertion Rules
  - Feature Adjustment Rules
  - Styles of Direct Speech
  - Target Tense/Aspect/Mood Rules
  - Relativization Structures, Strategies, and Hierarchy
  - Collocation Correction Rules
  - Genitival Thing-Thing Relationships
  - Theta Grid Adjustment Rules
  - Structural Adjustment Rules
- Synthesizing Grammar
  - Feature Copying Rules
  - Spellout Rules
  - Clitic Rules
  - Movement Rules
  - Phrase Structure Rules
  - Pronoun and Switch Reference Rules
  - Word Morphophonemic Rules
  - Find/Replace Rules

Figure 4-1. The Transfer and Synthesizing Grammars in TTA

The figure above shows that there are two main sections in TTA's grammar: a transfer grammar and a synthesizing grammar. The transfer grammar has nine different types of rules, and the synthesizing grammar has eight types of rules. The capabilities of these seventeen rule types will be presented below. However, the target lexicon must be described first because the rules in the grammar access the lexical features and forms that the linguist defines in his lexicon for his particular target language.

4.2 The Target Lexicon

This section will discuss TTA's target lexicon. The target lexicon serves as a repository for the target language stems. The target lexicon also enables linguists to define the features and forms that are pertinent to their target languages. The features will be described in section 4.2.1, and the forms will be described in section 4.2.2.

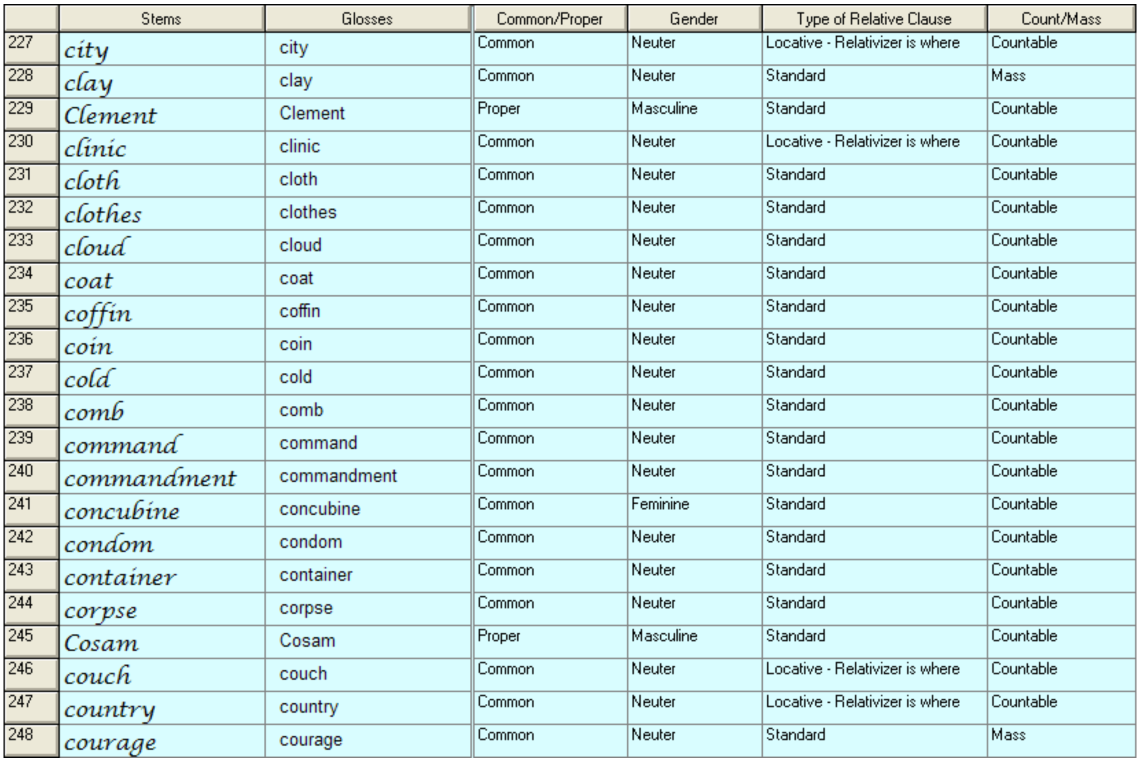
The target lexicon has seven predefined syntactic categories: 1) Nouns, 2) Verbs, 3) Adjectives, 4) Adverbs, 5) Adpositions, 6) Conjunctions, and 7) Particles. For these predefined syntactic categories, linguists are able to define the features and forms that are relevant to their stems. Not all languages will have lexemes in each of these seven categories, but the categories are available for languages that need them. These seven categories were selected for three reasons: 1) lexemes in these categories generally have semantic content as opposed to grammatical content, 2) the lexemes in these categories frequently require features and forms, and 3) these categories correspond to the seven semantic categories in the ontology (objects, events, object attributes, event attributes, relations, conjunctions, and particles). Linguists are also able to define additional syntactic categories that are pertinent to their languages; common examples include Complementizers, Relativizers, Articles, Demonstratives, etc.

4.2.1 Target Language Lexical Features

Target language lexical features serve to subcategorize the stems in a particular syntactic category in any way that is significant for the target language. Linguists are able to define and enter any feature values that are pertinent to their languages. Common features for nouns include Gender, Animacy, Countable, etc. A set of English nouns and their features are shown below in figure 4-2. As seen in that figure, English nouns require the following four features:

- *Common-Proper* – proper nouns don't take the definite article *the*.
- *Gender* – used when generating the singular pronouns *he, she,* and *it.*
- *Type of Relative Clause* – the generic relativizer is *that* (e.g., *The book that I read last week is very interesting.*), but locative nouns require the relativizer *where* (e.g., *The hospital where I was born is in Oregon.*), and temporal nouns require the relativizer *when* (e.g., *The year when astronauts first walked on the moon was 1969.*)
- *Count/Mass* – mass nouns are treated as singular even when they're plurl (e.g., *This wheat is good. *These wheats are good*).

The nouns in the table shown below have these four features set to the appropriate value for each particular noun.



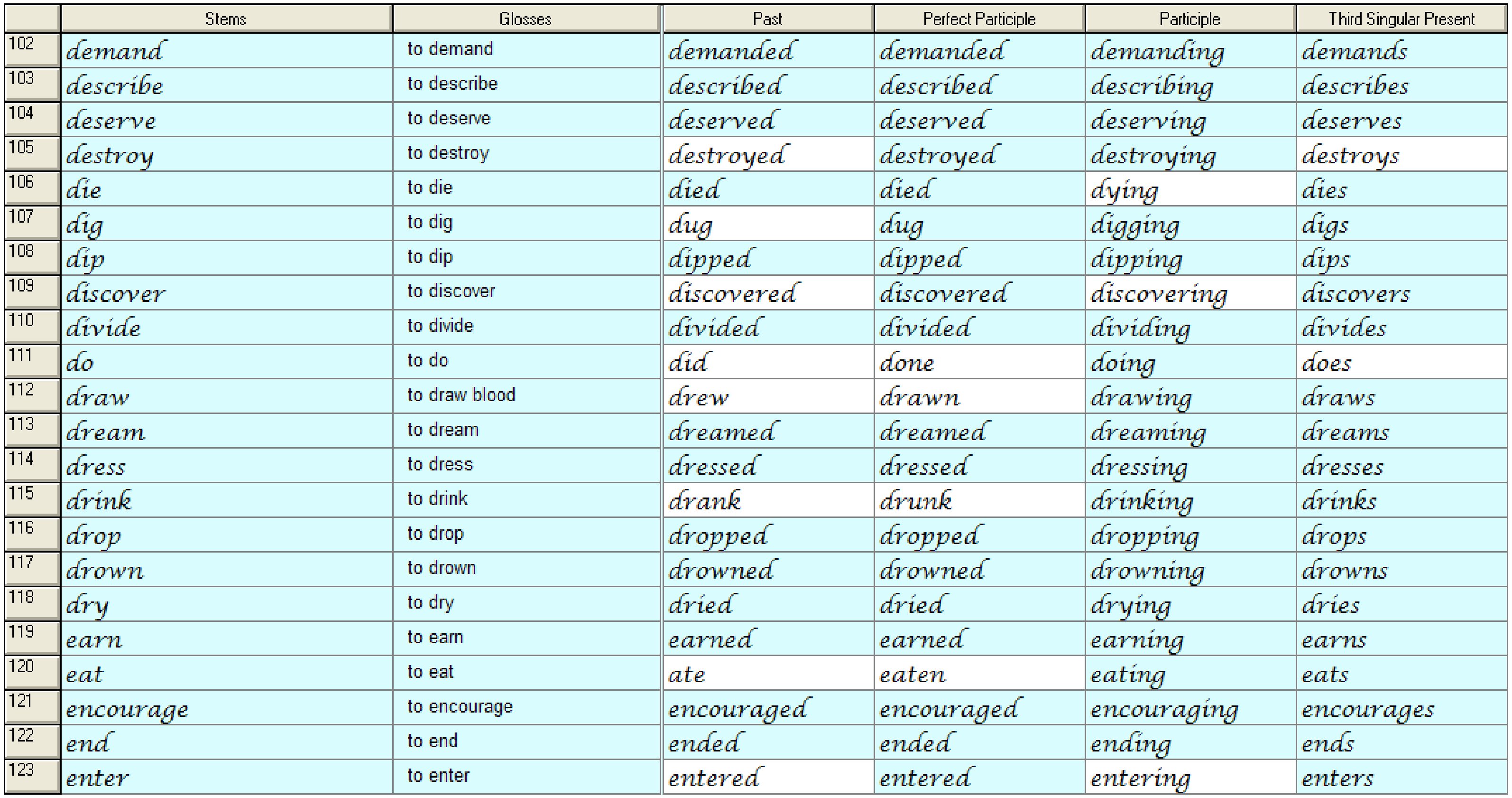
	Stems	Glosses	Common/Proper	Gender	Type of Relative Clause	Count/Mass
227	city	city	Common	Neuter	Locative - Relativizer is where	Countable
228	clay	clay	Common	Neuter	Standard	Mass
229	Clement	Clement	Proper	Masculine	Standard	Countable
230	clinic	clinic	Common	Neuter	Locative - Relativizer is where	Countable
231	cloth	cloth	Common	Neuter	Standard	Countable
232	clothes	clothes	Common	Neuter	Standard	Countable
233	cloud	cloud	Common	Neuter	Standard	Countable
234	coat	coat	Common	Neuter	Standard	Countable
235	coffin	coffin	Common	Neuter	Standard	Countable
236	coin	coin	Common	Neuter	Standard	Countable
237	cold	cold	Common	Neuter	Standard	Countable
238	comb	comb	Common	Neuter	Standard	Countable
239	command	command	Common	Neuter	Standard	Countable
240	commandment	commandment	Common	Neuter	Standard	Countable
241	concubine	concubine	Common	Feminine	Standard	Countable
242	condom	condom	Common	Neuter	Standard	Countable
243	container	container	Common	Neuter	Standard	Countable
244	corpse	corpse	Common	Neuter	Standard	Countable
245	Cosam	Cosam	Proper	Masculine	Standard	Countable
246	couch	couch	Common	Neuter	Locative - Relativizer is where	Countable
247	country	country	Common	Neuter	Locative - Relativizer is where	Countable
248	courage	courage	Common	Neuter	Standard	Mass

Figure 4-2. A Set of English Nouns with their Features

The features that are defined for each syntactic category in the lexicon will be available in the subsequent grammatical rules. For example, many languages add a plural morpheme only to the nouns that are animate. Therefore an animacy feature may be added to the target nouns, and each noun's value set appropriately. Then a rule in the synthesizing grammar will inspect a noun's animacy value and add the plural morpheme only to those nouns that are marked as animate.

4.2.2 Target Language Lexical Forms

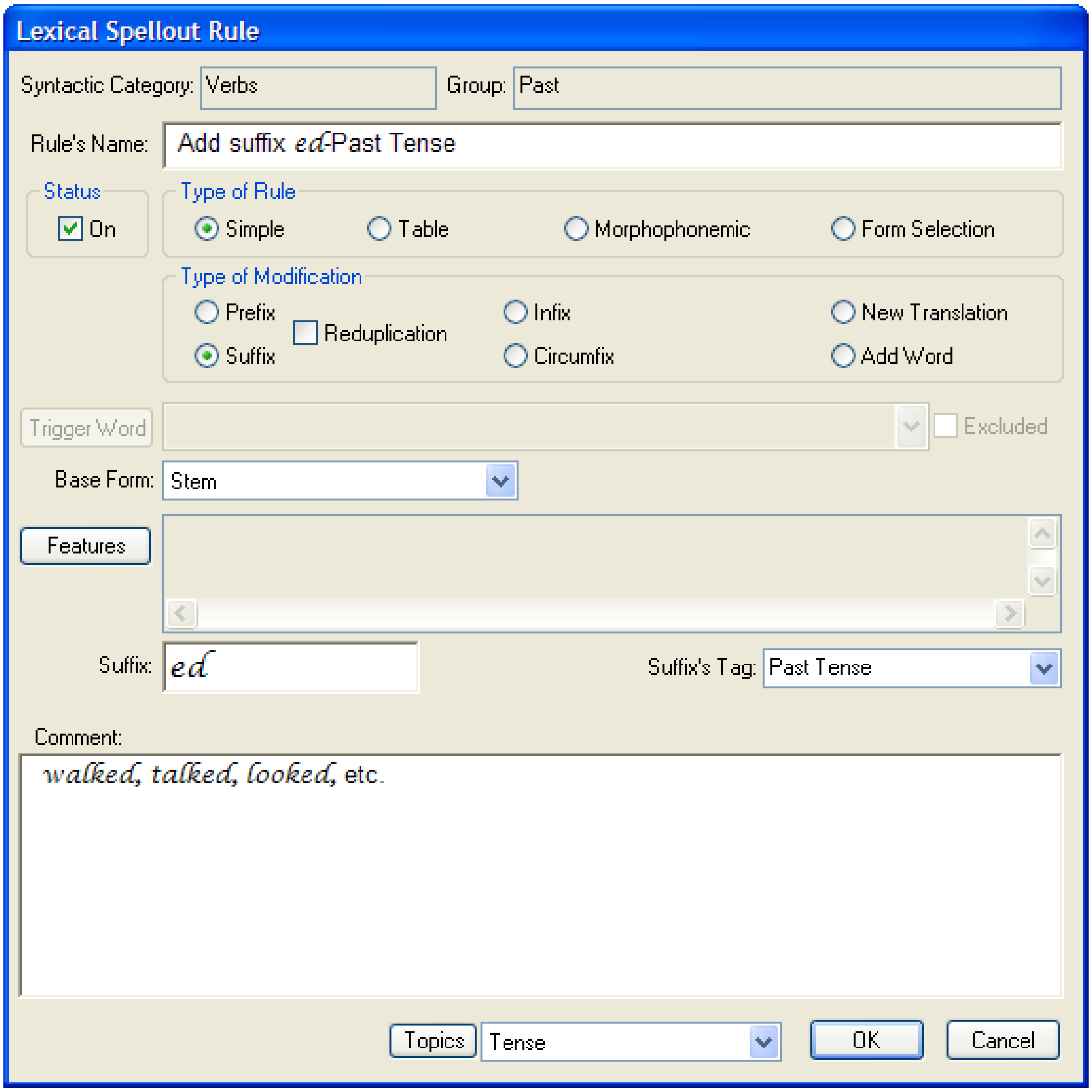
After linguists have entered several target stems into a syntactic category and defined the features that are relevant to those stems, they are able to specify the pertinent forms. Forms always consist of stems that are modified in some way such as a prefix, suffix, circumfix, etc. For example, English nouns have a singular form and a plural form. English verbs have a past tense form, a perfect participle form, a participle, and a third singular present tense form. A sample of verbs in the English lexicon with their forms is shown below in figure 4-3.



	Stems	Glosses	Past	Perfect Participle	Participle	Third Singular Present
102	demand	to demand	demanded	demanded	demanding	demands
103	describe	to describe	described	described	describing	describes
104	deserve	to deserve	deserved	deserved	deserving	deserves
105	destroy	to destroy	destroyed	destroyed	destroying	destroys
106	die	to die	died	died	dying	dies
107	dig	to dig	dug	dug	digging	digs
108	dip	to dip	dipped	dipped	dipping	dips
109	discover	to discover	discovered	discovered	discovering	discovers
110	divide	to divide	divided	divided	dividing	divides
111	do	to do	did	done	doing	does
112	draw	to draw blood	drew	drawn	drawing	draws
113	dream	to dream	dreamed	dreamed	dreaming	dreams
114	dress	to dress	dressed	dressed	dressing	dresses
115	drink	to drink	drank	drunk	drinking	drinks
116	drop	to drop	dropped	dropped	dropping	drops
117	drown	to drown	drowned	drowned	drowning	drowns
118	dry	to dry	dried	dried	drying	dries
119	earn	to earn	earned	earned	earning	earns
120	eat	to eat	ate	eaten	eating	eats
121	encourage	to encourage	encouraged	encouraged	encouraging	encourages
122	end	to end	ended	ended	ending	ends
123	enter	to enter	entered	entered	entering	enters

Figure 4-3. A Set of English Verbs with their Forms

When a form of a particular stem is suppletive and cannot be generated by rules, the linguist must enter the form into the lexicon. These irregular forms are indicated with a white background as seen above in figure 4-3. The forms that are regular and generated by rules are displayed with a blue background. Linguists are able to write lexical spellout rules in order to generate the various forms. These lexical spellout rules corrrespond to the word formation rules proposed in Chomsky's lexicalist hypothesis (Chomsky 1970), but they are less powerful than Chomsky's word formation rules. The lexical spellout rules in TTA are only able to add inflectional affixes; they are not able to add derivational affixes. For example, the past tense form of most English verbs consists of the stem modified by the suffix *–ed* (e.g., *walked, blinked, climbed, weighed,* etc.). Therefore a lexical spellout rule is used to generate the past tense form by adding the suffix *–ed* to the stem. This lexical spellout rule is shown below in figure 4-4.


Lexical Spellout Rule

Syntactic Category: Verbs

Group: Past

Rule's Name: Add suffix *ed*-Past Tense

Status: ☑ On

Type of Rule: ● Simple ○ Table ○ Morphophonemic ○ Form Selection

Type of Modification: ○ Prefix ● Suffix □ Reduplication ○ Infix ○ Circumfix ○ New Translation ○ Add Word

Trigger Word □ Excluded

Base Form: Stem

Suffix: *ed*

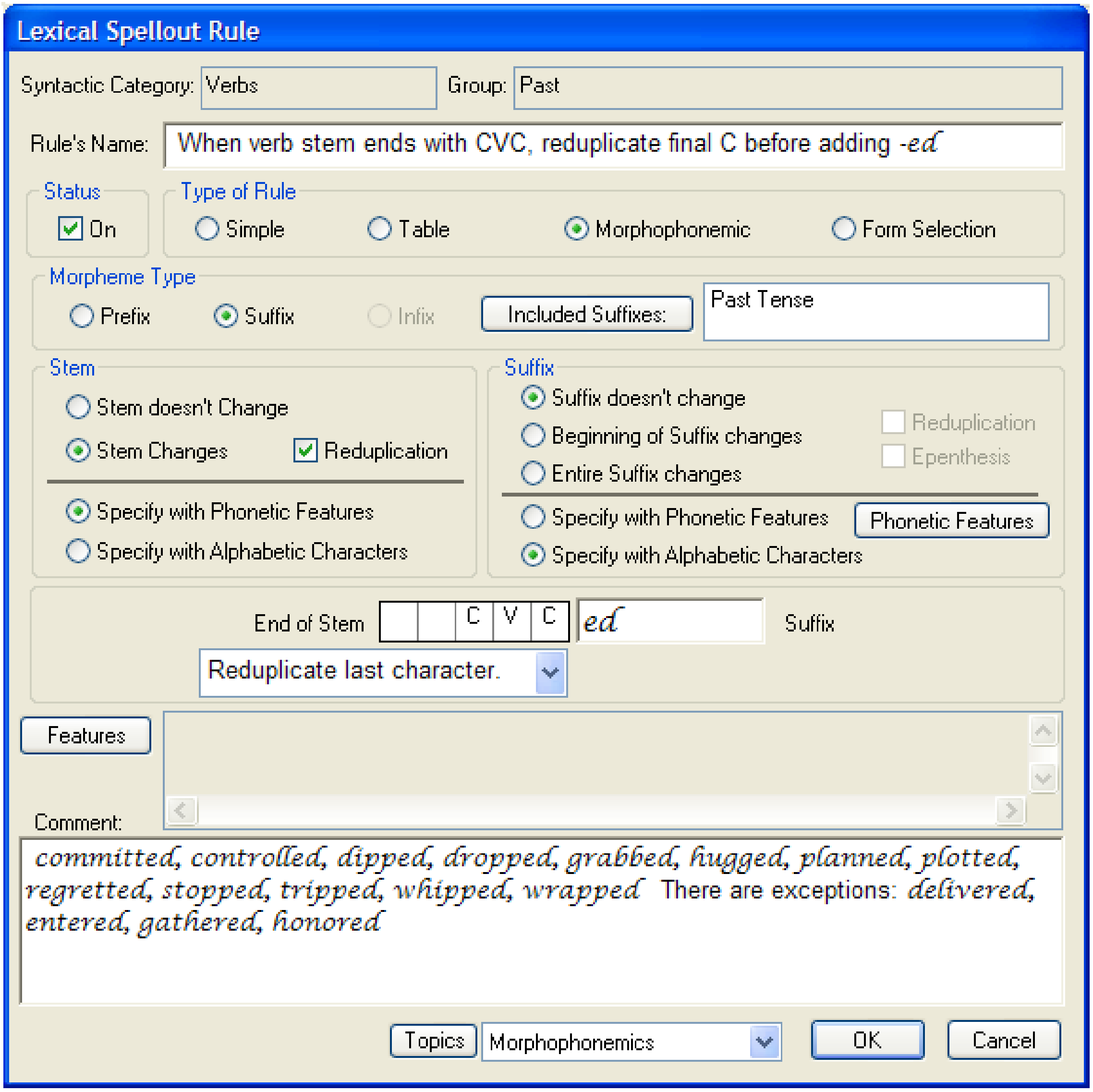
Suffix's Tag: Past Tense

Comment: *walked, talked, looked,* etc.

Topics: Tense

Figure 4-4. A Lexical Spellout Rule that adds the suffix –*ed* to English Verb Stems

As seen near the top of figure 4-4 above, there are four different types of lexical spellout rules: Simple, Table, Morphophonemic, and Form Selection. The rule template used for lexical spellout rules is identical to the rule template used for the standard spellout rules in the synthesizing grammar. Therefore these four different types of spellout rules will be discussed in the synthesis grammar section below. However, there are no morphophonemic rules for either of the test languages in the spellout section, so an English lexical morphophonemic rule is shown below in figure 4-5.



Lexical Spellout Rule

Syntactic Category: Verbs

Group: Past

Rule's Name: When verb stem ends with CVC, reduplicate final C before adding -*ed*

Status: ☑ On

Type of Rule: ○ Simple ○ Table ● Morphophonemic ○ Form Selection

Morpheme Type: ○ Prefix ● Suffix ○ Infix

Included Suffixes: Past Tense

Stem: ○ Stem doesn't Change ● Stem Changes ☑ Reduplication

● Specify with Phonetic Features ○ Specify with Alphabetic Characters

Suffix: ● Suffix doesn't change ○ Beginning of Suffix changes ○ Entire Suffix changes □ Reduplication □ Epenthesis

○ Specify with Phonetic Features ● Specify with Alphabetic Characters

End of Stem: C V C *ed* Suffix

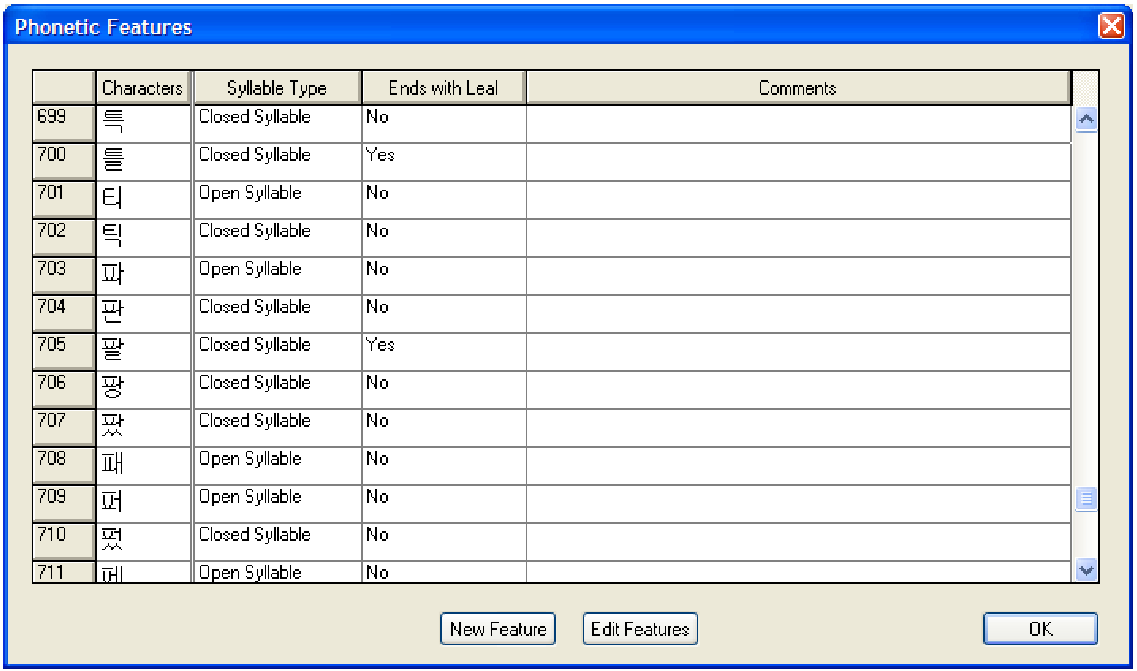
Reduplicate last character.

Comment: *committed, controlled, dipped, dropped, grabbed, hugged, planned, plotted, regretted, stopped, tripped, whipped, wrapped* There are exceptions: *delivered, entered, gathered, honored*

Topics: Morphophonemics

Figure 4-5. An English Lexical Morphophonemic Rule

The rule shown above applies when generating the past tense form of English verbs that end with CVC. For those verbs, the final consonant is reduplicated before adding the past tense suffix –*ed* (e.g., *controlled, grabbed, hugged, stopped*, etc.). In morphophonemic rules like the one shown above, linguists are able to specify the stem and affix using either alphabetic characters or phonetic features. Linguists using TTA are able to define the phonetic features that are relevant to their particular target language. Similar to the lexical features that were described in section 4.2.1 above, linguists are able to define and add any phonetic features and feature values that are pertinent to their language. Shown below in figure 4-6 is the phonetic feature dialog for Korean. Because each glyph in a Korean font represents a syllable rather than an individual character, the phonetic features that were defined for the Korean grammar are quite different from the phonetic features defined for the English grammar.



	Characters	Syllable Type	Ends with Leal	Comments
699	특	Closed Syllable	No	
700	틀	Closed Syllable	Yes	
701	티	Open Syllable	No	
702	틱	Closed Syllable	No	
703	파	Open Syllable	No	
704	판	Closed Syllable	No	
705	팔	Closed Syllable	Yes	
706	팡	Closed Syllable	No	
707	팠	Closed Syllable	No	
708	패	Open Syllable	No	
709	퍼	Open Syllable	No	
710	펐	Closed Syllable	No	
711	페	Open Syllable	No	

Figure 4-6. Phonetic Features for Korean[20]

Whether or not a particular syllable is open or closed is very significant in Korean morphophonemic operations as will be illustrated lated in this chapter. Therefore that feature was used extensively throughout the Korean grammar.

Lexical forms may be generated either by the lexical spellout rules or by the spellout rules in the synthesizing grammar. During the development of the grammars for this project, a guideline for this issue emerged as follows: If there are no irregularities in a particular target form, then that form can be generated by a spellout rule in the synthesizing grammar. In those cases, lexical forms are not necessary. For example, the Korean plural morpheme is –들 [deul].

[20] The third column in this table is entitled “Ends with Leal”. The word “leal” refers to the Korean character ‘ㄹ’.

Generally the animate nouns are much more likely to be marked with the plural morpheme[21], so a feature was added to the Korean nouns indicating their animacy. When an animate noun is marked as plural, it is always suffixed with –들 [deul]; there are no exceptions or irregularities to this pattern. Therefore, a plural form for Korean nouns was not needed in the lexicon, and a spellout rule in the synthesizing grammar adds the plural morpheme –들 [deul] to all of the animate nouns that have a Number value of Plural. Opposed to this are English plural nouns. English plural nouns have many irregularities. English noun stems are generally modified by the suffix –*s* in order to generate the plural form, and if there were no exceptions to this pattern, then a plural form would not be necessary in the English lexicon. However, since there are many suppletive English plural nouns such as *man/men, person/people, foot/feet, deer/deer,* etc., a plural form is required in the English lexicon. During this project it was found that Korean does not have any suppletive forms anywhere; the surface form for every word can be generated by rules. Therefore no lexical forms were needed in the Korean lexicon; all the surface forms are generated by the spellout rules in the synthesizing grammar. Since the spellout rules are used much more extensively in the synthesizing grammar than they are in the target lexicon, they will be more thoroughly described in section 4.4.2 below.

4.3 The Transfer Grammar

This section will describe the rules that are in TTA's transfer grammar. The transfer grammar has nine different types of rules, and each rule type will be discussed and illustrated. As was mentioned above, the transfer grammar is responsible for performing the transfer step of the translation process. Therefore the transfer grammar adjusts the English influenced semantic representations so that they become new underlying representations that are

[21] The rules that determine when the Korean plural morpheme –들[deul] should be inserted are quite complex, and context must be taken into consideration. Because the goal of this project is to generate texts that are easily understandable, grammatically correct, semantically equivalent to the source documents, and at approximately a sixth grade reading level, a simplified set of rules was adopted for inserting the Korean plural morpheme. For the texts that have been generated to date, these rules have worked well. Undoubtedly as more texts are generated in Korean, these rules will need to be refined.

appropriate for the target language. When the transfer grammar was designed, the fundamental question that had to be answered was as follows: What capabilities must the transfer grammar possess in order to transform the semantic representations into new underlying representations that are appropriate for a very wide variety of languages? Several of the most common requirements of the transfer grammar were listed above; these requirements include generating grammatical relations from semantic roles, performing theta grid adjustments, etc. However, many more adjustments are certainly required. When designing the transfer grammar, each of the tasks had to be identified, and then a type of rule had to be designed to perform each task. A model of the resulting transfer grammar is shown below in figure 4-7.

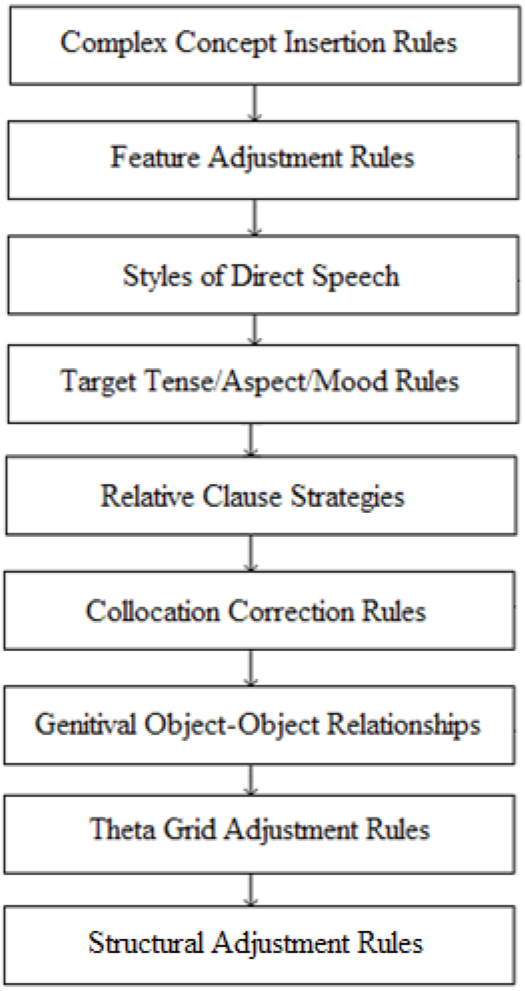
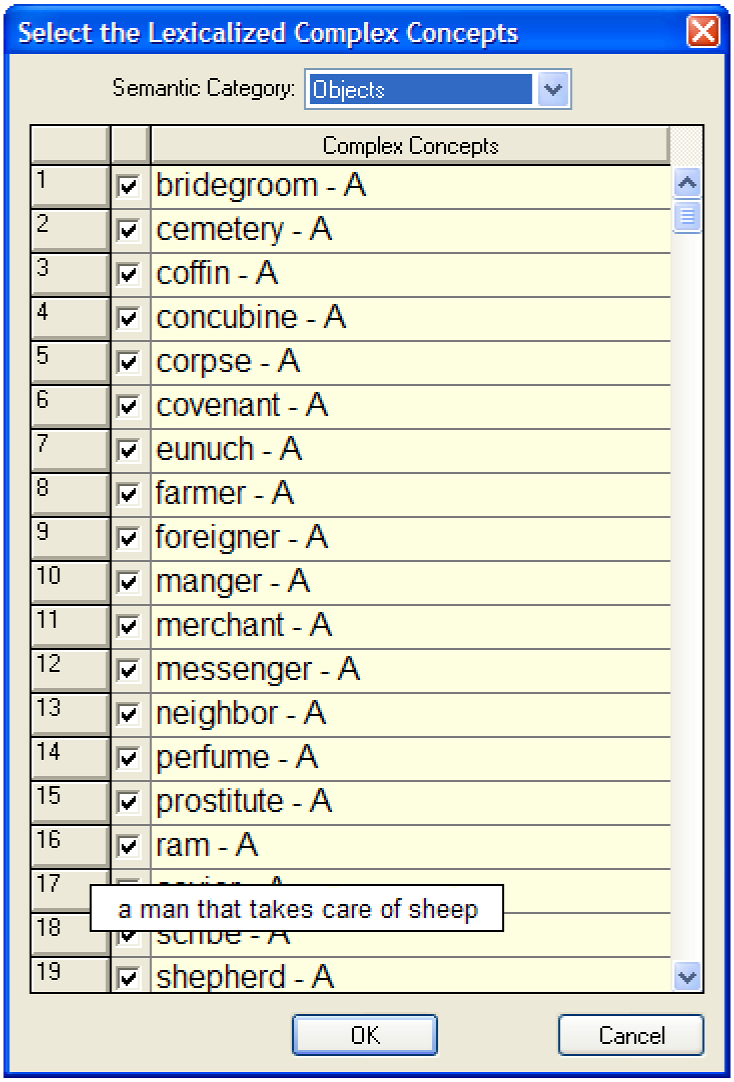

Complex Concept Insertion Rules
Feature Adjustment Rules
Styles of Direct Speech
Target Tense/Aspect/Mood Rules
Relative Clause Strategies
Collocation Correction Rules
Genitival Object-Object Relationships
Theta Grid Adjustment Rules
Structural Adjustment Rules

Figure 4-7. The Transfer Grammar in The Translator's Assistant

As seen in figure 4-7 above, there are nine different types of rules in the transfer grammar. The arrows in the figure indicate the sequence in which the various rule types are executed. The rules are executed from top to bottom, so the Complex Concept Insertion rules are executed first, then the Feature Adjustment rules, etc. Each of these nine rule types will be described below.

4.3.1 Complex Concept Insertion Rules

Section 3.3 in the previous chapter discussed the composition of the semantic representations. The semantic representations consist of semantically simple concepts in order to increase the probability that other target languages will have good lexical equivalents. However, if the generated texts consist of only very simple words, a problem arises because the texts become monotonously long and the intended message is not faithfully communicated. Section 3.3.1 discussed the situation with *veterinarian* as found in the text that describes how to prevent the spread of Avian Influenza. That section also discussed the concepts SCOLD and SCREAM, both of which are semantically complex and inserted into the semantic representations only if the user activates the associated complex concept insertion rules. Users of TTA are able to activate complex concept insertion rules by checking the appropriate boxes as shown below in figure 4-8.



Select the Lexicalized Complex Concepts

Semantic Category: Objects

		Complex Concepts
1	☑	bridegroom - A
2	☑	cemetery - A
3	☑	coffin - A
4	☑	concubine - A
5	☑	corpse - A
6	☑	covenant - A
7	☑	eunuch - A
8	☑	farmer - A
9	☑	foreigner - A
10	☑	manger - A
11	☑	merchant - A
12	☑	messenger - A
13	☑	neighbor - A
14	☑	perfume - A
15	☑	prostitute - A
16	☑	ram - A
17	☑	savior - A
18	☑	scribe - A
19	☑	shepherd - A

a man that takes care of sheep

Figure 4-8. TTA's Complex Concept Activation Dialog

The figure above shows some of the semantically complex objects that have prewritten *Complex Concept Insertion rules*. The concept in the last row of the dialog is SHEPHERD, and the popup shows how that concept has been explicated in the semantic representations: a man that takes care of sheep. SHEPHERD is a semantically complex concept, and most languages probably will not have a lexical equivalent. The languages that do not have a lexical equivalent will not activate the complex concept insertion rule for SHEPHERD. As was mentioned in section 3.3.1, often the complex concepts are paired with semantic molecules or primitives in the semantic representations. That is the case with SHEPHERD; it usually occurs in the semantic representations as man/shepherd for reasons that will be presented below. The semantically complex objects, events, and object attributes are each handled in different ways in the semantic representations, so they will each be discussed separately.

4.3.1.1 Semantically Complex Objects

Most semantically complex objects that occur in the source texts have been explicated for this project by using a relative clause to modify an object that is a semantic primitive or molecule. For example, SHEPHERD is a semantically complex object, so it has been explicated in the semantic representations as “man that takes care of sheep.” Therefore, when *shepherd* occurs in a source text, it is usually replaced with MAN-A [MAN-A CARE-A SHEEP-A] in the semantic representations. However, when a semantically complex object occurs repeatedly in a source text, a problem arises if the explication is used to replace every occurrence of that concept. For example, *shepherd* occurs repeatedly in the source text for Luke 2:8-10:

> Luke 2:8 *That night some shepherds were in a field that was near Bethlehem. Those shepherds were protecting their sheep.*
> Luke 2:9 *Suddenly an angel appeared in front of those shepherds. The Lord's glory was shining on those shepherds. So those shepherds became very afraid.*
> Luke 2:10 *But the angel said to the shepherds, "Do not be afraid. ..."*

Because SHEPHERD is semantically complex, it cannot appear directly in the semantic representations. However, when developing the semantic representation for a particular source text, it is not a simple matter of substituting “man that takes care of sheep” for each occurrence of “shepherd”. If each occurrence of "shepherd" in the source text were replaced with “man that takes care of sheep”, the semantic representations for Luke 2:8-10 would become:

> Luke 2:8 That night some men who take care of sheep were in a field that was near Bethlehem. Those men who take care of sheep were protecting them.
> Luke 2:9 Suddenly an angel appeared in front of those men who take care of sheep. The Lord's glory was shining on those men who take care of sheep. So those men who take care of sheep became very afraid.
> Luke 2:10 But the angel said to the men who take care of sheep, "Do not be afraid. ..."

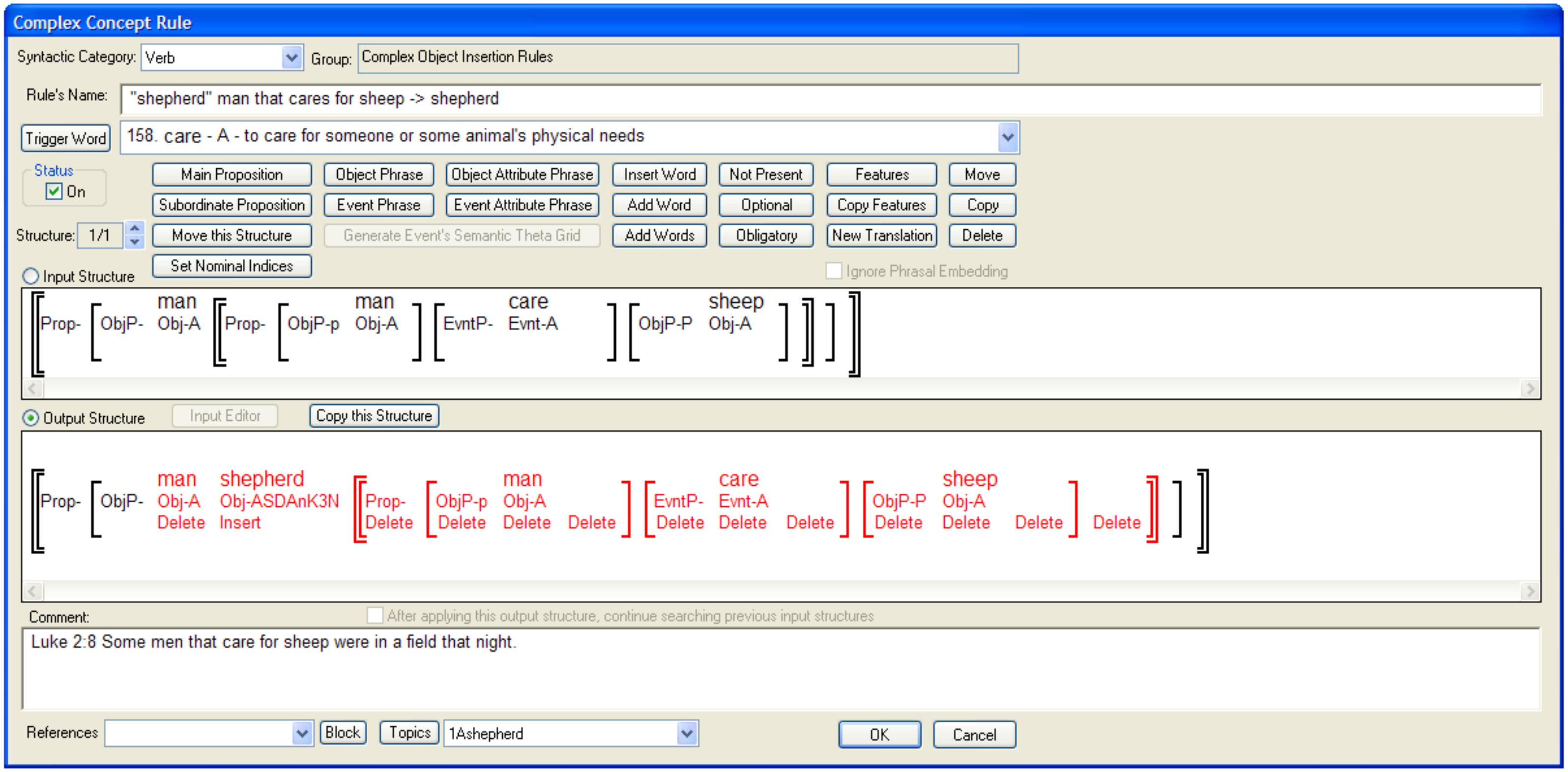
By constantly repeating the relative clause "who take care of sheep," the message becomes distorted and the text becomes excessively repetitive. Therefore, when a particular complex object such as SHEPHERD appears repeatedly in a passage, the first occurrence of the concept is explicated, in this case with “man who takes care of sheep”. Then the subsequent occurrences of the complex concept in the source text are combined with a semantic molecule or primitive. In the case with SHEPHERD, it is paired with the semantic primitive MAN. Therefore the semantic representations for Luke 2:8-10 are as follows:

> Luke 2:8 That night some men who take care of sheep were in a field that was near Bethlehem. Those men/shepherds were protecting their sheep.
> Luke 2:9 Suddenly an angel appeared in front of those men/shepherds. The Lord's glory was shining on those men/shepherds. So those men/shepherds became very afraid.
> Luke 2:10 But the angel said to the men/shepherds, "Do not be afraid. ..."

The first occurrence of *shepherd* in the source text was explicated as seen in Luke 2:8 above. But then all the subsequent occurrences of *shepherd* in the source text were changed to MAN/SHEPHERD. When TTA’s grammar is executed, wherever it finds MAN/SHEPHERD in the semantic representation, it will check to see if the complex concept insertion rule for SHEPHERD has been activated. If that rule is active, TTA will remove MAN and leave SHEPHERD in the semantic representation. If the rule is not active, TTA will remove SHEPHERD and leave MAN in the semantic representation. Similarly when TTA’s grammar is executing the complex concept insertion rules, if the rule for SHEPHERD has been activated, then all occurrences of “man who takes care of sheep” in the semantic representations will be replaced with SHEPHERD. Therefore, if the user does not activate the complex concept insertion rule for SHEPHERD, these verses become:

> Luke 2:8 That night some men who were taking care of sheep were in a field that was near Bethlehem. Those men were protecting their sheep.
> Luke 2:9 Suddenly an angel appeared in front of those men. The Lord's glory was shining on those men. So those men became very afraid.
> Luke 2:10 But the angel said to the men, "Do not be afraid. ..."

For languages that do not have a lexical equivalent for SHEPHERD, the text above is easily understandable, and it is not distorted by constantly repeating the relative clause “that take care of sheep”. The actual rule that replaces MAN-A [MAN-A CARE-A SHEEP-A] in the semantic representations with SHEPHERD-A is shown below in figure 4-9. That rule consists primarily of an input structure and output structure. If the rule is activated, the input structure searches for occurrences of MAN-A modified by the relative clause MAN-A CARE-A SHEEP-A. Wherever that structure is found in the semantic representations, the output structure will be applied. The output structure of this rule indicates that the concept MAN-A must be deleted, then the concept SHEPHERD-A must be inserted into the semantic representation, and finally the relative clause MAN-A CARE-A SHEEP-A must be deleted. This rule is fairly complex, but it is prewritten for the user. The user only needs to activate or deactivate this rule by checking or unchecking the box shown above in figure 4-8.



Complex Concept Rule

Syntactic Category: Verb

Group: Complex Object Insertion Rules

Rule's Name: "shepherd" man that cares for sheep -> shepherd

158. care - A - to care for someone or some animal's physical needs

Status ☑ On

Structure: 1/1

□ Ignore Phrasal Embedding

○ Input Structure

[Prop- [ObjP- man Obj-A [Prop- [ObjP-p man Obj-A] [EvntP- care Evnt-A] [ObjP-P sheep Obj-A]]]]

● Output Structure

[Prop- [ObjP- man Obj-A Delete shepherd Obj-ASDAnK3N Insert [Prop- Delete [ObjP-p Delete man Obj-A Delete Delete] [EvntP- Delete care Evnt-A Delete Delete] [ObjP-P Delete sheep Obj-A Delete Delete] Delete]]]

Comment:

□ After applying this output structure, continue searching previous input structures

Luke 2:8 Some men that care for sheep were in a field that night.

References

Topics: 1Ashepherd

Figure 4-9. Complex Concept Insertion Rule for SHEPHERD

The two primary components in the rule shown above are the input structure and the output structure. When the rule is executed, TTA will search for places in the semantic representations that match the structure specified in the rule's input structure. Whenever a match is found, the semantic representations will be changed according to the specifications of the output structure. The output structure highlights in red all the constituents that will be changed by the rule. In this case, MAN-A will be deleted, and the relative clause that modifies MAN-A will also be deleted. Then the object SHEPHERD-A will be inserted into the semantic representation. When SHEPHERD-A is inserted into the text, it will copy all the feature values that were associated with MAN-A. Therefore, if MAN-A had been tagged with a Number value of Plural, SHEPHERD-A will also be tagged as Plural. All of the buttons above the input structure enable linguists to build the input and output structures. Linguists are able to insert object phrases, event phrases, propositions, concepts, etc., into the input and output structures. They are also able to set features on each of the constituents in the input structure, and they are able to change the feature values of the constituents in the output structure. All of the rules in TTA follow this general pattern: linguists construct structures with features and concepts, and when those structures are found in the semantic representations, the rule's output structure is applied.

4.3.1.2 Semantically Complex Events

Semantically complex events which occur in the source texts must be replaced in the semantic representations with a semantically simple event along with the necessary arguments and modifiers. For example, all occurrences of the semantically complex event TO SIGN (e.g., *John signed the letter.*) in the source texts are replaced with "X write X's name on Y" in the semantic representations. Unlike the complex objects which are paired with a semantic primitive or molecule (e.g., MAN/SHEPHERD), every occurrence of a complex event in a source text must be replaced with the explication in the semantic representation. The complex concept insertion rule that searches the semantic representations for occurrences of "X write X's name on/in Y" is shown below in figure 4-10. The popup below IN-H in the input structure shows that the rule applies when the relation is either IN-H ‘to be in something that is written such as a book or letter’ or ON-C ‘generic location on something’. The popup above the final object phrase in the output structure shows that the semantic role of that phrase will be changed to ‘Most Patient-like’. If the linguist activates this rule, it will change propositions such as *John wrote his name on the letter* to *John signed the letter*.

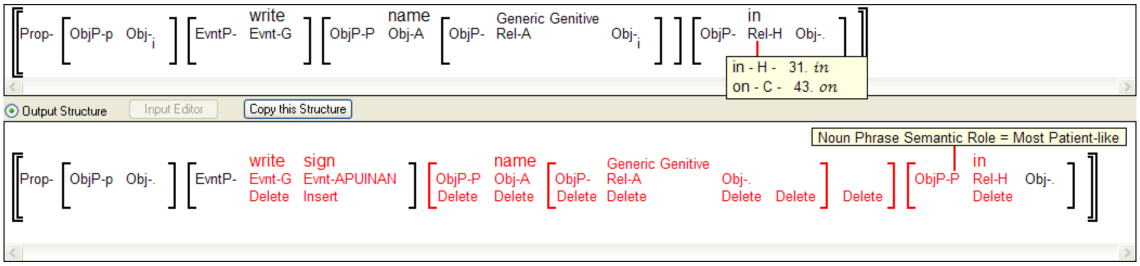


Figure 4-10. Complex Concept Insertion Rule for TO SIGN

The rule shown above in figure 4-10 is somewhat similar to the rule for MAN/SHEPHERD. As seen in the output structure, this rule deletes the event WRITE, the Patient NP which contains NAME, and the relation IN/ON from the final NP. Then it inserts the complex event SIGN, and all the features that were on WRITE are copied to SIGN. Therefore, if WRITE had a Time value of ‘Discourse’, SIGN will also have a Time value of ‘Discourse’. The semantic role of the final NP in the output structure is changed to ‘Most Patient-like’, thereby completing the transformation of the simple event WRITE to the complex event SIGN.

4.3.1.3 Semantically Complex Object Attributes

When a semantically complex object attribute occurs in a source text, it is generally converted to a relative clause in the semantic representations. If a complex object attribute occurs multiple times in a particular passage, the first occurrence will be explicated with a relative clause, but the subsequent occurrences will be allowed to remain in the semantic representations. For example, the semantically complex object attribute "paralyzed" is explicated as "X that is not able to move X's legs.[22]" This complex concept occurs repeatedly in the source text for Luke 5:18:

> Luke 5:18 *Some men decided to bring a paralyzed man to Jesus. So these men put the paralyzed man on a mat. And the men carried the paralyzed man to the house where Jesus was teaching people. These men tried to carry the paralyzed man into the house so that Jesus could see the paralyzed man.*

If every occurrence of *paralyzed* in the source text were explicated as “X that is not able to move X’s legs”, the semantic representation for this verse would become:

> Luke 5:18 Some men decided to bring a man that was not able to move his legs to Jesus. So these men put the man that was not able to move his legs on a mat. And the men carried the man that was not able to move his legs to the house where Jesus was teaching people. These men tried to carry the man that was not able to move his legs into the house so that Jesus could see the man that was not able to move his legs.

By constantly repeating the relative clause *that is not able to move his legs,* the message becomes distorted. Therefore, the approach that was adopted for dealing with semantically complex object attributes in the source texts is to explicate the first occurrence of the concept, and then for the subsequent occurrences, insert the complex object attribute into the semantic representation. Following this guideline, the semantic representation for Luke 5:18 becomes:

> Luke 5:18 Some men decided to bring a man that was not able to move his legs to Jesus. So these men put the paralyzed man on a mat. And the men carried the paralyzed man to the house where Jesus was teaching people. These men tried to carry the paralyzed man into the house so that Jesus could see the paralyzed man.

When the grammar in TTA is executed and a semantically complex object attribute is found in the semantic representations, the generator will check to see if the associated complex concept insertion rule has been activated. If the linguist activated the rule indicating that the target language has a lexical equivalent for the complex object attribute, then the complex object attribute will be allowed to remain in the semantic representation. If the rule is not active, the

[22] The word “paralyzed” is defined as “unable to move part or all of your body or feel it” in Longman’s Dictionary of Contemporary English, 2003, p. 1194. For the purposes of this project and for the source documents that have been analyzed to date, the explication presented above sufficiently conveys the message.

complex object attribute will be removed from the semantic representation. So if the user does not activate the complex concept insertion rule for PARALYZED, this verse will become:

> Luke 5:18 Some men decided to bring a man that was not able to move his legs to Jesus. So these men put the man on a mat. And the men carried the man to the house where Jesus was teaching people. These men tried to carry the man into the house so that Jesus could see the man.

This text is easily understandable and the message is not distorted by constantly repeating the relative clause *that is not able to move his legs*. The rule that looks for occurrences of “X be not able to move X’s legs” in the semantic representations, and then replaces the relative clause with the complex object attribute PARALYZED is similar to figures 4-9 and 4-10, so it will not be shown here.

Although TTA already has many prewritten complex concept insertion rules, additional rules will certainly be added in the future as more semantic representations are developed. Brief examples of several additional complex concept insertion rules are shown below:

Semantically complex objects:

- ‘man that will marry a woman soon’ -> ‘bridegroom’
- ‘place where people bury other people who died’ -> ‘cemetery’
- ‘person who grows crops’ -> ‘farmer’
- ‘person who sells things’ -> ‘merchant’
- ‘person who delivers messages’ -> ‘messenger’
- ‘oil that smells good’ -> ‘perfume’

Semantically complex events:

- ‘to die in water’ -> ‘drown’
- ‘to say false things’ -> ‘to lie’
- ‘to do all the things that someone tells you to do’ -> ‘to obey’
- ‘to plan to do bad things to a person’ -> ‘to plot’

Semantically complex object attributes:

- ‘person who is not able to hear things’ -> ‘deaf person’
- ‘person who has much money’ -> ‘rich person’

Certain complex concepts in the source texts have not been explicated because their explications have proven problematic. For example, animal names such as *camel, sheep, bear,* *lion,* etc., certainly are not lexical universals, but accurately explicating them using semantic primitives or molecules results in a long and distorted text. Certain artifacts are easily explicated such as *perfume* = 'oil that smells good'*, coffin* = 'box that people carry a dead body in'*, manger* = 'box that contains food for animals'*,* etc. However, accurately explicating other artifacts such as *cross, helmet, sandal,* etc., also distorts the texts. At this time no guidelines have been developed to determine which artifacts should be explicated, and which should not.

A few semantically complex events have been used in the semantic representations, but rather than explicating them, they have simply been paired with semantic primitives or molecules. For example, *to wonder* is semantically complex, but a succinct explication has not been developed. Therefore it is always paired with the semantic primitive THINK in the semantic representations as THINK-B/WONDER-A. If the target language has a lexical equivalent for WONDER, the linguist will check the associated box, and then whenever THINK-B/WONDER-A is found in the semantic representations, WONDER-A will be used and THINK-B will be discarded. If the linguist does not check the box for WONDER, then TTA will discard WONDER-A and use THINK-B. Similarly a succinct explication for *to pursue* has not been developed, so it is always paired in the semantic representations with *to follow* as FOLLOW-A/PURSUE-A.

In addition to the prewritten complex concept insertion rules, linguists are able to write their own rules to insert complex concepts that have been lexicalized by their target languages. For example, Korean has a word 우기 [u gi] which means 'rainy season,' so a complex concept insertion rule looks for occurrences of 'season of rain' in the semantic representations, and when that construction is found, the rule replaces 'season' with 우기 [u gi], and deletes the embedded phrase 'of rain.' Korean has another word 소경 [so gyeong] which means 'blind person.' So another complex concept insertion rule for Korean changes 'blind person' to 소경 [so gyeong].

4.3.2 Feature Adjustment Rules

The second type of rule in the transfer grammar is called *Feature Adjustment rules*. These rules allow a linguist to change the features in the semantic representations to new values that are appropriate for a specific target language. These rules make three specific types of adjustments to the feature system that was described in section 3.3.2:

- hide unused or irrelevant features,
- collapse feature values that are not morphologically distinguished in a language, and
- generate additional features from prewritten rules.

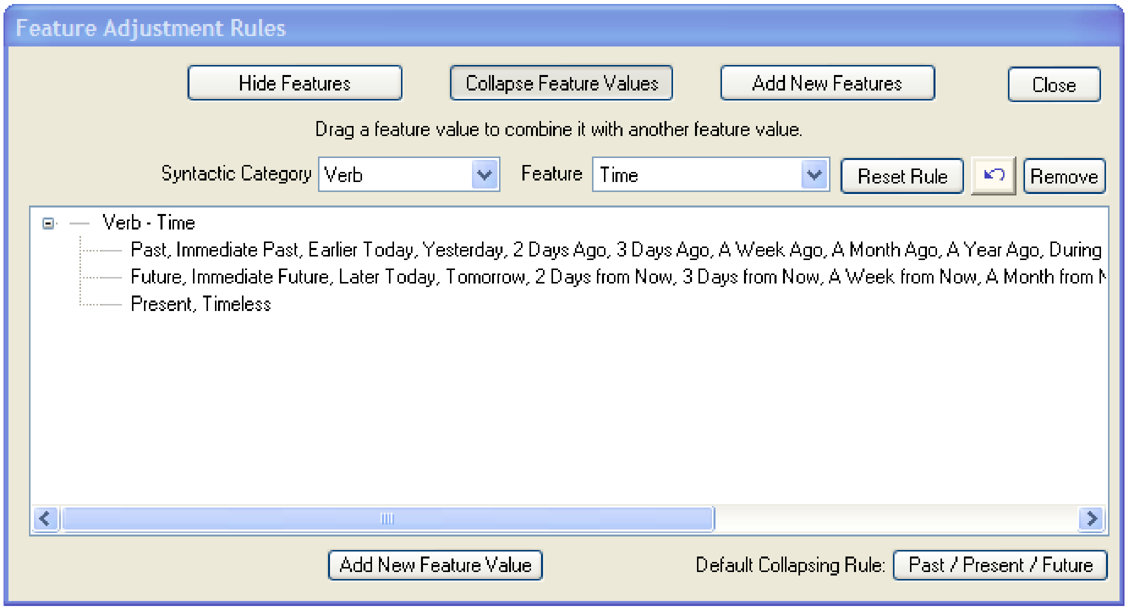
These three tasks will each be described.

4.3.2.1 Hide Unused or Irrelevant Features

The feature system that was described in the previous chapter includes features which are necessary for some languages, but irrelevant to other languages. For example, the five propositional features that describe each speech situation (e.g., Speaker, Listener, Speaker's Attitude, Speaker's Age, and Speaker to Listener's Age) are necessary for languages like Korean that include honorifics, but those features are unnecessary for languages like English that do not encode honorifics. Therefore, when a feature is irrelevant to a language, that feature can be hidden so that the linguist does not have to deal with it when developing his grammar. The dialog that enables linguists to hide features is very simple so it will not be shown here.

4.3.2.2 Collapse Unused Feature Values

As was mentioned in the previous chapter, many of the features include far more values than any particular language will use. For example, very few languages morphologically distinguish singular, dual, trial, quadrial, plural, and paucal. No language will distinguish all the past Time values and future Time values that are included in the semantic representations for the events. Therefore linguists using TTA are able to collapse the feature values so that only the values that are relevant to their language will be displayed in the semantic representations and in the grammar rules. Shown below in figure 4-11 is the rule for English that collapses all of the past Time values to a new value called Past, and all of the future Time values to a new value called Future. English encodes timeless events (e.g., *The sun rises in the east*.) with present tense, so the Timeless value has been collapsed with Present.



Drag a feature value to combine it with another feature value.

Syntactic Category Verb

Feature Time

Verb - Time
- Past, Immediate Past, Earlier Today, Yesterday, 2 Days Ago, 3 Days Ago, A Week Ago, A Month Ago, A Year Ago, During
- Future, Immediate Future, Later Today, Tomorrow, 2 Days from Now, 3 Days from Now, A Week from Now, A Month from N
- Present, Timeless

Default Collapsing Rule: Past / Present / Future

Figure 4-11. The Feature Adjustment Dialog – Collapse Features

By collapsing the values that are irrelevant to a language, the user is able to focus on the values that are significant to his language when writing the grammar rules. For example, if a target language only distinguishes singular and plural, then the rules in the synthesizing grammar should not have to deal with Dual, Trial, and Quadrial. Therefore the feature adjustment rules change the feature values that occur in the semantic representations to new values that are appropriate for the target language.

4.3.2.3. Add New Feature Values

There are many features which are common in the world's languages, but those features do not belong in the semantic representations. For example, perhaps all of the world's languages use grammatical relations to some degree, but semantic roles rather than grammatical relations belong in the semantic representations. Grammatical relations can easily be generated by a rule from the semantic roles, so a rule has been prewritten to do that. If the linguist wants grammatical relations to be added to the semantic representations, he simply activates the appropriate rule. That rule generates generic grammatical relations; if a language requires grammatical relations that are different from the ones generated by the rule, then the linguist is able to add the grammatical relations required by his language, and then write his own rule to generate them appropriately. Other features that are common and useful when writing a grammar can also be generated by rules and then added to the semantic representations. This section of the feature adjustment dialog allows users to activate prewritten rules which generate new features that are added to the semantic representations. When the linguist writes his grammar, he is then able to use these generated features. This dialog is very simple and will not be shown here.

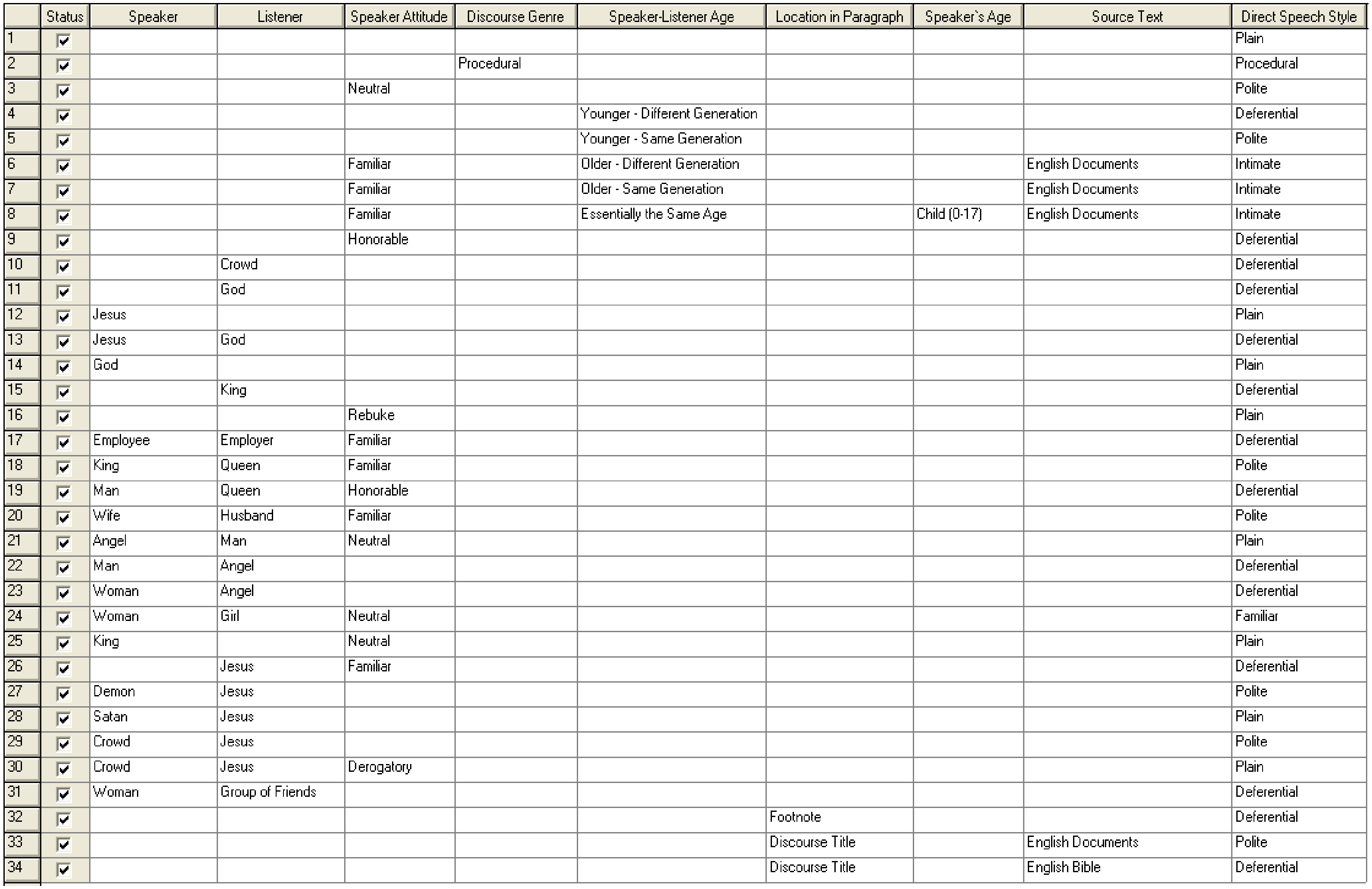
4.3.3 Styles of Direct Speech Rules

Many languages employ a variety of techniques to indicate relative status when two people talk to one another. Therefore, as was described in section 3.3.2.5, all propositions that are direct quotes in the semantic representations are tagged with five features indicating the general category of the speaker, the listener, the speaker's attitude, the speaker's approximate age, and the age of the speaker relative to the listener. These features may then be used to set another feature called Direct Speech Style. Linguists using TTA may add as many direct speech styles as are relevant to their target languages. For example, it is well documented that Korean has at least six styles of speech (Cho et al. 2000:9):

- deferential speech indicated by the verbal suffixes –습니다 [seup ni da]/ㅂ니다 [p ni da],
- polite speech indicated by the suffixes –어요 [eo yo]/아요 [a yo],
- blunt style indicated by suffixes –소 [so]/오 [o],
- familiar style indicated by the suffix –네 [ne],
- intimate speech indicated by the suffixes –어 [eo]/아 [a], and

- plain speech indicated by –는다 [neun da]/ㄴ다 [n da]/다 [da].

Shown on the following page in Figure 4-12 are the rules that generate the appropriate style of direct speech for Korean. As seen in that figure, there are columns for each of the five speech features: ‘Speaker’, ‘Listener’, ‘Speaker’s Attitude’, ‘Speaker-Listener Age’, and ‘Speaker’s Age’. The final column of each row sets the speech style appropriately for the situation described in that row by the earlier columns. The first row sets the default value to ‘Plain’ speech because that is the style used in most situations.



	Status	Speaker	Listener	Speaker Attitude	Discourse Genre	Speaker-Listener Age	Location in Paragraph	Speaker`s Age	Source Text	Direct Speech Style
1	☑									Plain
2	☑				Procedural					Procedural
3	☑			Neutral						Polite
4	☑					Younger - Different Generation				Deferential
5	☑					Younger - Same Generation				Polite
6	☑			Familiar		Older - Different Generation			English Documents	Intimate
7	☑			Familiar		Older - Same Generation			English Documents	Intimate
8	☑			Familiar		Essentially the Same Age		Child (0-17)	English Documents	Intimate
9	☑			Honorable						Deferential
10	☑		Crowd							Deferential
11	☑		God							Deferential
12	☑	Jesus								Plain
13	☑	Jesus	God							Deferential
14	☑	God								Plain
15	☑		King							Deferential
16	☑			Rebuke						Plain
17	☑	Employee	Employer	Familiar						Deferential
18	☑	King	Queen	Familiar						Polite
19	☑	Man	Queen	Honorable						Deferential
20	☑	Wife	Husband	Familiar						Polite
21	☑	Angel	Man	Neutral						Plain
22	☑	Man	Angel							Deferential
23	☑	Woman	Angel							Deferential
24	☑	Woman	Girl	Neutral						Familiar
25	☑	King		Neutral						Plain
26	☑		Jesus	Familiar						Deferential
27	☑	Demon	Jesus							Polite
28	☑	Satan	Jesus							Plain
29	☑	Crowd	Jesus							Polite
30	☑	Crowd	Jesus	Derogatory						Plain
31	☑	Woman	Group of Friends							Deferential
32	☑						Footnote			Deferential
33	☑						Discourse Title		English Documents	Polite
34	☑						Discourse Title		English Bible	Deferential

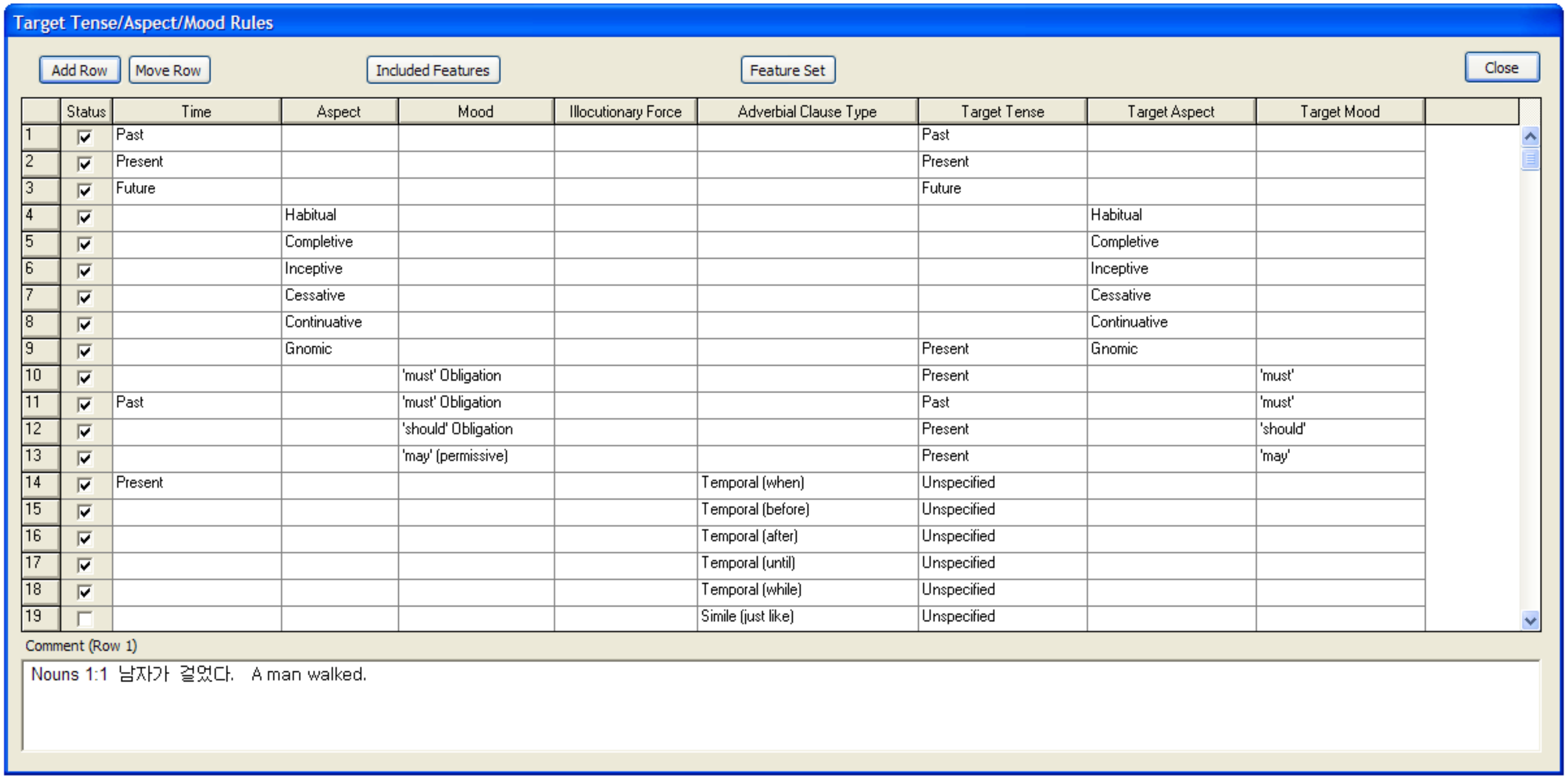
Figure 4-12. Styles of Direct Speech for Korean

When the rules shown above are executed, TTA begins with the first row and checks if the conditions specified in that row match the conditions in the semantic representation. If the conditions match, the style of direct speech specified in the last column of that row will be saved. Since no conditions are specified in the first row, the first row always matches every proposition in the semantic representations, so the 'Speech Style' will be set to 'Plain' which is the default value. Then the subsequent rows are examined. Whenever the conditions specified by a particular row match the conditions in the semantic representation, the direct speech value specified in the last column of that row will be saved. After all the rows have been examined, the last saved value will be added to the proposition's features in the semantic representation. Then subsequent rules in the synthesizing grammar will look at that value and add the appropriate morphology to signal each particular type of speech.

4.3.4 Target Tense, Aspect, Mood Rules

As was described in section 3.3.2.2, the TAM system employed in the semantic representations is not language specific. Most languages will have very different views of time, aspect, and mood, so their TAM systems will not necessarily correlate with TTA's TAM system. Therefore converting the TAM system of the semantic representations to the TAM system of the target language may not be completely possible. The TAM rules are responsible for converting as closely as possible the TAM system of the semantic representations to the TAM system of the target language. Shown on the following page in figure 4-13 is the Target Tense/Aspect/Mood Rule dialog for Korean. Notice that this dialog refers to Time values of Past, Present, and Future. This is because these rules are executed after the feature collapsing rules which were described above in section 4.3.2.2. Therefore this dialog only deals with the Time values that are pertinent to the target language rather than all the Time values that are possible in the semantic representations.

The execution of the target TAM rules is very similar to the execution of the Direct Speech Style rules. The conditions of the first row are examined to see if they match the conditions of the current proposition in the semantic representation. If the conditions match, then the values specified in the Target Tense, Target Aspect, and Target Mood columns are copied into the semantic representation. Then the next row in the rule is checked to see if its conditions match the semantic representation. Whenever a row's conditions match the conditions in the semantic representation, the Target Tense, Target Aspect, and Target Mood values specified by that row are copied into the semantic representation.



	Status	Time	Aspect	Mood	Illocutionary Force	Adverbial Clause Type	Target Tense	Target Aspect	Target Mood
1	☑	Past					Past		
2	☑	Present					Present		
3	☑	Future					Future		
4	☑		Habitual					Habitual	
5	☑		Completive					Completive	
6	☑		Inceptive					Inceptive	
7	☑		Cessative					Cessative	
8	☑		Continuative					Continuative	
9	☑		Gnomic				Present	Gnomic	
10	☑			'must' Obligation			Present		'must'
11	☑	Past		'must' Obligation			Past		'must'
12	☑			'should' Obligation			Present		'should'
13	☑			'may' (permissive)			Present		'may'
14	☑	Present				Temporal (when)	Unspecified		
15	☑					Temporal (before)	Unspecified		
16	☑					Temporal (after)	Unspecified		
17	☑					Temporal (until)	Unspecified		
18	☑					Temporal (while)	Unspecified		
19	☐					Simile (just like)	Unspecified		

Comment (Row 1)

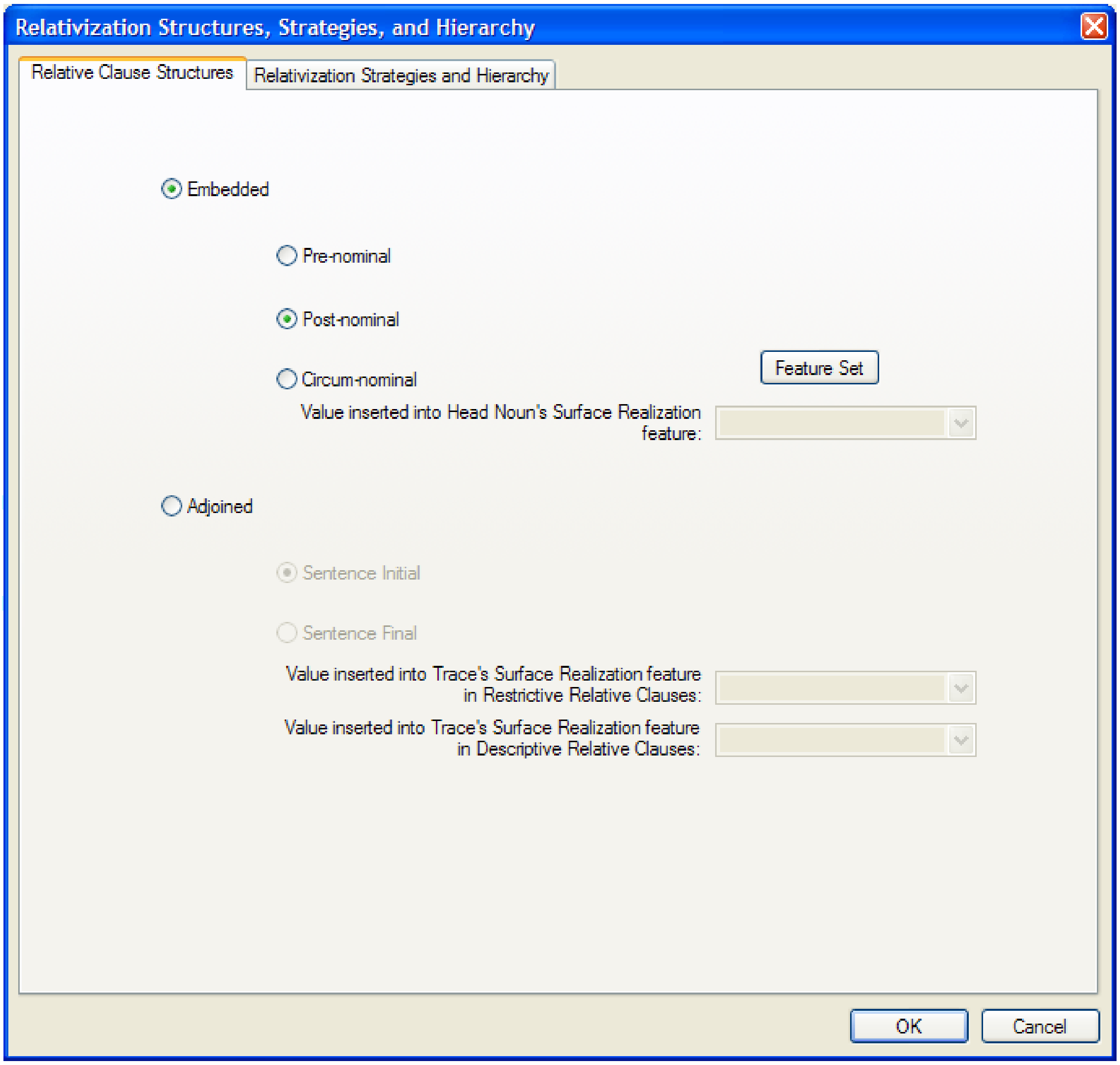
Nouns 1:1 남자가 걸었다. A man walked.

Figure 4-13. The Tense/Aspect/Mood Rule Dialog for Korean

In many languages certain types of adverbial clauses have non-finite verbs (e.g., *John worked extra hours in order to earn more money. John studied instead of watching TV.*). Similarly some moods have non-finite verbs (e.g., *John must go to the doctor.*). This table provides a convenient method of indicating when a target verb in a particular environment should be unspecified for tense. Linguists using TTA are able to enter whatever values are pertinent to their languages in the columns marked Target Tense, Target Aspect, and Target Mood. Linguists are also able to specify which features of the event and the proposition are relevant to determining the values for Target Tense, Target Aspect, and Target Mood. For example, in the semantic representations, imperatives are usually marked with a Time value of Immediate Future, but English imperatives use bare stem verbs. By including the Illocutionary Force feature in the TAM table, the rule provides a very convenient method of specifying that English imperatives are unmarked for tense. To whatever degree possible, these TAM rules transform the TAM system in the semantic representations into the TAM system of the target language. Subsequent rules then provide the appropriate morphology.

4.3.5 Relative Clause Structures, Strategies, and the Relativization Hierarchy

Extensive typological research has been done regarding relative clauses and the various strategies employed by the world's languages when constructing relative clauses (Comrie 1989:138, Givón 1990:645). Not all languages permit relative clauses, but the languages which do use relative clauses apply a limited number of strategies to a limited number of grammatical relations in what is commonly called the NP Accessibility Hierarchy (Keenan & Comrie 1977, Comrie 1989:156). Linguists have found that relative clauses may be either embedded or adjoined. If a language uses embedded relative clauses, then they may be pre-nominal, post-nominal, or circum-nominal. If a language uses adjoined relative clauses, then they are either sentence initial or sentence final. These options are all available in the Structures tab of TTA's relativization dialog which is shown below in figure 4-14.



Relativization Structures, Strategies, and Hierarchy

Relative Clause Structures | Relativization Strategies and Hierarchy

● Embedded

○ Pre-nominal

● Post-nominal

○ Circum-nominal

Value inserted into Head Noun's Surface Realization feature:

○ Adjoined

● Sentence Initial

○ Sentence Final

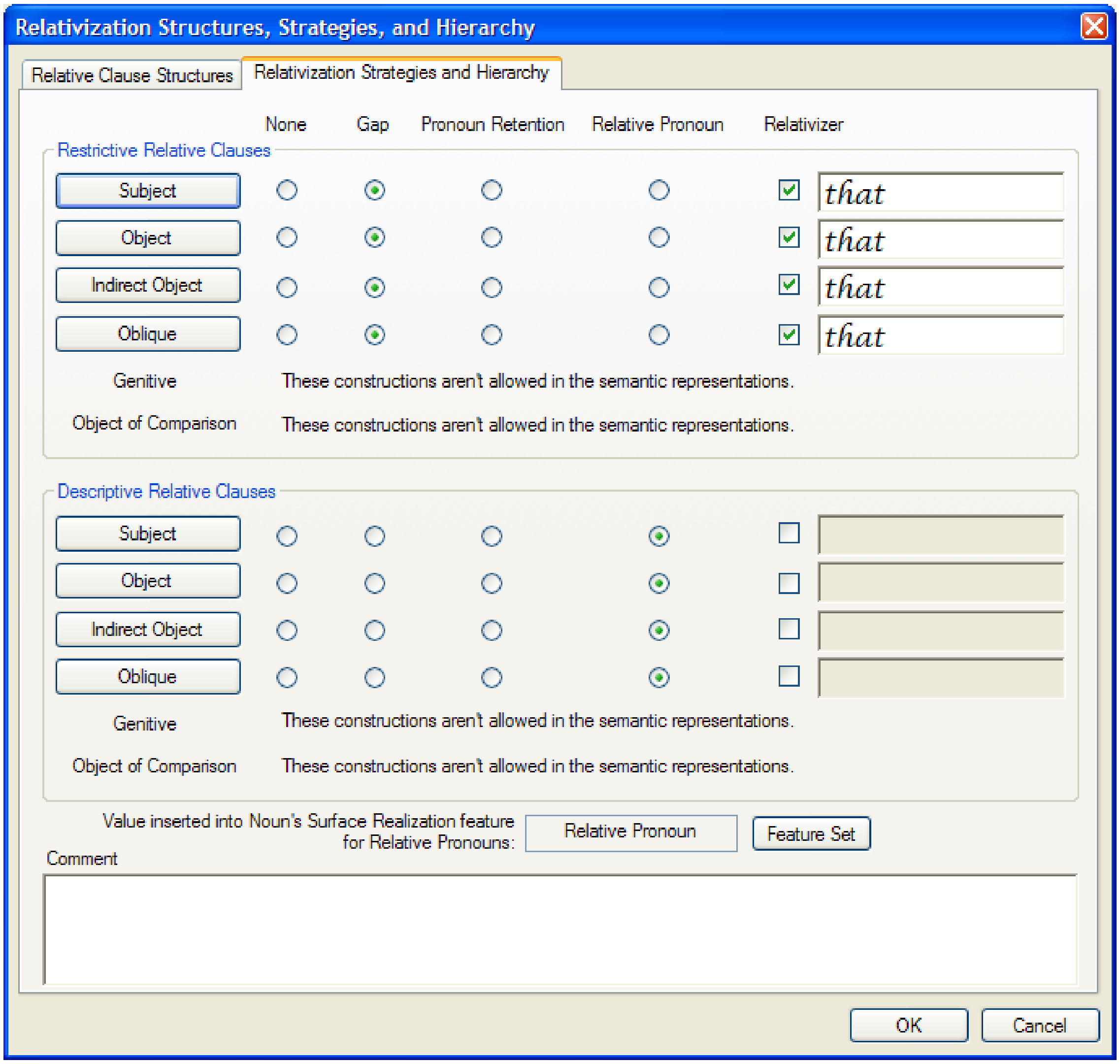
Value inserted into Trace's Surface Realization feature in Restrictive Relative Clauses:

Value inserted into Trace's Surface Realization feature in Descriptive Relative Clauses:

Figure 4-14. The Structures Tab of the Relativization Dialog

The dialog shown above has values that have been entered for English relative clauses. As can be seen in this figure, the ‘Embedded’ option has been selected because English uses embedded relative clauses, and the ‘Post-Nominal’ option has been selected because English positions its relative clauses after the head noun. After a linguist has specified a structure for the relative clauses, he is able to specify the strategies by selecting the ‘Relativization Strategies and Hierarchy’ tab seen at the top of the dialog.

The relativization hierarchy and strategies are shown in the Strategies and Hierarchy tab of the relativization dialog, as seen in figure 4-15 below. Similar to the dialog above, the dialog below has the pertinent values selected for English relative clauses.



Relativization Structures, Strategies, and Hierarchy

Relative Clause Structures | Relativization Strategies and Hierarchy

Restrictive Relative Clauses

	None	Gap	Pronoun Retention	Relative Pronoun	Relativizer	
Subject	○	●	○	○	☑	that
Object	○	●	○	○	☑	that
Indirect Object	○	●	○	○	☑	that
Oblique	○	●	○	○	☑	that
Genitive	These constructions aren't allowed in the semantic representations.					
Object of Comparison	These constructions aren't allowed in the semantic representations.					

Descriptive Relative Clauses

	None	Gap	Pronoun Retention	Relative Pronoun	Relativizer	
Subject	○	○	○	●	□	
Object	○	○	○	●	□	
Indirect Object	○	○	○	●	□	
Oblique	○	○	○	●	□	
Genitive	These constructions aren't allowed in the semantic representations.					
Object of Comparison	These constructions aren't allowed in the semantic representations.					

Value inserted into Noun's Surface Realization feature for Relative Pronouns: Relative Pronoun

Figure 4-15. The Strategies and Hierarchy Tab of the Relativization Dialog

As seen in the figure above, there are two categories of relative clauses: restrictive relative clauses which serve to identify the nominal being modified, and descriptive relative clauses (also known as non-restrictive relative clauses) which provide additional information about a referent that has already been identified. Both types of relative clauses can be applied to nominals in the relativization hierarchy which is shown below:

Subject > Object > Indirect Object > Oblique > Possessor > Object of Comparison

The semantic representations do not permit relative clauses to be formed on possessors (e.g., *The man whose house I bought lives in Dallas.*) or objects of comparison (e.g., *John is the man who I am taller than.*) because many of the world's languages do not permit those constructions.

The typological research done with respect to relative clauses indicates that languages employ four different strategies for encoding $\mathrm{NP}_{\mathrm{Rel}}$:

- Non-Reduction strategy,
- Gap strategy,
- Pronoun Retention strategy, and
- Relative Pronoun strategy.

Linguists using TTA are able to specify which strategy is used at each position of the hierarchy for the two types of relative clauses, as seen in figure 4-15 above. The dialog above also permits linguists to insert a relativizer into each relative clause. In certain cases additional rules are necessary to make very specific changes to particular relative clauses. For example, English relative clauses that modify a place use *where* as the relativizer as in *The hospital <u>where</u> I was born is in Eugene, Oregon.* Similarly English relative clauses that modify a time use *when* as the relativizer as in *The year <u>when</u> Neil Armstrong walked on the moon was 1969.* The dialog shown above inserts only the most generic relativizer *that*, then subsequent rules change the generic relativizer to a specific relativizer. For languages such as Korean which combine the relativizer with a tense morpheme in a portmanteau morph, the dialog above cannot be used; instead a separate set of rules must generate those relativizers. If a language uses the Relative Pronoun strategy at any point on the relativization hierarchy, the linguist must specify the value that will be inserted into the nominal's Surface Realization feature ("Relative

Pronoun" shown above in figure 4-15). Then when the spellout rules are executed, that feature can be inspected and the rule will generate the appropriate surface form of the relative pronoun.

4.3.6 Collocation Correction Rules

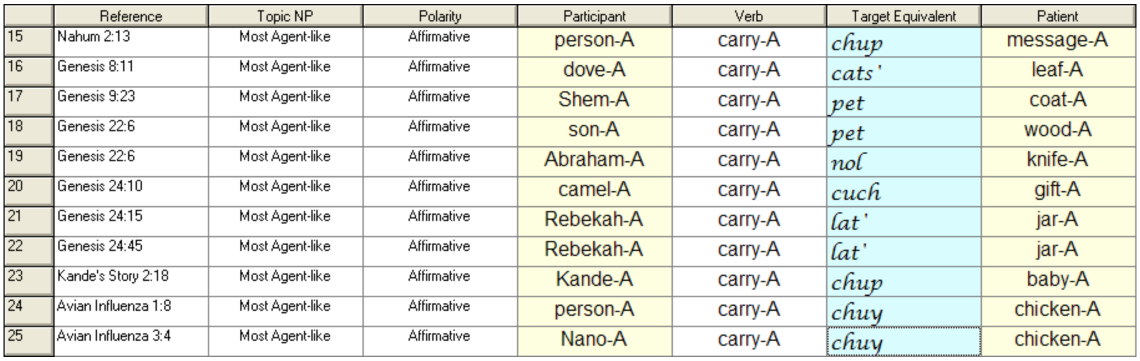
Collocation deals with how words go together (Larson 1984:141) and how words and phrases attract other words and co-occur with certain grammatical choices (Sinclair 1991:112). Collocational clash is often raised as an issue for machine translation and natural language generation projects. Even when the semantic representations consist of semantically simple words, collocation is a problem that must be dealt with. For example, HAVE is a semantic molecule, and in TTA's ontology there are ten specific senses of HAVE. HAVE sense D is used when the state object is an abstract such as a dream (e.g, *John had a dream.*), problem (e.g, *John has a problem.*), authority (e.g, *John has the authority to close the school.*), trouble (e.g., *John had some trouble yesterday.*), hope, courage, etc. Even though HAVE-D is semantically simple, other languages probably will not use their lexical equivalent of HAVE-D with all of those patient objects. The reason for this is that every word in every language has its own collocational range and restrictions. In other languages people may *see dreams* as in Russian, or *dream dreams* as in Korean, or they may *drink trouble* (Larson 1984:141), *feel courage*, etc. Collocation correction rules are designed to handle this particular problem. When translating a document from one language to another, there are many cases where a source word is best translated with a particular target word, but in certain specific environments, that source word must be translated with a different target word. Choosing the other target word is strictly a collocational issue rather than a grammatical issue. Therefore TTA's collocation correction rules change one target word to another target word in an environment that consists solely of other concepts. Each source concept has its own collocation correction rule which lets a linguist specify which target word should be used to translate that particular concept in a specific environment.

Collocation correction rules are able to reduce a problem that arises when the source language has only a generic term, but the target language has several specific terms. As was stated in section 3.3.1, it is impossible to build an ontology that will work well for every language, one of the reasons being that some languages have lexemes that have much more specific meanings than does the associated lexeme in the source language. For example, English has the very generic verb *to carry*, and that verb is used whether a person carries an object in his hands, on his back, over his shoulders, in his pocket, or on his head. Other languages such as Tzeltal (Larson 1984:89) have specific verbs for each of these situations, and they do not have a generic verb meaning 'to carry.' The Tzeltal verbs and their associated meanings are listed below in Table 4-1.

Table 4-1. Tzeltal Verbs Meaning 'to carry' (Larson 1984:89-90)

jelup'in	to carry across the shoulders
nol	to carry in the palm of the hand
chup	to carry in a pocket or pouch
chuy	to carry in a bag
lats'	to carry under the arm
pach	to carry on the head
toy	to carry aloft
yom	to carry different items together
lut'	to carry with tongs
pet	to carry in the arms
cats'	to carry between one's teeth
lup	to carry on a spoon
lat'	to carry in a container
cuch	to carry on the back

When translating a text into a language such as Tzeltal, the proper target language verb must be used in each situation in order to communicate the message. Collocation correction rules are able to significantly reduce the problem of translating a generic source term with a specific target term. Shown below in figure 4-16 is part of the collocation correction rule for CARRY-A. As seen in the figure, that collocation correction rule includes a column for the agent and another column for the object that is being carried. A Tzeltal speaker was not available for this experiment, but several of the verbs listed in Table 4-1 above have been entered into the collocation correction rule for CARRY-A shown below.



	Reference	Topic NP	Polarity	Participant	Verb	Target Equivalent	Patient
15	Nahum 2:13	Most Agent-like	Affirmative	person-A	carry-A	*chup*	message-A
16	Genesis 8:11	Most Agent-like	Affirmative	dove-A	carry-A	*cats'*	leaf-A
17	Genesis 9:23	Most Agent-like	Affirmative	Shem-A	carry-A	*pet*	coat-A
18	Genesis 22:6	Most Agent-like	Affirmative	son-A	carry-A	*pet*	wood-A
19	Genesis 22:6	Most Agent-like	Affirmative	Abraham-A	carry-A	*nol*	knife-A
20	Genesis 24:10	Most Agent-like	Affirmative	camel-A	carry-A	*cuch*	gift-A
21	Genesis 24:15	Most Agent-like	Affirmative	Rebekah-A	carry-A	*lat'*	jar-A
22	Genesis 24:45	Most Agent-like	Affirmative	Rebekah-A	carry-A	*lat'*	jar-A
23	Kande's Story 2:18	Most Agent-like	Affirmative	Kande-A	carry-A	*chup*	baby-A
24	Avian Influenza 1:8	Most Agent-like	Affirmative	person-A	carry-A	*chuy*	chicken-A
25	Avian Influenza 3:4	Most Agent-like	Affirmative	Nano-A	carry-A	*chuy*	chicken-A

Figure 4-16. Tzeltal Collocation Correction Rule for CARRY-A

The first row in the rule above indicates that in the semantic representations, a person carried a message: PERSON-A CARRY-A MESSAGE-A. Since people usually carry messages in a pocket or pouch, the Tzeltal verb *chup* was entered into that row of the collocation correction rule. The second row indicates that a dove carried a leaf. Since doves usually carry leaves in their beaks, the Tzeltal verb *cats'* could be used for that situation[23]. The sixth row indicates that a camel carried gifts. Since camels always carry things on their backs, the verb *cuch* "to carry on the back" could be used. The final row indicates that a man named Nano carried a chicken somewhere. If in the Tzeltal culture people carry chickens in bags, the verb *chuy* could be entered into that row of the collocation correction rule. So these collocation correction rules are able to insert target language lexemes into very particular contexts, and the target lexemes may have a much more restricted meaning than does the source concept. Thus these collocation correction rules are able to reduce the problem of translating a generic source term with a very particular target term.

4.3.7 Genitival Object-Object Relationships

The proposal for this project stated that very precise object-object relationships will be identified and inserted into the semantic representations (proposal p. 4). For example, *the king of Babylon* will be represented as shown below in figure 4-17.

[23] Each row in this table has a reference indicating where that particular occurrence of CARRY-A occurs. For example, the second row in this table has a reference of Genesis 8:11. That reference is strictly for the user's benefit; it is not part of the collocation correction rule. At every occurrence in the semantic representations where a dove carries a leaf, this rule will insert the Tzeltal verb *cats'*.

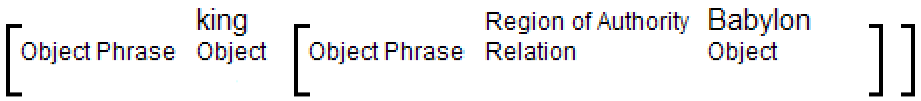


Figure 4-17 Semantic Representation of *king of Babylon*

The relation Region-of-Authority is a semantic marker which precisely indicates the relationship between the two objects KING and BABYLON. However, this approach did not work well for the vast majority of genitival object-object relationships because the number of relationships encoded by the English genitive is virtually unlimited. Therefore, as was mentioned in section 3.3.3.7, that approach was abandoned and the exact opposite approach was adopted. Instead of identifying very precisely the relationships that exist between object-object pairs that are encoded with a genitive in English, a generic genitival marker is inserted into the semantic representations between the two objects. For example, *John's work* is represented in the semantic representations as shown below in figure 4-18; that figure shows that the relation "Generic Genitive" is inserted into the semantic representations to indicate the relationship between the two objects.

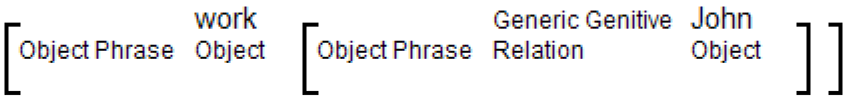


Figure 4-18 Semantic Representation of *John's work*

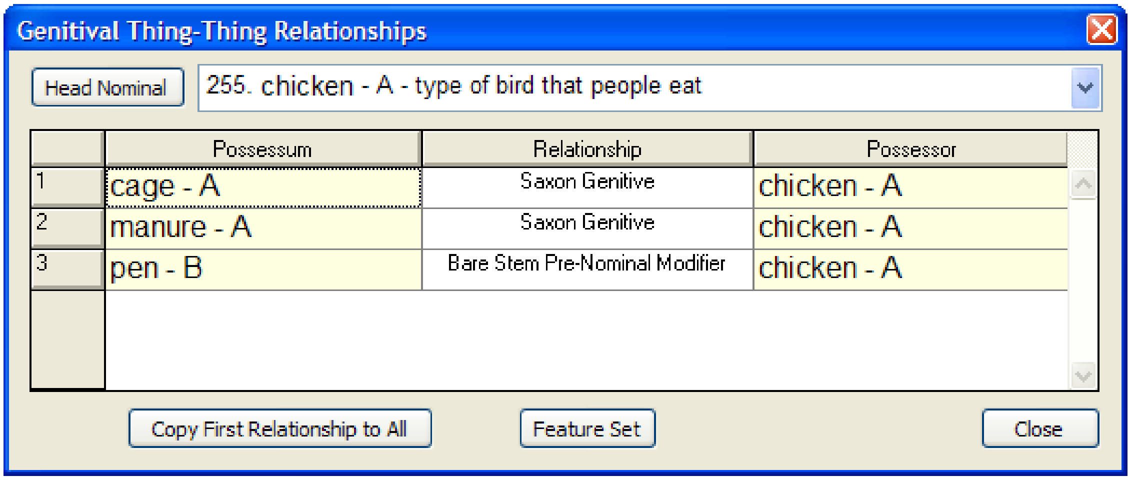
Then a section was added to TTA's transfer grammar to deal with the "Generic Genitive" marker. That section of the grammar allows a linguist who is building a grammar for a language to specify how each particular object-object pair should be encoded in the target language. This approach has worked quite well. As was stated in section 3.3.3.7, ten very specific object-object relationships were identified, and specific semantic markers are inserted into the semantic representations to indicate each of them:

- Body-Part (e.g., *Melissa's eye*),
- Made-Of (e.g., *house of cards*),
- Group (e.g., *herd of sheep*),

- Kinship (e.g., *Mary's mother*),
- Name (e.g., *man named John*),
- Owner (e.g., *John's book*),
- Quantity (e.g., *two liters of oil*),
- Region-of-Authority (e.g., *king of Egypt*),
- Part-Whole (e.g., *the back of the boat),* and
- Nationality (e.g., *an Egyptian prince*).

All other object-object relationships that are encoded with a genitive in English are marked with "Generic Genitive" in the semantic representations, and the linguist is then able to specify how each particular object-object pair should be encoded in the target text.

Shown below in figure 4-19 is the dialog showing the English object-object relationships that occur in the semantic representations with CHICKEN-A. English has three methods of encoding these genitival object-object relationships: Saxon Genitive (e.g., *the man's hat*), Norman Genitive (e.g., *the roof of the house*), and Bare Stem Pre-nominal (e.g, *garage door, chicken pen,* etc.).


Genitival Thing-Thing Relationships
Head Nominal
Copy First Relationship to All
Feature Set
Close

255. chicken - A - type of bird that people eat

	Possessum	Relationship	Possessor
1	cage - A	Saxon Genitive	chicken - A
2	manure - A	Saxon Genitive	chicken - A
3	pen - B	Bare Stem Pre-Nominal Modifier	chicken - A

Figure 4-19. Generic Genitival Object-Object Relationships involving CHICKEN-A

Linguists are able to enter the various methods that their target languages use to encode object-object relationships, and then in this dialog they specify the particular method used for each object-object pair. Subsequent rules then provide the necessary morphology for each method of encoding. As can be seen in the figure above, for each head nominal in the semantic representations there is a list of all the nominals that occur in a generic genitival relationship with that head nominal. The dialog above shows that in the semantic representations, there are occurrences of the following:

[NP CAGE-A [NP Generic-Genitive CHICKEN-A]]

[NP MANURE-A [NP Generic-Genitive CHICKEN-A]]

[NP PEN-B [NP Generic-Genitive CHICKEN-A]]

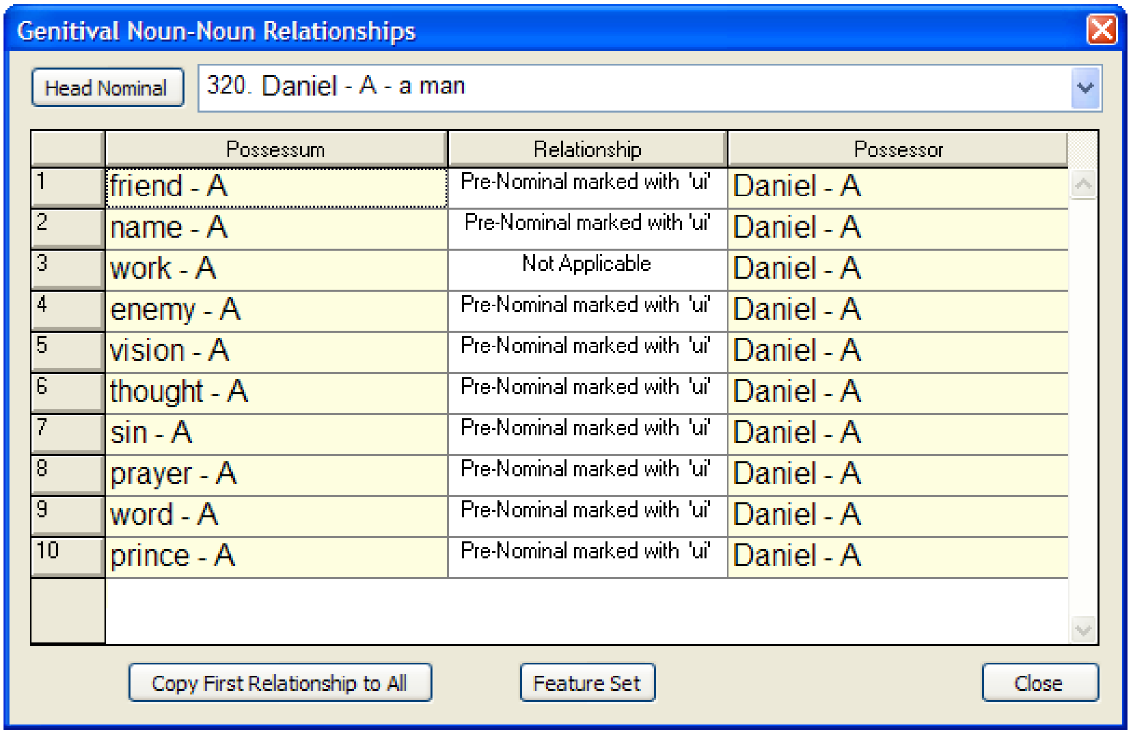
In English the CHICKEN-CAGE and CHICKEN-MANURE combinations are marked with a Saxon Genitive, so they will be realized in English as *chicken's cage* (e.g., Avian Influenza 1:8 *The people will clean the chickens' cages.*) and *chicken's manure* (e.g., Avian Influenza 3:4 *The people burned all the chickens' manure.*). The CHICKEN-PEN combination is marked with a Bare Stem Pre-nominal in English, so it will be realized as *chicken pen* (e.g., Avian Influenza 4:7 *The people helped Nano clean his chicken pen.*).

Certain object-object combinations that are encoded with a genitive relationship in English cannot be encoded with a genitive in other languages. For example, English speakers can say *Daniel's work is better than John's work.* However, the Korean equivalent is *The work that Daniel did is better than the work that John did*. In such cases, the linguist can use the dialog shown above to specify that a particular object-object combination cannot be handled with a genitive construction, and it must be dealt with in a different way. The generic genitive object-object relationships involving DANIEL-A are shown below in figure 4-20. Most genitival relationships are marked in Korean with the postposition –의 [ui] as in example 4-1.

(4-1)나-는 요한-의 책-을 봤-다.
I-Topic John-Possessor book-Object see.Past-Declarative
'I saw John's book.' (Korean example provided by JungAe Lee)

Thus, most of the genitival object-object relationships involving DANIEL are marked with "Pre-Nominal marked with 'ui'" as seen in the figure below. But figure 4-20 shows that WORK-A

Generic Genitive DANIEL-A must be handled by a separate rule, so that object-object combination is set to 'Not Applicable' in row 3.


Genitival Noun-Noun Relationships
Head Nominal
320. Daniel - A - a man
Copy First Relationship to All
Feature Set
Close

	Possessum	Relationship	Possessor
1	friend - A	Pre-Nominal marked with 'ui'	Daniel - A
2	name - A	Pre-Nominal marked with 'ui'	Daniel - A
3	work - A	Not Applicable	Daniel - A
4	enemy - A	Pre-Nominal marked with 'ui'	Daniel - A
5	vision - A	Pre-Nominal marked with 'ui'	Daniel - A
6	thought - A	Pre-Nominal marked with 'ui'	Daniel - A
7	sin - A	Pre-Nominal marked with 'ui'	Daniel - A
8	prayer - A	Pre-Nominal marked with 'ui'	Daniel - A
9	word - A	Pre-Nominal marked with 'ui'	Daniel - A
10	prince - A	Pre-Nominal marked with 'ui'	Daniel - A

Figure 4-20. DANIEL and WORK cannot occur in a Genitival Relationship in Korean

A subsequent rule will convert "WORK-A Generic-Genitive DANIEL-A" to WORK-A modified by the relative clause DANIEL-A DO-A WORK-A.

4.3.8 Theta Grid Adjustment Rules

As has been stated several times, the semantic representations are significantly influenced by English. The events in the ontology are events that have been lexicalized by English, and the theta grids for those events generally correspond to the English perspective of those events. If the semantic representations were being developed in a different language, the ontology would certainly contain different events, and the theta grids for those events would also be different. The purpose of the Theta Grid Adjustment rules is to change the theta grids of the events in the ontology so that they correspond with the theta grids of the corresponding target verbs. Every event in the ontology has a specified theta grid, some arguments being obligatory, others being optional. Every event in the ontology also has a prewritten theta grid adjustment rule. Those rules have access to the argument structure of each event, and whether an argument is obligatory or optional, so that information is reflected in the rules' input structures. For example, one event in the ontology is TRADE-A as in *John*-Agent *traded a chicken*-Patient *for a duck*-Destination *with Mary*-Source, and its theta grid is shown below:

Event: TRADE-A 'to trade one thing for something else with someone'

TRADE-A	Agent	Patient	Destination	(Source)

As seen in the theta grid shown above, the agent, patient, and destination NPs are obligatory in the semantic representations, but the source NP is optional. Every occurrence of TRADE-A in the semantic representations will follow this pattern. The semantic representation of *John traded a chicken for a duck with Mary* is shown below in figure 4-21.

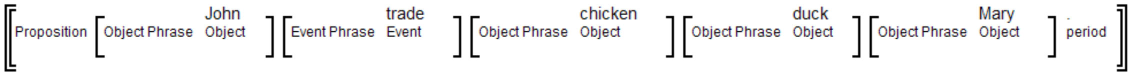


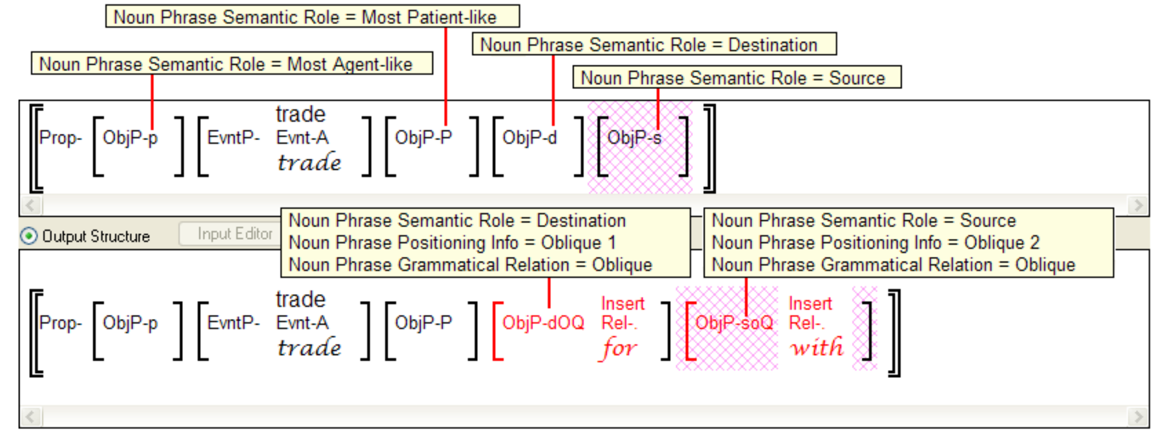
Figure 4-21. Semantic Representation of *John traded a chicken for a duck with Mary.*

As seen in the figure above, there are no relations in the argument object phrases[24]; the relations must be supplied by rules. The purpose of the theta grid adjustment rules is to adjust this pattern to fit each particular target language. Generally theta grid adjustment rules perform two tasks: 1) insert case markers into the object phrases, and 2) specify the relative order of the oblique NPs. Note that all constituent ordering is done by phrase structure rules which will be discussed later in the synthesis section, but the theta grid adjustment rules set features on the oblique NPs so that the phrase structure rules will be able to order them properly. For example, in the English sentence *John traded a chicken for a duck with Mary,* the oblique NP *for a duck* must precede the oblique NP *with Mary*. The English case markers for TRADE-A and the relative ordering of the oblique NPs is shown in the grid below.

[24] There are very few events in the ontology which have relations in any of their argument object phrases. Examples of events which do include relations in their argument phrases include BE-F which has a relation in its State object phrase (e.g., *John is under/in/beside the car*.) and LIVE-A which has a relation in its Destination object phrase (e.g., *John lives under/near/beside the bridge.*)

trade-A	Agent	Patient	Destination	(Source)
	John	chicken	duck	Mary
English Case Markers			*for*	*with*
English Ordering			Oblique NP-1	Oblique NP-2

Other languages that have a lexical equivalent for TRADE-A will quite possibly view the event differently and therefore require other modifications to the proposition's structure. The English theta grid adjustment rule for TRADE-A is shown below in figure 4-22.



Noun Phrase Semantic Role = Most Patient-like

Noun Phrase Semantic Role = Destination

Noun Phrase Semantic Role = Most Agent-like

Noun Phrase Semantic Role = Source

Noun Phrase Semantic Role = Destination
Noun Phrase Positioning Info = Oblique 1
Noun Phrase Grammatical Relation = Oblique

Noun Phrase Semantic Role = Source
Noun Phrase Positioning Info = Oblique 2
Noun Phrase Grammatical Relation = Oblique

Figure 4-22. The English Theta Grid Adjustment Rule for TRADE-A

As can be seen in the figure above, the event TRADE-A takes four arguments which are indicated by the four object phrases in the input structure. The final argument has a Semantic Role value of Source, and it is optional as indicated by the purple cross-hatch. The output structure indicates that English prepositions will be inserted into two of the object phrases, and the features of those phrases will be set to indicate how they should be ordered by the phrase structure rules. The English preposition *for* will be inserted into the Destination object phrase, and that phrase's features will be set to 'Grammatical Relation' = 'Oblique' and 'Positioning Info' = 'Oblique 1'[25]. Similarly the preposition *with* will be inserted into the Source object phrase if it

[25] If there are multiple object phrases that have the same semantic role, another section of the grammar allows linguists to specify whether the Theta Grid Adjustment rules should insert adpositions into only the first phrase of a particular semantic role, the last phrase of that semantic role, or all the phrases with that semantic role. English inserts these prepositions into only the first phrase with the specified semantic role (e.g., *John traded a duck for a chicken, a frog, and a mouse with Mary. *John traded a duck for a chicken, for a frog, and for a mouse with Mary.*). Korean inserts postpositions into only the last phrase of a specified semantic role.

is present in the semantic representation, and that phrase's features will be set to 'Grammatical Relation' = 'Oblique' and 'Positioning Info' = 'Oblique 2'. The popup above the final object phrase in the output structure shows that its grammatical relation has been set to 'Oblique', and its position is set to 'Oblique 2'. Therefore the phrase structure rules will be able to order these oblique phrases appropriately using the features Grammatical Relation and Positioning Info[26].

The Korean equivalent of *John traded a chicken for a duck with Mary* is literally *John exchanged his chicken for Mary's duck.* Therefore the Korean Theta Grid Adjustment rule is considerably more complex than the English rule shown above. The rule that restructures all propositions that contain TRADE-A in the semantic representations so that they conform to the Korean perspective is shown below in figure 4-23.

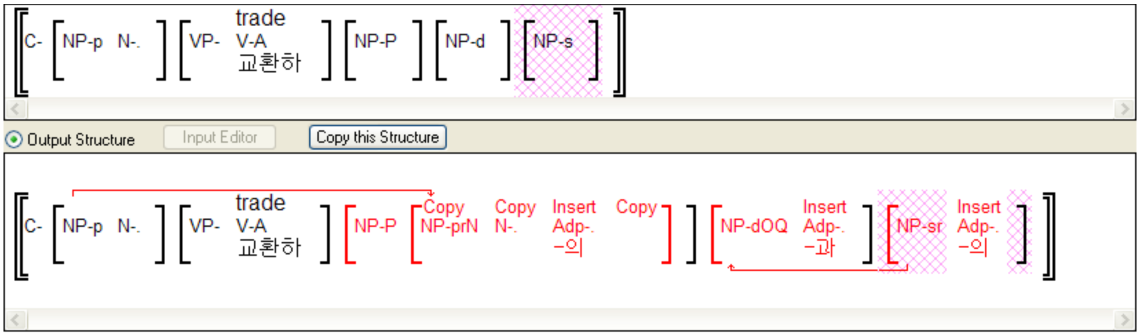


Figure 4-23. The Korean Theta Grid Adjustment Rule for TRADE-A

Example (4-2) below shows the details of the Korean equivalent of *John traded a duck for a chicken with Mary*.

(4-2) 요한-은 마리아-의 닭-과 자기 오리-를 교환했-다.
John-Topic Mary-Genitive chicken-and his duck-Object exchange.Past-Declarative
John exchanged his duck for Mary's chicken. Or *John traded a duck for a chicken with Mary.* (Korean text generated by TTA)

The theta grid adjustment rule shown above in figure 4-23 copies the topic nominal into the direct object phrase and marks it as the possessor of the object. That rule also moves the source phrase into the destination phrase, and marks the source object as the possessor of the

[26] The ordering of the oblique phrases will be described thoroughly in the section below that discusses the phrase structure rules.

destination object. The rule also inserts the conjunction -과 [gwa] 'and' into the destination phrase. Whenever this rule finds the event TRADE-A in the semantic representations, it will restructure the proposition to conform to the Korean perspective.

As was stated above, the primary purpose of the Theta Grid Adjustment rules is to change the theta grid of a particular event in the semantic representations to match the target language's requirements. When an event includes a patient proposition as one of its arguments, the rule must specify the structure of that patient proposition. Extensive typological research has been done on the structures of patient propositions (Noonan 2007:42-140; Givón 1990:515; Payne 1997:313), and this research indicates that the matrix verb dictates the structure of its patient proposition. All complement clauses in a particular language can be positioned on a continuum ranging from 'fully propositional' to 'highly merged.' There are three indicators in every object complement clause that indicate where it should be placed on this continuum:

- the presence or absence of a complementizer,
- whether the subject of the complement clause is coded as the subject of the complement's verb or as the object of the matrix verb, and
- whether the verb in the object complement is finite or non-finite.

Several examples of English object complement clauses are shown below in examples (4-3i) through (4-3iii).

(4-3i) *John thinks* [*that she might have read a book*].

(4-3ii) *John told* [*her to read a book*].

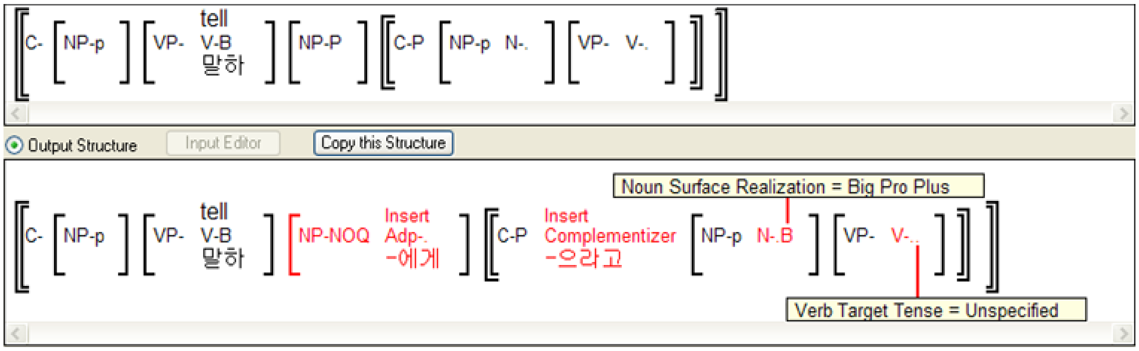
(4-3iii) *John made* [*her read a book*].

In example (4-3i) above, there is a complementizer *that,* the pronoun *she* is in the nominative case indicating that it is coded as the subject of the verb *read* rather than as the object of the verb *think*, and the verb phrase *might have read* is finite. Therefore that object complement clause is categorized as fully prepositional. In (4-3ii) above, there is no complementizer, the pronoun *her* is in the accusative case indicating that it is coded as the object of the verb *tell* rather than the subject of the verb *read*, and the verb in the object complement is coded as a *to*-infinitive. Therefore that object complement is positioned toward the 'highly merged' end of the continuum. In the final example (4-3iii), there is not a complementizer, the pronoun *her* is again coded as the object of *make* rather than the subject of *read*, and the verb is a bare stem infinitive. Therefore that object complement must be positioned at the 'highly merged' end of the continuum. Each language has its own techniques for encoding object complement clauses, and the theta grid adjustment rules are used to specify the structure of the object complement clause for each particular matrix verb. When TTA builds the input structures for the theta grid adjustment rules, it has access to the theta grid for each event in the ontology. If an event takes a patient proposition, TTA will include the patient proposition in the input structure, and the agent object phrase and the event phrase will be included in the patient proposition. This makes it very easy for the linguist to specify the structure of the patient proposition using the theta grid adjustment rule for the matrix event. An event that takes a patient proposition is TELL-B as in *John told Mary to read a book.* The theta grid for TELL-B is shown below.

Event: TELL-B 'one person tells another person to do something'

TELL-B	Agent	Patient	Patient Proposition

The Korean theta grid adjustment rule for the event TELL-B is shown in figure 4-24 below.



Noun Surface Realization = Big Pro Plus

Verb Target Tense = Unspecified

Figure 4-24. The Korean Theta Grid Adjustment Rule for TELL-B

In the semantic representations, TELL-B always has a patient NP and a patient proposition as seen in the input structure of the rule shown above in figure 4-24. The Korean equivalent of *John told Mary to read a book* is 요한은 마리아에게 책을 읽으라고 말하였다 as shown in example (4-4) below.

(4-4) 요한-은 마리아-에게 책-을 읽-으라고 말하-였-다
John-Topic Mary-to book-Object read-Complementizer tell-Past-Declarative
John told Mary to read a book. (Korean text generated by TTA)

The rule shown above in figure 4-24 indicates that when the matrix verb is TELL-B, Korean marks the patient nominal with the postposition –에게 [e ge] which generally signals an indirect object. The rule also inserts the complementizer –으라고 [eu ra go], and specifies that the patient proposition's verb is unmarked for tense as indicated by the yellow popup showing the features of the patient proposition's verb. Therefore in the Korean structure, *Mary* is encoded as an indirect object of the matrix verb, there is a complementizer, and the verb in the complement clause is non-finite. This rule will enable the synthesizing grammar to produce the proper surface structure for the complement clause of 말하다 [mal ha da] 'to tell'.

When a target language does not have a lexical equivalent for a particular source event, the theta grid adjustment rule for that source event must restructure the proposition so that it matches the target language's requirements. For example, it was stated in the introduction to this chapter that Korean does not have a lexical equivalent for WEIGH-A as in *The box weighed ten pounds*, or for WEIGH-B as in *John weighed the box.* For cases like these, the theta grid adjustment rules must restructure the source proposition into a new underlying representation that uses the target language's lexemes and structures. The Korean equivalent of *The dog weighed fifteen kilograms* is shown below in example (4-5)[27].

[27] All weights, distances, and volumes are specified with metric units in the semantic representations because they're much more common than the English units. When generating text in English, a structural adjustment rule is used to convert the metric units to English units.

(4-5) 개-의 무게-는 십오 키로-이-였-다.
dog-Genitive weight-Topic fifteen kilo-be-Past-Declarative
The dog's weight was fifteen kilos. or *The dog weighed 15 kilograms.* (Korean example provided by JungAe Lee)

As seen in example (4-5), Korean has a noun meaning 'weight' and it uses the copula 이다 [i da] 'to be.' The rule that performs these adjustments for Korean is shown below in figure 4-25.

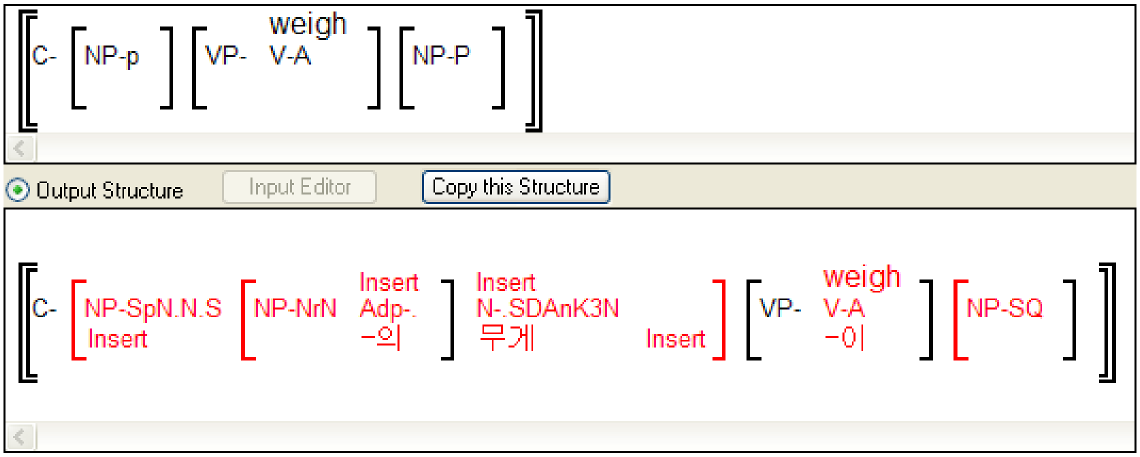
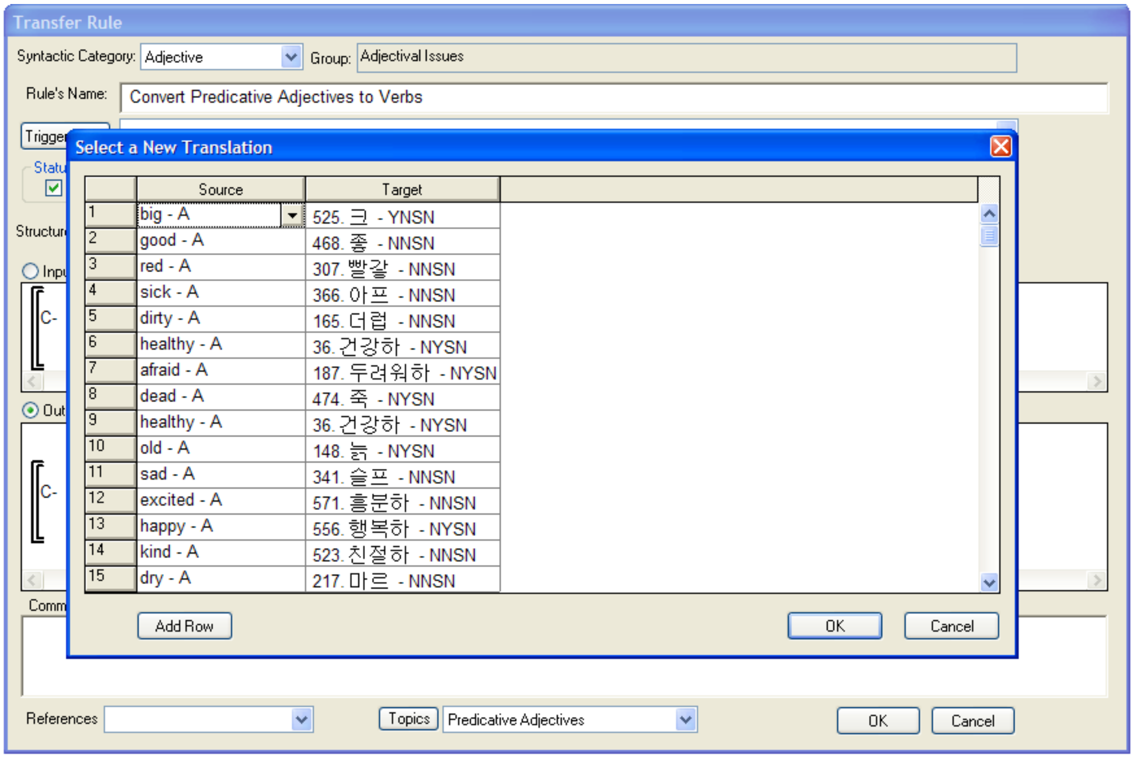


Figure 4-25. The Korean Theta Grid Adjustment Rule for WEIGH-A

The rule shown above in figure 4-25 inserts the noun 무게 [mu ge] 'weight' as the agent, and changes the original agent object so that it becomes the possessor of 무게 [mu ge]. The source verb WEIGH-A is then linked to the Korean verb 이다 [i da] 'to be'. Thus this rule restructures all the propositions in the semantic representations that contain the event WEIGH-A so that they conform to the Korean equivalent.

4.3.9 Structural AdjustmentRules

The final type of rule in the transfer grammar is called *Structural Adjustment rules*. These rules perform the remaining tasks that are necessary in order to adjust the structure of the semantic representations into an appropriate underlying representation for the target language. Common tasks performed by the structural adjustment rules include *syntactic aggregation*, *inserting aspectual, modal, and polarity auxiliaries*, *converting predicative adjective constructions to verbs*, *converting distance, weight, and volume measurements from metric units to the units used by the target language*, etc. For example, English uses predicative adjective constructions such as *The apple is red*, so there are many predicative constructions in the semantic representations. But many languages will use a verb that means *be red* rather than a copula plus an adjective. A single structural adjustment rule can convert all of the predicative constructions in the semantic representations to verbs as shown below in figure 4-26.



Syntactic Category: Adjective

Group: Adjectival Issues

Rule's Name: Convert Predicative Adjectives to Verbs

	Source	Target
1	big - A	525. 크 - YNSN
2	good - A	468. 좋 - NNSN
3	red - A	307. 빨갈 - NNSN
4	sick - A	366. 아프 - NNSN
5	dirty - A	165. 더럽 - NNSN
6	healthy - A	36. 건강하 - NYSN
7	afraid - A	187. 두려워하 - NYSN
8	dead - A	474. 죽 - NYSN
9	healthy - A	36. 건강하 - NYSN
10	old - A	148. 늙 - NYSN
11	sad - A	341. 슬프 - NNSN
12	excited - A	571. 흥분하 - NNSN
13	happy - A	556. 행복하 - NYSN
14	kind - A	523. 친절하 - NNSN
15	dry - A	217. 마르 - NNSN

References

Topics: Predicative Adjectives

Figure 4-26. Structural Adjustment Rule that Converts Predicative Adjective Constructions to Korean Verbs

The dialog in figure 4-26 above shows that many adjectives are used predicatively in the semantic representations, and each of them must be converted to a particular Korean verb. So the linguist is able to set up a conversion table for each adjective that is used predicatively in the semantic representations, and that table produces the appropriate Korean verb.

Another common task performed by the structural adjustment rules is syntactic aggregation. There are many situations where syntactic aggregation may be performed: 1) Combining two simple propositions that have the same subject into a single proposition: *John* *went to the library. Then John studied for the test. -> John went to the library and studied for the test.* 2) Combining two propositions that have the same agent and event, but different patients: *John put on his clothes. Then John put on his coat. -> John put on his clothes and coat.* 3) Combining oblique phrases that have similar constituents: *John told Mary about all the things that Steve did and about all the things that Steve said. -> John told Mary about all the things that Steve did and said.* etc. A rule for English that combines two simple propositions that have the same subject is shown below in figure 4-27. If all the circumstances are satisfied, the output structure of that rule will move the second proposition into the first, and set the Surface Realization feature on the subject of the second proposition to PRO. However, there are many restrictions to combining propositions in English; a few of them are listed below:

- Both subjects must have Affirmative Polarity. *No man has climbed that mountain. And no man has seen the other side of that mountain. *No man has climbed that mountain and seen the other side of it.*
- The second proposition must not have a preposed adverbial clause. *John studied diligently for the test. And if John passes the test, he will go on a vacation. *John studied diligently for the test and if he passes, he will go on a vacation.*
- Both propositions must have the same illocutionary force. *John went to the store. Did John buy some bread? *John went to the store and did he buy some bread?*

These restrictions and numerous others are all specified in the rule shown below in figure 4-27.

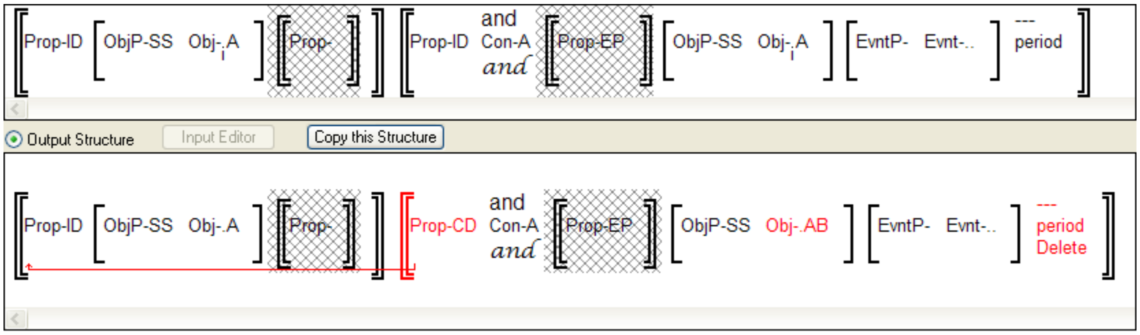


Figure 4-27. Structural Adjustment Rule that Combines Two Simple Propositions into a Single Sentence for English

Another common task performed by the structural adjustment rules is to convert genitival object-object relationships so that they are appropriate for the target language. As was stated in section 4.3.7 above, not all genitival object-object relationships in the semantic representations can be expressed with a genitive in the target language. Some of the genitival object-object relationships in the semantic representations must be converted to new structures that are appropriate for the target language. The genitival object-object relationship shown below in figure 4-28 cannot be expressed with a genitive construction in Korean, so a structural adjustment rule converts that structure to a noun modified by a relative clause; that structural adjustment rule is shown below in figure 4-29.

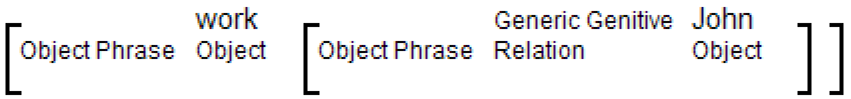


Figure 4-28. Semantic Representation of *John's work*

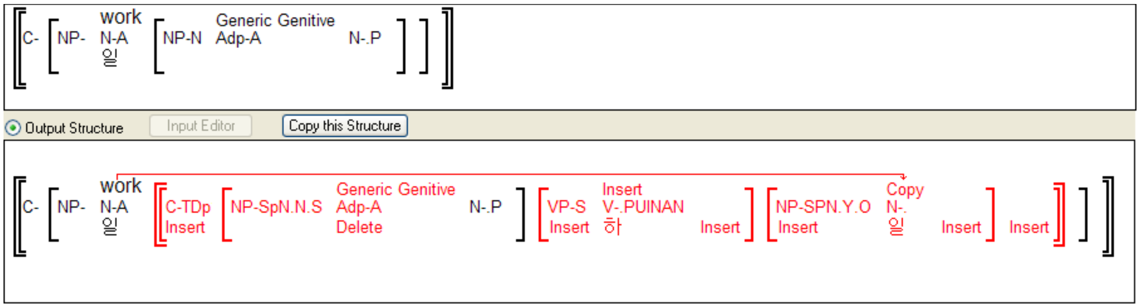


Figure 4-29. Structural Adjustment Rule that Converts "X's work" to "work that X did" for Korean

The rule shown above in figure 4-29 inserts a relative clause to modify the object WORK-A. The object that was in the genitival relationship with WORK-A becomes the agent of the relative clause, and the Generic Genitive relation is deleted. The rule also inserts the Korean verb 하다 [ha da] meaning 'to do' into the relative clause, and copies the object WORK-A so that it becomes the patient of the verb 하다 [ha da]. The result is that all occurrences of X Generic Genitive WORK-A in the semantic representations are converted to *the work that X did.*

The structural adjustment rules, combined with all of the other types of rules in the Transfer Grammar, are able to adjust the semantic representations so that they contain the target language's features, structures, and lexemes. Then the synthesizing grammar can begin synthesizing the surface forms. Developing the transfer grammar for a language is generally much more complex than developing the synthesizing grammar. In the two test languages for this dissertation, and in the other test languages as well, the synthesizing grammars were generally very easy to develop. However, developing the transfer grammars was much more difficult and time consuming. More work needs to be done to make the development of the transfer grammars practical and efficient.

4.4 The Synthesizing Grammar

This section will describe TTA's synthesizing grammar. The synthesizing grammar has eight different types of rules, and each of those rule types will be described and illustrated. When the synthesizing grammar for TTA was designed, it was considered desirable to make it resemble as closely as possible the descriptive grammars that linguists routinely write. Linguists have spent decades identifying the various tasks that must be performed during synthesis, so that research was integrated into TTA's synthesizing grammar. After examining many descriptive grammars written by field linguists, a list of the most common tasks performed during synthesis was compiled:

- indicate agreement amongst constituents,
- modify stems with affixes,
- perform morphophonemic operations across morpheme and word boundaries,
- add clitics to phrases,
- move constituents from one location to another,
- order the constituents, and
- identify where pronouns may be used.

After identifying these common tasks, a model of TTA's synthesizing grammar was developed. Then rules were designed to accomplish each of the common tasks listed above. The synthesizing grammar in TTA is not an operational model of any one particular linguistic theory, and it is intentionally generic so that grammars may be developed using a variety of linguistic models. However, the synthesizing grammar most closely parallels the transformational-generative model. Therefore it has feature copying rules, spellout rules, clitic rules, movement rules, phrase structure rules, and pronoun identification rules. A model of TTA's synthesizing grammar is shown below in figure 4-30.

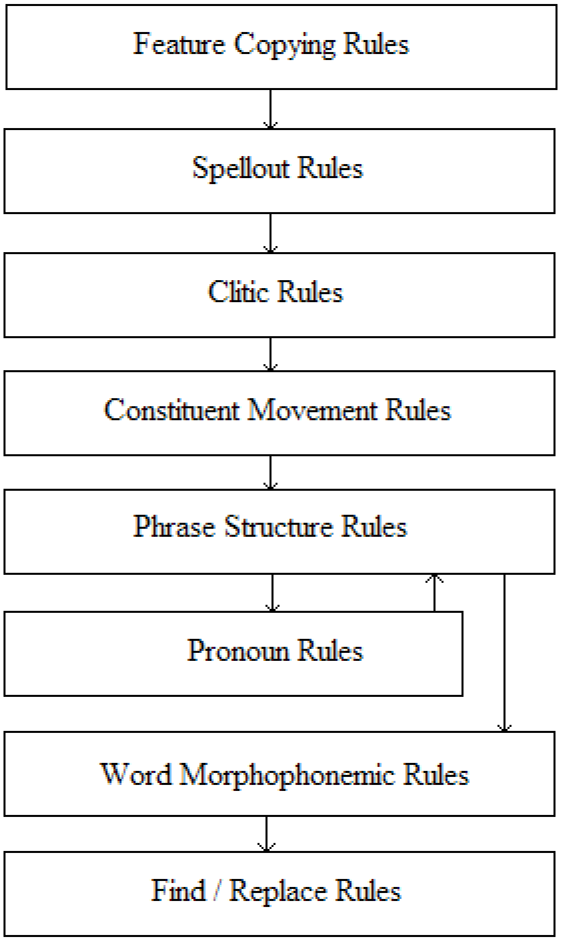


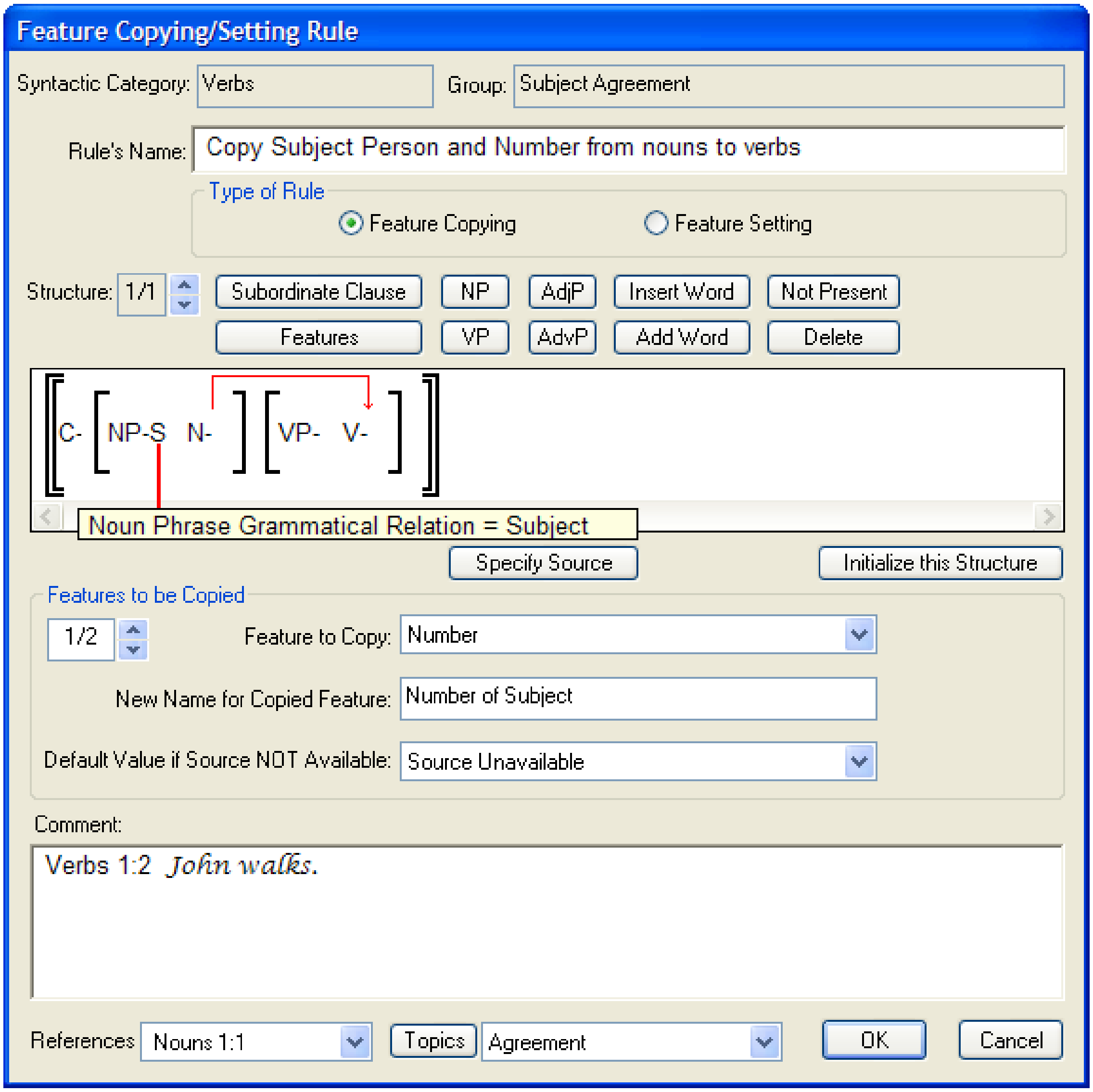
Figure 4-30. The Synthesizing Grammar in The Translator's Assistant

As seen in the figure above, the synthesizing grammar has eight different types of rules. The rules are generally applied from top to bottom, but there is one exception to this pattern. After the Pronoun Rules have been applied, the Phrase Structure Rules are executed again because sometimes the pronoun rules insert new constituents into the text, and those constituents must be positioned properly.

As was mentioned above, the synthesizing grammar is responsible for performing the synthesis step of the translation process. It is the synthesizing grammar that takes the adjusted underlying representation created by the transfer grammar, and then generates the target surface structure texts. Each of the types of rules shown in the model above in figure 4-30 and their capabilities will now be discussed in detail.

4.4.1 Feature Copying Rules

Across languages there are many dimensions of agreement in surface structure. For example, English verbs agree in person and number with their subjects, Korean verbs agree with the honorability of their subjects, Greek articles and adjectives agree in case, number, and the declension of their head noun (Summers 1950:15, 27), etc. In many languages all the constituents that modify a noun must agree with the noun’s gender or class, and this is often called *concord*. The feature copying rules in TTA copy one or more feature values on one constituent and paste them on another constituent so that the subsequent spellout rules can add the required morphology to indicate the agreement. Shown below in figure 4-31 is the feature copying rule for English which copies the person and number of the subject noun to the verb.



Feature Copying/Setting Rule

Syntactic Category: Verbs

Group: Subject Agreement

Rule's Name: Copy Subject Person and Number from nouns to verbs

Type of Rule: ● Feature Copying ○ Feature Setting

Structure: 1/1

[[C- [NP-S N-] [VP- V-]]]

Noun Phrase Grammatical Relation = Subject

Features to be Copied: 1/2

Feature to Copy: Number

New Name for Copied Feature: Number of Subject

Default Value if Source NOT Available: Source Unavailable

Comment:

Verbs 1:2 *John walks.*

References: Nouns 1:1

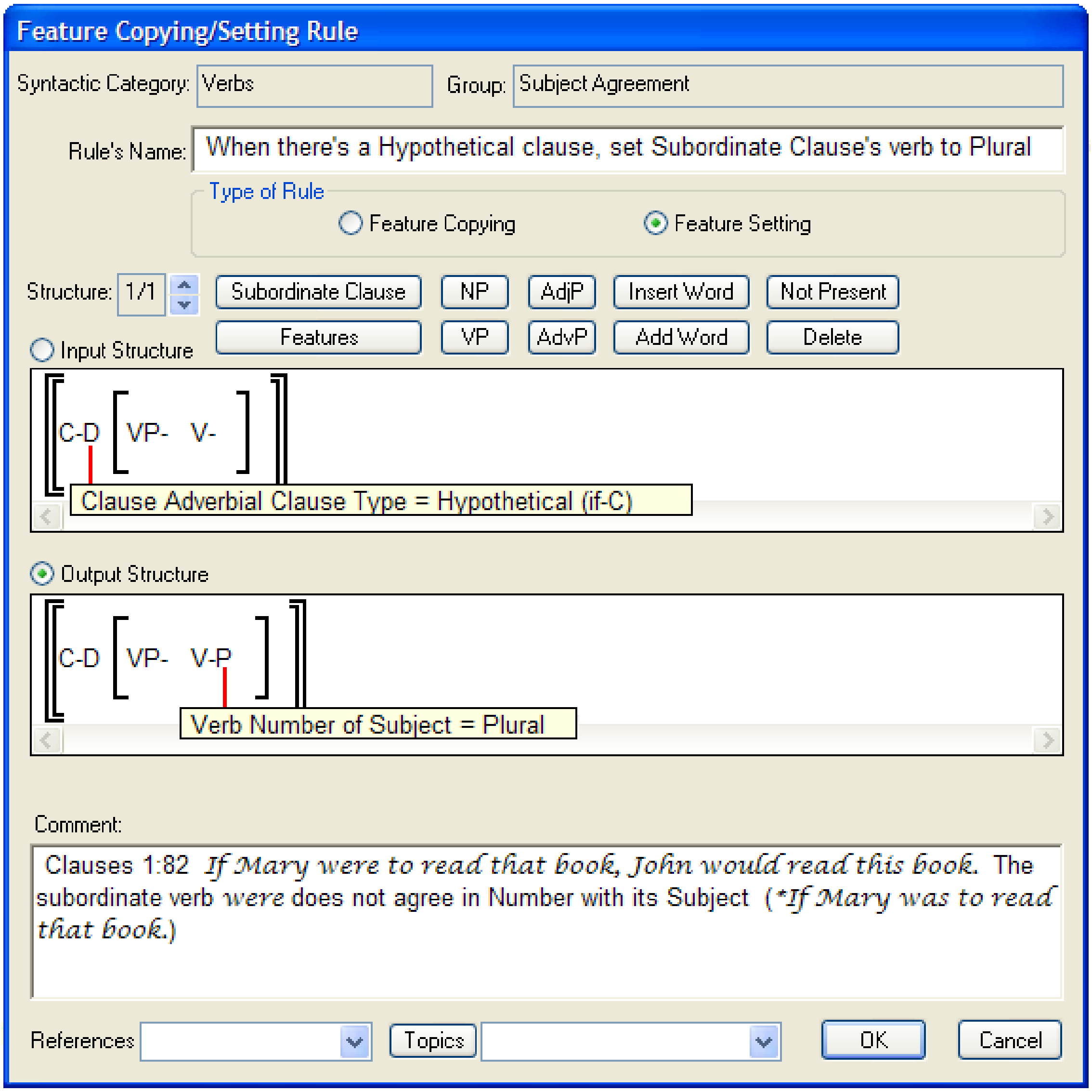
Topics: Agreement

Figure 4-31. A Feature Copying Rule for English

As seen in figure 4-31 above, linguists are able to build an input structure which specifies both the source and the destination of the feature that is to be copied. The rule in the figure above copies the Number value from the nominals in subject noun phrases to the verb. The yellow popup below the input structure shows that the NP must be a subject NP. Whenever a feature copying rule copies the noun feature called Number, the rule finds all of the nominals that match the criteria, and sums their number values. For example, if there are two subject NPs, and each contains a singular noun, then Singular + Singular = Dual. Similarly, if there are three subject

NPs, Singular + Dual + Singular = Quadrial, Singular + Dual + Trial = Plural, etc. After the summation of the Number values, the grammar applies the Number collapsing rule described above in section 4.3.2.2 to the result of the summation. Therefore, in English, Singular + Singular + Singular = Trial (e.g., *John, Mary, and Steve went to the movie.*), and Trial is changed to Plural. Whenever one constituent in a sentence agrees in some way with another constituent, a feature copying rule is used to copy the relevant features from the source of the agreement to the destination of the agreement.

Occasionally a grammar needs to set a particular feature to a certain value in order to override the agreement system. For example, in English hypothetical adverbial clauses such as *If Mary were to read that book,* the auxiliary *were* is always in the plural form even when the subject is singular (**If Mary was to read that book, …*). Therefore the feature copying rules also have the ability to set a copied feature to a particular value after the agreement system has set it to the typical value. A feature copying rule for English copies the number of the subject to the verb, and that feature on the verb is called 'Number of Subject'. But that feature must be set to 'Plural' for the example mentioned above. The rule shown below in figure 4-32 is the rule for English that sets a verb's 'Number of Subject' feature to 'Plural' if the clause is a hypothetical adverbial clause.



Feature Copying/Setting Rule

Syntactic Category: Verbs

Group: Subject Agreement

Rule's Name: When there's a Hypothetical clause, set Subordinate Clause's verb to Plural

Type of Rule: ○ Feature Copying ● Feature Setting

Structure: 1/1

○ Input Structure

[[C-D [VP- V-]]]

Clause Adverbial Clause Type = Hypothetical (if-C)

● Output Structure

[[C-D [VP- V-P]]]

Verb Number of Subject = Plural

Comment:

Clauses 1:82 _If Mary were to read that book, John would read this book._ The subordinate verb _were_ does not agree in Number with its Subject (_*If Mary was to read that book._)

Figure 4-32. A Feature Setting Rule for English

The figure above shows that the rule looks for propositions that have a ‘Type’ value of ‘Hypothetical’. When a hypothetical proposition is found, the output structure will set the verb’s ‘Number of Subject’ feature value to ‘Plural’ as indicated in the yellow popup below the output structure. Then the spellout rule that generates the surface form of that auxiliary verb will generate *were* rather than *was*.

4.4.2 Spellout Rules

The spellout rules are the most versatile type of rule in TTA's grammar. They occur in the target lexicon in order to generate the lexical forms as was described in section 4.2.2 above. They also occur here in the synthesizing grammar, and they occur again later in the synthesizing grammar in order to generate the surface forms of pronouns and switch reference markers. These rules are generally responsible for producing the final surface form of each individual target word. The spellout rules look at the features of particular constituents, and then modify those constituents with the appropriate morphology. Therefore spellout rules have the following capabilities:

- add prefixes, suffixes, infixes, and circumfixes to stems,
- reduplicate a specified number of characters or an entire word,
- insert words into phrases or sentences,
- select a form of a target word from the target lexicon,
- replace one target word with another word,
- perform morphophonemic operations across syllable boundaries[28].

Spellout rules generally look at a small environment such as a word, its phrase, and its clause, and then perform one of the operations listed above based on that environment. For example, if a particular noun in a semantic representation has a Number value of Plural, a spellout rule can add the plural morpheme.

As was mentioned above in section 4.2.2, there are four types of spellout rules:

- simple spellout rules which supply a single affix or word,
- table spellout rules which supply a table of morphemes or words,

[28] Morphophonemic operations which occur across word boundaries must be executed later after the phrase structure rules have put all the constituents in their proper order. For example, the English indefinite article *a* changes to *an* when it precedes words that begin with a vowel. These morphophonemic operations can't be performed until all the constituents are in their proper order.

- morphophonemic rules which perform morphophonemic operations or spelling corrections, and
- form selection rules which select a form of a word from the target lexicon and insert it into the generated text.

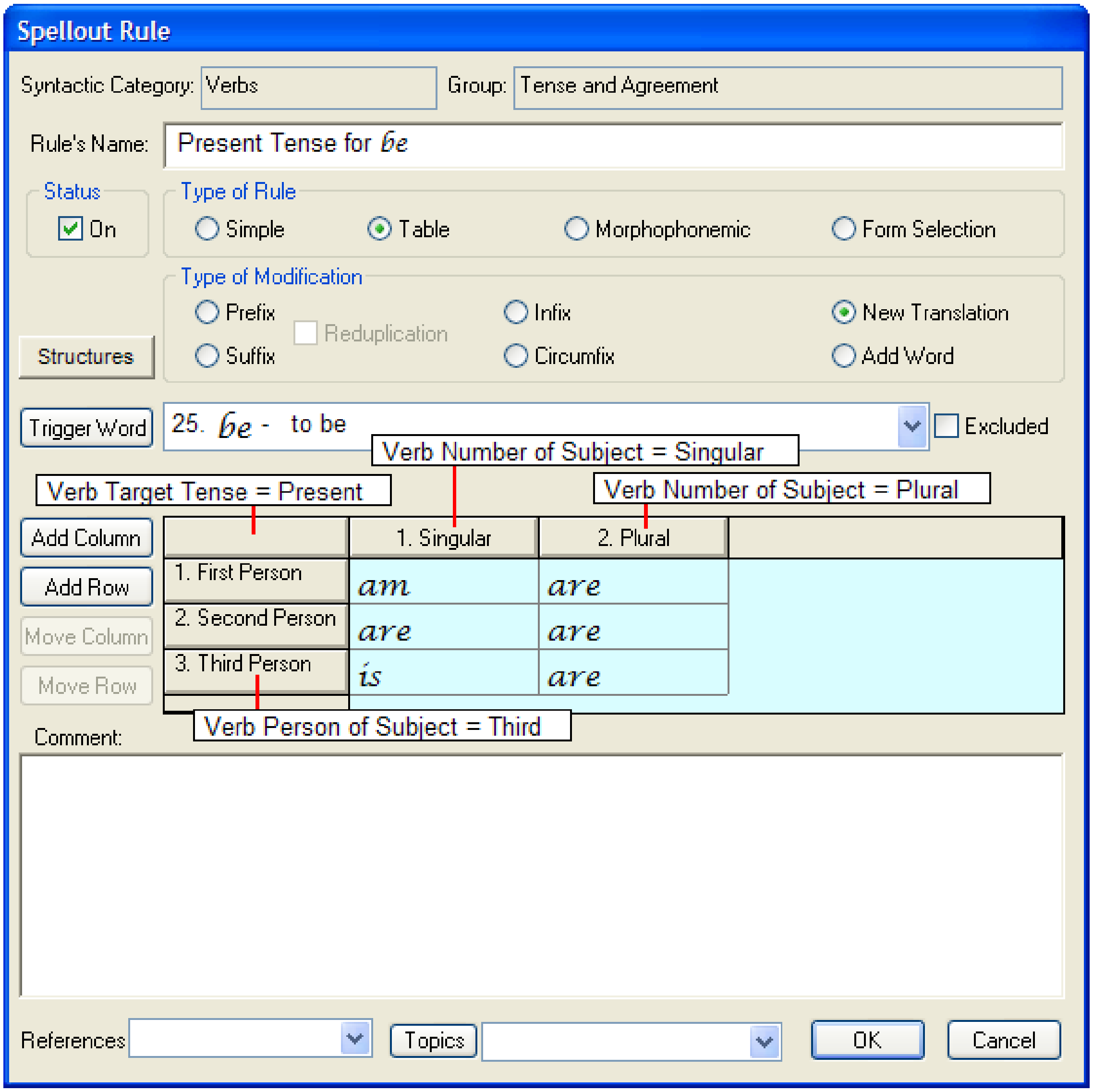
Each of these four types of spellout rules will now be described.

4.4.2.1 Simple Spellout Rules

Simple spellout rules add a single morpheme or word to the specified environment, as was shown above in figure 4-4. That rule is a simple lexical spellout rule which adds the suffix –*ed* to form the past tense form of English verbs.

4.4.2.2 Table Spellout Rules

Table spellout rules let linguists build a table of morphemes or words that are related in some way, and each column and row in the table have particular feature values. For example, the rule shown below in figure 4-33 inserts the proper present tense form of *to be*.



Spellout Rule

Syntactic Category: Verbs

Group: Tense and Agreement

Rule's Name: Present Tense for *be*

Status: ☑ On

Type of Rule: ○ Simple ● Table ○ Morphophonemic ○ Form Selection

Type of Modification: ○ Prefix ○ Suffix □ Reduplication ○ Infix ○ Circumfix ● New Translation ○ Add Word

Trigger Word: 25. *be* - to be □ Excluded

Verb Target Tense = Present

Verb Number of Subject = Singular

Verb Number of Subject = Plural

	1. Singular	2. Plural
1. First Person	*am*	*are*
2. Second Person	*are*	*are*
3. Third Person	*is*	*are*

Verb Person of Subject = Third

Comment:

Figure 4-33. An English Table Spellout Rule that Inserts the Proper Form of *to be*

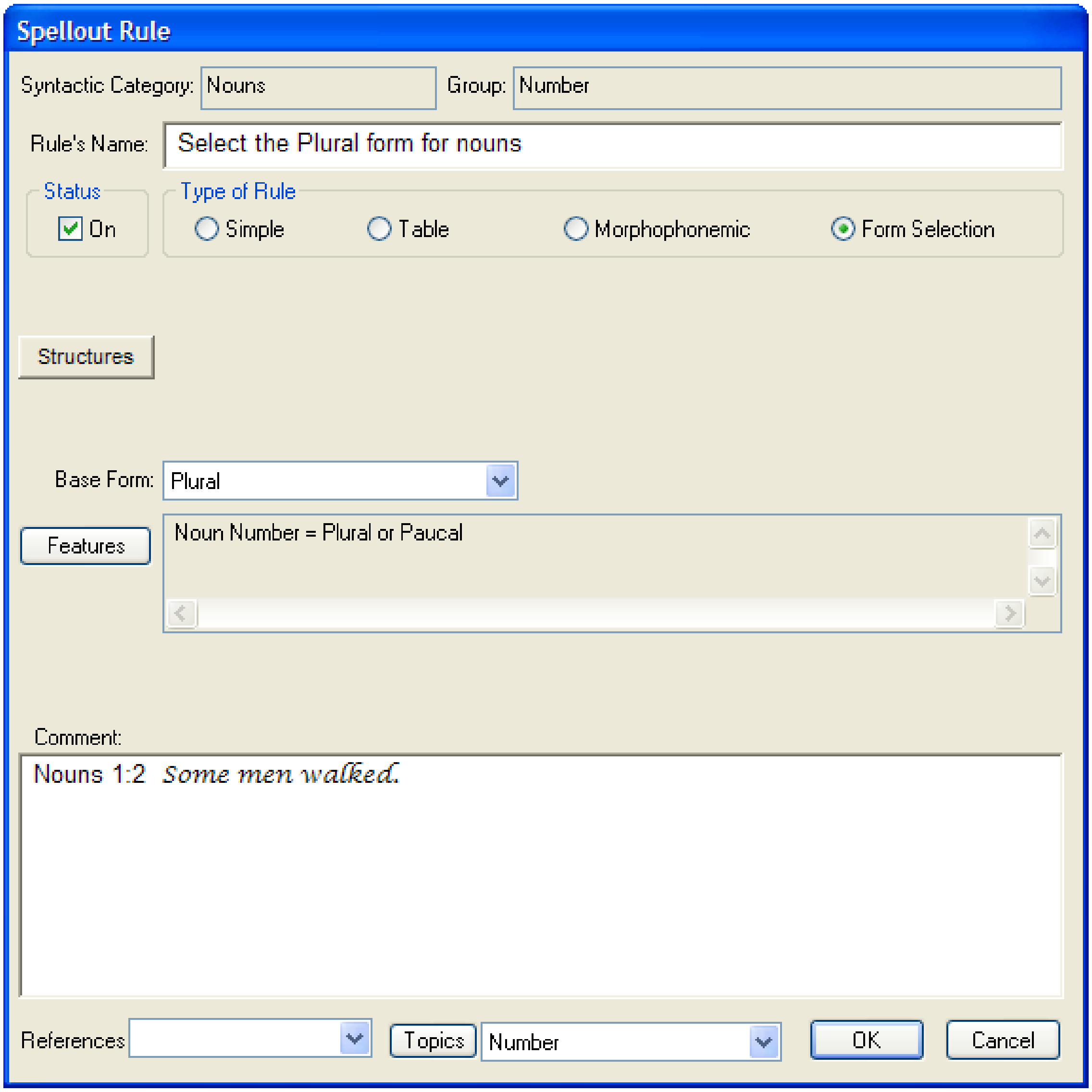
The table in the rule above has a column for singular subjects, and another column for plural subjects. It alsot has a row for each person of the subject noun. When this rule is executed, TTA will first check to see if the current verb in the semantic representation is mapped to the target verb *to be*. If the target verb is not *to be,* then this rule won't apply to the verb. If the target verb is *to be*, then this rule will walk through the table searching for cells where both the row's features and the column's features match the features of the current verb in the semantic representation. The form of *to be* in the last cell of the table where all the features match will be inserted into the text.

4.4.2.3 Morphophonemic Spellout Rules

After either a table spellout rule or a simple spellout rule has added an affix to a stem, quite often morphonemic operations must be applied to either the affix, the stem, or both. Niether of the test languages required any spellout morphophonemic rules, but English required many lexical morphophonemic rules as was shown above in figure 4-5.

4.4.2.4 Form Selection Spellout Rules

As was described above in section 4.2.2, linguists are able to enter their target language words into TTA's lexicon. If they decide that lexical forms are necessary because there are words with suppletive forms, then spellout rules are used to select those lexical forms in particular environments. The spellout rule that selects the plural form for English nouns is shown below in figure 4-34.



Spellout Rule

Syntactic Category: Nouns

Group: Number

Rule's Name: Select the Plural form for nouns

Status: ☑ On

Type of Rule: ○ Simple ○ Table ○ Morphophonemic ● Form Selection

Base Form: Plural

Noun Number = Plural or Paucal

Comment:

Nouns 1:2 *Some men walked.*

References:

Topics: Number

Figure 4-34. The English Spellout Rule that Selects the Plural Lexical Form for Nouns

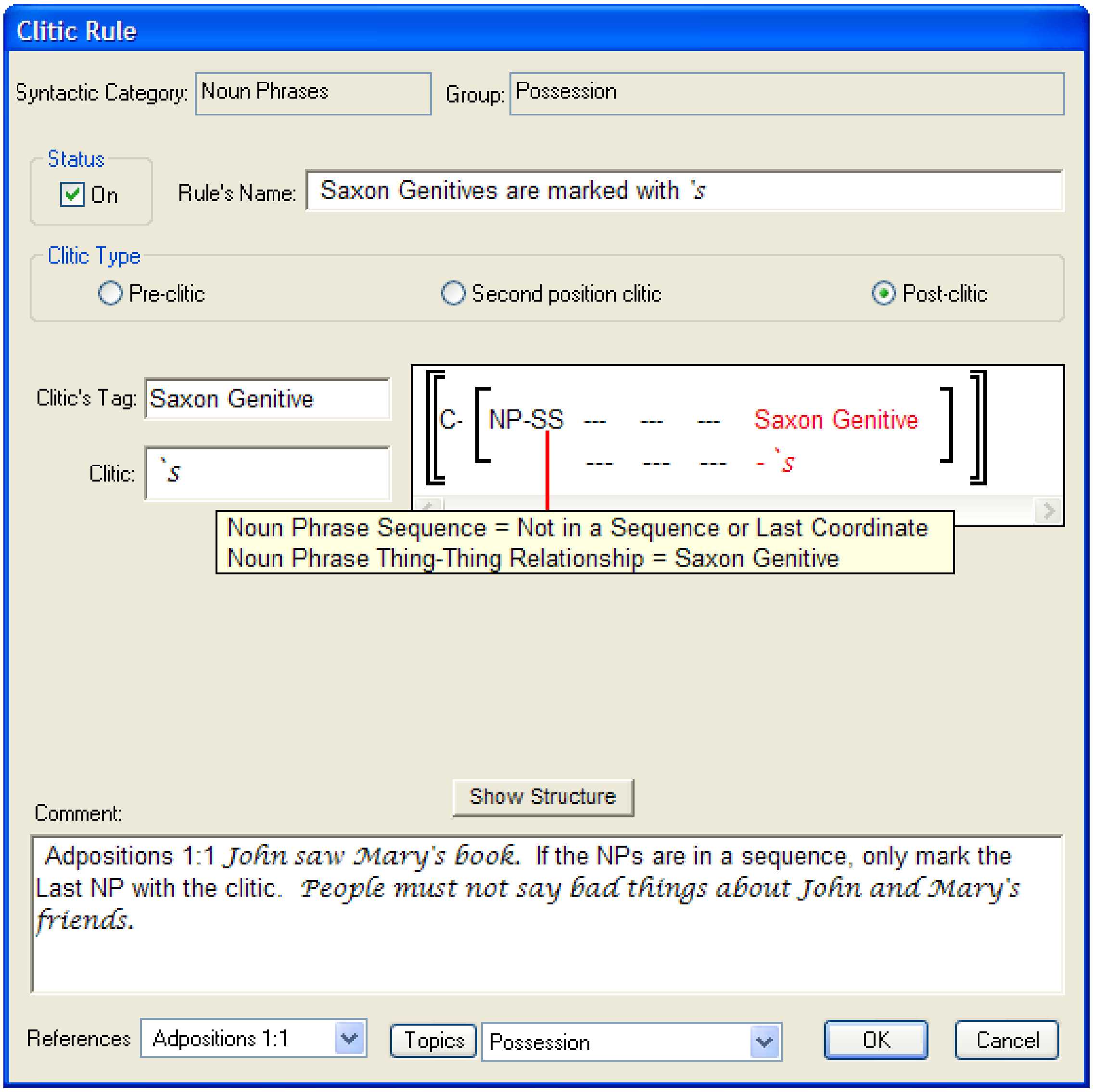
In the English lexicon, all of the plural forms were generated by rules, and all of the suppletive forms were manually entered. When this rule in figure 4-34 is executed, it looks for a noun in the semantic representations that has a Number value of either Plural or Paucal. When such a noun is found, this rule selects the Plural form of the specified target noun from the lexicon and inserts it into the generated text.

4.4.3 Clitic Rules

A clitic is defined as a morpheme that functions at a phrasal or clausal level, but which binds phonologically to another word (Payne 1997:22). Extensive typological research indicates that clitics may attach in three different locations:

- Pre-clitics attach at the beginning of the first word in the phrase or clause,
- Second position clitics attach at the end of the first word in the phrase or clause, and
- Post-clitics attach at the end of the last word in the phrase or clause (Payne 1997:22).

The rules in TTA's synthesizing grammar which insert clitics take advantage of this typological research and allow linguists to enter the clitic, its type, and a tag for the clitic. Shown below in figure 4-35 is an English clitic rule which adds the post-clitic –*'s* to NPs which are marked with a 'Thing-Thing Relationship' value of 'Saxon Genitive.' If there is a sequence of NPs marked with 'Saxon Genitive', this rule will insert the clitic into only the last NP of the sequence.



Syntactic Category: Noun Phrases

Group: Possession

Status: ☑ On

Rule's Name: Saxon Genitives are marked with 's

Clitic Type: ○ Pre-clitic ○ Second position clitic ● Post-clitic

Clitic's Tag: Saxon Genitive

Clitic: `s

[[C- [NP-SS --- --- --- Saxon Genitive]]
--- --- --- - `s

Noun Phrase Sequence = Not in a Sequence or Last Coordinate
Noun Phrase Thing-Thing Relationship = Saxon Genitive

Comment: Adpositions 1:1 *John saw Mary's book.* If the NPs are in a sequence, only mark the Last NP with the clitic. *People must not say bad things about John and Mary's friends.*

References: Adpositions 1:1

Topics: Possession

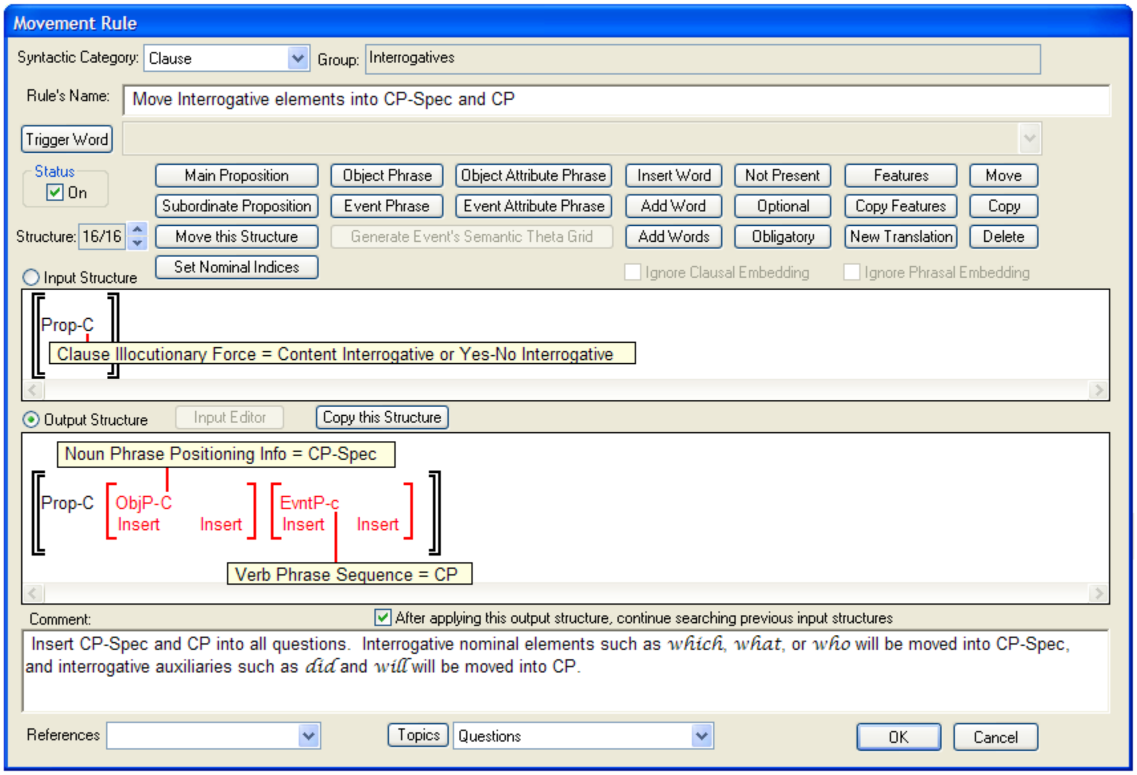
Figure 4-35. A Clitic Rule for English that Inserts the Post-clitic –*'s*

As seen in the rule above, TTA permits linguists to specify the three different types of clitics in accord with the typological research.

4.4.4 Movement Rules

Not all grammatical theories ascribe to movement[29], but many theories do, so movement rules were added to TTA's synthesizing grammar to accommodate linguists who ascribe to a theory that includes movement. Movement rules are able to move one or more constituents from one location in a proposition to some other location in that proposition. If the destination of the movement is not currently available in the proposition, then the movement rule is able to insert the destination. For example, the grammar that was developed for English loosely resembles the Principles and Parameters model described by Haegeman (Haegeman 1994). Therefore, in sentences such as *Will John read that book?*, the auxiliary *will* is generated under the INFL node of I', but then moved to the empty position dominated by C (Haegeman 1994:301). Since the propositions in the semantic representations do not contain CP or CP-Spec, those nodes must first be inserted into all questions by the English grammar. It was decided for this project to make the CP-Spec node an object phrase because they always contain words that are associated with objects (e.g., *which, what, who, to whom,* etc.). The constituents in the CP-Spec node will later be ordered by the phrase structure rule for NPs, so making CP-Spec an object phrase was a reasonable approximation. Similarly it was decided to make the CP node an event phrase because they always contain words that are associated with events (e.g., *will, did, are, should,* etc.). Shown below in figure 4-36 is the movement rule that inserts CP-Spec and CP into the semantic representations for all propositions that are questions.

[29] Generalized Phrase Structure Grammar (Gazdar et al. 1985:138), Role and Reference Grammar (Van Valin 2001:209), Functional-Typological Grammar (Givón 1990), and other grammatical theories do not posit transformations, movement, or traces; therefore these theories do not include movement.



Syntactic Category: Clause Group: Interrogatives

Rule's Name: Move Interrogative elements into CP-Spec and CP

Status ☑ On

Structure: 16/16

○ Input Structure

Prop-C

Clause Illocutionary Force = Content Interrogative or Yes-No Interrogative

● Output Structure

Noun Phrase Positioning Info = CP-Spec

Verb Phrase Sequence = CP

Comment: ☑ After applying this output structure, continue searching previous input structures

Insert CP-Spec and CP into all questions. Interrogative nominal elements such as *which*, *what*, or *who* will be moved into CP-Spec, and interrogative auxiliaries such as *did* and *will* will be moved into CP.

References

Topics Questions

Figure 4-36. Movement Rule for English that inserts CP-Spec and CP into all Questions

When the rule shown above finds a proposition that has an Illocutionary Force value of either Content Interrogative or Yes-No Interrogative, it will insert an object phrase that is tagged as CP-Spec, and an event phrase tagged as CP into the proposition. Then other movement rules are able to move specified constituents into these phrases. An English rule that moves the interrogative nominal elements from object phrases to the CP-Spec phrase and the interrogative verbal elements from the event phrase to CP is shown below in figure 4-37.

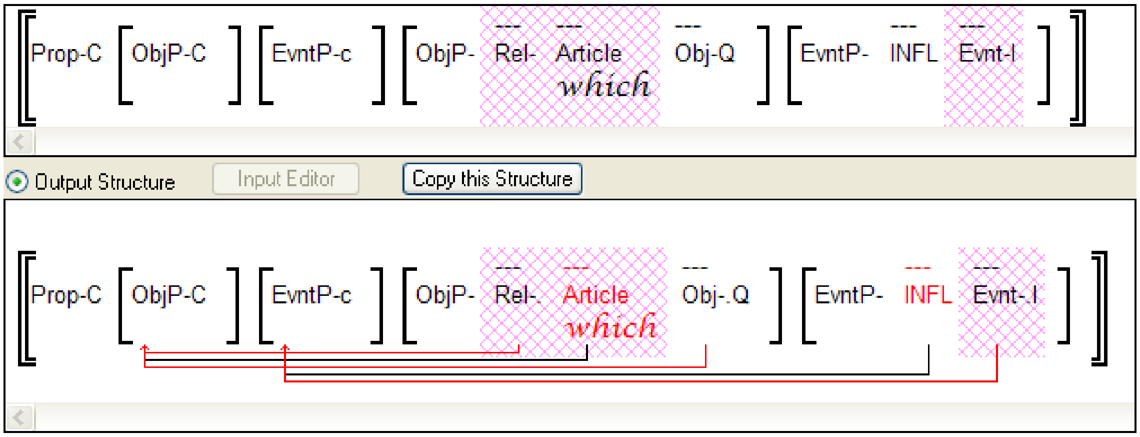
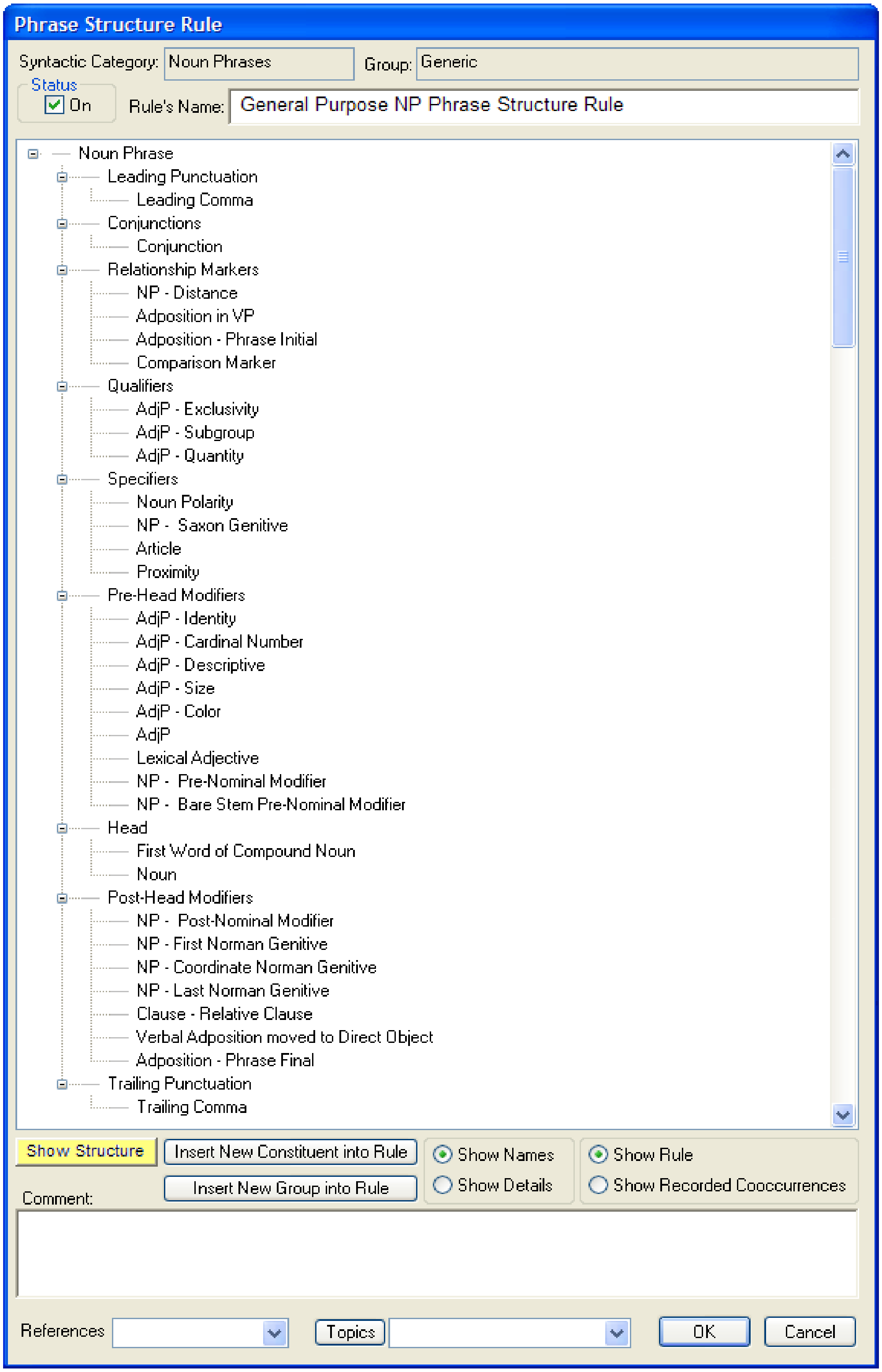


Figure 4-37. A Movement Rule for English

The rule shown above deals with propositions such as *What did John read?, Which book did John read?,* and *To whom did John give a book?*. The details of the many different movements required by English interrogatives will be discussed thoroughly in the next chapter.

4.4.5 Phrase Structure Rules

The phrase structure rules (PSRs) are responsible for positioning all of the target constituents in their proper order. Each target language grammar initially has five empty phrase structure rules: one for ordering all of the constituents in NPs, another for ordering all the constituents in VPs, another for ordering the constituents in Adjective phrases, another for Adverb phrases, and another for clauses. If there are special situations which require unique ordering, then linguists are able to add additional phrase structure rules. The phrase structure rules in TTA do not reflect the phrase structure rules of the Principles and Parameters model. Instead a very simple approach was adopted. The phrase structure rules in TTA list in the order required by the target language all of the constituents that might occur in each particular type of phrase or in a proposition. For example, the NP phrase structure rule for English is shown below in figure 4-38. That NP phrase structure rule applies to every NP, whether the NP is at the clause level or is embedded within another NP. Every constituent in that PSR is considered optional; no NP will contain all of the elements listed in that rule, but the elements that are in a particular NP in the target text will be ordered as specified by the PSR.



Syntactic Category: Noun Phrases

Group: Generic

Status: ☑ On

Rule's Name: General Purpose NP Phrase Structure Rule

- Noun Phrase
  - Leading Punctuation
    - Leading Comma
  - Conjunctions
    - Conjunction
  - Relationship Markers
    - NP - Distance
    - Adposition in VP
    - Adposition - Phrase Initial
    - Comparison Marker
  - Qualifiers
    - AdjP - Exclusivity
    - AdjP - Subgroup
    - AdjP - Quantity
  - Specifiers
    - Noun Polarity
    - NP - Saxon Genitive
    - Article
    - Proximity
  - Pre-Head Modifiers
    - AdjP - Identity
    - AdjP - Cardinal Number
    - AdjP - Descriptive
    - AdjP - Size
    - AdjP - Color
    - AdjP
    - Lexical Adjective
    - NP - Pre-Nominal Modifier
    - NP - Bare Stem Pre-Nominal Modifier
  - Head
    - First Word of Compound Noun
    - Noun
  - Post-Head Modifiers
    - NP - Post-Nominal Modifier
    - NP - First Norman Genitive
    - NP - Coordinate Norman Genitive
    - NP - Last Norman Genitive
    - Clause - Relative Clause
    - Verbal Adposition moved to Direct Object
    - Adposition - Phrase Final
  - Trailing Punctuation
    - Trailing Comma

● Show Names ○ Show Details

● Show Rule ○ Show Recorded Cooccurrences

Comment:

References

Figure 4-38. The Phrase Structure Rule for English NPs

In the phrase structure rule shown above, linguists are able to enter groups, and then enter constituents into the groups.

When the rule above is applied to a particular NP in the target text, all of the constituents in that NP are removed and inserted into a buffer area. Then TTA walks through the NP phrase structure rule looking for the constituents specified in the rule. So when the rule shown above in figure 4-38 is applied to a particular NP, TTA will first search for all the constituents in the buffer area that are labeled "Leading Comma." If there are one or more constituents labeled as "Leading Comma" in the buffer area, they will be moved back into the NP and placed at the beginning. Then TTA will search the buffer area for all constituents labeled as "Conjunction." If there are one or more conjunctions in the buffer area, they will be moved back into the NP and placed after the Leading Commas. Then TTA will search through the buffer area for each of the remaining constituents specified in the rule. Whenever it finds a constituent in the buffer area that matches the constituent specified in the rule, that constituent will be moved from the buffer area back into the NP and placed at the end of the NP. After this process has been completed, all of the constituents in the NP will be in their proper order as specified by the rule. After the rule has been completed, if there are still constituents remaining in the buffer area because they were not specified in the rule, those constituents will be inserted at the end of the NP. Then this process will be repeated for the next NP in the text. After all of the NPs have had their constituents ordered properly, the VP phrase structure rule will be applied to all the VPs. Then the adjective phrase PSR will be applied to all the adjective phrases, and the adverb phrase PSR will be applied to all the adverb phrases. Finally the clause PSR will be applied to all the clauses. The result will be that all the constituents in the target text are in their proper order as specified by the various PSRs.

4.4.6 Pronoun and Switch Reference Rules

Every language has its own set of rules for determining where pronouns may or may not be used. Therefore there are no pronouns in the semantic representations. As was mentioned in section 3.3.2.1.5, every object in the semantic representations has a feature called ‘Person’, and two of the values are ‘First’ and ‘Second’. However, even when it comes to first and second person pronouns, each language has its own rules and forms for pronouns. For example, in Korean when someone is talking to an older person, he must use a title rather than a second person pronoun. Many languages have two different forms for the first person singular pronoun, one indicating deferential speech, the other being neutral. For these reasons linguists must develop rules to determine where pronouns may be used and the proper form for each pronoun. In order to facilitate this process, the synthesizing grammar in TTA has two sections under the Pronoun and Switch Reference Rules. The first section allows linguists to write rules which identify where pronouns may be used. The second section has spellout rules to insert the proper form of each pronoun into the text. Since third person pronouns cannot be determined until after all the constituents are in their proper order, these pronoun generation rules must follow the PSRs.

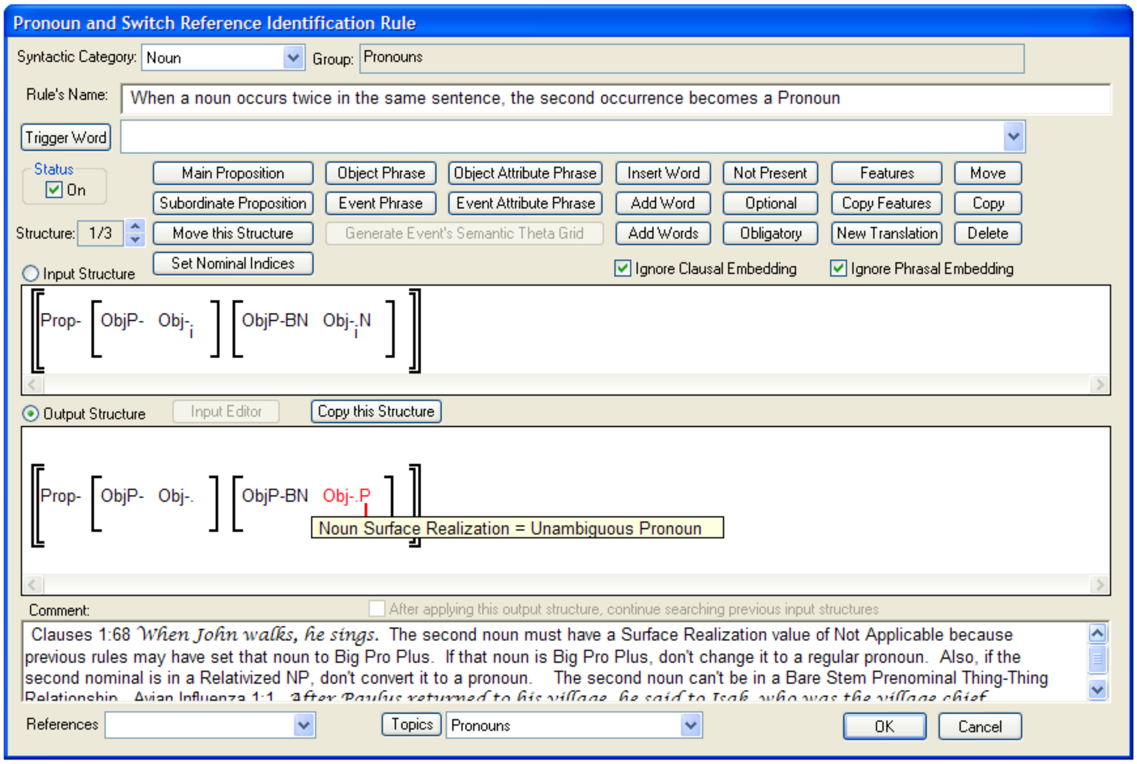
Switch reference markers are somewhat similar to pronouns in that they cannot be determined until after all the constituents are in their final order. The process for determining where to use switch reference morphemes is essentially identical to the process for determining where to use third person pronouns, so these two processes are combined into one section in TTA’s grammar. After identifying all the locations where switch reference markers should be inserted, the process of adding the morphemes is again essentially identical to the process of inserting pronouns. Therefore pronouns and switch reference markers are both added in the same two step process:

- identify where the pronouns and switch reference markers should be inserted, and
- insert the proper surface form for each pronoun and switch reference marker.

The rules that identify where to use pronouns and switch reference markers are identical to the structural adjustment rules that were described above in section 4.3.9. The rules that insert the pronouns and switch reference morphemes are identical to the spellout rules that were described in section 4.4.2 above. These two processes will be described separately here.

4.4.6.1 Pronoun and Switch Reference Identification Rules

A typical third person pronoun identification rule is shown below in figure 4-39. That rule looks for two occurrences of the same nominal occurring in a sentence, and then the output structure sets the feature called Surface Realization of the second nominal to Unambiguous Pronoun.



Pronoun and Switch Reference Identification Rule

Syntactic Category: Noun
Group: Pronouns
Rule's Name: When a noun occurs twice in the same sentence, the second occurrence becomes a Pronoun
Status ☑ On
Structure: 1/3
○ Input Structure
☑ Ignore Clausal Embedding ☑ Ignore Phrasal Embedding
● Output Structure
Noun Surface Realization = Unambiguous Pronoun
Comment: □ After applying this output structure, continue searching previous input structures

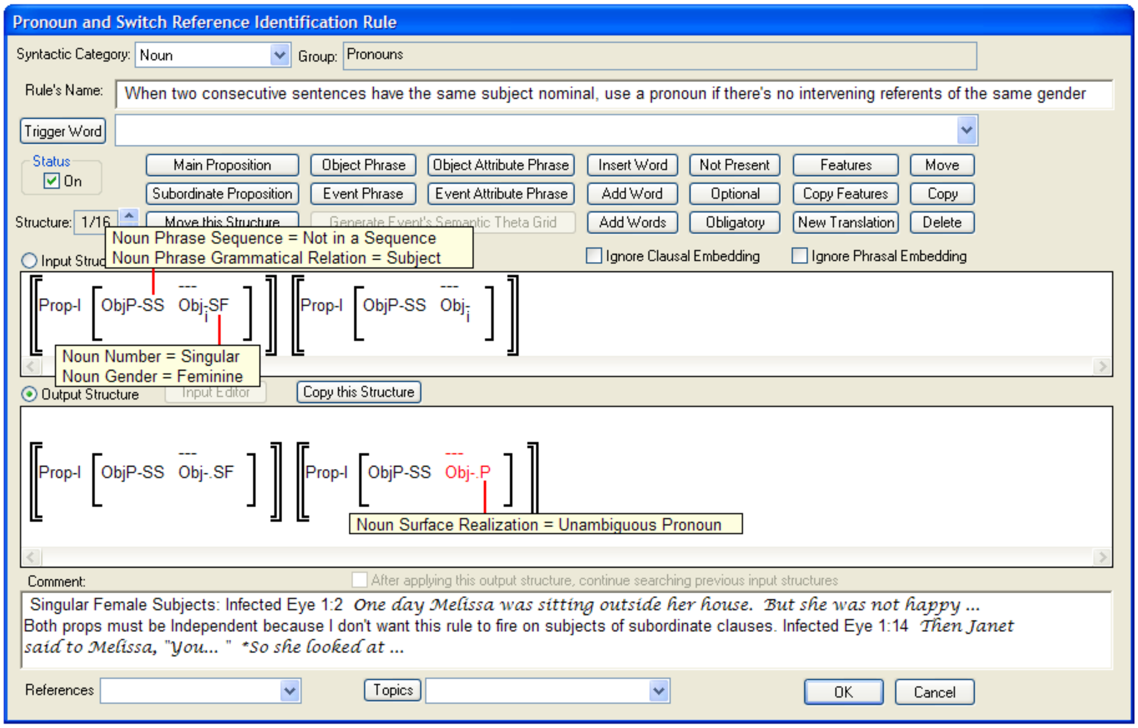
Clauses 1:68 *When John walks, he sings.* The second noun must have a Surface Realization value of Not Applicable because previous rules may have set that noun to Big Pro Plus. If that noun is Big Pro Plus, don't change it to a regular pronoun. Also, if the second nominal is in a Relativized NP, don't convert it to a pronoun. The second noun can't be in a Bare Stem Prenominal Thing-Thing Relationship. Avian Influenza 1:1 *After Paulus returned to his village, he said to Isak, who was the village chief*

References:
Topics: Pronouns

Figure 4-39. A Third Person Pronoun Identification Rule that Searches within a Single Sentence

In the input structure of the rule shown above, note that the two nouns both have indexes of “i”. This indicates that they must have the same Object Index value as was described in section 3.3.2.1. Also note that the options “Ignore Clausal Embedding” and “Ignore Phrasal Embedding” located immediately above the input structure are both checked. These options indicate that when TTA is searching through the underlying representations, it may search embedded clauses and embedded phrases for a nominal that has the same index as in the first NP.

Another common third person pronoun identification rule is shown below in figure 4-40. That rule searches for two independent propositions in a sequence which have the same subject nominal.


Pronoun and Switch Reference Identification Rule

Syntactic Category: Noun Group: Pronouns

Rule's Name: When two consecutive sentences have the same subject nominal, use a pronoun if there's no intervening referents of the same gender

Status ☑ On

Structure: 1/16

○ Input Structure

Noun Phrase Sequence = Not in a Sequence
Noun Phrase Grammatical Relation = Subject

□ Ignore Clausal Embedding □ Ignore Phrasal Embedding

Noun Number = Singular
Noun Gender = Feminine

● Output Structure

Noun Surface Realization = Unambiguous Pronoun

Comment: □ After applying this output structure, continue searching previous input structures

Singular Female Subjects: Infected Eye 1:2 *One day Melissa was sitting outside her house. But she was not happy ...*
Both props must be Independent because I don't want this rule to fire on subjects of subordinate clauses. Infected Eye 1:14 *Then Janet said to Melissa, "You... " *So she looked at ...*

References Topics

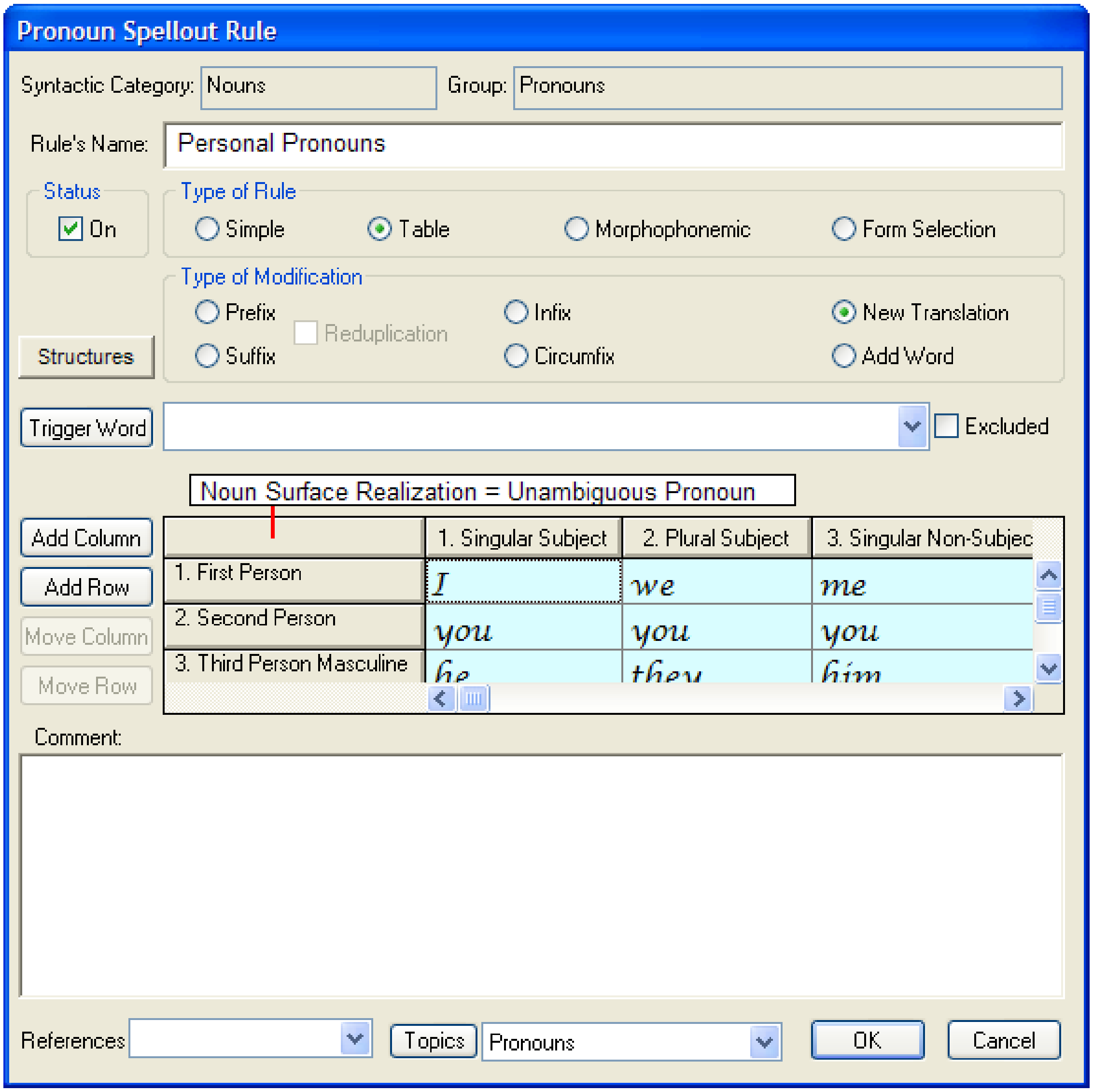
Figure 4-40. A Third Person Pronoun Identification Rule that Searches Two Sentences

In the rule shown above, there are sixteen different structures as indicated in the upper left corner. The first structure searches for singular female subjects that have the same Object Index value, another structure searches for singular masculine subjects with the same Object Index value, another searches for plural subjects, etc. When these situations are found in the underlying representations, the output structure will set the second nominal's Surface Realization feature to Unambiguous Pronoun. A subsequent rule will then search for an intervening noun with the same gender. If a noun with the same gender is found between the original two nouns, the second noun will have its Surface Realization feature set back to Not Applicable. Note that the pronoun spellout rules in the next section are set to only apply to the nouns that have a Surface Realization value of Unambiguous Pronoun.

Properly identifying where pronouns should be used in a particular language is a very difficult task, and rules are not always able to identify where mother-tongue speakers prefer or expect to use a pronoun. The guideline that has been adopted for this problem is to write rules that generate enough pronouns so that the texts are somewhat natural, but the rules must not generate too many pronouns, because that causes the texts to be misunderstood. Therefore these rules are used to identify and generate the obvious pronouns, but the texts must be edited by mother-tongue speakers to insert the less obvious pronouns.

4.4.6.2 Pronoun and Switch Reference Spellout Rules

The pronoun spellout rules are identical to the spellout rules that were described above in section 4.4.2, so they will not be described again here, but two examples will be provided. Shown below in figure 4-41 is the spellout rule for English personal pronouns. Note that in the upper left corner of the grid, the feature specifies that this rule only applies to nouns that have a Surface Realization feature value of Unambiguous Pronoun. The features in the upper left corner of the table must be satisfied before the rule will start searching the rows and columns for additional matching features.



Pronoun Spellout Rule

Syntactic Category: Nouns Group: Pronouns

Rule's Name: Personal Pronouns

Status: ☑ On

Type of Rule: ○ Simple ● Table ○ Morphophonemic ○ Form Selection

Type of Modification: ○ Prefix ○ Suffix □ Reduplication ○ Infix ○ Circumfix ● New Translation ○ Add Word

Trigger Word: □ Excluded

Noun Surface Realization = Unambiguous Pronoun

	1. Singular Subject	2. Plural Subject	3. Singular Non-Subjec
1. First Person	I	we	me
2. Second Person	you	you	you
3. Third Person Masculine	he	they	him

Comment:

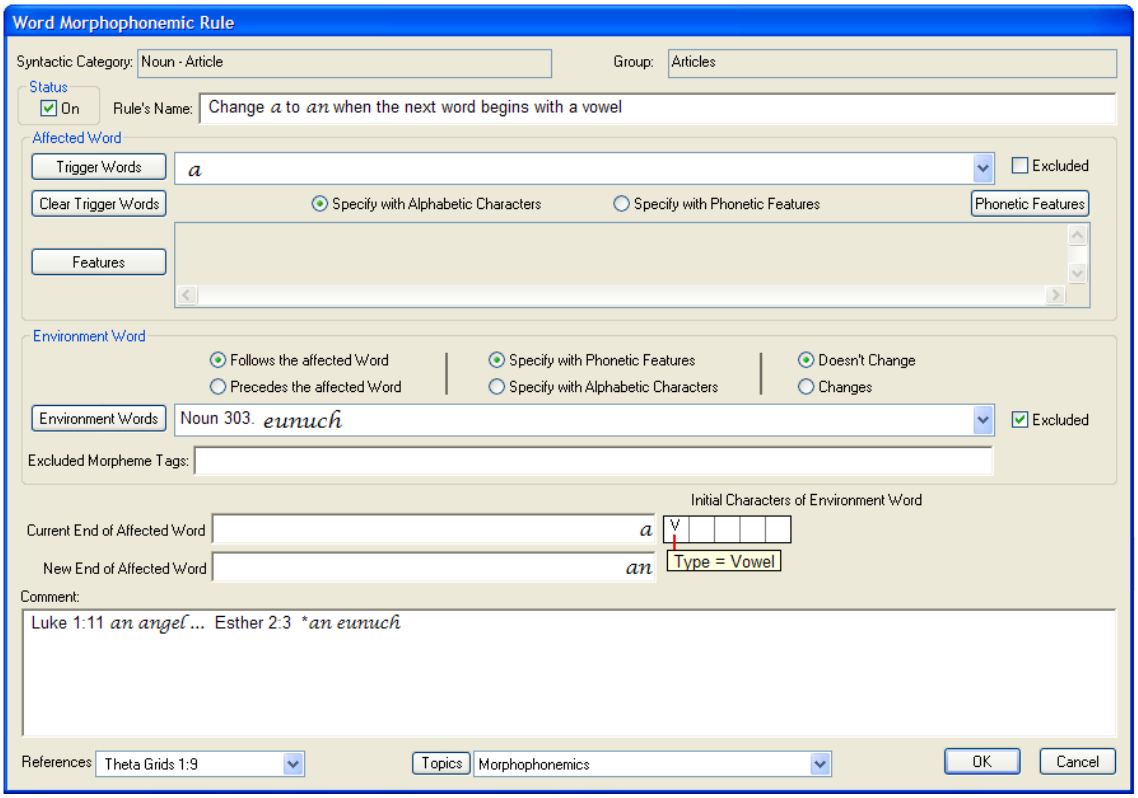
References Topics: Pronouns

Figure 4-41. A Pronoun Spellout Rule that Inserts English Personal Pronouns

Because the pronoun rules occasionally insert new constituents into the text, the phrase structure rules must be executed again to position the new constituents properly. Therefore, as was shown in figure 4-30 above, after the pronoun rules have been executed, the phrase structure rules are executed again.

4.4.7 Word Morphophonemic Rules

The word morphophonemic rules are somewhat similar to the spellout morphophonemic rules that were described in section 4.4.2.3 above. However, the word morphophonemic rules apply across word boundaries rather than morpheme boundaries. Therefore the word morphophonemic rules must be executed after the PSRs have ordered all of the target constituents properly. Word morphophonemic rules let linguists specify the category of the word that is affected by its environment. Then the rule describes how that word is to be changed using either alphabetic characters or phonetic features. A sample of a word morphophonemic rule is shown below in figure 4-42.



Word Morphophonemic Rule

Syntactic Category: Noun - Article
Group: Articles

Status: ☑ On
Rule's Name: Change *a* to *an* when the next word begins with a vowel

Affected Word

Trigger Words: *a*
□ Excluded
● Specify with Alphabetic Characters
○ Specify with Phonetic Features

Environment Word

● Follows the affected Word
○ Precedes the affected Word
● Specify with Phonetic Features
○ Specify with Alphabetic Characters
● Doesn't Change
○ Changes

Environment Words: Noun 303. *eunuch*
☑ Excluded
Excluded Morpheme Tags:

Initial Characters of Environment Word: V
Type = Vowel
Current End of Affected Word: *a*
New End of Affected Word: *an*

Comment:
Luke 1:11 *an angel …* Esther 2:3 **an eunuch*

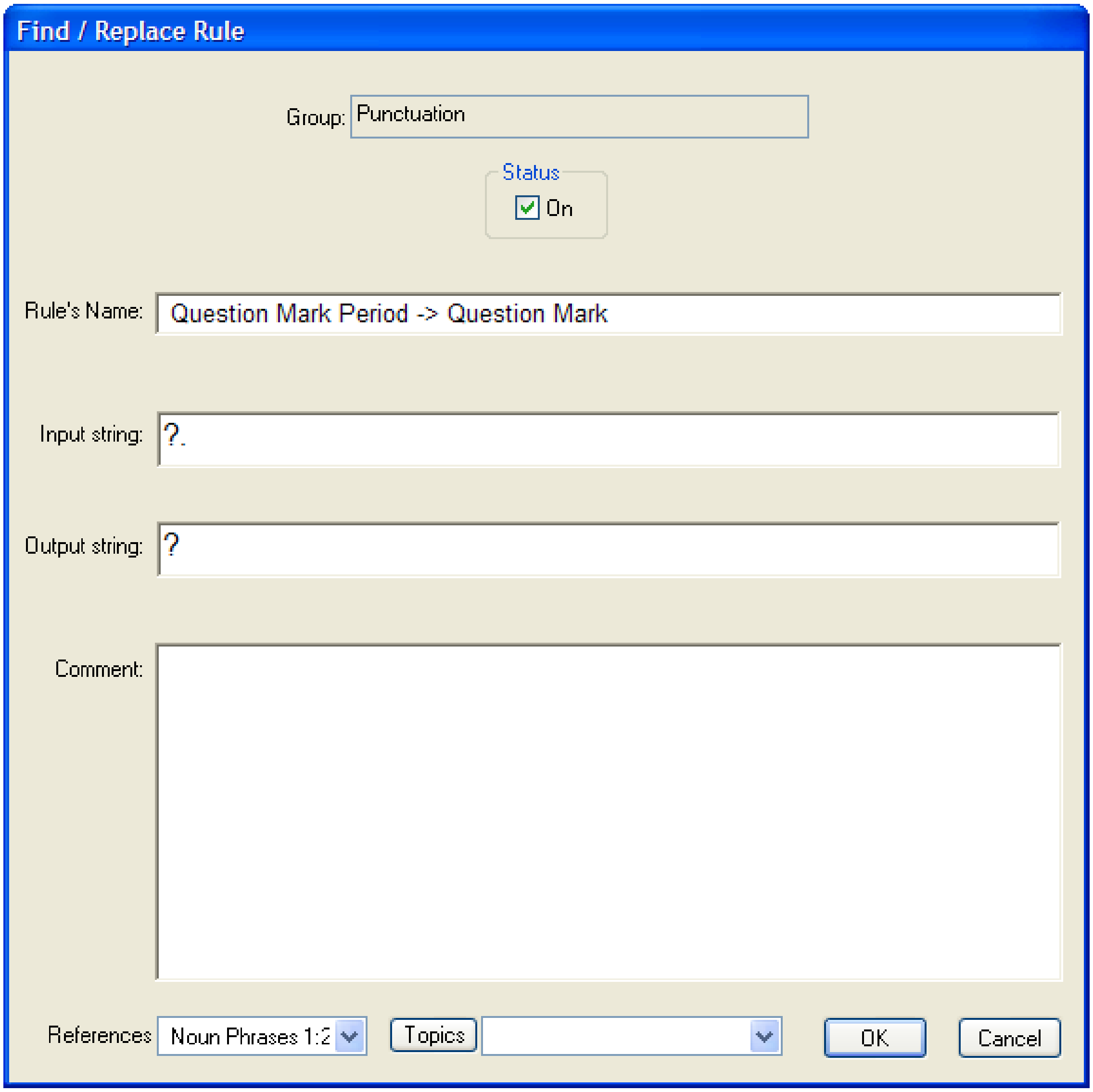
References: Theta Grids 1:9
Topics: Morphophonemics

Figure 4-42. An English Word Morphophonemic Rule that Changes *a* to *an* before Words that Start with a Vowel

As seen in the rule above, linguists are able to specify a syntactic category, in this case Articles. If necessary, linguists are able to specify one or more particular words in that category. The rule shown above only applies to the article *a*, it does not apply to the other English article *the*. When the article *a* occurs before a word that begins with a vowel, the word must be changed to *an*, as specified in the field labeled “New End of Affected Word.” Since the rule is applied to standard text rather than phonetically transcribed text, there may be exceptions. For example, the word *eunuch*, if it were written phonetically, would begin with a '*y*'. Since the text is not phonetically transcribed, the word *eunuch* begins with a vowel, but this rule must not apply to *a eunuch*. Therefore linguists are able to specify particular words that are not allowed in the environment, as shown in the rule above.

4.4.8 Find/Replace Rules

The find/replace rules are responsible for cleaning up little details in the generated text that cannot be done by the regular grammatical rules. These rules are used most frequently to correct the punctuation, but they also do other small tasks like performing English contractions such as changing *I am* to *I'm*. Punctuation marks in TTA are treated as separate words and are generally inserted by spellout rules. For example, if a language uses question marks at the end of its questions, a spellout rule must insert those question marks. Similarly if a language uses commas to separate nouns that are in a sequence such as *John, Mary, Bill, and Steve went to the store,* those commas must be inserted by rules. Therefore commas, quotation marks, question marks, and exclamation points are inserted and treated as separate words during the generation process. Since words are always followed by a space in the generated text, that space must be deleted if the next word is a punctuation mark. The find/replace rules function just like the find/replace option in word processors: they search for a specified string, and when that string is found in the generated text, it is replaced by the specified output string. The only punctuation marks that occur in the semantic representations are periods. Every proposition in the semantic representations ends with a period. If a language marks questions with a final question mark, then a rule must insert those question marks into questions, and then the PSRs must position those question marks at the end of the sentence. The period in the semantic representation is still in the generated text, so this results in a question mark followed by a period. English uses question marks at the end of their questions, so the English rule that changes the sequence "? ." to "?" is shown below in figure 4-43.



Group: Punctuation

Status

☑ On

Rule's Name: Question Mark Period -> Question Mark

Input string: ?.

Output string: ?

Comment:

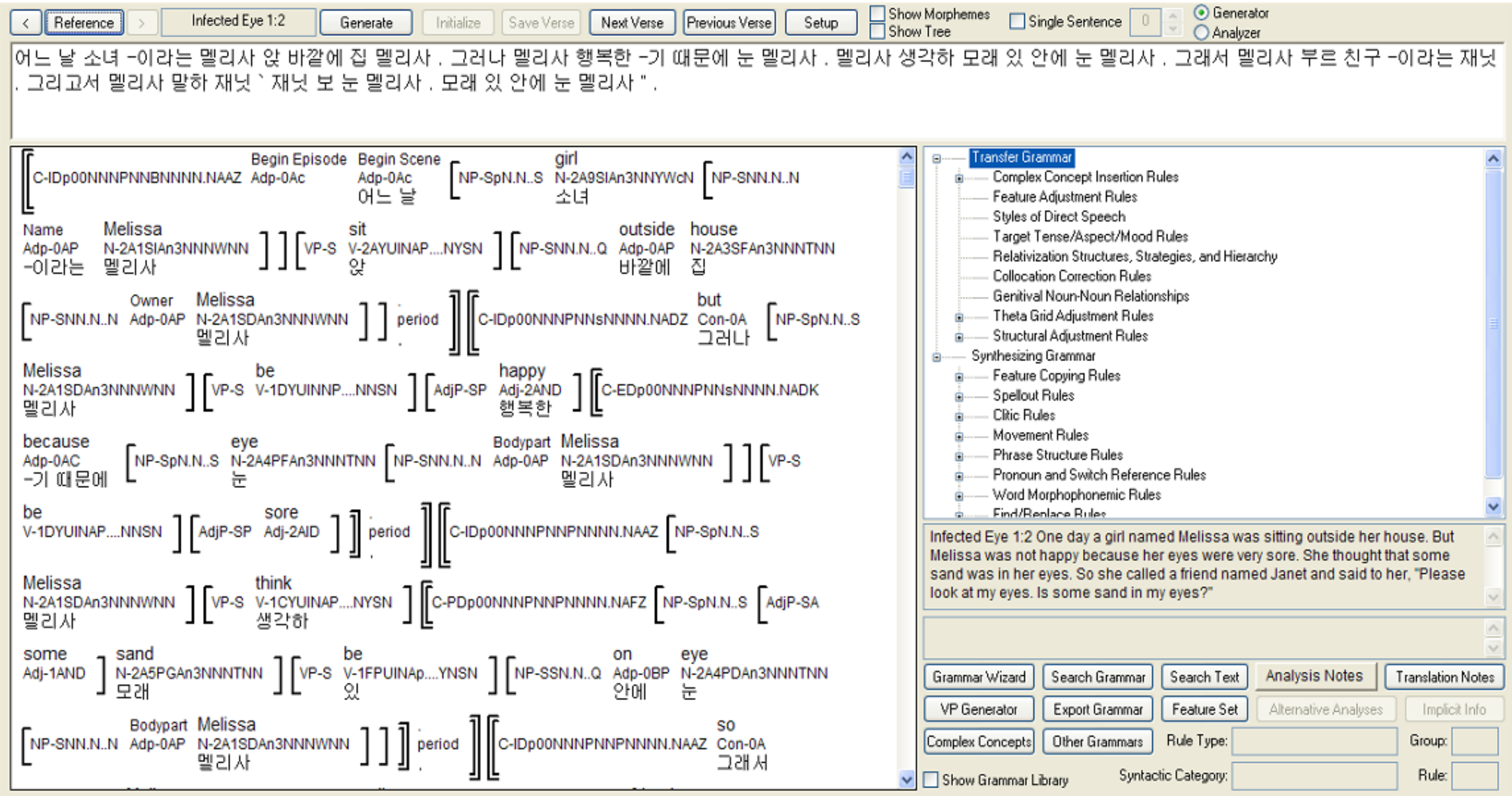
References: Noun Phrases 1:2

Figure 4-43. A Find/Replace Rule that Deletes Periods which are Preceded by Question Marks

Similar find/replace rules correct the punctuation wherever necessary.

4.5 Conclusions

After a linguist has built his target lexicon with its features and forms, and after he has developed the transfer and synthesizing grammars, TTA is able to generate target text. Figure 3-33 at the end of chapter three showed the semantic representation for Infected Eye 1:2. Shown below in figure 4-44 is the same semantic representation, but this time the source concepts have been linked to Korean words. Figure 4-44 shows the screen where linguists link source concepts to target words and build their grammars. The grammar tree is shown in the upper right part of the screen, and the ungenerated target text is shown at the top of the screen. That figure shows what the screen looks like before the linguist clicks the Generate button. After the necessary information has been entered into the target lexicon and the grammar, the linguist can click the Generate button located in the upper left corner of the screen, and TTA will execute the target grammar and display the generated text. When the Generate button is clicked, it takes approximately ten seconds to execute the Korean grammar. First the transfer grammar is executed as was described in section 4.3 above; then the synthesizing grammar is executed as described in section 4.4.


< Reference > Infected Eye 1:2 Generate Initialize Save Verse Next Verse Previous Verse Setup
□ Show Morphemes
□ Show Tree
□ Single Sentence 0
● Generator
○ Analyzer

어느 날 소녀 -이라는 멜리사 앉 바깥에 집 멜리사 . 그러나 멜리사 행복한 -기 때문에 눈 멜리사 . 멜리사 생각하 모래 있 안에 눈 멜리사 . 그래서 멜리사 부르 친구 -이라는 재닛 . 그리고서 멜리사 말하 재닛 ` 재닛 보 눈 멜리사 . 모래 있 안에 눈 멜리사 " .

```
                         Begin Episode Begin Scene         girl
[[C-IDp00NNNPNNBNNNN.NAAZ Adp-0Ac       Adp-0Ac      [NP-SpN.N..S N-2A9SIAn3NNYWcN [NP-SNN.N..N
                                        어느 날                    소녀

Name     Melissa                         sit                                outside   house
Adp-0AP  N-2A1SIAn3NNNWNN ] ] [VP-S V-2AYUINAP....NYSN ] [NP-SNN.N..Q Adp-0AP   N-2A3SFAn3NNNTNN
-이라는   멜리사                          앉                                  바깥에    집

             Owner    Melissa                   .                                 but
[NP-SNN.N..N Adp-0AP  N-2A1SDAn3NNNWNN ] ] period ]] [[C-IDp00NNNPNNsNNNN.NADZ Con-0A  [NP-SpN.N..S
                      멜리사                     .                                 그러나

Melissa                   be                         happy
N-2A1SDAn3NNNWNN ] [VP-S V-1DYUINNP....NNSN ] [AdjP-SP Adj-2AND ] [[C-EDp00NNNPNNsNNNN.NADK
멜리사                                                 행복한

because              eye                                Bodypart Melissa
Adp-0AC [NP-SpN.N..S N-2A4PFAn3NNNTNN [NP-SNN.N..N Adp-0AP  N-2A1SDAn3NNNWNN ] ] [VP-S
-기 때문에            눈                                          멜리사

be                              sore        .
V-1DYUINAP....NNSN ] [AdjP-SP Adj-2AID ] ] period ]] [[C-IDp00NNNPNNPNNNN.NAAZ [NP-SpN.N..S
                                            .

Melissa                   think
N-2A1SDAn3NNNWNN ] [VP-S V-1CYUINAP....NYSN ] [[C-PDp00NNNPNNPNNNN.NAFZ [NP-SpN.N..S [AdjP-SA
멜리사                     생각하

some       sand                      be                                  on       eye
Adj-1AND ] N-2A5PGAn3NNNTNN ] [VP-S V-1FPUINAp....YNSN ] [NP-SSN.N..Q Adp-0BP  N-2A4PDAn3NNNTNN
           모래                      있                                   안에     눈

             Bodypart Melissa                   .                                  so
[NP-SNN.N..N Adp-0AP  N-2A1SDAn3NNNWNN ] ] ] period ]] [[C-IDp00NNNPNNPNNNN.NAAZ Con-0A
                      멜리사                     .                                  그래서
```

Transfer Grammar
Complex Concept Insertion Rules
Feature Adjustment Rules
Styles of Direct Speech
Target Tense/Aspect/Mood Rules
Relativization Structures, Strategies, and Hierarchy
Collocation Correction Rules
Genitival Noun-Noun Relationships
Theta Grid Adjustment Rules
Structural Adjustment Rules
Synthesizing Grammar
Feature Copying Rules
Spellout Rules
Clitic Rules
Movement Rules
Phrase Structure Rules
Pronoun and Switch Reference Rules
Word Morphophonemic Rules
Find/Replace Rules

Infected Eye 1:2 One day a girl named Melissa was sitting outside her house. But Melissa was not happy because her eyes were very sore. She thought that some sand was in her eyes. So she called a friend named Janet and said to her, "Please look at my eyes. Is some sand in my eyes?"

Figure 4-44. The Lexicon and Grammar Development Screen before clicking the Generate Button

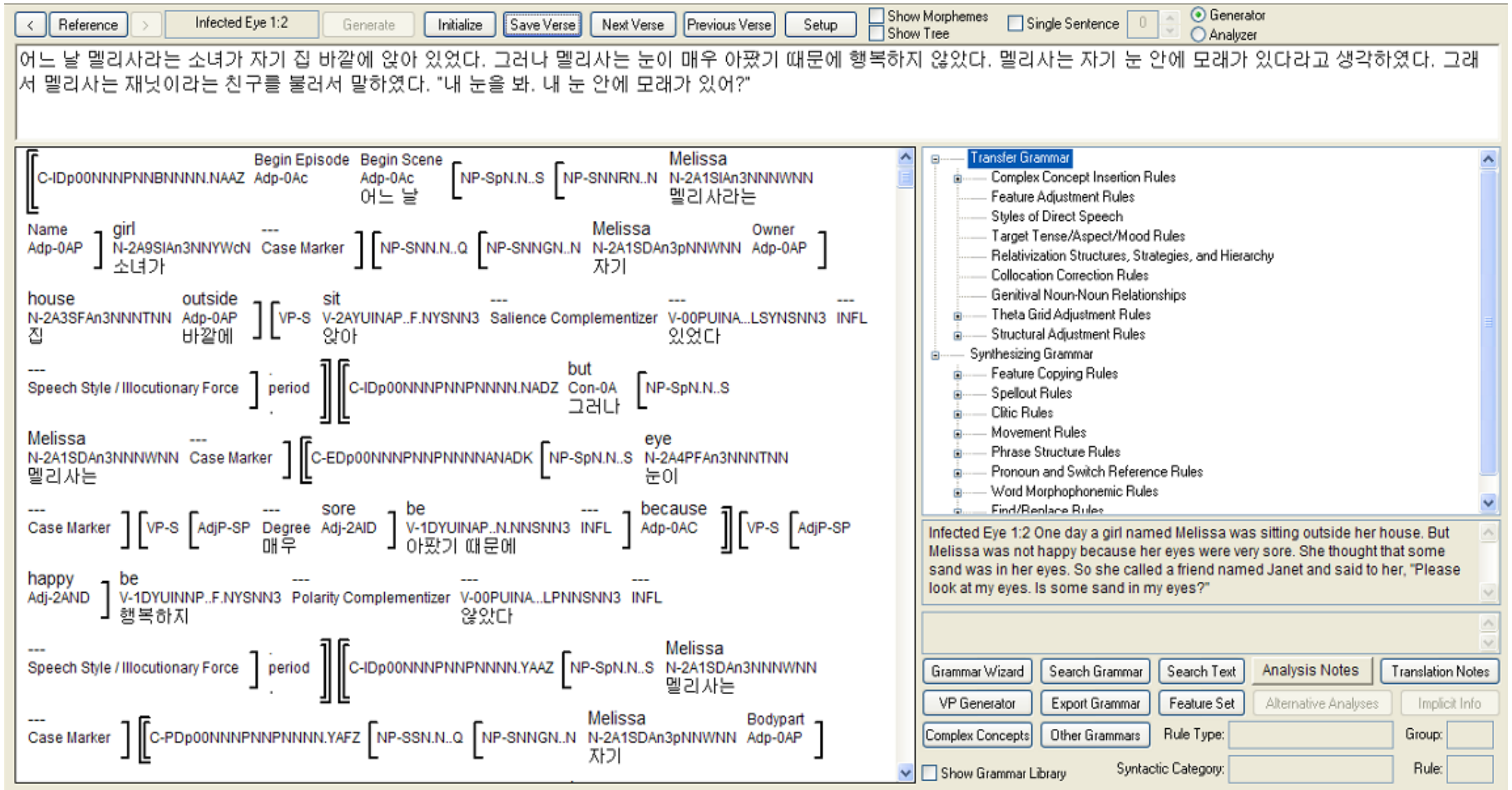
Figure 4-45 on the following page shows the results after the Korean grammar has been executed. As seen in the large window, the semantic representation has been restructured so that it consists of the target language's lexemes, structures, and features. For example, *be sore* has been converted from the predicative adjective construction [VP BE-D] [AdjP SORE-A] that occurs in the semantic representation to the Korean verb 아프다 [a peu da] 'to be sore or sick.' Similarly *be happy* has been converted from a predicative adjective construction to the Korean verb 행복하다 [haeng bok ha da] 'to be happy.' The constituents have been ordered as specified by the Korean PSRs, and the pronoun 자기 [ja gi] has been inserted where appropriate. When TTA generates this verse in English, the final result is:

> One day a girl named Melissa was sitting outside her house. But Melissa was not happy because her eyes were very sore. She thought that some sand was in her eyes. So she called a friend named Janet and said to her, "Please look at my eyes. Is some sand in my eyes?"

When TTA generates this verse in Korean, the final result is:

> 어느 날 멜리사라는 소녀가 자기 집 바깥에 앉아 있었다. 그러나 멜리사는 눈이 매우 아팠기 때문에 행복하지 않았다. 멜리사는 자기 눈 안에 모래가 있다라고 생각하였다. 그래서 멜리사는 재닛이라는 친구를 불러서 말하였다. "내 눈을 봐. 내 눈 안에 모래가 있어?"

This result is seen in the window at the top of the screen in figure 4-45 below.



어느 날 멜리사라는 소녀가 자기 집 바깥에 앉아 있었다. 그러나 멜리사는 눈이 매우 아팠기 때문에 행복하지 않았다. 멜리사는 자기 눈 안에 모래가 있다라고 생각하였다. 그래서 멜리사는 재닛이라는 친구를 불러서 말하였다. "내 눈을 봐. 내 눈 안에 모래가 있어?"

```
[[                         Begin Episode  Begin Scene                          Melissa
  C-IDp00NNNPNNBNNNN.NAAZ  Adp-0Ac        Adp-0Ac      [NP-SpN.N..S [NP-SNNRN..N  N-2A1SIAn3NNNWNN
                                          어느 날                                  멜리사라는

Name       girl                ---                           Melissa              Owner
Adp-0AP ]  N-2A9SIAn3NNYWcN    Case Marker ][NP-SNN.N..Q [NP-SNNGN..N  N-2A1SDAn3pNNWNN  Adp-0AP ]
           소녀가                                            자기

house              outside            sit                    ---                       ---                  ---
N-2A3SFAn3NNNTNN   Adp-0AP  ][VP-S    V-2AYUINAP..F.NYSNN3   Salience Complementizer   V-00PUINA...LSYNSNN3 INFL
집                 바깥에              앉아                                              있었다

---                                   .                                     but
Speech Style / Illocutionary Force ]  period  ]][[C-IDp00NNNPNNPNNNN.NADZ   Con-0A   [NP-SpN.N..S
                                      .                                     그러나

Melissa            ---                                                      eye
N-2A1SDAn3NNNWNN   Case Marker ][[C-EDp00NNNPNNPNNNNANADK [NP-SpN.N..S      N-2A4PFAn3NNNTNN
멜리사는                                                                     눈이

---                                  ---      sore          be                   ---        because
Case Marker ][VP-S [AdjP-SP          Degree   Adj-2AID  ]   V-1DYUINAP..N.NNSNN3 INFL  ]    Adp-0AC  ]][VP-S [AdjP-SP
                                     매우                    아팠기 때문에

happy       be                     ---                       ---                    ---
Adj-2AND ]  V-1DYUINNP..F.NYSNN3   Polarity Complementizer   V-00PUINA...LPNNSNN3   INFL
            행복하지                                           않았다

---                                   .                                                       Melissa
Speech Style / Illocutionary Force ]  period  ]][[C-IDp00NNNPNNPNNNN.YAAZ [NP-SpN.N..S       N-2A1SDAn3NNNWNN
                                      .                                                       멜리사는

---                                                                          Melissa            Bodypart
Case Marker ][[C-PDp00NNNPNNPNNNN.YAFZ [NP-SSN.N..Q [NP-SNNGN..N             N-2A1SDAn3pNNWNN   Adp-0AP ]
                                                                             자기
```

Infected Eye 1:2 One day a girl named Melissa was sitting outside her house. But Melissa was not happy because her eyes were very sore. She thought that some sand was in her eyes. So she called a friend named Janet and said to her, "Please look at my eyes. Is some sand in my eyes?"

Rule Type:

Group:

Syntactic Category:

Rule:

Figure 4-45. The Lexicon and Grammar Development Screen after clicking the Generate Button

This chapter described TTA's target lexicon, the transfer grammar, and the synthesizing grammar. The target lexicon enables linguists to enter their stems, and then define features and forms for those stems. The non-suppletive lexical forms are generated by lexical spellout rules, while the suppletive forms are entered into the lexicon manually. The transfer grammar consists of nine different rule types, and each of those rule types was described and illustrated in section 4.3. Those rules convert the semantic representations to a new underlying representation that is appropriate for a particular target language. The synthesizing grammar was described in section 4.4. The synthesizing grammar contains eight different types of rules, and each of those rule types was described and illustrated. Those rules are responsible for synthesizing the final surface forms of the target text. This chapter concluded with an example showing the English and Korean drafts that were generated by TTA for Infected Eye 1:2. English and Korean are vastly different languages, but in both cases TTA was able to generate texts that are easily understandable, grammatically correct, and semantically equivalent to the source document. More examples of the English and Korean drafts generated by TTA are included in appendix A.

In order to test the capabilities of TTA's grammars, many chapters of text were generated in English and Korean. The next chapter of this dissertation will discuss some of the more complex issues that were encountered during the development of the grammars for these two languages. Specifically, the next chapter will describe the generation of Korean direct speech honorifics and English questions.

CHAPTER 5

LEXICON AND GRAMMAR DEVELOPMENT:

GENERATING TEXT IN TWO TEST LANGUAGES

5.1 Introduction

This chapter will describe how TTA generates target language text from the semantic representations that were described in chapter 3. Substantial amounts of text were generated in the two test languages: Korean and English. During the development of the lexicons and grammars for these languages, challenging issues were encountered and resolved. This chapter will describe some of the more complex problems that were dealt with using the grammatical apparatus that was described in chapter 4.

As was stated in chapter 1, the fundamental question that this research proposes to answer is as follows: if the semantic representations contain sufficient information, and if the grammar possesses sufficient capabilities, then will TTA be able to generate texts of sufficient quality that they improve the productivity of experienced mother-tongue translators? In order to answer this question, a sizeable amount of text was generated in two unrelated languages: Korean and English. This chapter will present an overview of one particular issue that was encountered during the grammar development process for each of the test languages. In particular, the generation of the Korean honorific system will be described in section 5.2, and English question formation will be described in section 5.3. Finally, the results of the grammar building process will be discussed. Graphs will be presented at the end of this chapter illustrating how the number of new grammatical rules required for each chapter of generated text decreases rapidly, thereby demonstrating that TTA's grammatical rules are genuinely capturing the significant linguistic generalizations of each test language.

5.2 Korean Honorifics

This section will describe the process of generating the Korean honorific morphemes. First an analysis of the honorific morphemes will be presented in section 5.2.1, and then the pertinent grammatical rules that generate these morphemes will be presented in section 5.2.2. Korean is a chaining language with an elaborate system of honorifics. These honorifics are indicated in five specific ways (Cho et al. 2000:8-9):

- six different speech levels, the appropriate level being determined by the relationship between the speaker and listener and the social context,
- two sets of first and second person pronouns, the proper choice being determined by the relationship between the speaker and listener,
- an array of titles, the proper choice being determined by the speaker and listener's genders and relative ages, the addressee's profession, etc.,
- plain and honorific vocabulary, the proper choice being determined by a particular referent's social status, and
- plain and honorific grammatical relation markers and verbal suffixes.

For this dissertation a Korean lexicon and grammar were developed that were sufficient to generate the Grammar Introduction, three community development articles, and the following biblical texts: Luke 1-10, Ruth, Esther, Daniel, and Nahum. Section 5.2.1 will discuss the five ways that honorifics are encoded in Korean, and section 5.2.2 will discuss how the honorifics are generated by TTA's grammar.

5.2.1 Analysis of Korean Honorifics

As was indicated above, Korean has five distinct methods of encoding honorifics. Each of these methods will now be discussed.

5.2.1.1 Speech Levels

Korean has six different speech levels, each speech level being indicated by the sentence final suffix on the verb. These suffixes indicate both the speech level and the illocutionary force, and are listed below in Table 5-1 (Cho et al. 2000:9).

Table 5-1 The Six Styles of Korean Speech

	Declarative	Interrogative	Imperative
Plain	-다 [da]	-느냐 [neu nya]	-어라 [eo ra]
Intimate	-어 [eo]	-어 [eo]	-어 [eo]
Familiar	-네 [ne]	-니 [ni]	-어[eo]
Blunt	-소 [so]	-소 [so]	-어[eo]
Polite	-어요 [eo yo]	-어요 [eo yo]	-으세요 [eu se yo]
Deferential	-습니다 [seup ni da]	-습니까 [seup ni kka]	-으세요 [eu se yo]

The forms listed in the table above are used after closed syllables; most of these suffixes have a morphophonemic variant which is required after an open syllable. The appropriate speech level is determined by the relationship between the speaker and listener, the social context, and the speaker’s attitude. *Plain speech* is typically used when adults talk to children, older siblings talk to younger siblings, and in written text (Cho et al. 2000:10). *Intimate speech* is used when close friends talk to one another, when preschool children talk to family members, and when teachers talk to their students. The *familiar style* is used in very casual situations between two people who know each other well. *Blunt style* is rarely used by speakers these days, but it was occasionally used by speakers of previous generations. The *polite style* is the most commonly used style; it is used between close adult friends, social equals, and when children speak to adults. The *deferential style* is used in formal situations such as news reports, public lectures, and when a social inferior talks to a social superior.

5.2.1.2 Deferential and Polite Pronouns

Korean has plain and deferential first person pronouns as shown below in Table 5-2 (Rogers et al. 1992:54).

Table 5-2. Plain and Deferential First Person Pronouns

	Plain	Deferential
First Person Singular	나 [na]	저 [jeo]
First Person Plural	우리 [u ri]	저희들 [jeo hui deul]

The first person deferential pronouns are used when talking to someone older, or when speaking to an audience. The plain pronouns are used in all other circumstances, but Korean speakers frequently drop the pronouns (Rogers et al. 1992:54). Korean also has plain and honorific second person pronouns as listed in Table 5-3 (Rogers et al. 1992:54,321,334).

Table 5-3. Plain and Honorific Second Person Pronouns

	Plain	Honorable
Second Person Singular	너 [neo]	당신 [dang sin]
Second Person Plural	너희들 [neo hui deul]	여러분 [yeo reo bun]

5.2.1.3 Titles

Korean uses a vast array of titles when one person addresses another person. These titles are used in situations where English speakers would typically use either the person's name or the second person singular pronoun (Cho et al 2000:9). A few of these titles include: 언니 [eon ni] 'older sister' (used by a female who is talking to an older female friend), 오 빠 [o ppa] 'older brother' (used by a female who is talking to an older male friend), 누나 [nu na] 'older sister' (used by a male who is talking to an older female friend), 형 [hyeong] 'older brother' (used by a male who is talking to an older male friend), 어머니 [eo meo ni] 'mother' (used when talking to one's mother), 아버지 [a beo ji] 'father' (used when talking to one's father), 아저씨 [a jeo ssi] 'uncle' (used when talking to an adult male stranger or an adult male that does not have a more specific title), 아주머니 [a ju meo ni] 'aunt' (used when talking to an adult female stranger or an adult female that does not have a more specific title), 목사님 [mok sa nim] 'pastor', 사모님 [sa mo nim] 'pastor's wife' or 'professor's wife', 교수님 [gyo su nim] 'professor', 선생님 [seon saeng nim] 'teacher', 의사선생님 [ui sa seon saeng nim] 'doctor', 대왕 [dae wang] '(great) king', etc.

5.2.1.4 Honorary Lexical Items

Korean also has a small number of commonly used words that have two forms; one form is plain and the other form indicates honor to the person or thing being talked about. Several of these word pairs are listed below in Table 5-4 (Cho et al. 2000:8).

Table 5-4. Plain and Honorable Vocabulary

Plain	Honorific	
밥 [bap]	진지 [jin ji]	rice, meal
집 [jip]	댁 [daek]	house
이름 [i reum]	성함 [seong ham]	name
나이 [na i]	연세 [yeon se]	age
먹다 [meok da]	잡수시다 [jap su si da]	eat
자다 [ja da]	주무시다 [ju mi si da]	sleep
있다 [it da]	계시다 [gye si da]	stay

For example, when talking about someone's 'house', the usual word is 집 [jip]. However, when talking about a professor's house or the house of someone who deserves respect, the word that must be used is 댁 [daek] as shown in the table above. The entries in the last three rows of the table are verbs, and the honorable form must be used when the subject of the sentence is a referent that deserves respect. There are two other verbs that have a plain form and a deferential form that indicates respect for a senior. These two verbs are shown in table 5-5 (Cho et al. 2000:283) below. The deferential form of the first verb, 드리다 [deu ri da], signals respect for the recipient, while the deferential form of the second verb, 뵙다 [boep da], signals respect for the object.

Table 5-5. Plain and Deferential Vocabulary

Plain	Deferential
주다 [ju da] 'to give'	드리다 [deu ri da] 'to give to a senior'
보다 [bo da] 'to see'	뵙다 [boep da] 'to see a senior'

There are also verbs which have a common meaning, but in particular contexts these verbs have an alternate meaning and are intended to show respect. For example, the Korean verb meaning *to die* is 죽다 [juk da]. The verb 돌아가다 [dor a ga da] generally means *to go back, to return*. However, when someone talks about a parent who died, the verb 돌아가다

[dor a ga da] is used rather than 죽다 [juk da] as a way of indicating respect. Example (5-1i) occurs in Kande's Story 1:12 and illustrates this use of 돌아가다 [dor a ga da].

(5-1i) 만일 아버지-께서 돌아가-시-면
if.A father-Subj.Honor die-Honor-if.B

저희들-은 어떻게 음식-을 사-ㄹ 것-입니까?
we-Subj how food-Obj buy-Fut-question.Honor
'If father dies, how will we buy food?' (Korean text generated by TTA)

In this example, Kande is a young girl, and she is asking her mother a question. In the question's protasis, Kande is talking about her father dying, so she uses 돌아가다 [dor a ga da] rather than 죽다 [juk da].

5.2.1.5 Honorary Grammatical Relations

Korean indicates its various grammatical relations using suffixes. When marking the subject NP[30], there are two sets of suffixes: -께서(는) [kke seo (neun)] which is used when the subject of the sentence deserves honor, and –은/는 [eun/neun] which is used when the subject of the sentence does not require honor. For example, in Kande's Story 1:5, Kande asks her mother, "Do you know a secret?"

(5-1ii) 어머니-께서는 비밀-을 알-고 계시-ㅂ니까?
Mother-Subj.Honor secret-Obj know-Imperf.Honor-interrogative.Honor
'Do you(mother) know a secret?' (Korean text generated by TTA)

Kande must show respect when speaking to her mother, so the sentence above includes three methods of indicating honor:

- The honorable subject marker –께서는 [kke seo neun] is used rather than the standard marker –는[neun].

[30] Korean is a topic-comment language, but the grammatical relation of the topic is often called 'Subject'. There are also Korean verbs which are called double subject verbs. For example, 필요하다 means *to need.* The sentence 철수가 이 책이 필요하다 means *Chulsoo needs this book.* Both Chulsoo and book are marked with the subject marker, so this verb is called a 'double subject' verb rather than a 'double topic' verb.

- The verb ‘to know’ is in the imperfective, so usually the imperfective auxiliary 있다 [it da] would be used. However, the verb 있다 [it da] has a lexical honorific as was shown in the final row of table 5-4 above, so the lexical honorific 계시다 [gye si da] is used.
- The deferential interrogative marker –ㅂ니까 [b ni kka] is used rather than the plain or intimate interrogative markers.

Whenever the honorable subject marker –께서(는) [kke seo (neun)] is used, the verb also includes the honorable suffix marker –으시 [eu si], which changes to –시 [si] after open syllables. Example (5-1ii) above uses the lexical honorific 계시다 [gye si da], and the lexical honorifics always include the –으시 [eu si] suffix in the stem. There are also cases where –으시 [eu si] must be used even when the honorable subject marker –께서(는) [kke seo (neun)] is not used. An example illustrating this is shown below in (5-1iii). The subject in the protasis does not use the honorable subject marker, and there is no lexical honorific for 보다 [bo da] ‘to see’, but the honorable verbal suffix –으시 [eu si] is used as seen below in (5-1iii).

(5-1iii) 만일 여러분-이 아픈 닭-을 보-시-면
if.A you-Subj sick chicken-Obj see-Honor-if.B

특별한 새장 안에 그 닭-을 넣-으세요.
special cage in that chicken-Obj put-Imperative.Honor
‘*If you see a sick chicken, put it in a special cage.*’ (Korean text generated by TTA)

Example (5-1iii) above is from Avian Influenza 3:12 where a doctor is speaking to a group of people who live in a particular village. The verb 보-시-면 [bo si myeon] at the end of the protasis includes the honorable suffix –으시 [eu si] which was changed to –시 [si] because it follows the open syllable 보 [bo].

Another grammatical relation that may be marked with either an honorable suffix or a plain suffix is the indirect object. Most indirect objects are marked with the neutral suffix –에게

[e ge]. However, if the indirect object requires the honorific morphemes, then it is marked with –께 [kke]. These two indirect object markers are illustrated in examples (5-1iv) and (5-1v) below.

(5-1iv) 철수-는 민수-에게 말하-였-다. “…”
Chulsoo-Subj Minsu-IndObj say-Past-Declarative
‘Chulsoo said to Minsu, “…”’ (Korean example provided by JungAe Lee)

(5-1v) 철수−는 어머니−께 말하−였−다. “…”
Chulsoo-Subject mother-IndObj say-Past-Declarative
‘Chulsoo said to his mother, “…”’ (Korean example provided by JungAe Lee)

In example (5-1iv), Chulsoo and Minsu are peers, so the plain indirect object marker –에게 [e ge] is used. In example (5-1v), 어머니 [eo meo ni] ‘mother’ must be marked with the honorable indirect object marker –께 [kke].

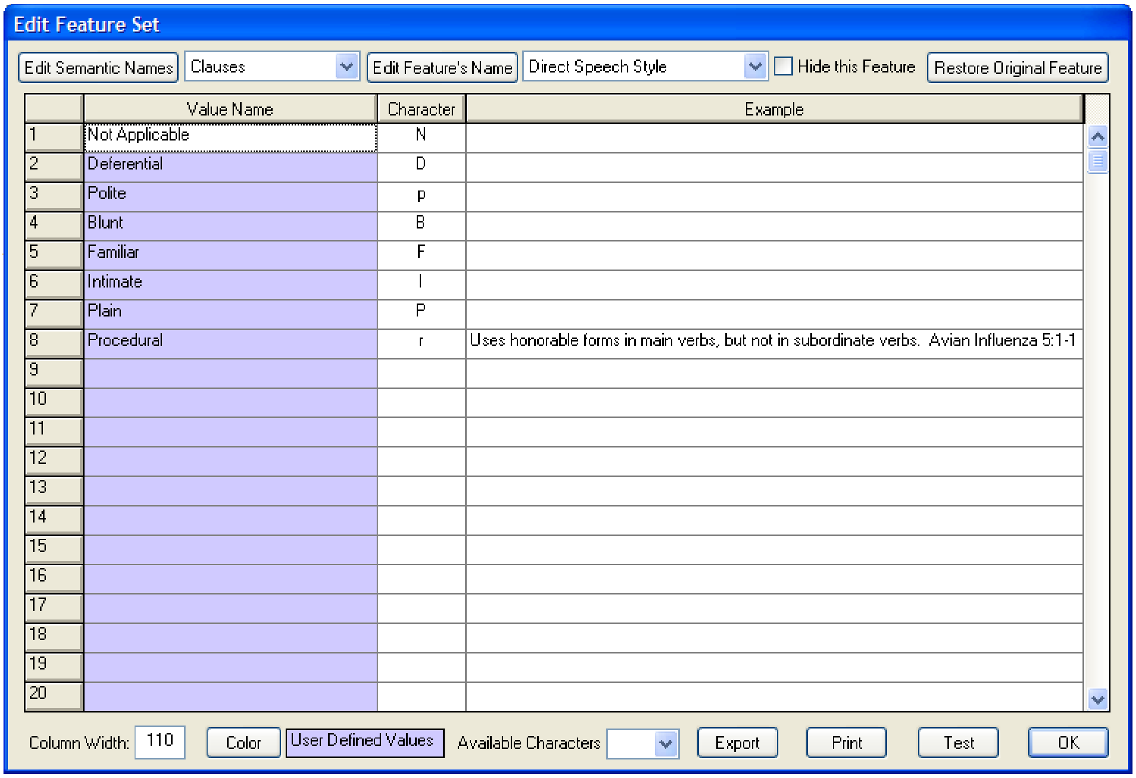
5.2.2 Generating Korean Honorifics

5.2.2.1 Generating the Six Speech Levels

In order to generate the speech levels required by Korean, the propositional Direct Speech features that were described in sections 3.3.2.5.5 through 3.3.2.5.9 were used. These features and a few of their values are repeated below:

- ‘Speaker’ – ‘Boy’, ‘Brother’, ‘Crowd’, ‘Daughter’, Employee’, ‘Employer’, ‘Father’, etc.
- ‘Listener’ – ‘Boy’, ‘Brother’, ‘Crowd’, ‘Daughter’, Employee’, ‘Employer’, ‘Father’, etc.
- ‘Speaker’s Attitude’ – ‘Neutral’ (speaker and listener do not know each other), ‘Familiar’ (speaker and listener know each other, and no particular emotions are involved in the speech), ‘Anger’, ‘Rebuke’, etc.
- ‘Speaker to Listener’s Age’ – ‘Older – different generation’, ‘Older – same generation’, ‘Essentially the same age’, ‘Younger – different generation’, ‘Younger – same generation’
- ‘Speaker’s Age’ – ‘Child (0-17)’, ‘Young Adult (18-24)’, ‘Adult (25-49)’, ‘Elder (50+)’

The Styles of Direct Speech rules use the features listed above and other propositional features such as ‘Discourse Genre’, ‘Illocutionary Force’, etc., and then specify a value for another feature called ‘Direct Speech Style’. These rules were introduced in section 4.4.3, and a small section of the Korean Styles of Direct Speech dialog was shown in figure 4-12. In order to use these rules, the linguist must first define the styles of speech that are pertinent to his language. This is done in the dialog shown below in figure 5-1.



Clauses

Direct Speech Style

☐ Hide this Feature

	Value Name	Character	Example
1	Not Applicable	N	
2	Deferential	D	
3	Polite	p	
4	Blunt	B	
5	Familiar	F	
6	Intimate	I	
7	Plain	P	
8	Procedural	r	Uses honorable forms in main verbs, but not in subordinate verbs. Avian Influenza 5:1-1
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			

Column Width: 110

Figure 5-1. Dialog where Speech Styles are Defined for Korean

Initially this dialog contains only one value: ‘Not Applicable’ which is seen in row one. Every proposition in the semantic representations, including subordinate propositions, titles, footnotes, etc., has a feature called ‘Direct Speech Style’, and initially that feature has a value of ‘Not Applicable’. Linguists using TTA are able to define any values for ‘Direct Speech Style’ that are relevant to their language. The dialog shown above includes the six Korean speech styles that were mentioned earlier in table 5-2, but it also includes an additional style called ‘Procedural’ seen in row eight. The final chapter of the Avian Influenza text is procedural; a few sentences of the computer generated English draft of that text follow:

> Avian Influenza 5:1 *You must protect your chickens and your animals from this disease. You must work with the other people who live in your village. You and the other people who live in your village must learn about this disease. If you prevent this disease from spreading, your animals will be healthy.*
> Avian Influenza 5:2 *You must do these things in order to prevent Avian Influenza from killing you and your animals.*
> Avian Influenza 5:3 *1) When you buy chickens and ducks at the market, you must be very careful. Chickens and ducks have Avian Influenza often. When you cut the meat, use a special board. You must put only raw meat on that board. You must not put on that board meat that you cooked. After you cook the meat, wash your hands with soap thoroughly.*
> Avian Influenza 5:4 *2) When you buy eggs at the market, you also must be careful. Before you boil the eggs, wash them thoroughly. After you touch the eggs, wash your hands with soap.*

When this text was generated in Korean, it was found that texts of this type sound best if the deferential speech style is used with the final verb, but the subordinate verbs must not be marked with the honorific morpheme –으시 [eu si] which generally occurs in deferential text. For example, the final sentence of Avian Influenza 5:1 is shown below in example (5-1vi).

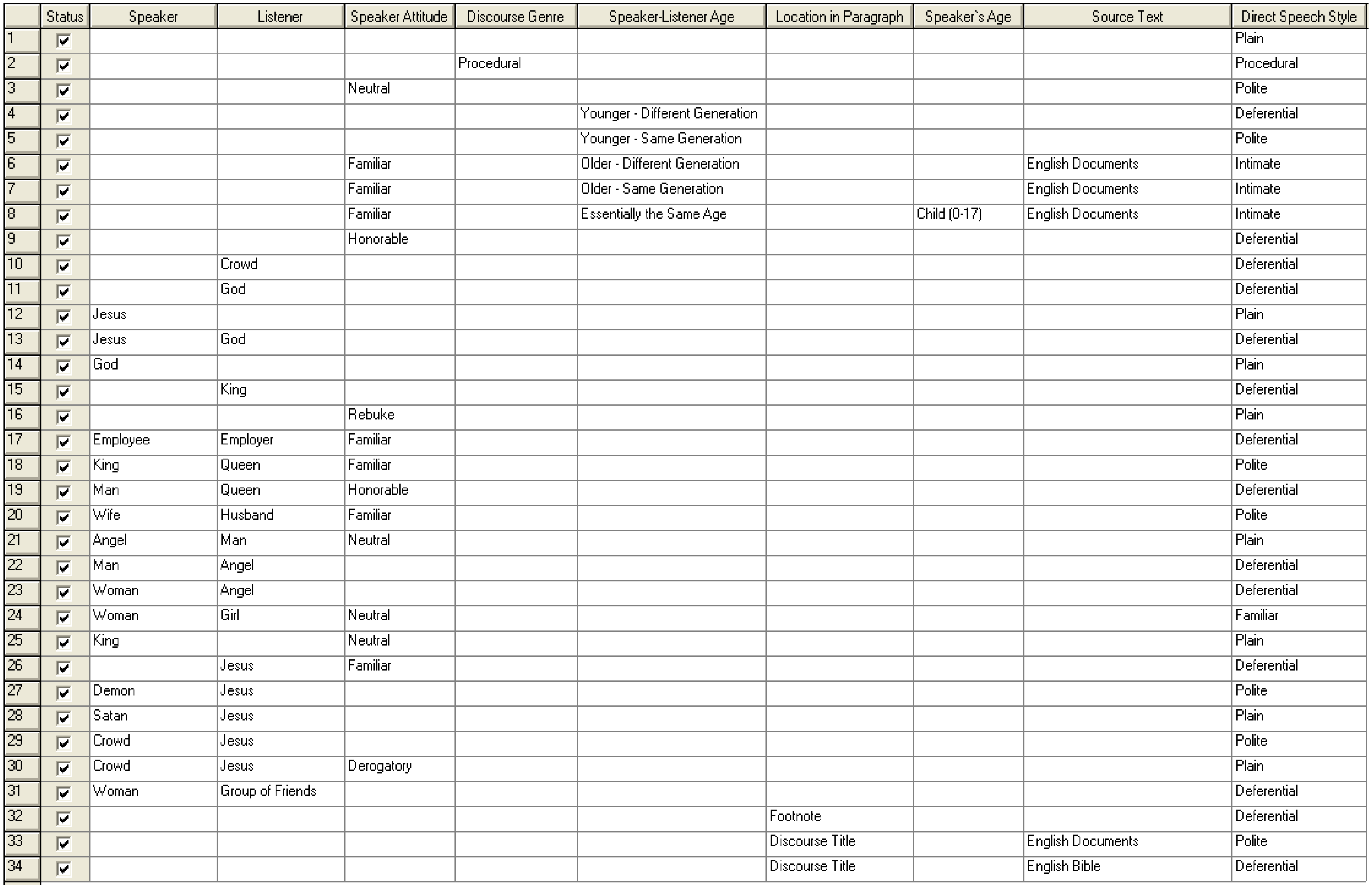
(5-1vi)	만일	여러분-이	이	병-이	퍼지-는 것을	막-으면
	if.A	you.Honor-Subj	this	disease-Subj	spread-Comp	block-if.B

여러분-의	동물-들-은	건강하-ㄹ 것이-ㅂ니다.
You.Honor-Pos	animal-Plural-Subj	healthy.be-Future-Declarative.Deferential

‘*If you prevent this disease from spreading, your animals will be healthy.*’ (Korean text generated by TTA)

If this example were standard deferential speech, the verb at the end of the protasis would include the honorific morpheme –으시 [eu si]. However, 막으시면 [mag eu si myeon] in this context makes the text sound strange, so an additional speech style called ‘Procedural’ was added to the standard list of Korean speech styles. That value was then used to trigger the addition of the deferential morphemes on main verbs, but that speech style blocks the addition of 으시 [eu si] on subordinate verbs.

After the speech styles were defined, the speech style rules were used to specify when each particular style should be used. The complete set of speech style rules for Korean is shown below in figure 5-2.



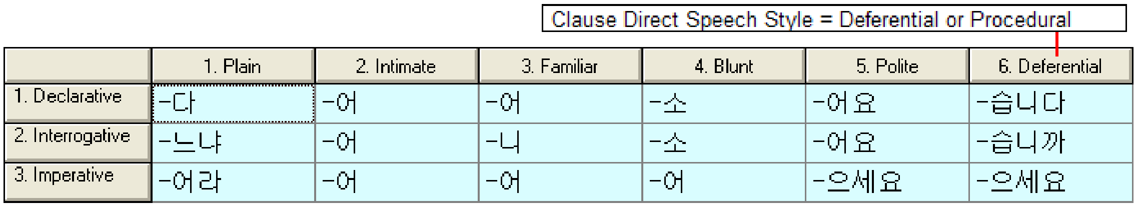
	Status	Speaker	Listener	Speaker Attitude	Discourse Genre	Speaker-Listener Age	Location in Paragraph	Speaker`s Age	Source Text	Direct Speech Style
1	☑									Plain
2	☑				Procedural					Procedural
3	☑			Neutral						Polite
4	☑					Younger - Different Generation				Deferential
5	☑					Younger - Same Generation				Polite
6	☑			Familiar		Older - Different Generation			English Documents	Intimate
7	☑			Familiar		Older - Same Generation			English Documents	Intimate
8	☑			Familiar		Essentially the Same Age		Child (0-17)	English Documents	Intimate
9	☑			Honorable						Deferential
10	☑		Crowd							Deferential
11	☑		God							Deferential
12	☑	Jesus								Plain
13	☑	Jesus	God							Deferential
14	☑	God								Plain
15	☑		King							Deferential
16	☑			Rebuke						Plain
17	☑	Employee	Employer	Familiar						Deferential
18	☑	King	Queen	Familiar						Polite
19	☑	Man	Queen	Honorable						Deferential
20	☑	Wife	Husband	Familiar						Polite
21	☑	Angel	Man	Neutral						Plain
22	☑	Man	Angel							Deferential
23	☑	Woman	Angel							Deferential
24	☑	Woman	Girl	Neutral						Familiar
25	☑	King		Neutral						Plain
26	☑		Jesus	Familiar						Deferential
27	☑	Demon	Jesus							Polite
28	☑	Satan	Jesus							Plain
29	☑	Crowd	Jesus							Polite
30	☑	Crowd	Jesus	Derogatory						Plain
31	☑	Woman	Group of Friends							Deferential
32	☑						Footnote			Deferential
33	☑						Discourse Title		English Documents	Polite
34	☑						Discourse Title		English Bible	Deferential

Figure 5-2. Styles of Direct Speech for Korean

Each row in the dialog shown above is considered a separate rule. When these rules are executed, TTA will look at the features associated with a proposition, and then look at the first rule and see if the features specified in that row match the features of the current proposition. If the proposition's features match the features specified in the first row, then the proposition's 'Direct Speech Style' feature will be set to the value specified in the last column of that row. Since the first row in the table shown above in figure 5-2 does not specify any values for 'Speaker', 'Listener', 'Speaker's Attitude', etc., that row will match every proposition in the semantic representations. That row specifies that the 'Direct Speech Style' must be set to 'Plain'. Therefore every proposition, including those that are not direct speech, will have its 'Direct Speech Style' set to 'Plain' by the first rule in this dialog. Then the next row in this table will be examined to see if its features match the current proposition's features. The second row says that if the 'Discourse Genre' is 'Procedural', then the 'Direct Speech Style' must be set to 'Procedural'. All of the propositions in chapter 5 of Avian Influenza have their 'Discourse Genre' feature set to 'Procedural', so the second row of this dialog will set the 'Direct Speech Style' to 'Procedural' for each of those propositions.

The rows at the top of this dialog are generally very generic and apply to many propositions; the rows toward the bottom of the dialog tend to be very specific and apply in only select situations. For example, as was mentioned above, when a young person speaks to an older person of a different generation, the deferential style must be used, and this is indicated in row four. That row specifies that when the speaker is younger than the listener and of a different generation, the speech style must be deferential. So that row handles all the situations where children speak to their parents, their teachers, and all other adults. Also mentioned above was the fact that when someone speaks to a group of people, the speaker always uses the deferential style regardless of his social status. So professors, pastors, political leaders, etc., will use the deferential style when addressing a group of people. The tenth row of this dialog states that when the 'Listener' is 'Crowd', the 'Deferential' speech style must be used regardless of the speaker. However, in Korean translations when Jesus addresses a crowd, he traditionally is portrayed using the ‘Plain’ speech style. Therefore row twelve of this dialog indicates that when the speaker is ‘Jesus’, he uses plain speech regardless of the audience. So even if the ‘Listener’ is a crowd, when Jesus speaks, the plain speech style will be generated. The one exception to the rule in row twelve occurs when Jesus speaks to God. When the ‘Speaker’ is ‘Jesus’ and the ‘Listener’ is ‘God’, row thirteen indicates that the ‘Speech Style’ must be ‘Deferential’. Similarly when a king addresses a crowd, he will use plain speech, and this is indicated in row 25.

After the ‘Direct Speech Style’ feature has been set by the table of rules shown above, a table spellout rule supplies the appropriate suffix to the final verb. The table for that spellout rule is shown below in figure 5-3.



Clause Direct Speech Style = Deferential or Procedural

	1. Plain	2. Intimate	3. Familiar	4. Blunt	5. Polite	6. Deferential
1. Declarative	-다	-어	-어	-소	-어요	-습니다
2. Interrogative	-느냐	-어	-니	-소	-어요	-습니까
3. Imperative	-어라	-어	-어	-어	-으세요	-으세요

Figure 5-3. Table Spellout Rule that Supplies the Speech Style and Illocutionary Force Markers

The table shown above has a column for each of the six speech styles mentioned earlier, and a row for each of the illocutionary forces. Since procedural texts use the deferential style suffixes, the last column of this table applies when the ‘Direct Speech Style’ is either ‘Deferential’ or ‘Procedural’. One cell in the table above contains a different value than the corresponding cell in Table 5-2. The Declarative/Familiar cell in the spellout rule has -어 [eo], while that same cell in Table 5-2 above contains –네 [ne]. The reason for this difference is that the –네 [ne] value is considered old fashioned; when people use the familiar style today, they use –어 [eo] rather than –네 [ne]. All of the morphemes listed in this table are used after closed syllables. When these morphemes follow open syllables, morphophonemic rules are required to convert them appropriately, but those rules will not be presented here. Examples (5-1vii) through (5-1x) below illustrate the application of the Styles of Direct Speech rules and this table spellout rule.

(5-1vii)	만일	누군가-가	아프-게	되-면
	if.A	someone-Subj	sick-Comp	become-if.B

나-를	즉시	불러-라.
1st.Sing-Obj	immediately	call-Imp.Plain

‘*If someone becomes sick, call me immediately*.’ (Korean text generated by TTA)

Example (5-1vii) comes from Avian Influenza 4:5 in which a man named Paulos is talking to his friend named Nano. In the semantic representation of this proposition, the ‘Speaker’ is ‘Man’, the ‘Listener’ is ‘Man’, the ‘Speaker Attitude’ is ‘Familiar’ meaning that the speaker and listener know each other, the ‘Speaker-Listener Age’ is ‘Essentially the Same Age’, and the ‘Speaker’s Age’ is ‘Adult’. The first row of the Styles of Direct Speech dialog set the ‘Direct Speech Style’ to ‘Plain’, and no other rows in that dialog matched this proposition. Since the ‘Illocutionary Force’ value is ‘Imperative’, the Plain/Imperative cell in the spellout table added –어라 [eo ra] to the verb stem 부르 [bu reu]. A morphophonemic rule changed the verb stem 부르 [bu reu] to 불러 [bul leo], and then the suffix –어라 [eo ra] was changed to –라 [ra] by another morphophonemic rule. The final result is 불러라 [bul leo ra].

(5-1viii)	저-에게	진실-을	말씀하-여	주-세요.
	1st.Sg.Deferential-to	truth-Obj	say.Honor-Comp	Request-Imp.Deferential

‘*Please tell me the truth*.’ (Korean text generated by TTA)

Example (5-1viii) comes from Kande’s Story 1:10 in which a girl named Kande is talking to her mother. The proposition’s ‘Speaker’ is ‘Daughter’, ‘Listener’ is ‘Mother’, ‘Speaker Attitude’ is ‘Familiar’, ‘Speaker-Listener Age’ is ‘Younger – Different Generation’, and the ‘Speaker’s Age’ is ‘Child’. The first row of the Speech Styles dialog set ‘Direct Speech Style’ to ‘Plain’, but then the fourth row changed ‘Direct Speech Style’ to ‘Deferential’. Since the ‘Illocutionary Force’ value is ‘Imperative’, the Deferential/Imperative cell in the table spellout rule added –으세요 [eu se yo] to the verb stem 주 [ju]. Then a morphophonemic rule changed –으세요 [eu se yo] to –세요 [se yo] because it follows an open syllable. The final result is 주세요 [ju se yo].

(5-1ix)	우리-는	돈-을	벌-기 위해	일-하-기	시작해-야만	해요.
	1st.Pl.Plain-Subj	money-Object	earn-in.order.to	work-do-Comp	start-must	Decl.Polite

‘*We must start working in order to earn money*.’ (Korean text generated by TTA)

Example (5-1ix) comes from Kande’s Story 3:11 in which Kande’s younger sister named Falala is speaking to Kande. The proposition’s ‘Speaker’ is ‘Sister’, ‘Listener’ is ‘Sister’, ‘Speaker Attitude’ is ‘Familiar’, ‘Speaker-Listener Age’ is ‘Younger – Same Generation’, and the ‘Speaker’s Age’ is ‘Child’. The first row of the Speech Styles dialog set ‘Direct Speech Style’ to ‘Plain’, and then the fifth row changed ‘Direct Speech Style’ to ‘Polite’. The proposition’s ‘Illocutionary Force’ value is ‘Declarative’, so the Polite/Declarative cell in the spellout rule added the suffix –어요 [eo yo] to the verb stem 하 [ha]. Then a morphophonemic rule changed 하-어요 [ha eo yo] to 해요 [hae yo].

(5-1x)	만일	너희–들–이	이	남자–들–과	자–면
	if.A	2nd-Pl-Subj	these	man-Pl-with	sleep-if.B

아마도	이	남자–들–에게서	인체 면역 결핍	바이러스–와
maybe	these	man-Pl-from	HIV	virus-and

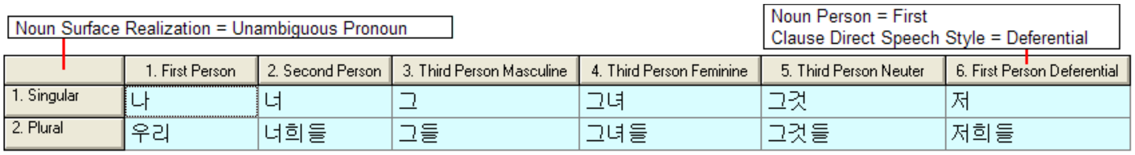
다른	병–을	옮–을 거–야.
other	disease-Obj	catch-Fut-Decl.Familiar

‘*If you sleep with those men, you might catch HIV and other diseases from them*.’
(Korean text generated by TTA)

Example (5-1x) comes from Kande’s Story 3:14 in which a nurse is talking to Kande and her younger sister. The proposition’s ‘Speaker’ is ‘Woman’, ‘Listener’ is ‘Girl’, ‘Speaker Attitude’ is ‘Neutral’ meaning that the nurse does not know Kande or her sister, ‘Speaker-Listener Age’ is ‘Older – Different Generation’, and the ‘Speaker’s Age’ is ‘Adult’. The first row of the Speech Styles dialog set ‘Direct Speech Style’ to ‘Plain’, then the third row changed ‘Direct Speech Style’ to ‘Polite’, and finally row 24 changed ‘Direct Speech Style’ to ‘Familiar’. The proposition’s ‘Illocutionary Force’ value is ‘Declarative’, so the Familiar/Declarative cell in the spellout rule added the suffix –어 [eo]. A morphophonemic rule changed –어 [eo] to –야 [ya] so the final form is 옮을 거야 [olm eur geo ya].

5.2.2.2 Generating the Plain, Deferential and Polite Pronouns

The deferential pronouns are used whenever the 'Direct Speech Style' is 'Deferential'. Therefore, after the Styles of Direct Speech rules were executed as was described in the previous section, the deferential pronouns are generated by a table spellout rule. The spellout rule that inserts the plain and first person deferential pronouns is shown below in figure 5-4.



Noun Surface Realization = Unambiguous Pronoun

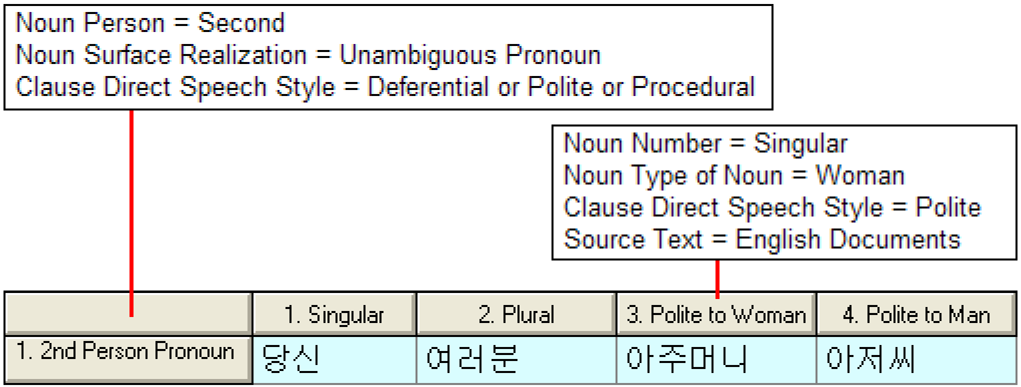
Noun Person = First
Clause Direct Speech Style = Deferential

	1. First Person	2. Second Person	3. Third Person Masculine	4. Third Person Feminine	5. Third Person Neuter	6. First Person Deferential
1. Singular	나	너	그	그녀	그것	저
2. Plural	우리	너희들	그들	그녀들	그것들	저희들

Figure 5-4. Table Spellout Rule that Inserts the Plain and Deferential Pronouns

The deferential pronouns are in the last column of the table, and the features of that column indicate that they will be used whenever a nominal's 'Person' feature is 'First' and the proposition's 'Direct Speech Style' is 'Deferential'.

The polite second person pronouns are inserted by a similar table spellout rule which is shown below in figure 5-5.



Noun Person = Second
Noun Surface Realization = Unambiguous Pronoun
Clause Direct Speech Style = Deferential or Polite or Procedural

Noun Number = Singular
Noun Type of Noun = Woman
Clause Direct Speech Style = Polite
Source Text = English Documents

	1. Singular	2. Plural	3. Polite to Woman	4. Polite to Man
1. 2nd Person Pronoun	당신	여러분	아주머니	아저씨

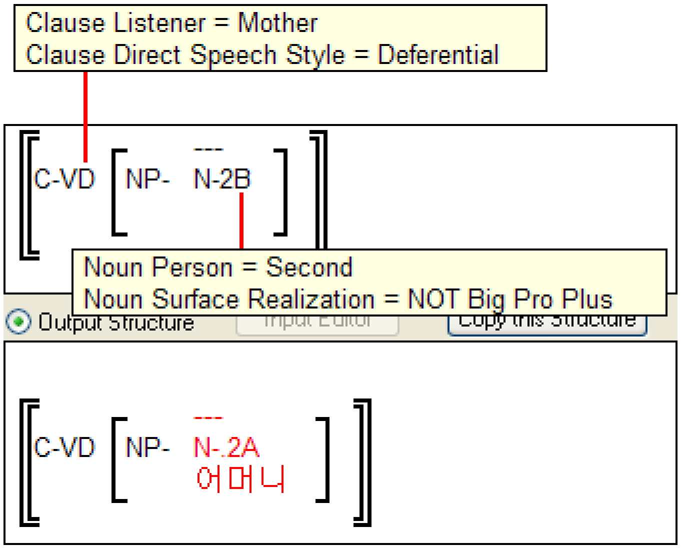
Figure 5-5. Table Spellout Rule that Inserts the Polite Second Person Pronouns[31]

[31] The last two entries in this table aren't actually pronouns; they are terms of address which literally mean 'aunt' and 'uncle'. These terms are used as substitutes for a second person pronoun when using polite speech.

The rule shown above inserts only the second person pronouns. The last two columns in the table have a feature that specifies that they only apply when the ‘Source Text’ is ‘English Documents’. There are multiple source text databases for this project. One source text database includes all the community development articles such as *Infected Eye* published by World Vision, *Kande’s Story* published by Shell Publishing, and *Avian Influenza* published by the Indonesian branch of The Summer Institute of Linguistics. Another source text database includes all the biblical books, another database includes the Grammar Introduction texts, etc. The rule shown above does not insert 아주머니 [a ju meo ni] or 아저씨 [a jeo ssi] unless the ‘Source Text’ is one of the community development articles because these titles were not considered appropriate in the biblical texts[32].

5.2.2.3 Generating the Titles

The titles that were mentioned above in section 5.2.1.3 are generally inserted by Pronoun Identification Rules. For example, the rule that inserts 어머니 [eo meo ni] when someone is speaking to his mother is shown below in figure 5-6.



Clause Listener = Mother
Clause Direct Speech Style = Deferential

Noun Person = Second
Noun Surface Realization = NOT Big Pro Plus

Figure 5-6. Pronoun Identification Rule that Inserts 어머니 [eo meo ni]

[32] For example, in Luke 2:35 a man is talking politely to Mary and says “You(Mary) will be sad.” The generated text is 당신은 매우 슬플 것이요. If the term of address 아주머니 were used rather than 당신, the text would sound strange.

The rule shown above applies whenever a proposition's 'Listener' is 'Mother', the 'Direct Speech Style' is 'Deferential', the nominal's 'Person' value is 'Second', and the nominal's 'Surface Realization' value is anything other than 'Big Pro Plus'. Whenever that situation is found, the word 어머니 [eo meo ni] 'mother' is inserted into the text. This rule applied in many places, one of them being Ruth 1:10 where Naomi's daughters-in-law said to her, "We want to go to your people with you."

(5-1xi) 저희-들-은 어머니-와 함께 어머니-의 사람-들-에게 가-고 싶-습니다
1st.Pl.Def-Pl-Subj mother-with mother-Pos person-Pl-to go-Comp want-Decl.Def
'*We want to go to your people with you*.' (Korean text generated by TTA)

This rule applied twice in this particular sentence, the first time was in the phrase 어머니와 함께 [eo meo ni wa ham kke] 'with you(mother)', and the second time in the phrase 어머니의 사람들에게 [eo meo ni ui sa ram deur e ge] 'to your(mother's) people'. Similar rules insert most of the other titles mentioned earlier. The rule that inserts 언니 [eon ni] 'older sister' when a younger sister talks to an older sister is shown below in figure 5-7.

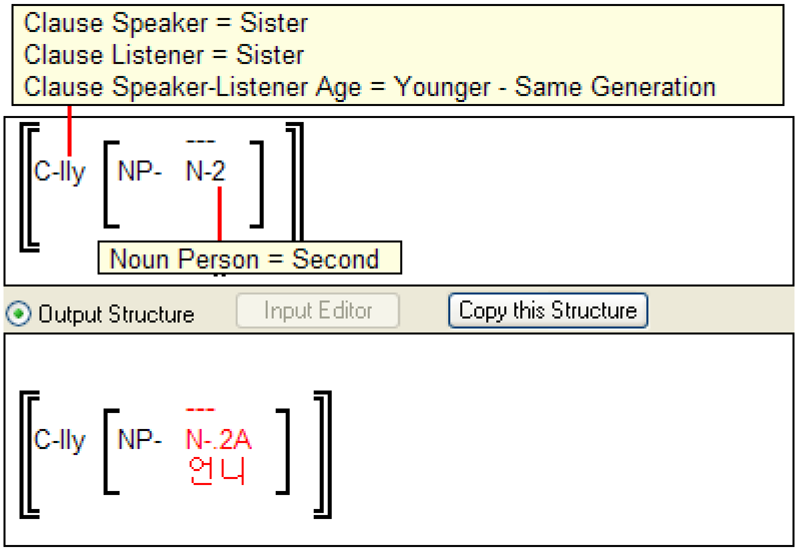

Clause Speaker = Sister
Clause Listener = Sister
Clause Speaker-Listener Age = Younger - Same Generation
Noun Person = Second

Figure 5-7. Pronoun Identification Rule that Inserts 언니 [eon ni]

The rule shown above applied in Kande's Story 1:2 as shown below in example (5-1xii).

(5-1xii) 언니-는 어머니-의 비밀-을 알-고 있-어요?
Older.Sister-Subj mother-Pos secret-Obj know-Imperf-Interrogative.Polite
'*Do you know mother's secret?*' (Korean text generated by TTA)

Although many of the titles used in Korean can be generated by TTA, there are certainly many others which cannot be generated. For example, one of the titles mentioned above is 사모님 [sa mo nim] 'teacher's wife' and is used when talking to the wife of one's teacher or pastor. Since the feature called 'Listener' does not have a value called 'Wife of Teacher' or 'Wife of Pastor'[33], TTA is not able to generate that particular title.

5.2.2.4 Generating the Honorific Vocabulary

When a particular Korean concept has both a plain form and an honorific form such as 집/댁 [jip/daek] 'house' mentioned above, the source concept is linked to the Korean plain form because the plain forms are used much more frequently than the honorific forms. Then structural adjustment rules look at the environment and decide whether or not an honorific form should be used. For example, as was mentioned above, the usual Korean word for *to die* is 죽다 [juk da]. However, when someone talks about his mother or father dying, the verb 돌아가다 [dor a ga da] must be used. The rule shown below in figure 5-8 looks for the verb 죽 [juk] and changes it to 돌아가 [dor a ga] if the topic of the sentence is either FATHER-A or MOTHER-A.

[33] A complete list of all the values for the feature called 'Listener' can be seen in table 3-20 in section 3.3.2.5.6.

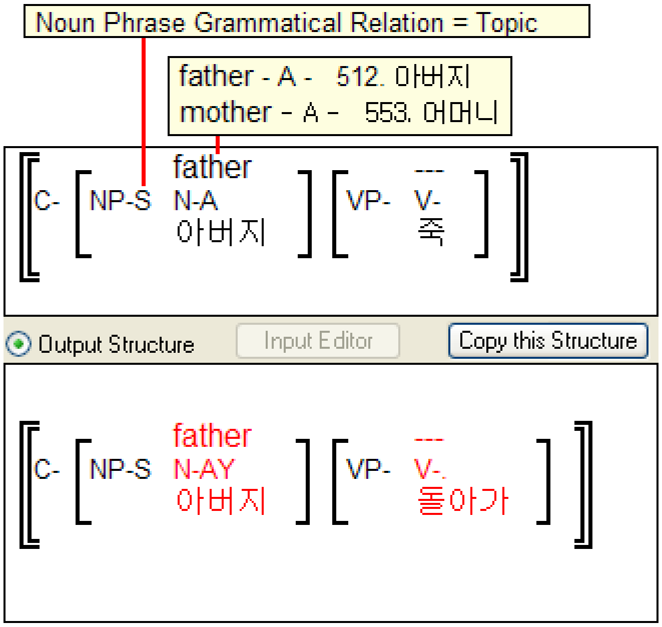


Figure 5-8. Structural Adjustment Rule that Changes 죽 [juk] to 돌아가 [dor a ga] if the Subject is FATHER-A or MOTHER-A

The rule shown above applied in many places, one of them being in Kande’s Story 1:12 mentioned earlier in example (5-1xii), and also in Kande’s Story 2:17 which is shown below.

(5-1xiii) 며칠 후에 어머니-께서는 돌아가-시-었-다. (시-었 -> 셨)
a.few.days after mother-Subj.Honor die-Honor-Past-Decl.Plain
‘*A few days later mother died*.’ (Korean text generated by TTA)

In order to change the plain form of a verb to the honorific form, the Theta Grid Adjustment rules are used. Shown below in figure 5-9 is one of the structures in the Theta Grid Adjustment rule for TALK-A. That rule changes the plain verb 말하다 [mal ha da] ‘to speak’ to the honorific form 말씀드리다 [mal sseum deu ri da] which signals respect to the person that will be talked to. This verb is required whenever someone says that he will speak to a father or mother.

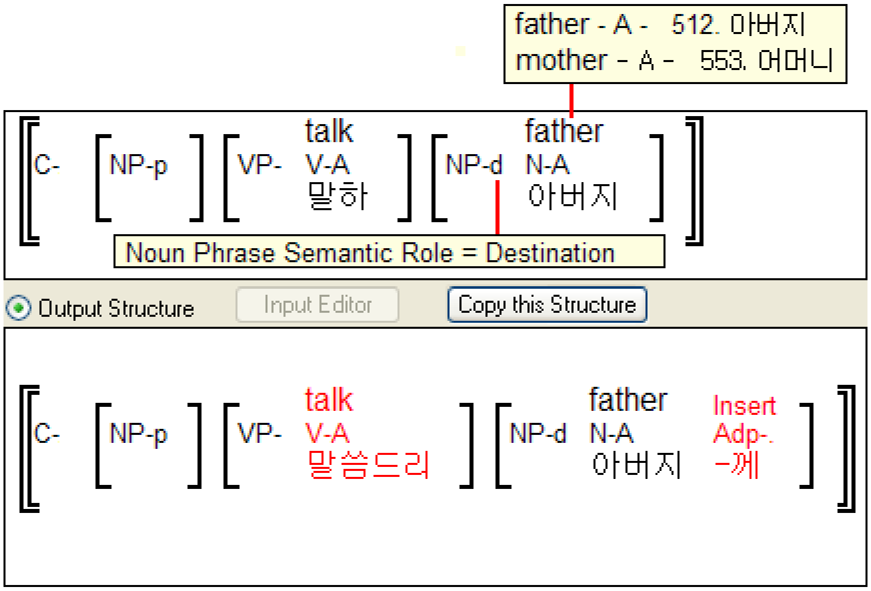


Figure 5-9. Theta Grid Adjustment Rule that Changes 말하다 [mal ha da] to 말씀드리다 [mal sseum deu ri da] when a Child Speaks to a Parent

The rule shown above applies when someone says something like, “I will talk to my father.” The event must be TALK-A and the ‘Destination’ NP must contain either MOTHER-A or FATHER-A. An example illustrating the application of this rule is in Kande’s Story 1:3.

(5-1xiv) 우리-는 어머니-께 말씀드려-야 해.
1st.Pl-Subj mother-to talk.Honor-must-Decl.Intimate
‘*We must talk to our mother*.’ (Korean text generated by TTA)

In this example Kande is speaker to her younger sister, so the ‘Direct Speech Style’ is ‘Intimate’. Since Kande is talking about *talking to her mother*, she must use the verb that shows respect, 말씀드리다 [mal sseum deu ri da], rather than the plain form 말하다 [mal ha da]. Another verb 말씀하다 [mal sseum ha da] is used when someone says something like, “Did our uncle say …” The Theta Grid Adjustment rule for SAY-A is used to change 말하다 [mal ha da] to 말씀하다 [mal sseum ha da] as shown below.

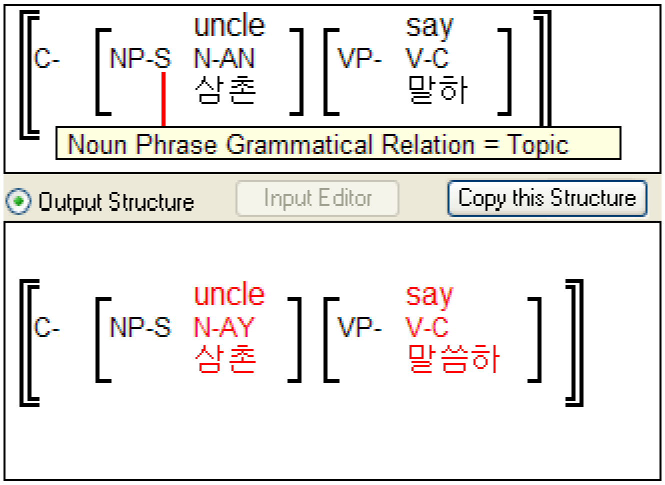


Figure 5-10. Structural Adjustment Rule that Changes 말하다 [mal ha da] to 말씀하다 [mal sseum ha da]

The rule shown above applied in Kande's Story 3:10.

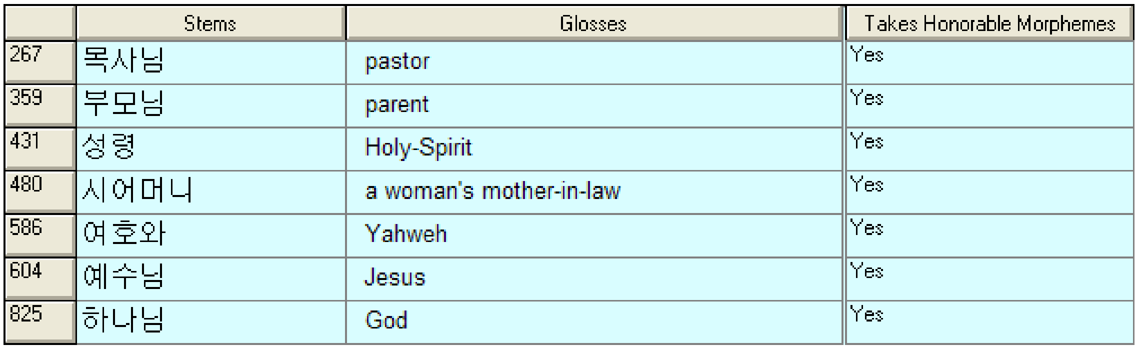
(5-1xv) 팔라라-는 칸디-에게 "삼촌-께서는 우리-가 다른 집-으로
Falala-Subj Kande-to "Uncle-Subj.Honor 1st.Pl-Subj other house-to

이사해-야만 한-다고 말씀하-시-었-어요?" 라고 물-었다. (시-었->셨)
move-must-Comp say-Honor-Past-Inter.Polite Comp ask-Past-Decl.Plain
'*Falala said to Kande, "Did our uncle say that we must move to another house?"*
(Korean text generated by TTA)

The rules shown here and many others are generally able to express the various degrees of honorifics to the relevant arguments.

5.2.2.5 Generating the Honorific Grammatical Relation Markers

In the Korean lexicon a feature was defined for Korean nouns to distinguish the nouns that require the honorable morphemes from the nouns that do not take the honorable morphemes. Figure 5-11 below shows a subset of the Korean nouns in the lexicon that was developed for this project. This subset shows most of the nouns that require honorable morphemes.



	Stems	Glosses	Takes Honorable Morphemes
267	목사님	pastor	Yes
359	부모님	parent	Yes
431	성령	Holy-Spirit	Yes
480	시어머니	a woman's mother-in-law	Yes
586	여호와	Yahweh	Yes
604	예수님	Jesus	Yes
825	하나님	God	Yes

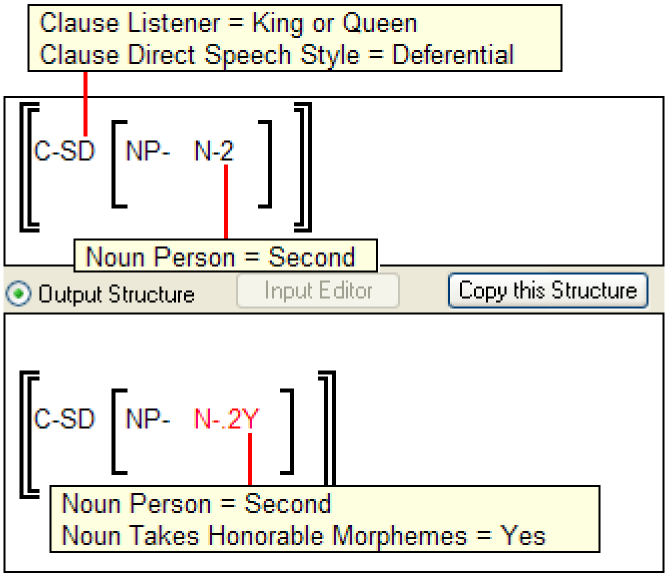
Figure 5-11 Nouns in the Korean Lexicon that Require Honorable Morphemes

Noticeably absent from the list above are the words 어머니 [eo meo ni] 'mother', 아버지 [a beo ji] 'father', 왕 [wang] 'king', and 여왕 [yeo wang] 'queen'. For this project it was decided that when a narrative talks about a mother, father, king or queen, the honorable suffixes should not be used, but when someone talks directly to a mother, father, king or queen, the honorable suffixes certainly must be used. Examples illustrating this distinction are found in Esther 1:1 and 1:19, which are shown below.

(5-1xvi) 크셀크세스 왕-은 수사-라는 도시-에 있-는 자기 궁궐-에서 살-았-다.
Xerxes king-Subj Susa-named city-at be-Rel he palace-in live-past-Decl.Plain
'*King Xerxes lived in his palace that was in a city named Susa.*' (Korean text generated by TTA)

(5-1xvii) 따라서 대왕-께서는 새 법-을 정하-셔-야 하-ㅂ니다.
therefore king-Subj.Honor new law-Obj decide-Honor-must-Decl.Deferential
'*Therefore you(king) must decide(write) a new law.*' (Korean text generated by TTA)

Example (5-1xvi) above comes from Esther 1:1, and it describes King Xerxes. In the Korean text, no honorific morphemes are used; the standard subject marker –은 [eun] is used after 왕 [wang] 'king', and the honorific morpheme –으시 [eu si] does not occur after the verb stem 살 [sal]. In example (5-1xvii) above which comes from Esther 1:19, a man is talking directly to King Xerxes. The direct speech includes the honorific subject marker –께서는 [kke seo neun] as well as the honorific marker –으시 [eu si] after the verb stem 정하 [jeong ha] 'to decide'. Also in that sentence the nominal that is realized with a second person pronoun in English is realized in Korean with 대왕 [dae wang] 'great king'. The structural adjustment rule which marks second person pronouns so that they take the honorable morphemes when the ‘Listener’ is ‘King’ or ‘Queen’ is shown below in figure 5-12.



Clause Listener = King or Queen
Clause Direct Speech Style = Deferential

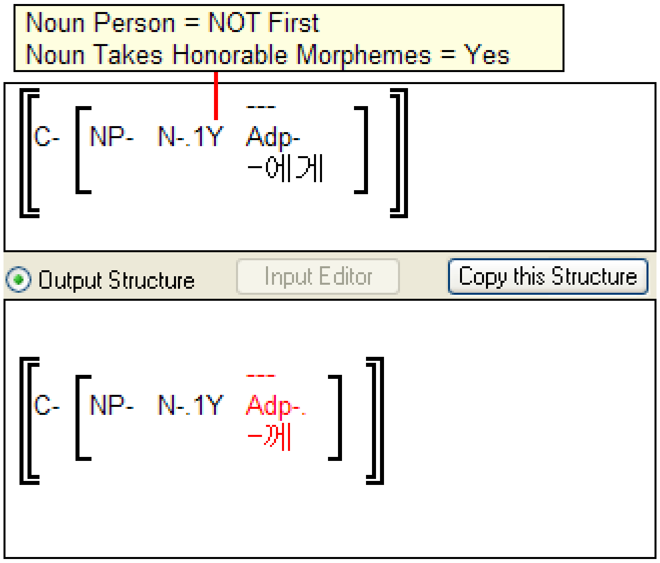
Noun Person = Second

Noun Person = Second
Noun Takes Honorable Morphemes = Yes

Figure 5-12. Honorable Morphemes must be Used when Talking to a King or Queen

Another structure in the same structural adjustment rule applies whenever the ‘Listener’ is ‘Father’ or ‘Mother’, and the ‘Speaker’ is ‘Adult Daughter’, ‘Adult Son’, ‘Daughter’, or ‘Son’, and the ‘Speech Style’ is ‘Deferential’.

When a noun that takes honorable morphemes is marked with the indirect object marker –에게 [e ge], that marker must be changed to the honorable form –께 [kke] unless the noun is realized with a first person pronoun. The structural adjustment rule that changes this grammatical relation marker from the plain form to the honorable form is shown below in figure 5-13.



Noun Person = NOT First
Noun Takes Honorable Morphemes = Yes

Figure 5-13. Structural Adjustment Rule that Changes –에게 [e ge] to –께 [kke] for Honorable Nouns

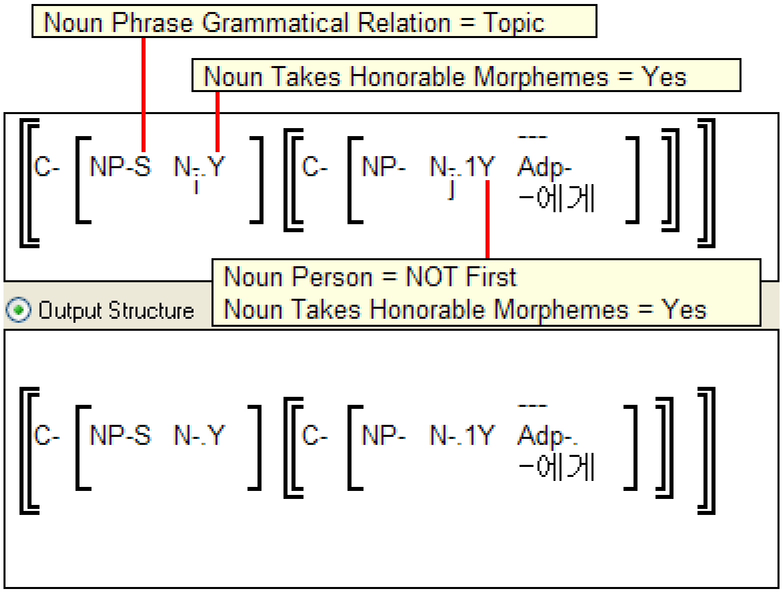
The rule shown above applied in many places, one of them being Kande's Story 1:6 which is shown below in example (5-1xviii).

(5-1xviii)	이니꼬-는	"저-는	아버지-께	비밀-에 대해서	말하-고
	Iniko-Subj	"1st.Sg.Def-Subj	father-to.Honor	secret-about	tell-Comp

싶-습니다!"	라고	말하-였-다.
want-Decl.Deferential	Comp	say-Past-Decl.Plain

'*Iniko said, "I want to tell our father about the secret.'* (Korean text generated by TTA)

In the example shown above, 아버지 [a beo ji] 'father' is marked with the honorable indirect object marker –께 [kke] rather than the plain marker –에게 [e ge]. However, the rule shown above must not apply if the nominal that takes the honorable indirect object marker –께 [kke] is in a subordinate clause, and the subject nominal in the matrix clause also takes the honorable morpheme. Therefore when this situation is found in the semantic representations, a structure bleeds the rule shown above and does not let it apply. That structure is shown below in figure 5-14.



Noun Phrase Grammatical Relation = Topic

Noun Takes Honorable Morphemes = Yes

Noun Person = NOT First
Noun Takes Honorable Morphemes = Yes

Figure 5-14. Structural Adjustment Rule that Prevents Changing –에게 [e ge] to –께 [kke] if Subject of Matrix Clause also takes Honorable Morphemes

The rule shown above applied in many places, one of them being Daniel 4:24 which is shown below in example (5-1xix).

(5-1xix) 하나님-께서는 천사-에게 대왕-에게 어떤 일-들-을 하-도록
God-Subj-Honor angel-to king-to certain thing-Pl-Obj do-Comp

명령하-시-었-습니다. (시-었 ->셨)
order–Honor-Past-Decl.Deferential
'*God commanded the angels to do certain things to you(king).'* (Korean text generated by TTA)

In the example shown above, the indirect object in the object complement clause is 대왕 [dae wang] 'great king', and generally 대왕 [dae wang] 'great king' must be marked with the honorable indirect object marker –께 [kke]. However, since the subject of the matrix clause is 하나님 [ha na nim] 'God', and since 하나님 [ha na nim] 'God' requires the honorable affixes, 대왕 [dae wang] 'great king' in the subordinate clause must not be marked with the honorable indirect object marker –께 [kke]. So this rule bleeds the previous rule and does not allow 대왕 [dae wang] 'great king' to be marked with –께 [kke].

This section described the five methods of encoding Korean honorifics. Section 5.2.1 described each of the five methods of signaling honor and examples were provided. Then section 5.2.2 discussed the features in the semantic representations that are pertinent when generating Korean honorific morphemes, and the rules which generate those morphemes were presented. There are certainly some honorific forms which TTA is not able to generate because the necessary information isn't available in the semantic representations. The most notable example of this is the numerous titles that Korean speakers use in places where an English speaker would use the second person singular pronoun or the listener's name. Section 5.2.2.3 mentioned that TTA is unable to generate a title such as 사모님 [sa mo nim] 'teacher's wife' because the feature called 'Listener' does not have a value called 'Wife of Teacher'. However, TTA is able to generate many of the titles, as well as the other honorific morphemes.

5.3 English Questions

This section will discuss English question formation in detail. The model for question generation adopted here is essentially identical to the model proposed by the Principles and Parameters theory and described thoroughly by Haegeman (1994:297-306). Therefore the details of English question formation will not be discussed. Instead this section will present the concepts and features used in the semantic representations for questions, and then the English rules that generate the proper surface forms will be described. The movement required during English question construction is extensive, so it will be discussed in detail.

5.3.1 The Concepts and Features of Content Questions in the Semantic Representations

TTA's ontology currently includes five question concepts that are event attributes:

- WHERE-A,
- WHEN-A,
- WHY-A,
- HOW-A, and
- HOW-LONG(time)-A.

These event attributes always occur in event attribute phrases. There is an additional question concept in the ontology that is an object attribute:

- HOW-MUCH/MANY-A.

This concept always occurs in object attribute phrases which modify an object. The concept WHICH could have been another object attribute, but instead it is indicated by the object Participant Tracking value of 'Interrogative' as was described in section 3.3.2.1.2.

Noticeably absent from this apparatus are the question concepts WHO and WHAT. In TTA's semantic representations, the English concept WHO is seen as the equivalent of *which person*. Therefore there is no concept WHO in TTA's ontology. Instead WHO is indicated in the semantic representations by marking the object PERSON-A with a Participant Tracking value of Interrogative. For example, the semantic representation of *Who read this book?* is shown below in figure 5-15.

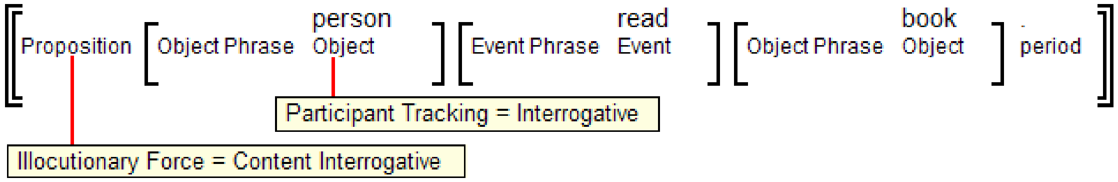


Figure 5-15. Semantic Representation of *Who read this book?*

Similarly there is no question concept WHAT in TTA's ontology. The concept WHAT is indicated in the semantic representations by marking the concepts THING-A or THING-B with a Participant Tracking value of Interrogative. THING-A is used for physical objects that can be held or seen; THING-B is used for abstract objects as in *John said many things.* An example illustrating THING-A marked with a Participant Value of Interrogative is shown below in figure 5-16. That figure shows the semantic representation of *What did John read?*

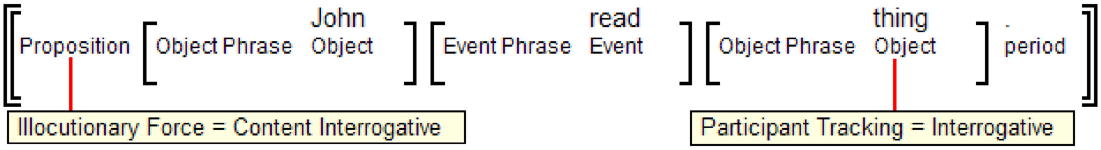


Figure 5-16. Semantic Representation of *What did John read?*

All objects other than PERSON-A, THING-A, and THING-B that have a Participant Tracking value of Interrogative require the insertion of the English interrogative article *which.* For example, the semantic representation for *Which book did John read?* is shown below in figure 5-17.

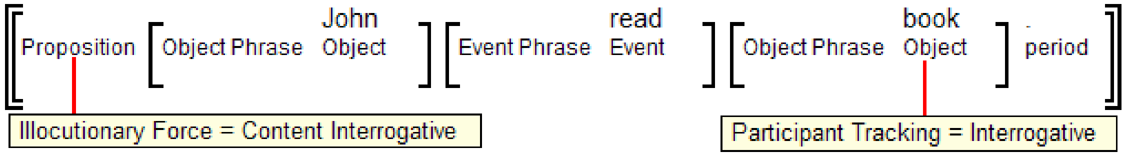


Figure 5-17. Semantic Representation of *Which book did John read?*

The semantic representations for questions that include an interrogative event attribute always include an event attribute phrase with the interrogative event attribute embedded. For example, the semantic representation of *When did John read a book?* is shown below in figure 5-18.

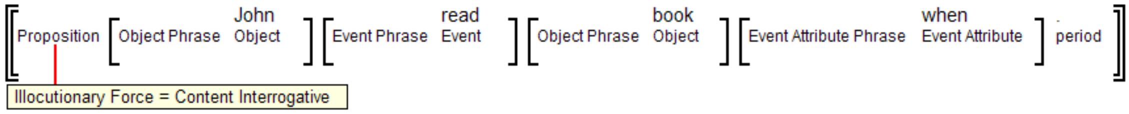


Figure 5-18. Semantic Representation of *When did John read a book?*

The semantic representations for *Where/Why/How/How long did John read a book?* are essentially identical to the semantic representation shown above in figure 5-18, except that the appropriate event attribute is inserted into the event attribute phrase.

5.3.2 The Concepts and Features of Yes/No Questions in the Semantic Representations

Yes/No questions are indicated in the semantic representations simply by marking a proposition's 'Illocutionary Force' with 'Yes/No Interrogative'; no additional concepts or features are required. For example, the semantic representation of *Did John read a book?* is shown below in figure 5-19.

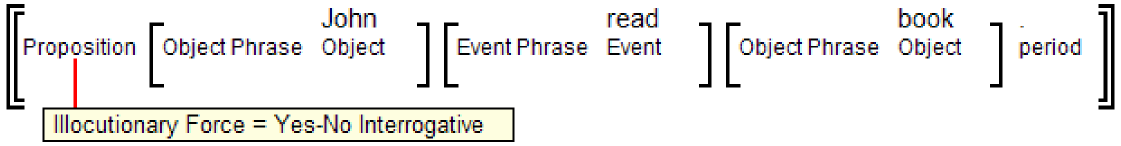


Figure 5-19. Semantic Representation for *Did John read a book?*

As seen in figure 5-19 above, yes/no questions are comprised of the same concepts as the declarative form, but the proposition's 'Illocutionary Force' value is set to 'Yes/No Interrogative' rather than 'Declarative'[34].

As was stated in section 4.4.4, the CP and CP-Spec nodes are not included in the semantic representations. Therefore a rule for English must insert those nodes into all propositions that have an Illocutionary Force value of either Content Interrogative or Yes/No Interrogative. The structure that performs that insertion was shown in figure 4.42. After those two nodes have been inserted into a question, the movement rules are able to move the appropriate constituents to the proper node. However, before the movement occurs, the necessary interrogative forms and the interrogative auxiliary must be generated. The process of generating English interrogatives will be described in the following section.

5.3.3 The Process of Generating English Surface Structure for Questions

The process of generating English questions can be divided into three steps: 1) insert the interrogative auxiliary *do* under certain conditions, 2) insert the interrogative article *which* under certain conditions, and 3) move the appropriate constituents to CP and CP-Spec. These three steps will be described in detail.

5.3.3.1 Generating the Interrogative Auxiliary *do*

In order to generate the proper surface forms for English questions, the interrogative auxiliary *do* must be inserted under the following conditions:

- The tense must be either past or present. If the tense is future as in *When will John read a book?*, then the interrogative auxiliary *do* is not required.
- The aspect must not be imperfective. If the aspect is imperfective as in *Why is John reading that book?*, then the imperfective auxiliary *be* is required rather than the interrogative auxiliary *do*, and *be* will be moved to CP rather than *do*. If the aspect is

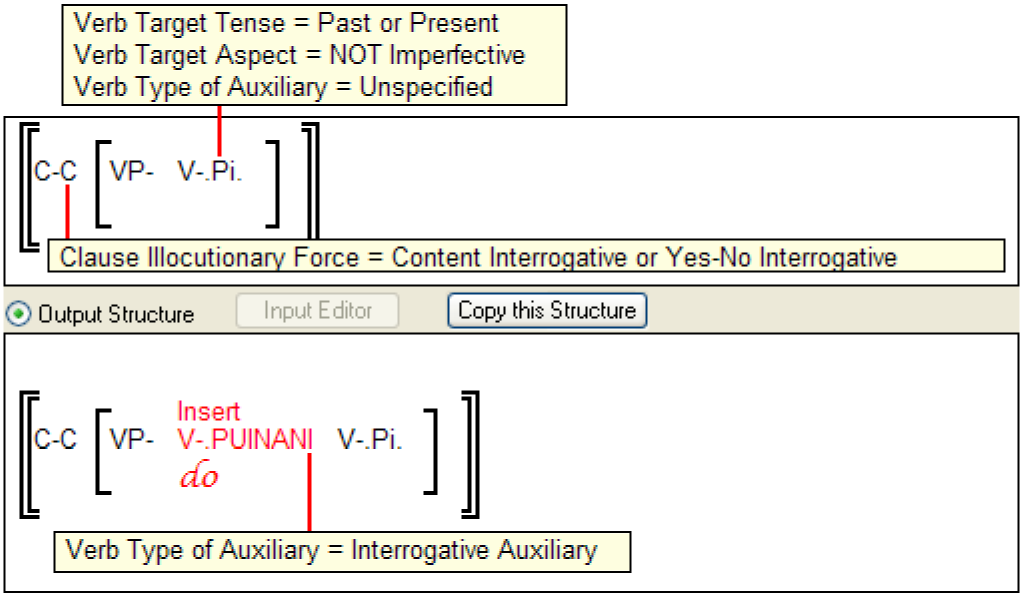
[34] TTA's semantic representations don't include facilities for tag questions.

anything other than imperfective as in *Why did John start/stop/finish reading that book?*, then the auxiliary *do* is required.

- The mood must be Indicative. If the mood is a value other than Indicative as in *When should/might/must John read the book?*, then the interrogative auxiliary *do* is not required.
- The subject nominal must not be the interrogative nominal. If the subject NP contains the interrogative element as in *Who read this book?* or *Which student read this book?*, then the auxiliary *do* must not be inserted.
- The main verb must not be *be*. If the main verb is *be* as in *Where is John?*, then the auxiliary *do* is not required.

The structural adjustment rule that looks for each of these situations and appropriately inserts the interrogative auxiliary *do* is shown below in figure 5-38. That rule has four structures. In all of the rules that have multiple structures[35] in TTA's grammar, the structures are always executed from last to first. The reason for this is so that the latter structures are able to bleed the earlier structures. The early structures do the real work intended by the rule, and the latter structures contain specific constructions where the rule must not be applied. In this particular structural adjustment rule, the last three structures bleed the first structure and prevent it from applying in particular situations. If none of the latter three structures match the current semantic representation, then the first structure will be executed and insert the auxiliary *do*. The first structure for this rule is shown below in figure 5-20.

[35] Each of the Theta Grid Adjustment rules, Structural Adjustment rules, Movement rules, and Pronoun Identification rules may have up to 99 structures.


Verb Target Tense = Past or Present
Verb Target Aspect = NOT Imperfective
Verb Type of Auxiliary = Unspecified

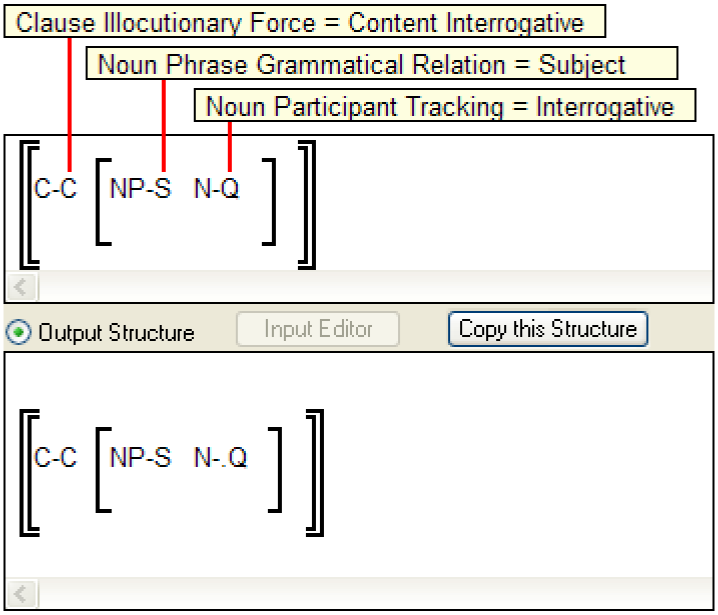
Clause Illocutionary Force = Content Interrogative or Yes-No Interrogative

Verb Type of Auxiliary = Interrogative Auxiliary

Figure 5-20. Structural Adjustment Rule that Inserts the English Interrogative Auxiliary *do*

The rule shown above looks specifically for verbs that have a tense value of past or present, an aspect value that is anything other than imperfective, and a proposition that has an illocutionary force value of either content interrogative or yes/no interrogative. When all those conditions are satisfied, this rule will insert the verb *do,* and set a feature to mark it as an interrogative auxiliary. Subsequent rules will move the interrogative auxiliary appropriately. This rule applied in many places when the English text was generated, a few of those places being: Clauses 1:92 *What did John read?* Clauses 1:93 *When did John read a book?* Clauses 1:97 *Which book did John read?* Clauses 1:98 *To whom did John give a book?* Clauses 1:99 *Which book did Mary want John to read?* Clauses 1:100 *Did John read a book?* Clauses 1:102 *Does John want to read a book?* Kande's Story 1:2 *Do you know mother's secret?* Avian Influenza 1:5 *How does this disease spread?*

This rule has three additional structures which bleed the first structure. The structure shown below is in the same rule that is shown in figure 5-20 above, but the second structure is displayed. The second structure bleeds the first structure and prevents the interrogative auxiliary *do* from being inserted when the subject of the sentence contains the interrogative nominal.



Clause Illocutionary Force = Content Interrogative

Noun Phrase Grammatical Relation = Subject

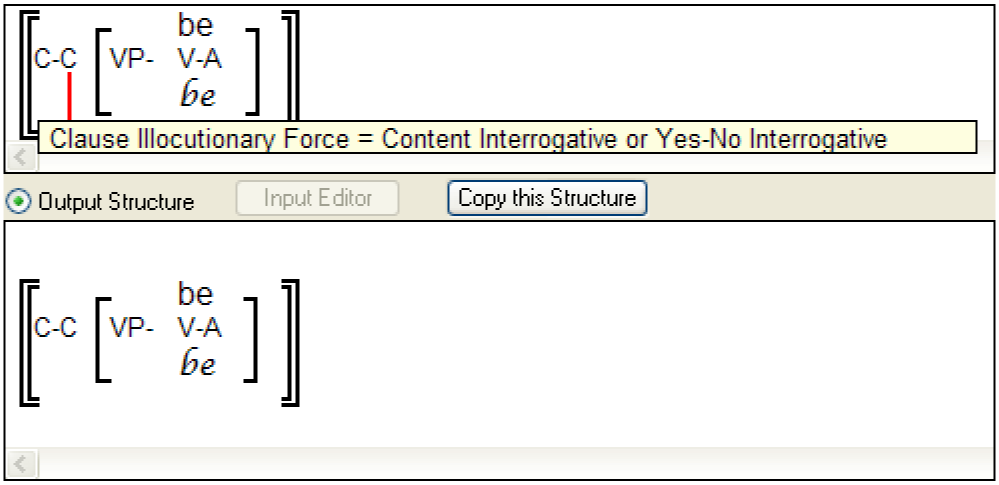
Noun Participant Tracking = Interrogative

Figure 5-21. Structural Adjustment Rule that Prevents *do* Insertion when Subject is *who, what* or *which* X

The structure shown above does not change anything in the proposition; it simply bleeds the first structure so that the interrogative auxiliary *do* will not be inserted when the subject nominal is marked as 'Interrogative'[36]. This structure applied in the following places: Clauses 1:91 *Who read a book?* Luke 8:45 *Who touched me?* Luke 10:36 *Which man acted like this man's neighbor?* Genesis 27:33 *Before you came into my tent, who brought food to me?*

[36] English speakers can say something like "Who DID read this book?", but that construction is unusual and requires a particular setting such as the following: A teacher is addressing a group of students who were supposed to read a book, but the teacher believes that very few if any of the students actually read the book. In that situation the teacher might ask the class, "So who did read this book?" TTA's semantic representations aren't able to capture the subtleties of situations such as these because the underlying presupposition is not overtly encoded in the semantics. Detailed discussions of contrastive stress and presupposition may be found in the following two articles: 1) Atlas, J. and Stephen Levinson. 1981. It-clefts, Informativeness, and Logical Form. In: P. Cole, Editor, Radical Pragmatics, Academic Press, New York. 2) Prince, E. 1978. A Comparison of wh-clefts and it-clefts in Discourse. Language, Journal of the Linguistic Society of America Baltimore, Maryland. 54:4 pp. 883-906.

The third structure of this rule is shown below in figure 5-22. Similar to the second structure, it does not change anything in the semantic representation; its only purpose is to prevent the first structure from applying when the main verb is *be*.

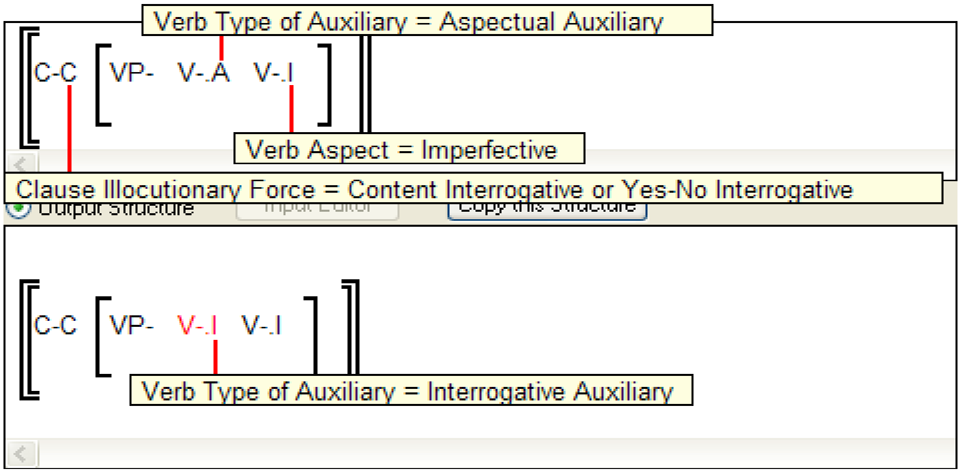


Clause Illocutionary Force = Content Interrogative or Yes-No Interrogative

Figure 5-22. Structural Adjustment Rule that Prevents *do* Insertion when the Main Verb is *be*

This structure applied in many places, a few of them being: Clauses 1:103 *Is John able to read a book?* Infected Eye 1:2 *Is some sand in my eyes?* Ruth 1:19 *Is this woman Naomi?* Ruth 2:5 *Who is that woman?* Ruth 2:10 *Why are you kind to me?* Esther 4:5 *Why are you sad?*

The final structure of this rule is shown below in figure 5-23. It checks if the imperfective auxiliary has been inserted into a content question. If so, it changes the feature on the auxiliary from Aspectual Auxiliary to Interrogative Auxiliary.



Verb Type of Auxiliary = Aspectual Auxiliary

Verb Aspect = Imperfective

Clause Illocutionary Force = Content Interrogative or Yes-No Interrogative

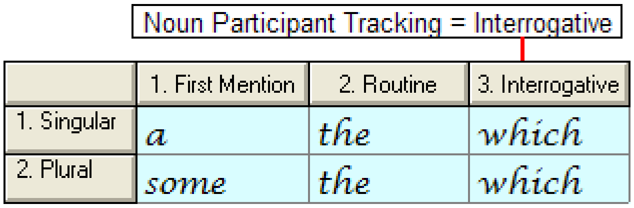
Verb Type of Auxiliary = Interrogative Auxiliary

Figure 5-23. Structural Adjustment Rule that Marks the Imperfective Aspectual Auxiliary *be* as the Interrogative Auxiliary

The structure shown above simply changes the feature value called ‘Type of Auxiliary’ from ‘Aspectual Auxiliary’ to ‘Interrogative Auxiliary’. The aspectual auxiliary is *be* rather than *do*, and it was inserted by an earlier rule that deals with aspect. However, when the aspect of a question is imperfective, English requires *be* in the question rather than *do*. Subsequent interrogative movement rules will move the aspectual auxiliary *be* because it is now marked as Interrogative Auxiliary. This structure applied in the following places: Luke 1:21 *Why is Zechariah staying in the holy room for a long time?* Genesis 21:17 *Hagar, why are you crying?* Genesis 32:17 *Where are you going?* Daniel 4:35 *Why are you doing these things?*

5.3.3.2 Inserting the Interrogative Article *which*

Objects which have a Participant Tracking value of Interrogative require the word *which* to be inserted. This insertion is accomplished by the table spellout rule shown below in figure 5-24.



Noun Participant Tracking = Interrogative

	1. First Mention	2. Routine	3. Interrogative
1. Singular	a	the	which
2. Plural	some	the	which

Figure 5-24. Table Spellout rule that Inserts English Articles

The word *which* generally is not called an article by English grammarians, but since it occupies the same position as articles and is mutually exclusive with articles, this rule inserts *which* with a label of Article. Then the PSR for NPs positions it properly. Two additional spellout rules are necessary to convert *which person* to *who/whom*, and *which thing* to *what*. Those spellout rules are very simple and will not be shown here.

5.3.3.3 Movement Required for English Questions

As was mentioned above, the movement required by English questions is extensive. Fourteen different structures were required to deal with all the possible constructions. Ten structures deal with content questions, and four structures deal with yes/no questions. Since these fourteen structures are in a single movement rule, TTA's grammar will execute them from last to first. The discussion here will present them in the same order that they are executed by TTA's grammar – from last to first. As was mentioned above, the latter structures are often used to bleed the earlier structures. In this particular rule no bleeding is necessary, but the latter structures handle very specific constructions while the earlier structures handle very generic constructions.

5.3.3.3.1 Movement Required for English Yes/No Questions

The last structure for English yes/no question movement is shown below.

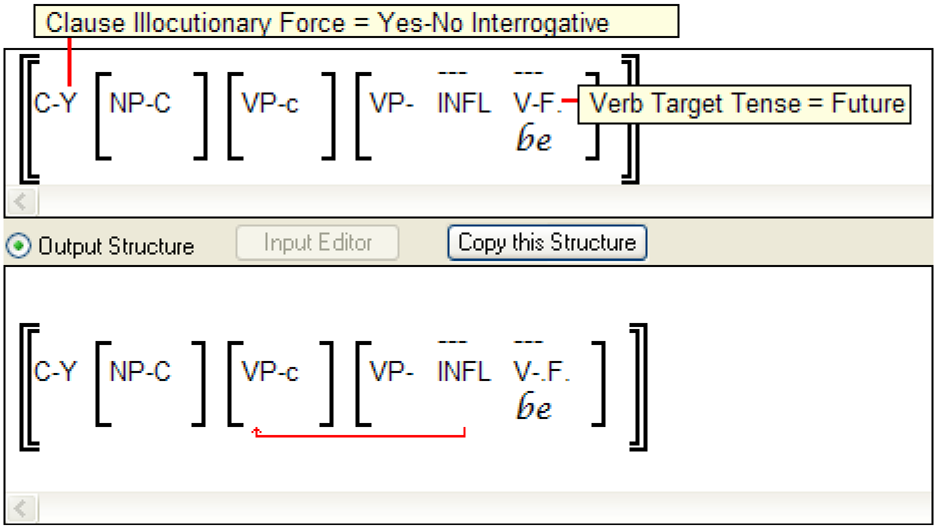


Figure 5-25. Structure 14 in the Interrogative Movement Rule

The structure above applies only to yes/no questions that have a main verb of *be* with future tense. When that structure is found, the INFL node will be moved to CP. This structure applied in Genesis 24:49 *Will you be kind to my master?*

When the main verb is *be* but the tense is present or past, then both INFL and *be* must be moved to CP. The structure that performs that task is shown below in figure 5-26.

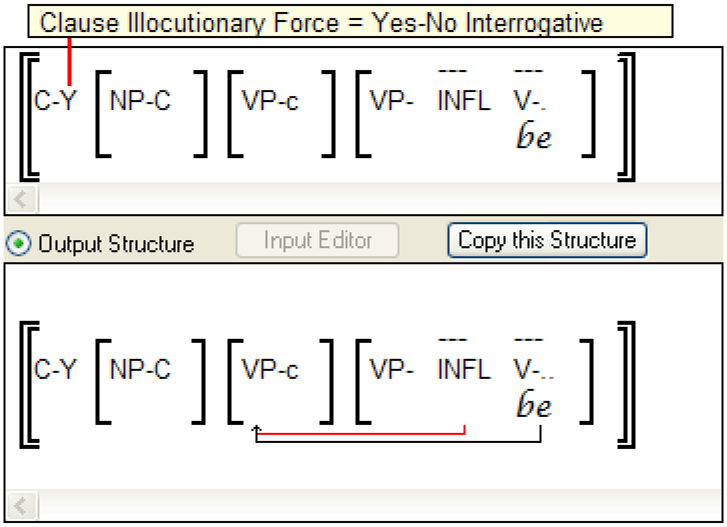
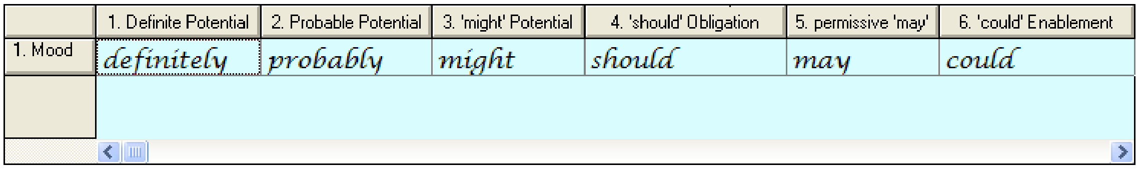


Figure 5-26. Structure 13 in the Interrogative Movement rule

The structure shown above is similar to the structure shown in figure 5-25 above, but it is more generic because the tense is not specified. Since the previous structure handled all the future tense cases, the tense of *be* does not need to be specified in this structure. This structure applied in places such as the following: Clauses 1:103 *Is John able to read a book?* Infected Eye 1:2 *Is some sand in my eyes?* Ruth 1:19 *Is this woman Naomi?*

The next structure looks specifically for yes/no questions that have a Mood word. The mood words are inserted by the table spellout rule shown below in figure 5-27.



	1. Definite Potential	2. Probable Potential	3. 'might' Potential	4. 'should' Obligation	5. permissive 'may'	6. 'could' Enablement
1. Mood	definitely	probably	might	should	may	could

Figure 5-27. Table Spellout Rule that Inserts English Mood Words

The table shown above inserts the appropriate mood word with a label of Mood. Then the structure shown below in figure 5-28 looks for yes/no questions that have a Mood word.

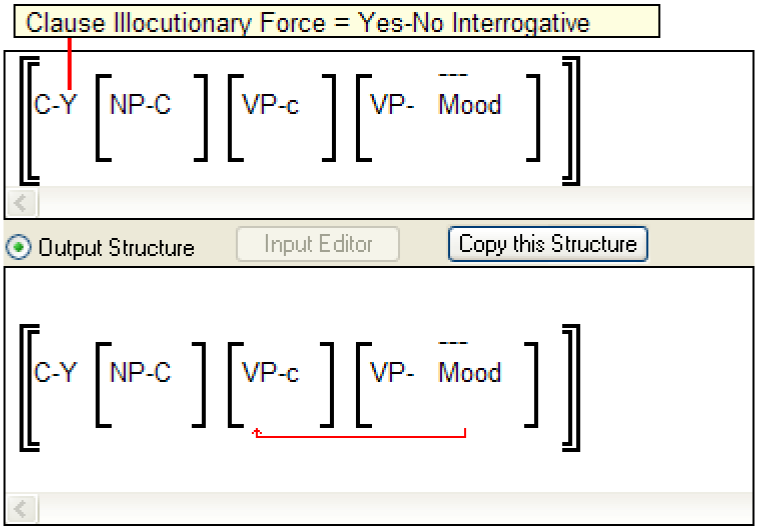


Figure 5-28. Structure 12 in the Interrogative Movement Rule

When the mood is ‘might’ as in *John might read this book*, ‘must’ as in *John must read this book*, ‘should’ as in *John should read this book,* or permissive ‘may’ as when a parent says *John may read this book*, the Target Tense/Aspect/Mood rules for English set the ‘Target Tense’ feature to ‘Unspecified’. Therefore the INFL node for these moods is blank. When this is the case, the movement rule shown above simply moves the Mood word from the VP to CP. The rule shown above applies in situations such as: Clauses 1:101 *Should John read a book?* Ruth 2:2 *May I go to a kind person’s field in order to glean barley?* Ruth 2:7 *May I glean barley behind the workers?* Genesis 4:9 *Must I always take care of my younger brother?*

The most generic movement rule for yes/no questions is shown below in figure 5-29.

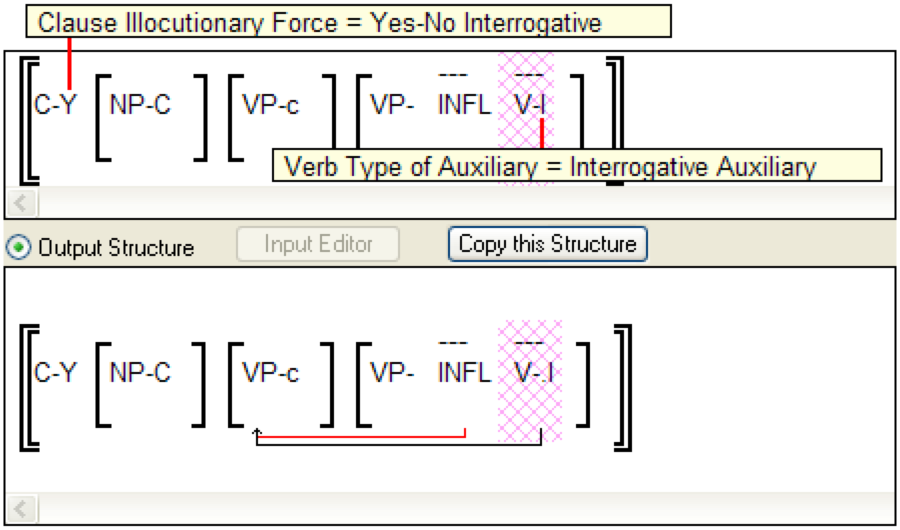
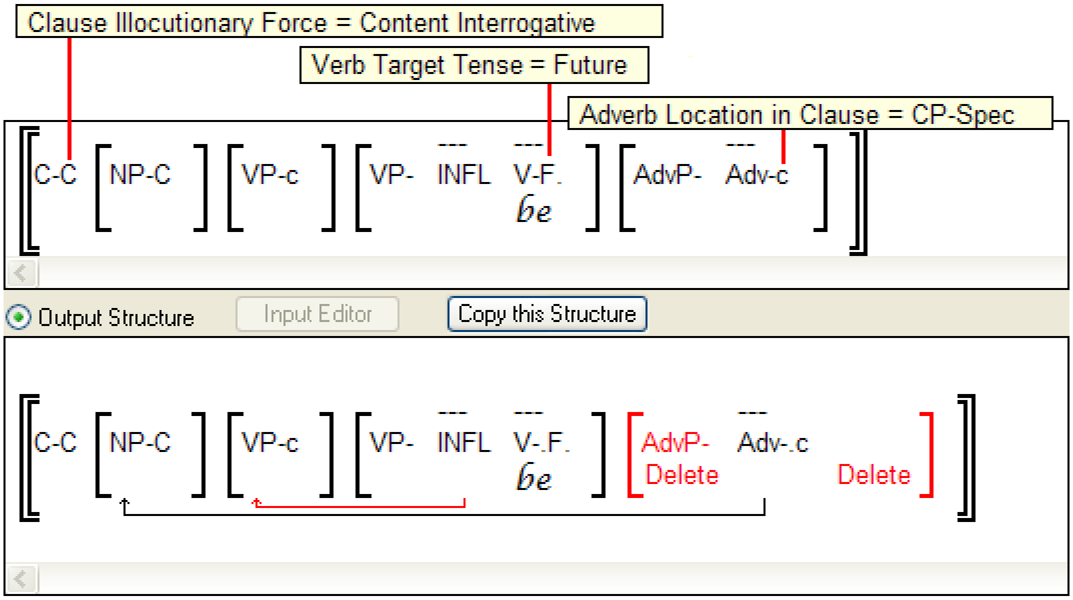


Figure 5-29. Structure 11 in the Interrogative Movement Rule

The structure shown above moves the INFL node to CP, and it also moves the optional Interrogative Auxiliary to CP if one is present. Examples of Yes/No questions that include the *do* auxiliary follow: Clauses 1:100 *Did John read a book?* Kande's Story 1:2 *Do you know mother's secret?* Kande's Story 1:10 *Does father have AIDS?* Kande's Story 3:10 *Did our uncle say that we have to move to another house?* Examples of Yes/No questions that do not include the *do* auxiliary include the following: Kande's Story 2:13 *If I touch mother, will I catch AIDS?* Esther 7:8 *Will you attack the queen while I'm with her in the palace?* Genesis 18:24 *If fifty righteous people live in that city, will you destroy it?*

5.3.3.3.2 Movement Required for English Content Questions

Structures 1 through 10 in the current movement rule deal with movement in content questions. These structures will be described in reverse order because that is how they are executed. As with the structures that deal with yes/no questions, these structures are ordered so that the more specific constructions are handled by the latter structures, and the more generic constructions are handled by the early structures. The most specific content question movement structure is shown below in figure 5-30.



Clause Illocutionary Force = Content Interrogative

Verb Target Tense = Future

Adverb Location in Clause = CP-Spec

Figure 5-30. Structure 10 in the Interrogative Movement Rule

The structure shown above in figure 5-30 looks for content questions that include a content question event attribute word such as *when, where, why, how* or *how long*, and also have a main verb of *be* with future tense. When that construction is found, INFL is moved to CP and the content question word is moved to CP-Spec. This structure has applied in only one location in the semantic representations that have been developed: Luke 1:34 *How will I be able to give birth to a son?*

The next structure is very similar to structure 10 shown above, but it is slightly more generic in that the tense of *be* is unspecified. Structure 9 of the interrogative movement rule is shown below in figure 5-31.

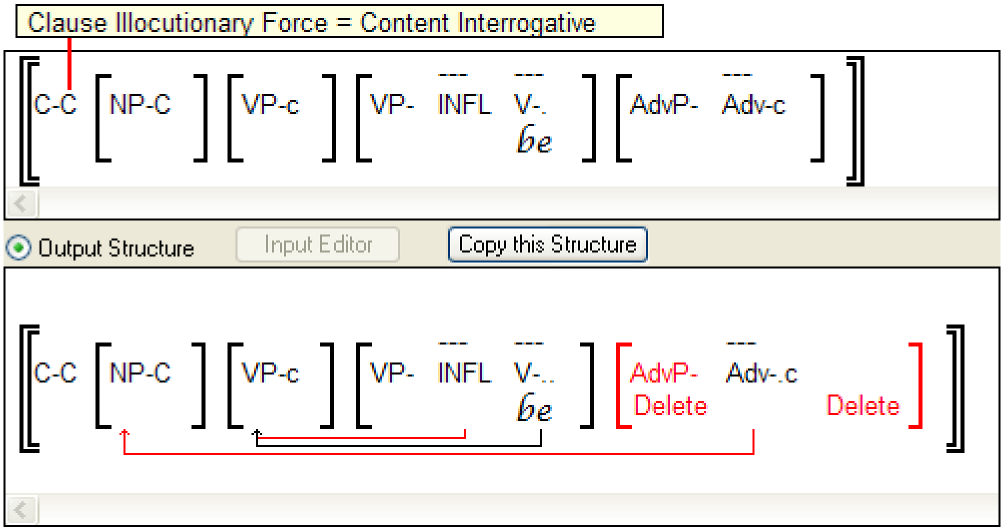
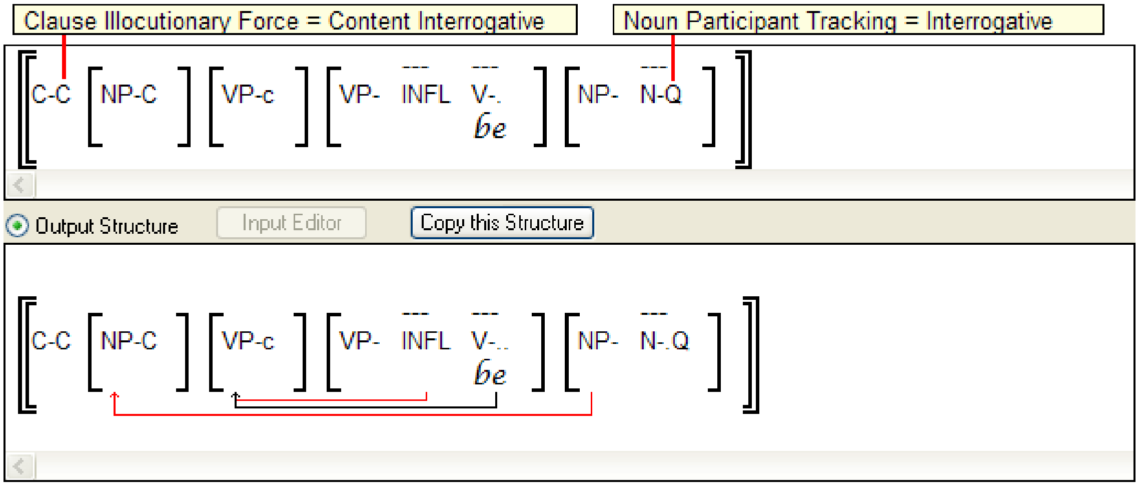

Clause Illocutionary Force = Content Interrogative

Figure 5-31. Structure 9 in the Interrogative Movement Rule

This structure moves both INFL and *be* to CP, and moves the interrogative event attribute to CP-Spec. This rule applied in many situations including the following: Ruth 2:10 *Why are you kind to me?* Genesis 3:9 *Where are you?* Genesis 4:9 *Where is your younger brother named Abel?*

The eighth structure is similar to the ninth structure, but it looks for propositions that have a nominal with a Participant Tracking value of Interrogative. That structure is shown below in figure 5-32.


Clause Illocutionary Force = Content Interrogative

Noun Participant Tracking = Interrogative

Figure 5-32. Structure 8 in the Interrogative Movement Rule

This structure applied in many places, a few of them being: Ruth 2:5 *Who is that woman?* Esther 6:4 *Who is in the courtyard?* Daniel 5:7 *Who is able to read this message?*

The seventh structure in this rule specifies a Mood word in the VP as seen below in figure 5-33.

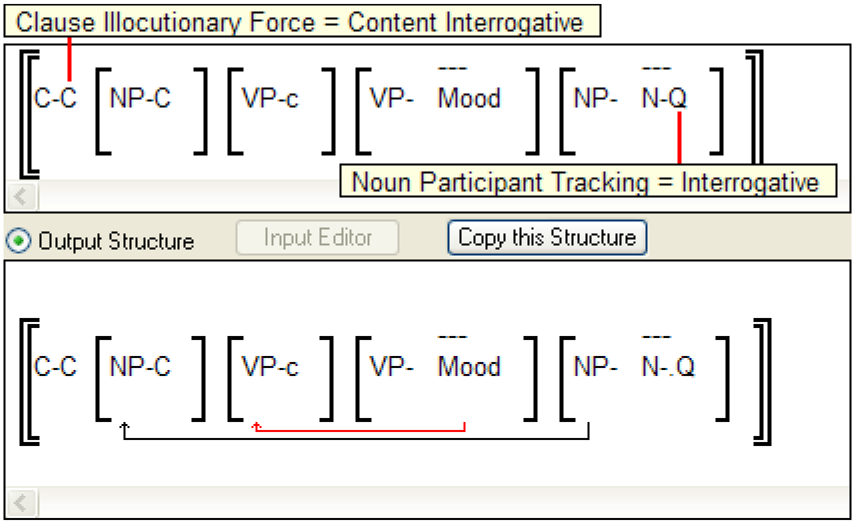
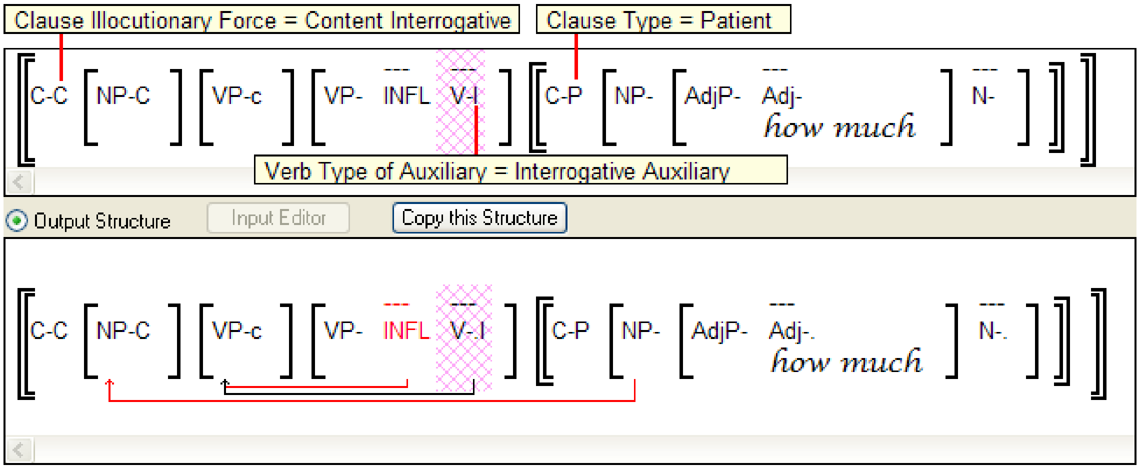


Figure 5-33. Structure 7 in the Interrogative Movement Rule

This structure looks for content questions that include a Mood word such as *should, might, must,* or *may,* and a nominal with a Participant Tracking value of Interrogative. When such a situation is found, the Mood word is moved to CP, and the entire NP that contains the interrogative nominal is moved to CP-Spec. Examples where this structure applied include: Luke 3:10 *What should we do?* Luke 10:25 *What must I do in order to live forever?* Genesis 30:31 *What should I give to you?*

The next structure looks specifically for the object attribute HOW-MUCH in a patient proposition that is embedded in a content question. When that situation is found, the entire NP containing HOW-MUCH is moved to the matrix CP-Spec node, and INFL from the matrix clause is moved to CP. If the matrix VP contains the interrogative auxiliary *do*, it also is moved to CP. This structure is shown below in figure 5-34.


Clause Illocutionary Force = Content Interrogative

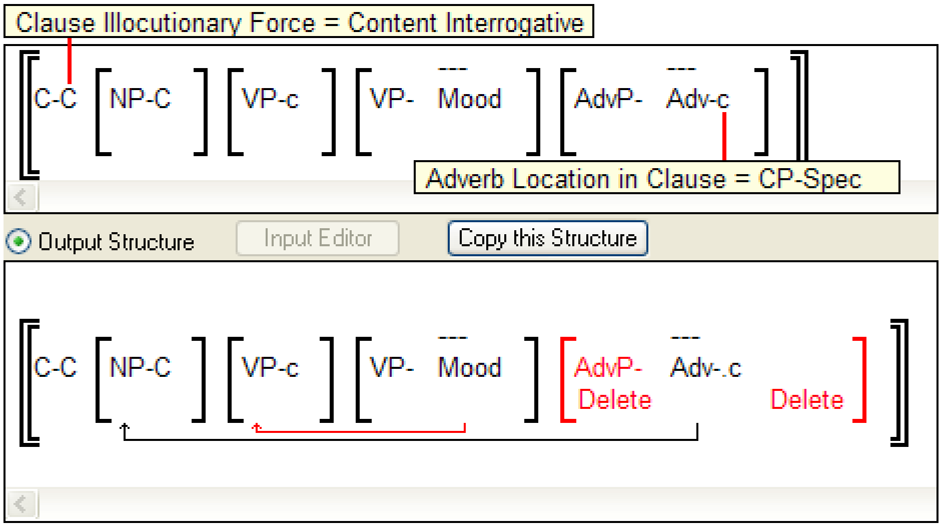
Clause Type = Patient

Verb Type of Auxiliary = Interrogative Auxiliary

Figure 5-34. Structure 6 in the Interrogative Movement Rule

This structure applied in only one location in the semantic representations developed thus far: Genesis 29:15 *How much money do you want me to pay you?*

The fifth structure of this rule looks for a Mood word and an interrogative event attribute. This structure is shown below in figure 5-35.


Clause Illocutionary Force = Content Interrogative

Adverb Location in Clause = CP-Spec

Figure 5-35. Structure 5 in the Interrogative Movement Rule

This structure applied in three locations: Esther 1:15 *How should I punish the queen?* Esther 6:6 *How should I honor that man?* Luke 9:41 *How long must I stay with you?*

The fourth structure in this rule is shown below in figure 5-36. It is similar to structure 9, but the main verb is unspecified so it may be anything other than *be*.

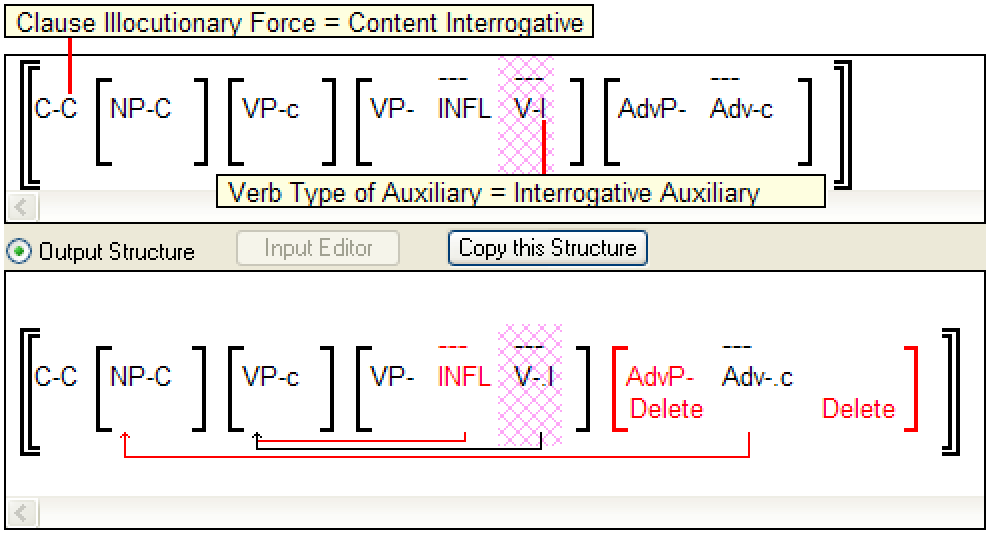
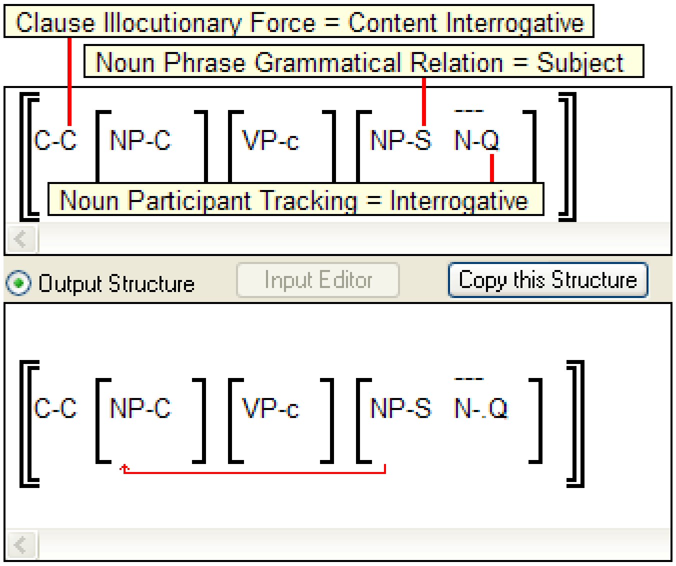


Figure 5-36. Structure 4 in the Interrogative Movement Rule

The structure shown above moves the content question event attribute to CP-Spec, and it moves INFL to CP. If the optional interrogative auxiliary *do* is present, then it will be moved to CP also. This structure is very generic and applied in many places, a few of them being: Clauses 1:93 *When did John read a book?* Kande's Story 1:12 *How will we live?* Kande's Story 3:8 *Where will my family live?* Avian Influenza 1:5 *How does this disease spread?* Luke 1:18 *How will my wife give birth to a baby?*

The third structure of this rule is shown below in figure 5-37. It looks for interrogative nominals in subject NPs. When that situation is found, it moves the entire subject NP into CP-Spec. When this is the case, moving INFL to CP is vacuous so it is not shown here.


Clause Illocutionary Force = Content Interrogative
Noun Phrase Grammatical Relation = Subject
C-C
NP-C
VP-c
NP-S

N-Q
Noun Participant Tracking = Interrogative
Output Structure
Input Editor
Copy this Structure
C-C
NP-C
VP-c
NP-S

N-.Q

Figure 5-37. Structure 3 in the Interrogative Movement Rule

This structure applied in the following places: Clauses 1:91 *Who read a book?* Luke 8:45 *Who touched me?* Luke 10:36 *Which man acted like this man's neighbor?* Nahum 3:7 *Who will mourn for Nineveh?* Genesis 27:33 *Before you came into my tent, who brought food to me?*

The second structure of this movement rule is shown below in figure 5-38. It looks for an interrogative nominal in an object complement clause. When this situation is found, the entire NP containing the interrogative nominal is moved to the matrix clause's CP-Spec, and the matrix INFL and optional interrogative auxiliary are moved to CP.

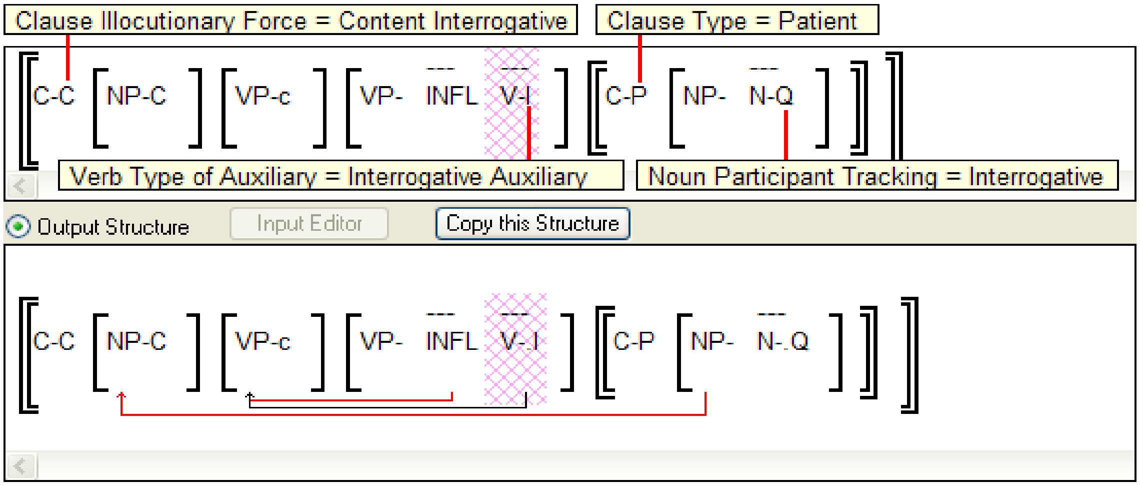
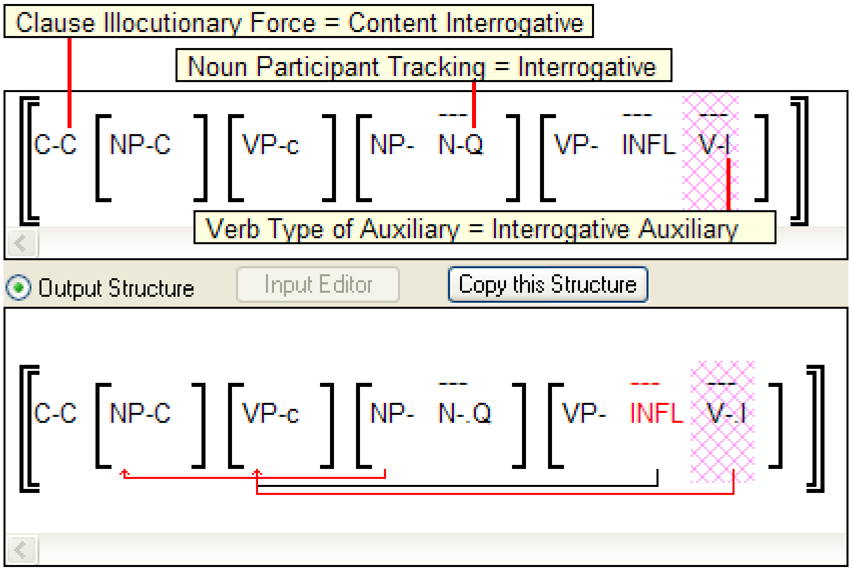


Figure 5-38. Structure 2 in the Interrogative Movement Rule

This structure applied in the following places: Clauses 1:99 *Which book did Mary want John to read?* Luke 7:24 *When you went to the desert, what did you want to see?* Luke 9:18 *Who do people say that I am?* Luke 9:20 *Who do you think that I am?* Luke 10:26 *What does the law command you to do?*

The first structure of this movement rule is shown below in figure 5-39. This structure looks for a nominal with a 'Participant Tracking' value of 'Interrogative' in a content question. The grammatical relation of the interrogative nominal is irrelevant so it is unspecified. The nominal may be modified by a preposition, relative clause, adjective, etc., but the modifiers are not relevant so they are unspecified. When this case is found, the entire NP containing the interrogative nominal is moved to CP-Spec, and INFL is moved to CP. If the optional interrogative auxiliary *do* is found, then it also is moved to CP.


Clause Illocutionary Force = Content Interrogative

Noun Participant Tracking = Interrogative

Verb Type of Auxiliary = Interrogative Auxiliary

Figure 5-39. Structure 1 in the Interrogative Movement Rule

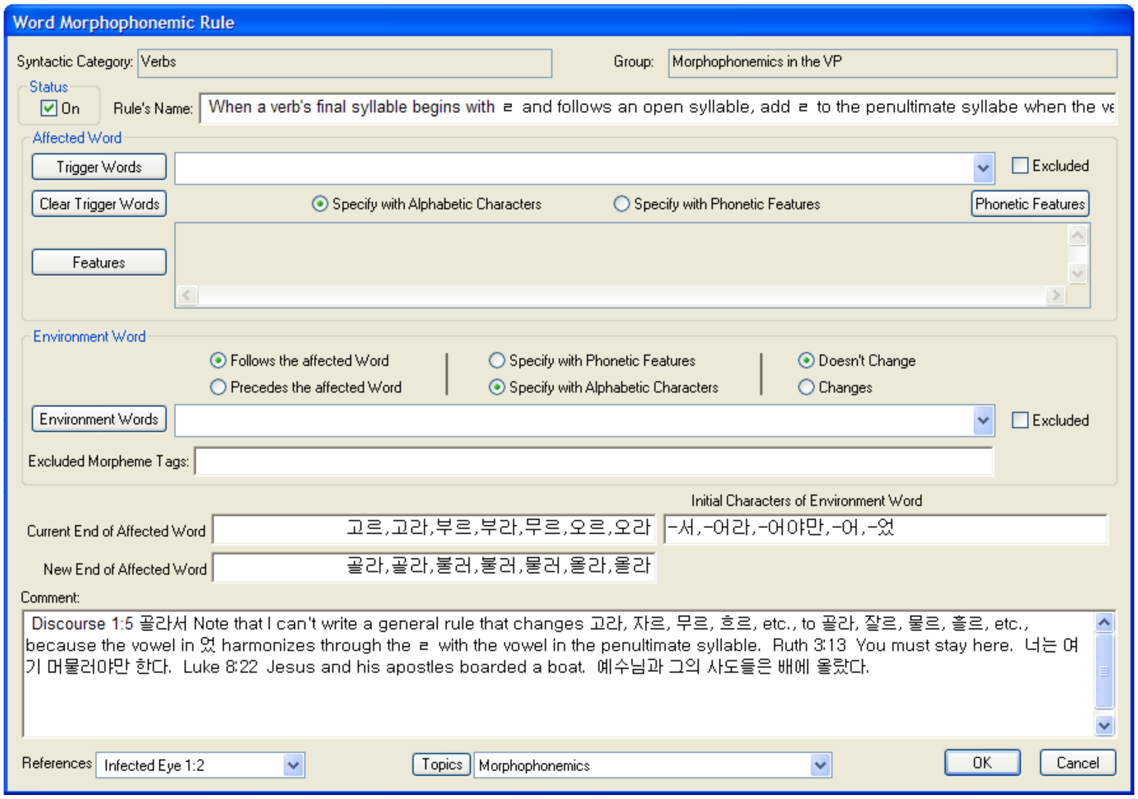
Several of the places where this structure applied include the following: Clauses 1:92 *What did John read?* Clauses 1:97 *Which book did John read?* Clauses 1:98 *To whom did John give a book?* Luke 10:26 *What do the books that Moses wrote say?* Ruth 3:16 *What did you do during the night?*

This section described in detail how Yes/No questions and content questions are constructed in the semantic representations. Then this section briefly presented the Principles and Parameters model for generating English questions. This model consists of three steps: 1) *do* insertion, 2) *which* insertion, and 3) movement to CP and CP-Spec. Finally this section showed the English rules in TTA that are responsible for generating English questions. These rules insert *do* and *which* where they are appropriate, and they also insert CP and CP-Spec into each question in the semantic representations. Then another rule moves the appropriate constituents to their final destinations.

5.6 Conclusions

This chapter has presented in detail some of the more intricate issues involved when generating text in English and Korean. This chapter presented an overview of the elaborate Korean system of honorifics, and it was shown how the rules in TTA are able to generate most of the honorific forms. This chapter also showed how a linguist is able to use TTA's grammatical apparatus to build a system resembling the Principles and Parameters model for English question generation.

There were certainly additional issues that had to be dealt with in each language. For example, Korean has numerous morphophonemic operations which required several dozen rules that all look essentially like the one shown below in figure 5-40.



Syntactic Category: Verbs

Group: Morphophonemics in the VP

Status: ☑ On

Rule's Name: When a verb's final syllable begins with ㄹ and follows an open syllable, add ㄹ to the penultimate syllabe when the ve

Affected Word

□ Excluded

● Specify with Alphabetic Characters ○ Specify with Phonetic Features

Environment Word

● Follows the affected Word ○ Precedes the affected Word

○ Specify with Phonetic Features ● Specify with Alphabetic Characters

● Doesn't Change ○ Changes

□ Excluded

Excluded Morpheme Tags:

Current End of Affected Word: 고르,고라,부르,부라,무르,오르,오라

Initial Characters of Environment Word: -서,-어라,-어야만,-어,-었

New End of Affected Word: 골라,골라,불러,불러,물러,올라,올라

Comment:
Discourse 1:5 골라서 Note that I can't write a general rule that changes 고라, 자르, 무르, 흐르, etc., to 골라, 잘르, 물르, 흘르, etc., because the vowel in 었 harmonizes through the ㄹ with the vowel in the penultimate syllable. Ruth 3:13 You must stay here. 너는 여기 머물러야만 한다. Luke 8:22 Jesus and his apostles boarded a boat. 예수님과 그의 사도들은 배에 올랐다.

References: Infected Eye 1:2

Topics: Morphophonemics

Figure 5-40. Example of a Korean Morphophonemic Rule

The rule shown above changes 고라-서 [go ra – seo] to 골라-서 [gol la – seo], 오라-었 [o ra – eot] to 올라-었 [ol la – eot], etc. Korean does not employ predicative constructions like English, so all the predicative constructions in the semantic representations had to be converted to Korean verbs. Korean also has a complex system of classifiers and two sets of numerals. Speakers of Korean must memorize which classifier and which set of numerals to use when counting different types of objects. Although there were many other issues to be dealt with in these test languages, the grammatical apparatus developed for this project was able to handle each of them. However, it must be repeated that TTA is intended to generate text that is at approximately a sixth grade reading level. TTA does not produce sophisticated, high quality literature, nor does it produce the final translation. Editing of TTA's texts by mother-tongue speakers is essential in order to make the texts sound natural.

Figures 5-41 and 5-42 shown below contain graphs indicating the number of new grammatical rules that were required for each chapter of text. These graphs clearly indicate that each subsequent chapter of text required less work from the linguists.

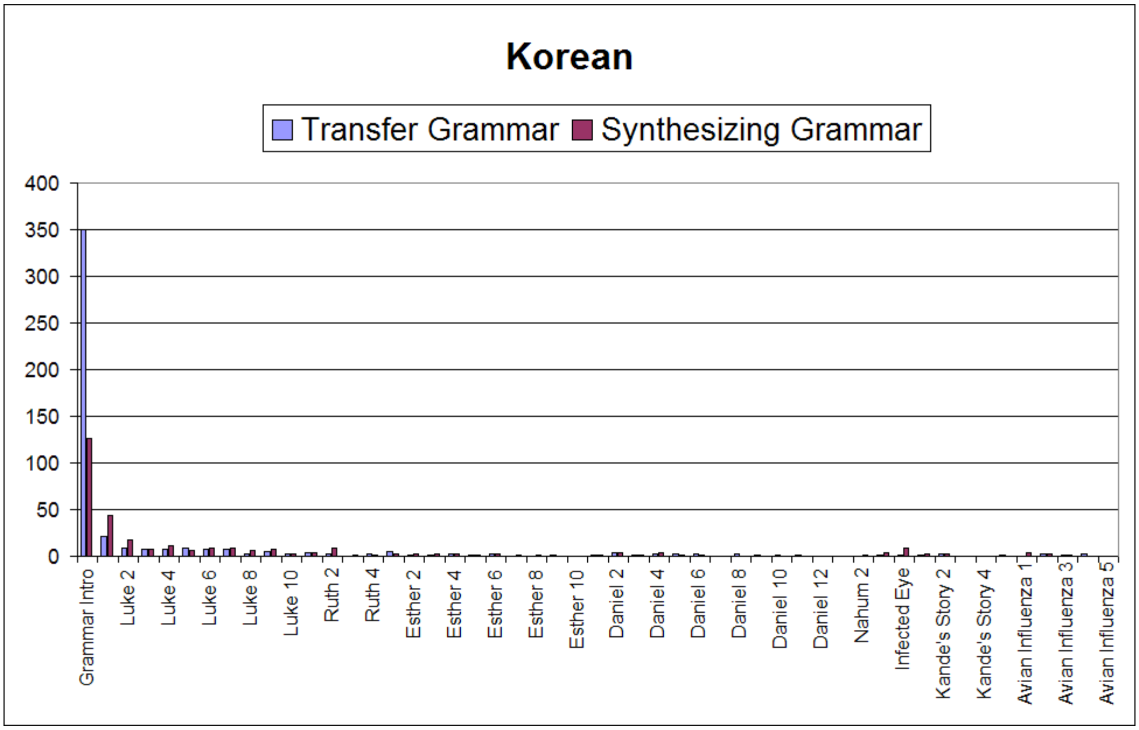


Figure 5-41. Number of New Rules Required for each Chapter of Korean Text

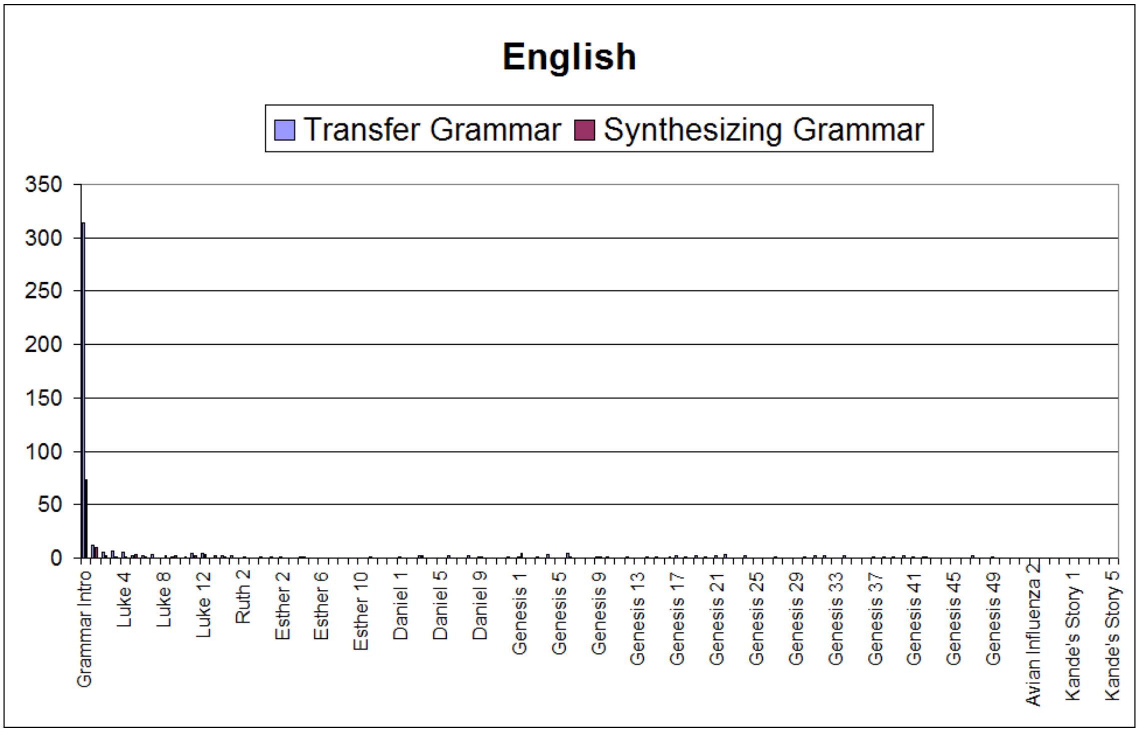


Figure 5-42. Number of New Rules Required for each Chapter of English Text

These graphs indicate that TTA's grammatical rules are genuinely capturing the significant linguistic generalizations of each language. These graphs also indicate that developing a language's transfer grammar is generally much more complex than developing the synthesizing grammar. In every case there were significantly more transfer rules required than synthesizing rules. The synthesizing grammars were generally developed much more quickly and easily than were the transfer grammars.

As was mentioned above, substantial amounts of text were generated in English and Korean. Then experiments were performed with the Korean texts in order to determine the quality of the generated drafts. The purpose of those experiments is to answer the following question: Are the generated texts of sufficient quality that they improve the productivity of experienced mother-tongue translators? The next chapter will describe those experiments and present the results. Appendix A contains the complete and unedited English and Korean drafts of Infected Eye, Avian Influenza, and Kande's Story as they were generated by TTA. Those drafts serve to illustrate the quality of the texts that can be expected from TTA.

CHAPTER 6

EVALUATING THE QUALITY OF THE TEXTS GENERATED BY TTA

6.1 Introduction

This chapter will describe the experiments that were performed to evaluate the quality of the texts generated by TTA. The primary purpose of these experiments was to determine whether or not the Korean drafts generated by TTA are of sufficient quality that they improve the productivity of experienced mother-tongue translators. The results of these experiments clearly indicate that when Korean mother-tongue translators use TTA's drafts, their productivity is typically quadrupled without any loss of quality.

In order to determine whether or not the texts generated by TTA are of sufficient quality that they improve the productivity of experienced mother-tongue translators, experiments were performed with the computer generated Korean drafts. The experiments consisted of the following two phases:

1) *Test for Increased Productivity*: Compare the quantity of text that an experienced mother-tongue translator is able to translate in a given amount of time with the quantity of computer generated text that the same mother-tongue translator is able to edit in the specified time.

2) *Test for Quality and Naturalness of Translation*: Ask other mother-tongue speakers to compare the texts edited in the first phase with professionally translated texts in order to determine if the edited computer drafts are of the same quality and naturalness as manually translated texts.

The details of these experiments will be presented below.

6.2 Evaluating the Korean Text

This section describes experiments that were performed in the U.S. to determine whether or not the Korean drafts generated by TTA are of sufficient quality that they improve the productivity of experienced mother-tongue translators. The results of these experiments indicate that the drafts generated by TTA approximately quadrupled the productivity of the experienced Korean translators. Additional experiments that were performed in both the U.S. and Korea indicate that the drafts generated by TTA and then edited by the translators are of a quality that is directly comparable to a professionally translated and published Korean text.

As was mentioned in section 5.2, a Korean lexicon and grammar were developed that were sufficient to generate the Grammar Introduction, three community development articles, and the following biblical texts: Luke 1-10, Ruth, Esther, Daniel, and Nahum. All of the generated texts were included in the following experiments except the draft of Nahum.

6.2.1 Test for Increased Productivity

In order to determine whether or not the generated texts are of sufficient quality that they increase the productivity of translators, eighteen experiments were performed to compare how much text an experienced mother-tongue translator could translate versus how much computer generated text the same person could edit in a given amount of time. All eighteen participants speak Korean as their first language. Sixteen of the participants were students at Southwestern Baptist Theological Seminary where they are working on either a Masters degree or a Ph.D. The other two participants have both completed advanced degrees at the University of Texas at Arlington. One of them completed a Masters degree in linguistics, and then worked in Papua New Guinea with the Summer Institute of Linguistics for approximately thirteen years in order to translate the New Testament into a Papuan language. After completing that translation, he returned to Dallas and is now the pastor of a church in north Dallas. The other participant completed a Ph.D. in environmental sciences at UTA, and then worked for Texas

Instruments for approximately twenty years. He now works as a certified, professional Korean translator and interpreter throughout the north Texas area.

Each of the eighteen participants described above performed the following experiment which consists of five steps:

Step 1: The participant was shown a computer generated Korean draft of a short story, and told that the story had been translated from English to Korean by a computer. The participant was then asked to read the story and edit the text to make it sound more natural and improve the quality. Then the following four steps were described. The participant was told that he should produce a presentable first draft for someone who has approximately a sixth grade reading level; he is not expected to produce a final draft of the translation. It was emphasized that his draft should be of sufficient quality that it could be read to a group of sixth grade students, and they would easily understand it and not be aware that it had been translated from English into Korean.

Step 2: The participant spent 15 minutes editing a computer generated Korean text to make it sound more natural and improve the quality.

Step 3: The participant spent 15 minutes manually translating an English text generated by TTA into Korean, again producing an initial rough draft suitable for people with a sixth grade reading level. The English text came from the same passage that the participant translated in step 2. For example, if the participant edited the computer's draft of Ruth 1:1 through 1:20 in step 2, then he began translating Ruth 1:21 into Korean during this step. After completing this step, the participant was asked if he had any questions or experienced any difficulties with either the English or Korean texts. After his questions were answered, the participant proceeded to the final two steps of the experiment.

Step 4: The participant spent 30 minutes editing a computer generated Korean text to make it sound more natural and improve the quality.

Step 5: The participant spent 30 minutes manually translating an English text generated by TTA into Korean. Again the participant was given an English text that began with the verse that came after the last verse that he edited in step 4.

These five steps were performed by eighteen experienced Korean mother-tongue translators, and each person was paid for his participation. Steps 2 and 3 were used to identify any problems and clarify any issues; the results of those two steps were not included in the final calculations. Generally these experiments were performed two at a time. In order to insure that there was not a bias toward either editing or manually translating, one of the two participants would perform the steps in the order listed above, and the other participant would perform the steps in the order 1, 3, 2, 5, 4. For example, Participant #1 in Table 6-4 below performed the steps in the order 1, 3, 2, 5, 4, and Participant #2 performed the steps in the order 1, 2, 3, 4, 5. Both Participants #1 and #2 did the experiments using the computer generated Korean and English drafts of Ruth. The results of these eighteen experiments are listed below in table 6-1. In order to determine the ratios of edited text to manually translated text, the Word Count tool in Microsoft Word was used to count the number of Korean words in each of the two texts.

Table 6-1. Ratios of Edited Words to Translated Words

	Date	Step 4	Step 5	Ratio
Participant #1	12/17/07	Translated Ruth 1:1 to 1:20	Edited Ruth 1:21 to 4:22	1827/470 =3.9
Participant #2	01/08/08	Edited Ruth 1:1 to 2:23	Translated Ruth 3:1 to 4:6	1222/675 =1.8
Participant #3	01/14/08	Translated Esther 1:1 to 1:20B	Edited Esther 1:20C to 5:14	2065/583 =3.5
Participant #4	01/14/08	Edited Esther 1:1 to 5:14	Translated Esther 6:1 to 7:3A	2684/361 =7.4
Participant #5	2/26/08	Translated Esther 6:1 to 7:3	Edited Esther 7:4 to 10:3 (18:50 minutes)	2432/428 =5.7
Participant #6	2/26/08	Edited Esther 6:1 to 9:21	Translated Esther 9:22 to 10:3 (26:40 minutes)	1626/455 =3.6
Participant #7	4/14/08	Edited Esther 1:1 to 2:22A	Translated Esther 2:23 to 3:15	1282/456 =2.8
Participant #8	5/15/08	Translated Luke 2:1 to 2:16	Edited Luke 2:17 to 4:13	1842/319 =5.8
Participant #9	5/15/08	Edited Luke 2:1 to 3:6	Translated Luke 3:7 to 3:32	1180/559 =2.1
Participant #10	5/15/08	Edited Luke 4:1 to 5:20	Translated Luke 5:21 to 5:39	1468/444 =3.3
Participant #11	5/15/08	Translated Luke 4:1 to 4:24A	Edited Luke 4:24B to 6:17	1794/424 =4.2
Participant #12	7/24/08	Translated Luke 1:1 to 19B	Edited Luke 1:20 to 2:18	1422/412 =3.5
Participant #13	7/24/08	Edited Luke 1:1 to 2:26	Translated Luke 2:27 to 2:41	1995/290 =6.9
Participant #14	8/7/08	Translated Luke 7:1 to 7:21	Edited Luke 7:22 to 9:14	2400/461 =5.2
Participant #15	8/7/08	Edited Luke 7:1 to 8:8	Translated Luke 8:9 to 8:27	1440/460 =3.1
Participant #16	10/17/08	Translated Daniel 1:1 to 21A	Edited Daniel 1:22 to 4:23	2811/535 =5.3
Participant #17	10/17/08	Edited Daniel 1:1 to 4:23	Translated Daniel 4:24 to 5:7	3379/640 =5.3
Participant #18	11/26/08	Edited Avian Influenza 1:1 to 3:11 (25 minutes)	Translated Avian Influenza 3:13 to 4:4C (25 minutes)	1166/136 =8.6
			Average Ratio	4.6

As was mentioned above, steps 4 and 5 generally took 30 minutes. However, due to a time constraint, participant #18 spent only 25 minutes performing each of these two steps. Participants #5 and #6 reached the end of their texts during step 5 before the 30 minutes had expired, so the time they spent on step 5 is recorded in the table. Then their ratios were calculated by extrapolating their work to 30 minutes. As seen in the table above, using TTA's drafts more than quadrupled the productivity of these experienced mother-tongue translators. Participant #18 is the certified professional translator that was described earlier. His ratio is particularly high because he types rather slowly. One criterion for participating in this experiment was that the person must be able to type Korean reasonably well, "reasonably well" being defined as using all his fingers while typing rather than just his index fingers, and he does not need to search for the characters. If a person had to search for the keys while typing, or if he used only his index fingers while typing, then he was not allowed to participate in this experiment. Participant #18 satisfied this criterion, but he typed slowly, so his ratio is rather high.

A paired *t*-test was performed to confirm that using the computer generated drafts increased the productivity of these translators much more than could be expected by chance. The mean number of words edited was 1891, and the mean number of words translated was 450. As seen in Table 6-1 above, the average ratio was 4.6 with a standard deviation of 1.8. The range of ratios was rather wide with a low value of 1.8 and a high of 8.6. Using an alpha criterion of .05 as the cutoff for statistical significance, the results are $t(17) = 9.89$, two-tail $p < .0001$. Since two-tail p is less than .05, the null hypothesis must be rejected, thereby confirming that using the computer generated drafts did greatly increase the productivity of these translators.

Participant #7 listed in table 6-1 above spent approximately thirteen years translating the entire New Testament into a language spoken in Papua New Guinea. Therefore he has the most experience working as a translator. The changes that he made to the computer generated draft of Esther 1:1 to 2:22A are summarized below:

1) Three times he deleted a clause initial conjunction (2:8, 11, 16).
2) Five times he changed 'man' to 'person' (1:14 twice, 15, 16, 21).
3) Once he changed a noun to a pronoun (2:20).

4) Once he changed a name to a title (1:5).

5) Once he deleted a nominal (1:11).

6) Twice he changed a sentence initial conjunction to a different conjunction (1:6, 2:4).

7) Five times he moved a word to a different location in the same sentence (1:8, 11, 2:11, 14, 22).

8) Twice he broke a long sentence into two sentences (1:1, 5).

9) Once he combined two sentences into a single sentence (1:13).

10) Five times he added a particle for emphasis (1:1, 5, 17, 20, 21).

11) Five times he changed the topic marker 은/는 to 이/가 (1:15, 2:3, 4, 7, 19)

12) Once he changed a future tense morpheme to a present tense morpheme (2:4).

13) Once he added a word for clarity (2:5)

14) Eleven times he changed a generic word to a more specific word (1:3, 4, 7 ‘a lot of wine’ to ‘enough wine’, 8, 10 ‘be happy’ to ‘be satisfied’, 13 ‘talk’ to ‘discuss’, 14 ‘talk’ to ‘seek advice’, 19, 2:3, 17 ‘declare’ to ‘proclaim’, 18 ‘declare’ to ‘proclaim’).

The changes above indicate that the computer generated draft was clearly understandable and grammatically correct. The majority of his changes were the type mentioned in (14) – changing a generic word to a more specific word. This is to be expected because the semantic representations contain semantically simple, generic words because they are more likely to have good lexical equivalents in other languages.

After performing these eighteen experiments, it is clear that the computer generated drafts more than quadrupled the productivity of these experienced mother-tongue translators. However, another set of experiments had to be performed to determine whether or not the participants had done a thorough job of editing the computer generated drafts.

6.2.2 Test for Quality of Translation

In order to determine whether the translators had done a thorough job of editing the computer generated texts, questionnaires were developed for each of the eighteen experiments that had been performed in the first phase. The questionnaires were each one page long so that they could be answered quickly by many people. At the top of the questionnaire was a simple statement requesting that the evaluator read the two sample texts. Then approximately one third of the page contained a sample from a participant's manually translated text, and then there was a red line across the page to indicate the end of that section. Then the next third of the page had a sample from the same participant's edited text, and another red line was drawn across the page to indicate the end of that section. Finally the bottom third of the page contained a question in Korean asking the evaluator to select one of the following three options which are translated into English as follows:

- The text in the first section is easier to understand than the text in the second section.
- The text in the second section is easier to understand than the text in the first section.
- The two texts are both equally easy to understand.

However, when these questionnaires were given to people, the respondents would select one of the options, and then give a reason such as the following: "The text in the first section is easier to understand because it has three participants in a single story, but the second text has ten participants and appears to be the end of one story and the beginning of a second story." After receiving numerous responses of this type, the questionnaire was revised so that the three options were:

- The text in the first section is a better translation than the text in the second section.
- The text in the second section is a better translation than the text in the first section.
- The two texts are both equally good translations.

When the questionnaires contained these options, the participants always wanted to see the source texts in order to determine the quality of the translations. After rewording the questions several more times, it became clear that the two samples of text must both cover the same passage; if the two texts were from two different passages, the responses would always be based on the content of the passages rather than the quality of the Korean prose. Therefore it was decided that the edited computer drafts would be compared with the corresponding passages from a published Korean Bible called 쉬운 성경 [swi un seong gyeong], which in English means 'The Easy Bible'. A group of ten Korean scholars began this translation project in 1994, and their target audience was specifically elementary school children (http://blog.naver.com/weddingbhc/30038624669). Before publishing this new translation, they asked elementary school teachers to edit it to insure that it was appropriate and understandable by school children. Then it was published in 2001 by Agape Publishing Company. This translation has become very popular in Korea with annual sales ranging from 100,000 to 200,000. Therefore it was decided that the edited computer drafts from the biblical texts produced by the seventeen participants in the first set of experiments described above would be compared with the same passages contained in 쉬운 성경 [swi un seong gyeong].

New questionnaires were developed and again they were each one page long so that they could be answered within a few minutes by many people. A sample questionnaire is included in appendix B. Approximately a third of each page contains a section of text from the manually edited computer draft for a particular passage such as Ruth 1:1 to 1:5, another third of the page has the same verses from 쉬운 성경 [swi un seong gyeong], and again there are red lines clearly distinguishing one section from the other section. At the top of the questionnaires are two sentences which are translated into English as follows: "Below are two short texts that were translated from English into Korean. Please read the two texts and answer the question below." Then at the bottom of the page is a question that is translated into English as follows: "Please read the two texts above, and then draw a circle around one of the following options:

- When a sixth grader who is unfamiliar with this story reads the two texts above, the first text is better than the second text.
- When a sixth grader who is unfamiliar with this story reads the two texts above, the second text is better than the first text.

- When a sixth grader who is unfamiliar with this story reads the two texts above, the two texts are approximately equal."

Six questionnaires of this format were developed for each of the seventeen experiments that were performed with the biblical texts described above. In three of the questionnaires, the edited computer draft occurs first on the page, and the text from 쉬운 성경 [swi un seong gyeong] occurs second; in the other three questionnaires, the text from 쉬운 성경 [swi un seong gyeong] appears first, and the edited computer draft appears second. Each questionnaire covers a different set of verses. For example, the first questionnaire for participant #3 covers Esther 2:1-4, the second questionnaire covers Esther 2:5-9, the third questionnaire covers Esther 2:10-14, etc. Therefore each questionnaire is unique. Six questionnaires were prepared for each of the seventeen experiments that dealt with biblical texts. Since there is not a professionally translated Korean text for sixth graders regarding the community development article entitled Avian Influenza, it was not possible to evaluate the edited computer draft from that experiment. These questionnaires were then distributed to adults at two Korean churches in the Dallas area. The evaluators all spoke Korean as their first language, and they were generally familiar with the biblical texts. The evaluators were not told how the two sample texts on the questionnaire had been produced. The results of the six evaluations for each of the seventeen biblical text experiments are shown below in table 6-2. In the final column of the table, the number after 'TTA' indicates the number or evaluators who chose the edited computer text as being better, the number after 'EB' (Easy Bible) indicates the number of evaluators who chose the 쉬운 성경 [swi un seong gyeong] text as being better, and the number after 'Equal' indicates the number of evaluators who said that the two texts are equal in quality.

Table 6-2. Adult Evaluations of the Korean Texts

	Evaluated Text	Evaluations
Participant #1	Ruth 1:21 to 4:22	TTA: 3 EB: 3 Equal: 0
Participant #2	Ruth 1:1 to 2:23	TTA: 3 EB: 2 Equal: 1
Participant #3	Esther 1:21 to 5:14	TTA: 1 EB: 4 Equal: 1
Participant #4	Esther 1:1 to 5:14	TTA: 4 EB: 2 Equal: 0
Participant #5	Esther 7:4 to 10:3	TTA: 2 EB: 2 Equal: 2
Participant #6	Esther 6:1 to 9:21	TTA: 1 EB: 2 Equal: 3
Participant #7	Esther 1:1 to 2:21	TTA: 2 EB: 4 Equal: 0
Participant #8	Luke 2:17 to 4:13	TTA: 2 EB: 3 Equal: 1
Participant #9	Luke 2:1 to 3:6	TTA: 4 EB: 1 Equal: 1
Participant #10	Luke 4:1 to 5:20	TTA: 2 EB: 3 Equal: 1
Participant #11	Luke 4:25 to 6:17	TTA: 1 EB: 1 Equal: 4
Participant #12	Luke 1:20 to 2:18	TTA: 4 EB: 2 Equal: 0
Participant #13	Luke 1:1 to 2:26	TTA: 3 EB: 1 Equal: 2
Participant #14	Luke 7:22 to 9:14	TTA: 5 EB: 1 Equal: 0
Participant #15	Luke 7:1 to 8:8	TTA: 2 EB: 3 Equal: 1
Participant #16	Daniel 1:22 to 4:23	TTA: 2 EB: 3 Equal: 1

Table 6-2 – continued

Participant #17	Daniel 1:1 to 4:23	TTA: 2 EB: 4 Equal: 0
	Total Evaluations	TTA: 43 EB: 41 Equal: 18

In the evaluations shown above, it is clear that adults who are familiar with the biblical text consider the edited computer drafts to be of a quality that is comparable with the professionally translated and published 쉬운 성경 [swi un seong gyeong]. Except for cases #3, #7, and #17 in the table above, at least half of the evaluators for each set of texts considered the edited computer drafts to be as good as, or better than the 쉬운 성경 [swi un seong gyeong] texts. For cases #3, #7, and #17, a third of the evaluators for each of those cases considered the edited computer drafts to be as good as, or better than the 쉬운 성경 [swi un seong gyeong] texts.

The number of evaluations for each translator's text was too small to perform individual χ^2 tests, but a χ^2 test was performed using the evaluation totals. Using 43, 41, and 18, $\chi^2(2)$ = 11.35, *p* = .0034. A *p* value less than .05 indicates that there is skewing among these three factors. Examining the data reveals that the reason for this skewing is because the number of equal evaluations is significantly lower than the evaluations from the other two options. Therefore a binomial distribution was performed between the TTA and EB evaluations. The two tail *p* binomial cumulative distribution probability is .91. Since this value is much higher than .05 which is the standard cutoff value, the difference between the TTA and EB evaluations is insignificant. Since the difference between the TTA and EB evaluations values is insignificant, the difference is almost certainly due to chance and is therefore not a reliable effect due to the method of testing or the difference in translation procedures. Therefore the two texts are essentially of equal quality.

Since TTA is intended to generate texts for people with approximately a sixth grade reading level, another set of evaluations was performed at a grade school in DeaGu, South

Korea. The questionnaires described above were modified so that the three options at the bottom of each page read as follows:

- Regarding the two texts above, the first text is better than the second text.
- Regarding the two texts above, the second text is better than the first text.
- Regarding the two texts above, both texts are approximately equal."

The results of the evaluations by sixth graders in Korea who are generally unfamiliar with the biblical texts are shown below in table 6-3.

Table 6-3. Sixth Graders' Evaluations of the Korean Texts

	Evaluated Text	Evaluations
Participant #1	Ruth 1:21 to 4:22	TTA: 5 EB: 1 Equal: 0
Participant #2	Ruth 1:1 to 2:23	TTA: 2 EB: 3 Equal: 1
Participant #3	Esther 1:21 to 5:14	TTA: 3 EB: 2 Equal: 1
Participant #4	Esther 1:1 to 5:14	TTA: 0 EB: 3 Equal: 3
Participant #5	Esther 7:4 to 10:3	This text was not evaluated because it is not suitable for children when taken out of context.
Participant #6	Esther 6:1 to 9:21	This text was not evaluated because it is not suitable for children when taken out of context.
Participant #7	Esther 1:1 to 2:21	TTA: 3 EB: 2 Equal: 1
Participant #8	Luke 2:17 to 4:13	TTA: 3 EB: 2 Equal: 1
Participant #9	Luke 2:1 to 3:6	TTA: 4 EB: 2 Equal: 0
Participant #10	Luke 4:1 to 5:20	TTA: 5 EB: 1 Equal: 0

Table 6-3 – continued

Participant #11	Luke 4:25 to 6:17	TTA: 2 EB: 3 Equal: 1
Participant #12	Luke 1:20 to 2:18	TTA: 4 EB: 1 Equal: 1
Participant #13	Luke 1:1 to 2:26	TTA: 2 EB: 3 Equal: 1
Participant #14	Luke 7:22 to 9:14	TTA: 4 EB: 2 Equal: 0
Participant #15	Luke 7:1 to 8:8	TTA: 3 EB: 1 Equal: 2
Participant #16	Daniel 1:22 to 4:23	TTA: 3 EB: 2 Equal: 1
Participant #17	Daniel 1:1 to 4:23	TTA: 2 EB: 2 Equal: 2
	Total Evaluations	TTA: 45 EB: 30 Equal: 15

In the evaluations shown above, it is very clear that sixth grade children who are unfamiliar with the biblical text consider the edited computer drafts to be of a quality that is comparable with the professionally translated and published 쉬운 성경 [swi un seong gyeong]. In every case at least half of the evaluators for each set of texts considered the edited computer drafts to be as good as, or better than the 쉬운 성경 [swi un seong gyeong] texts.

The number of evaluations done by sixth graders for each translator's text was too small to perform individual χ^2 tests, but a χ^2 test was performed using the evaluation totals. Using 45, 30, and 15, $\chi^2(2) = 15.00$, *p* = .0006. A *p* value less than .05 indicates that there is skewing among these three numbers. Examining the data reveals that there are two reasons for this skewing: 1) the number of students who said that the two texts are equal is significantly lower than the number of students who said that one text was better than the other, and 2) the number of students who said that the edited TTA drafts are better is significantly higher than the number of students who said that the 쉬운 성경 [swi un seong gyeong] texts are better or equal to TTA's edited drafts. Because there is skewing, a binomial distribution was performed between the TTA and EB evaluations. The two tail *p* binomial cumulative distribution probability is .1053. This indicates that there is only about a 10% probability that these two numbers are reliably the same. In other words, there is a 90% probability that the students genuinely prefer TTA's edited drafts over the 쉬운 성경 [swi un seong gyeong] texts. At this time it cannot be stated with statistical certainty that the students prefer the edited TTA drafts, but with a larger number of respondents, a significant preference for TTA's edited drafts might emerge.

6.2.3 Conclusions

This section described experiments that were performed to ascertain the quality of the Korean drafts generated by TTA. The results of these experiments lead to the following two conclusions:

1) The Korean drafts generated by TTA were of sufficient quality that they approximately quadrupled the productivity of eighteen experienced mother-tongue Korean translators.
2) The mother-tongue translators did a thorough job of editing the computer generated drafts because independent evaluations indicate that the edited texts were of a quality that is directly comparable to a professionally translated and published Korean text.

These experiments indicate that the computer generated Korean drafts are easily understandable and grammatically correct. Additionally, the generated texts are of sufficient quality that they approximately quadrupled the productivity of many experienced mother-tongue translators.

6.3 Conclusions

This chapter described experiments that were performed to ascertain the quality of the texts generated by TTA. These experiments were designed to accomplish two purposes:

1) test for increased productivity
2) test for quality

The most critical of these experiments is the first one which determined whether or not TTA's drafts increase the productivity of experienced mother-tongue translators.

The next chapter of this dissertation will summarize the results of this research. Then that chapter will discuss the significance of TTA, the situations in which it is appropriate to use this project, additional improvements that may be made to the system, and areas that require additional research.

CHAPTER 7

CONCLUSIONS

7.1 Introduction

This dissertation described the natural language generator called *The Translator's Assistant*. The fundamental question that this research answered is as follows: If the semantic representations contain sufficient information, and if the grammar possesses sufficient capabilities, then is The Translator's Assistant able to generate drafts of translations of texts that are of sufficient quality that they improve the productivity of experienced mother-tongue translators? The evidence presented in the previous chapters indicates that the answer is clearly yes. The drafts generated by TTA approximately quadruple the productivity of experienced mother-tongue translators without any loss of quality. The translators were able to edit the computer generated drafts in approximately one fourth the time they needed to manually translate the same quantity of text, and independent evaluators viewed their edited drafts as being of the same quality as manually translated texts. A myriad of sociolinguistic factors were certainly involved during the evaluation process, but those factors were beyond the scope of this research.

The fundamental purpose of TTA also deserves repeating: TTA generates texts that are easily understandable, grammatically correct, and semantically equivalent to the source texts. The generated drafts are at approximately a sixth grade reading level. The ontology, feature system, and grammar were designed to accommodate a very wide variety of target languages. The generated texts are by no means the final draft; they require editing by mother-tongue speakers in order to improve the naturalness and information flow. The generated texts certainly do not include all the specialized structures and vocabulary of the target language; the generated texts use the target language's simple, basic terms and structures in order to accurately convey the vast majority of the original text's meaning and content.

Much was learned during this process, and much remains to be done. This chapter will discuss several topics that require additional research in order to improve TTA in the future.

7.2 Topics Requiring Additional Research

TTA's ontology, semantic representational system, and grammar are all functional at the present time, and able to accommodate a very wide variety of target languages. However, as more experience is gained with additional languages, undoubtedly more features and structures will be added to the semantic representational system, and additional capabilities will be added to the grammar. Several examples of this were mentioned in chapter 3. While generating the Korean texts, all of the features associated with the speaker-listener relationship were added to propositions. The English grammar does not include any system of honorifics when one person speaks to another, so the features called Speaker, Listener, Speaker's Attitude, Speaker's Age, and Speaker to Listener's Age were not part of TTA's feature system at that time. But when the Korean project was started, these features, which were described in sections 3.3.2.5.5 through 3.3.2.4.9, were added in order to generate the necessary honorifics. These features are now available in the semantic representations for every target language that needs them. Generally features and structures aren't added to the semantic represenational system for one particular target language. However, if a structure or feature will be helpful to a variety of target languages, then the semantic representations may be modified to include the new structures or features. As more experience is gained with other languages, additional structures, features, and feature values will certainly be added to TTA's semantic representational system. Any structure or feature that will be useful to a wide variety of languages may be added to TTA's semantic representational system. The design of TTA's representational system allows it to be sufficiently flexible that it can accommodate these additions.

TTA is certainly functional at the present time, and usable by computational linguists. However, this research did not answer the question of whether or not field linguists with little computational experience are able to use TTA. It is still an open question as to whether or not TTA is a practical system that field linguists can readily use without help from computational linguists. Developing the lexicons and the synthesizing grammars for each of the test languages was very simple and required little time. Field linguists who have little computational experience are certainly able to develop their own lexicons and synthesizing grammars quickly and easily. However, developing the transfer grammars is considerably more complex and time consuming. The development of a good transfer grammar requires considerable experience dealing with ordering rules, feeding, bleeding, defining and setting features, etc. In order to facilitate the use of TTA by field linguists without help from computational linguists, the transfer grammar requires additional research.

Additional research could improve TTA in the following three ways: 1) improve the quality of the generated texts, 2) enable the system to cover an even broader range of target languages, and 3) make the system easier to use. In order to improve the quality of the texts generated by TTA, the ontology needs to be restructured, and the content of the semantic representations needs to be refined. In order to cover an even broader range of target languages, the feature system requires additional research. In order to make TTA easier to use, the transfer grammar needs to be simplified, and current typological research in areas such as semantic maps needs to be incorporated. Therefore there are five specific areas that require additional research: 1) the ontology, 2) the feature system, 3) the transfer grammar, 4) current typological studies in areas such as semantic maps, and 5) refinement of the content of the semantic representations.

7.2.1 The Ontology

As was stated in section 3.3.1, the ontology remains the most problematic issue in this project. However, there are two distinct ways to improve TTA's ontology: 1) develop a practical technique for explicating semantically complex objects, and 2) convert the ontology from simple lists to structured hierarchies.

7.2.1.1 Explicating Semantically Complex Objects

Chapter 3 presented several potential approaches to the problems associated with including semantically complex concepts in the ontology:

1) Use only the NSM primitives in the ontology.
2) Use semantically complex lexemes in the ontology and make very fine semantic distinctions between the various senses of each word. If the senses are sufficiently fine grained, they will be able to accommodate every lexeme in every language.
3) Develop a principled compromise between the two extreme options listed above. Therefore semantic primitives as well as semantic molecules are permitted in the ontology.

An ontology that consists solely of NSM primitives would theoretically be usable by all the world's languages. However, developing the semantic representations using only the NSM primitives is entirely impractical. The second option, using extremely fine grained semantics in order to accommodate every lexeme in every language is entirely impractical as well. Therefore this project adopted a principled compromise. Using the foundational principles of NSM theory, English semantic molecules were identified. The final result is that TTA's ontology includes concepts that have been placed in four semantic complexity levels: 1) the NSM primitives, 2) semantic molecules which are the words in Longman's defining vocabulary, 3) complex concepts that are explicated using the NSM primitives and semantic molecules, and then inserted into the semantic representations only if the linguist activates the associated complex concept insertion rule as was discussed in 4.3.1, and 4) complex concepts that are impractical to explicate. The number of entries in the ontology belonging to the fourth category has been kept to an absolute minimum.

The burden placed on the ontology was significantly reduced by the addition of the collocation correction rules which were described in 4.3.6. The collocation correction rules allow linguists to map a single source concept to multiple target words in accord with the other concepts that are in the environment. The example given in table 4-1 presented how the various Tzeltal words for ‘to carry’ may be inserted into the generated texts even though TTA’s ontology has only one sense of CARRY which refers to transporting an object from one place to another place.

Drastically reducing the number of semantically complex concepts in the ontology and using primarily the NSM primitives and the words in Longman’s defining vocabulary as the fundamental lexemes in the ontology had three effects: 1) it made the development of the semantic representations considerably more difficult, 2) it improved the quality of the generated texts, and 3) it reduced the demands placed on the grammar. Prior to adopting the NSM primitives and molecules, semantically complex concepts had been allowed in the semantic representations. Allowing complex concepts in the semantic representations made the development of the semantic representations very easy, but when a target language did not have a lexical equivalent for a complex concept, a rule in the transfer grammar was required to explicate the complex concept using the target language’s lexemes and structures. Those rules were complex, very difficult to write, and frequently distorted the intended message. Examples of this problem were provided in sections 4.3.1.1 to 4.3.1.3. Those examples demonstrated that when a rule replaces a complex concept with a target word that is modified by a phrase or clause, the message becomes distorted. This problem was virtually eliminated by explicating almost all of the complex concepts in the source texts using the NSM primitives and molecules. Succinctly and sufficiently explicating events, object attributes, and event attributes is generally possible. However, explicating semantically complex objects tends to be considerably more difficult. In certain cases, explicating semantically complex objects has proven impractical.

There are two particular categories of objects where this problem repeatedly arises: artifacts and animals.

Certain artifacts are sufficiently explicated quite easily by modifying an object with a descriptive proposition. Examples include the following:

- bandage: a piece of cloth that covers a wound
- cage: a structure that people put animals into
- perfume: oil that smells good
- manger: a box that contains animals' food

The explications listed above certainly are not complete, but they are generally sufficient to communicate the intended message. If the target culture is unfamiliar with one of these artifacts, these explications are generally able to communicate the message. Other artifacts are extremely difficult to succinctly explicate by modifying an object with a single proposition. Examples include the following:

- arrow: a long, thin, straight piece of wood with a point at one end and used as a weapon when shot from a bow.
- bow: a long, thin piece of wood held in a curve by a tight string and used as a weapon to shoot arrows.
- axe: a tool that has a sharp, heavy piece of metal at one end and a long wooden handle, and is generally used to cut down trees.

If a target culture is unfamiliar with one of these artifacts, these explications fail to communicate the message and also badly distort it. At the present time, many artifacts have been placed in the fourth category mentioned earlier – complex concepts that are impractical to explicate. No other solution has been identified for this problem, so it remains an issue that requires additional research.

The explication of animals has also proven problematic. Particularly in the biblical texts, animals are used because they possess certain qualities. If a particular animal is unknown in the target culture, it is impossible to succinctly explicate the animal in such a way that all the essential qualities are communicated. For example, sheep possess the following characteristics: 1) edible meat, 2) white wool, 3) need protection from predators, 4) follow other sheep without thinking, 5) get lost easily, and 6) not very intelligent. Capturing all of these characteristics in a succinct explication is impossible. At the present time there are two potential solutions to this problem: 1) have multiple senses of SHEEP in both the ontology and the semantic representations, or 2) have just single sense of SHEEP in the ontology, but in the semantic representations modify the first occurrence of SHEEP in each passage with a proposition 'which is an animal like a X'. TTA could be modified so that both of these solutions are incorporated, and then linguists could choose one solution or the other based on their particular situation. Both of these solutions communicate the message faithfully, but add complexity to the ontology and the development of the semantic representations.

The first solution requires multiple senses of SHEEP in both the ontology and the semantic representations. For example, SHEEP-A could refer to an animal that has edible meat, SHEEP-B refers to an animal that has white body hair, SHEEP-C refers to an animal that needs to be protected from predators, etc. Then when the semantic representations are being manually developed, SHEEP-A will be used whenever a source text has the word 'sheep', and the sheep's edible meat is in focus. Similarly SHEEP-B will be inserted into the semantic representations whenever a source text has the word 'sheep', and the sheep's white body hair is in focus. If a target culture is unfamiliar with sheep and does not have an animal that possesses all these characteristics, then a different animal may be linked to each sense of SHEEP. Then TTA will be able to insert the appropriate animal into the generated texts because each of these senses of SHEEP will be used appropriately during the manual development of the semantic representations.

The second solution requires only a single sense of SHEEP in the ontology, but additional concepts would be required in the ontology. The additional concepts would be ‘animal-with-edible-meat’, ‘animal-with-white-body-hair’, ‘animal-that-needs-protection-from-predators’, etc. Then when the semantic representations are being manually developed, whenever the word ‘sheep’ occurs in a source text and the sheep’s edible meat is in focus, the first occurrence of SHEEP in the passage would be modified with the proposition ‘which is an animal like a animal-with-edible-meat’. Linguists would then link the concept ‘animal-with-edible-meat’ to the name of an animal in the target culture that has edible meat. Similarly they would link the concept ‘animal-with-white-body-hair’ to the name of an animal that is known in the target culture and has white body hair. This solution could be used by translators who prefer to use a loan word for ‘sheep’, but still want to communicate the message in a way that is both historically accurate and culturally relevant.

7.2.1.2 Converting the Ontology from a Simple List to a Structured Hierarchy

As was discussed in section 3.3.1, the current version of TTA’s ontology consists of seven simple lists of words, each list containing the concepts for a particular semantic category, and the concepts in that list are sorted alphabetically. Converting these lists to structured hierarchies would serve two purposes: 1) help linguists develop their lexicons in a more natural way, and 2) improve the quality of the generated texts by using complex concepts which are not explicated if the target language has a lexical equivalent.

7.2.1.2.1 Improve the Target Lexicon Development

The concepts in TTA’s ontology could be converted to a hierarchy similar to the one shown in figure 3-4. The concepts themselves would not change; TTA’s ontology would still be comprised of 1) the NSM primitives, 2) semantic molecules which are the defining vocabulary in Longman’s Dictionary, 3) complex concepts which are explicated and inserted into the semantic representations only if the user activates the associated complex concept insertion rule, and 4) unexplicated complex concepts. However, by ordering the concepts in natural hierarchies, related concepts would be placed within the same group and their distinguishing features made clear. A possible high level view of these hierarchies is shown below:

Objects
- Proper Names
  - Men’s Names
  - Women’s Names
  - Geographical Names
  - Temporal Names
- Artifacts
- Animals
- Abstracts

Events (Levin, 1993:111ff.)
- Putting
- Removing
- Sending
- Carrying
- Change of Possession
- Learning
- Throwing
- Movement
- Speech Acts
- Mental Acts
- Preparation

Event Attributes
- Cardinal Numbers
- Ordinal Numbers
- Colors
- Size
- Quantity
- Emotions

For example, under Speech Act verbs, four of the entries would be SAY-A, SAY-B, SAY-C, and SAY-D as was described in section 3.3.1. Other concepts that would be placed in the Speech Act verbs under SAY-A include DECLARE, PROCLAIM, ORDER, etc. So the concepts themselves would not change, but the relationships between the concepts would be made clearer.

7.2.1.2.2 Improve the Quality of the Generated Texts

Chapter 6 presented the results of experiments that were done to determine the quality of the generated texts. In the Korean experiments, experienced mother-tongue translators were asked to edit the computer generated drafts and make them presentable first drafts. Then the changes that were made by the most experienced mother-tongue translator were examined and described. The most common type of change was to replace generic words with more specific words. This is to be expected because the semantic representations use very simple, generic words. The use of more complex concepts in the semantic representations would reduce this type of editing, but the complex concepts must be inserted into the semantic representations only when the target language has good lexical equivalents. The complex concepts would be inserted into the semantic representations using the technique described in section 4.3.1. That technique always pairs a complex concept with a simple concept such as MAN-A/SHEPHERD-A. When the semantic representation of a particular verse is loaded, TTA looks for these pairs, and then discards the simple concept if the complex concept has a target equivalent. If the complex concept does not have a target equivalent, then TTA automatically discards the complex concept and inserts the simple concept into the semantic representation. By organizing the concepts in TTA's ontology into a hierarchy, it would be easier for linguists to specify which complex concepts should be automatically inserted into the semantic representations, and which should not. The semantically complex concepts could be placed beneath the generic concepts in the ontological hierarchy, and if the user supplies a target equivalent for a complex concept, then that complex concept would automatically be used throughout all the semantic representations. But if the user does not supply a target equivalent for a particular complex concept, then TTA would automatically use the generic concept. For example, SAY-A is used for all direct speech. Complex concepts occurring under SAY-A could include DECLARE and PROCLAIM. During the development of the semantic representations, the pair SAY-A/DECLARE-A would be inserted wherever appropriate. When a high ranking government employee says something officially, DECLARE could be used if the target language has an equivalent. If the target language does not have an equivalent for DECLARE, then SAY would automatically be used. For another example, RED-A is a color which is a semantic molecule. The complex colors CRIMSON and SCARLET could be inserted beneath RED in the ontology. If the user supplies a target equivalent for CRIMSON or SCARLET, then those complex concepts would automatically be used where appropriate in the semantic representations. If the user does not supply a target equivalent, then the pairs RED-A/CRIMSON-A and RED-A/SCARLET-A in the semantic representations would automatically be collapsed to RED-A.

7.2.2 The Feature System

The feature system is another area that requires additional research. Undoubtedly more features will be added to TTA's feature system in order to accommodate additional languages. For example, many of the world's languages employ evidentiality markers. If typological research is able to identify an exhaustive list of the various types of evidential markers, then a feature called Evidentiality could be added to propositions, and the exhaustive list would serve as the possible values for that feature. Then each proposition in the semantic representations could be marked with the appropriate value of Evidentiality. Languages which employ evidential markers could examine this feature, and then spellout rules would insert the appropriate word or morpheme. Other candidates for additional features include mirativity, direction of motion, etc.

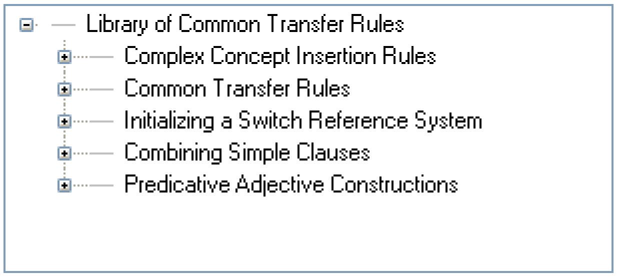
Another feature which requires additional research is the 'Time' value associated with events. Section 3.3.2.2.1 presented two options: absolute time and relative time. The relative time option was chosen, and that option fit the two test languages well. However, it is well documented that languages employ many different views of time, so this feature deserves additional research.

7.2.3 The Transfer Grammar

Figures 5-41 and 5-42 show the number of new transfer rules and the number of new synthesizing rules required for each chapter of text in the two test languages. Those graphs reveal two interesting facts: 1) for each test language, the number of new rules required for each subsequent chapter of text significantly decreases, and 2) for each test language, the number of transfer rules significantly exceeds the number of synthesizing rules. Even for English, the number of transfer rules far exceeds the number of synthesizing rules. This is somewhat surprising given that the semantic representations include concepts that have been lexicalized by English, structures that are employed by English, and an English world view. Therefore the English transfer grammar is not required to deal with lexical mismatch, perform any major restructuring, or accommodate a different world view. The main tasks performed by the English transfer grammar consist of theta grid adjustments, adding the appropriate auxiliaries, combining propositions where appropriate, and dealing with the special requirements of subordinate clauses. Generating the various propositions and texts in the Grammar Introduction required almost 320 transfer rules in the English grammar. Generating those same propositions and texts in Korean required only 30 additional rules. Many of the transfer rules in the Korean grammar are performing the same tasks as the corresponding rules in the English transfer grammar: theta grid adjustments, adding auxiliaries, combining propositions, etc. Although the underlying functions of the two transfer grammars are very similar, the actual rules for each language are quite different, thereby producing texts that are appropriate for the vastly different languages. Developing a transfer grammar for a language is a complex and time consuming process, but there are two ways that this problem could be reduced: 1) develop a library of common transfer rules, and 2) divide the transfer grammar into a two stage process, the first stage being developed by a computational linguist for a group of related languages, and the second stage being developed by a field linguist for a particular language within that group of related languages.

7.2.3.1 Develop a Library of Common Transfer Rules

An area that deserves additional research is the prospect of identifying common tasks required by the transfer grammars of various languages. If common tasks can be identified, then prewritten rules that perform those tasks could be added to the Grammar Library so that the linguist could simply activate or deactivate the rules. At the present time TTA’s Grammar Library has only a few entries as seen in figure 7-1 below.



Library of Common Transfer Rules

Complex Concept Insertion Rules

Common Transfer Rules

Initializing a Switch Reference System

Combining Simple Clauses

Predicative Adjective Constructions

Figure 7-1. TTA's Grammar Library

Areas where there is significant potential for developing rules which could be applicable to many different languages include: the identification of pronouns, the construction of clause chains and switch reference systems, combining propositions, inserting auxiliaries, etc.

7.2.3.2 Convert the Transfer Grammar to a Two-Stage Process

As was mentioned earlier, developing a good transfer grammar is a complex process that requires experience in rule ordering, feeding, bleeding, and defining and setting feature values. It is probably not practical to expect field linguists to develop their own transfer grammars without help from a computational linguist. In order to alleviate this problem, one possible approach is to convert the transfer grammar to a two-stage process. The first stage would be developed by a computational linguist for a group of related languages. That stage would perform common but complex tasks such as combining propositions, inserting auxiliaries, handling lexical mismatch, etc. Then field linguists with little computational experience could develop the second stage of the transfer grammar which would perform theta grid adjustments and other minor changes to the semantic representations. The field linguists would also develop their own lexicons and synthesizing grammars. If a group of related languages share similar sentence structures, closely related lexemes, common pronominal systems, common world views, etc., then it is conceivable that a single transfer grammar could be developed for the entire group, and then additional language specific changes could be performed by a second stage of the transfer grammar. This would significantly reduce the amount of work required from a field linguist when developing a grammar for his language.

7.2.4 Semantic Maps and other Current Typological Research

Typologists are presently developing semantic maps which simultaneously capture language universals and language-specific grammatical knowledge (Croft 2003:133). For example, the distinction between singular and plural is a functional domain, and linguists have found that many languages use the animacy hierarchy to determine which nominals should be marked as plural, and which should not be marked. Semantic maps of plural inflection for five languages are shown below in table 7-1.

Table 7-1 Semantic Maps of Plural Inflection (Croft 2003:134)

	1/2 Pronouns	3 Pronouns	Human Nouns	Animate Nouns	Inanimate Nouns
Guaraní	X				
Usan	X	X			
Tiwi	X	X	X		
Kharia	X	X	X	X	
English	X	X	X	X	X

The conceptual space underlying these particular semantic maps is the animacy hierarchy which is universal. The conceptual spaces for these and other semantic maps are empirically constructed using vast amounts of cross-linguistic data. As typologists develop additional conceptual spaces for various functional domains, those spaces and domains could be incorporated into TTA's grammar so that linguists could more quickly indicate the specific details of their particular language using the semantic maps.

7.2.5 The Semantic Representational System

In order to further improve the quality of the generated texts, the semantic representational system must be refined. For example, the current system does not include a method of indicating *let's* as in *Let's go to the store.* Whenever the source texts have a *let's* construction, the proposition is converted to *We should ….* Similarly, the semantic representation system does not currently have any method of representing *can* or *can't*. Whenever a source text has either *can* or *can't*, the proposition is converted to *be able* or *not be able*. *Can* and *cannot* are NSM primitives (Goddard 1998:58), so theoretically they are present in every language. While *can* and *be able* are semantically equivalent and generally communicate the same message, there are instances where changing a *can* construction to a *be able* construction slightly affects the message. For example, in Infected Eye 1:6 Melissa says, “I’m not able to see.” In that particular situation, it would be much more natural for her to say, “I can’t see” which is how her statement was worded in the original source text. However, in Kande’s Story 1:16 Kande’s mother says, “Sometimes I feel very weak. And sometimes I’m not able to work.” In the second proposition, changing *not able* to *can’t* as in “And sometimes I can’t work” seems to slightly diminish the condition of Kande’s mothers. There seems to be pragmatic distinctions between *can* and *be able*, but those distinctions are unclear at this time. However, it is clear that not every instance of *can* or *cannot* in a source text should be converted to *be able* or *not be able.* Refining the semantics to include these finer shades of meaning would increase the quality of the generated texts in target languages that have mechanisms which indicate these finer shades of distinction. However, there is always a trade-off between adding more content to the semantic representations and the amount of work required to develop the grammar. As finer shades of meaning are permitted in the semantic representations, additional work is required by the linguists to construct grammars that reflect these finer shades of meaning.

7.3 Final Conclusions

TTA in its present form is a system that vastly reduces the amount of time and effort required to translate documents from one language to another. With the additional research described above, TTA could potentially produce even better texts with less effort from the linguists. It is hoped that this system will 1) help improve the quality of people’s lives around the world by providing them with translations of vital information, 2) help preserve many of the endangered languages by providing texts in those languages and thorough descriptions of those languages, and 3) provide linguists with a research tool which enables them to simultaneously describe languages and generate texts for speakers of those languages.

APPENDIX A

KOREAN AND ENGLISH DRAFTS OF THREE SHORT STORIES GENERATED BY THE TRANSLATOR'S ASSISTANT

Shown below in tables A-1, A-2, and A-3 are the English and Korean drafts of three short stories that were generated by The Translator's Assistant. The drafts that are shown below have not been edited by mother-tongue speakers; they are shown here unedited in order to illustrate the quality of text that can be expected from The Translator's Assistant. All texts that are generated by The Translator's Assistant require post-editing by a mother-tongue speaker in order to improve their information flow and naturalness. However, as can be seen in the examples below, even without editing the texts are easily understandable and grammatically correct. The first story is published by Shell Publishing, and its purpose is to educate people about AIDS. The second story is published by World Vision, and its purpose is to help prevent eye infections caused by flies. The final story is published by the Indonesian branch of The Summer Institute of Linguistics, and its purpose is to help prevent the spread of Avian Infuenza.

Table A-1 below contains the original text[37] for Kande's Story in the first column, TTA's unedited English draft in the second column, and TTA's unedited Korean draft in the third column. The original text has been included in this table to illustrate how source texts are changed so that they conform to the controlled English influenced meta-language that is used in TTA's semantic representational system. As was discussed in this dissertation, when a semantic representation is developed for a source text, the following changes are made:

- Semantically complex concepts in the source text are generally explicated using the NSM primitives and Longman's defining vocabulary.
- Complex sentences in the source text are divided into multiple, simple propositions.
- Each significant participant in a story is introduced with a complete proposition in the semantic representation.
- Significant events that are implicit in the source text are made explicit in the semantic representation.

[37] The version of Kande's Story that appears here is an early version. Shell Publishing has edited this story, and a more recent version can be downloaded from http://www.scripture-engagement.org/node/57 accessed December 11, 2010.

- The events in the semantic representation must conform to their theta grids as specified in the ontology, so implicit arguments are made explicit in the semantic representation.

Although TTA's English draft of Kande's Story does not exactly match the original text, the vast majority of the original text's content is faithfully communicated by TTA's draft. The semantic content of the Korean draft shown in the final column is identical to the semantic content of the English draft.

Table A-1. The Original Text for "Kand'e Story" and TTA's English and Korean Drafts

Chapter 1: Mama's Secrets	1:1 Title: Kande's mother knows a secret.	1:1 제목: 칸디의 어머니는 비밀을 알고 있어요.
Kande sat under a tree reading. Her sister Teshi came running. "Kande, Kande! I heard the women talking. They said Mama has a secret! What can it be?"	1:2 One day a girl named Kande was sitting near a tree. Kande was reading a book. She had a younger sister named Teshi. Teshi ran to Kande. Teshi was very excited. She said to Kande, "Kande! Kande! I heard certain women talking to each other. Those women said that mother knows a secret! Do you know mother's secret?"	1:2 어느 날 칸디라는 소녀가 나무 가까이에 앉아 있었다. 칸디는 책을 읽고 있었다. 칸디는 태쉬라는 여동생이 있었다. 태쉬는 칸디에게 달려갔다. 태쉬는 매우 흥분하였다. 태쉬는 칸디에게 말하였다. "언니! 언니! 나는 어떤 여자들이 서로에게 말하는 것을 들었어요. 이 여자들은 어머니께서 비밀을 알고 계시다고 말하였어요! 언니는 어머니의 비밀을 알고 있어요?"
"I think I know, little sister," said Kande. "Let's go talk to Mama and find out. I'll race you!"	1:3 Kande said, "I might know mother's secret. We should go to our house and talk to our mother. Mother might tell us about her secret. I'll race you to our house!"	1:3 칸디는 말하였다. "나는 아마도 어머니의 비밀을 알 거야. 우리는 집으로 가서 어머니께 말씀드려야 해. 어머니께서는 아마도 우리에게 자기 비밀에 대해서 말씀하실 거야. 나는 집까지 너와 경주할 거야!"
Kande and Teshi arrived home laughing and out of breath. Their sisters Falala and Iniko and their brother	1:4 Kande and Teshi ran to their house quickly. When Kande and Teshi arrived at the house, they were laughing. They had	1:4 칸디와 태쉬는 자기 집으로 빨리 달려갔다. 칸디와 태쉬는 집에 도착하였을 때 웃고 있었다. 칸디와 태쉬는

Table A-1 - continued

Jumoke gathered too. Mama hushed them. "Be quiet and let your father sleep," she said as she herded them away.	two younger sisters. One younger sister's name was Falala. And the other younger sister's name was Iniko. Kande and Teshi also had a younger brother named Jumoke. Falala, Iniko, and Jumoke heard Kande and Teshi laughing. So they ran to the door to see Kande and Teshi. Then mother said to all the children, "Be quiet because your father has to sleep." Then she walked from the house with the children.	여동생 두 명이 있었다. 여동생 한 명의 이름은 팔라라였다. 그리고 다른 여동생의 이름은 이니꼬였다. 또한 칸디와 태쉬는 주목이라는 남동생이 있었다. 팔라라와 이니꼬와 주목은 칸디와 태쉬가 웃는 것을 들었다. 그래서 팔라라와 이니꼬와 주목은 칸디와 태쉬를 보기 위해 문으로 달려갔다. 그러자 어머니가 모든 아이들에게 "아버지께서 주무셔야만 하기 때문에 조용히 해" 라고 말하였다. 그리고서 어머니가 아이들과 함께 집으로부터 걸었다.
Teshi asked, "Mama, do you have a secret?" Mama put her hand on her middle and said, "Our family is growing bigger."	1:5 Teshi asked her mother, "Mother, do you know a secret?" Mother put her hands on her stomach. Then she said, "I'll give birth to a baby soon."	1:5 태쉬는 어머니에게 "어머니, 어머니께서는 비밀을 알고 계십니까?" 라고 물었다. 어머니가 배 위에 손을 얹었고 "나는 곧 아기를 낳을 거야" 라고 말하였다.
Three-year-old Iniko said, "I'm going inside to tell Father!" Mama picker her up before she could get to the door. "He already knows, little one. Let him rest," Mama said.	1:6 Iniko, who was three years old, said, "I want to tell our father about this secret!" But before Iniko ran into the house, mother picked her up quickly. Mother said, "Your father already knows about this secret. We have to allow your father to sleep."	1:6 세 살인 이니꼬는 "저는 아버지께 그 비밀에 대해서 말하고 싶습니다!" 라고 말하였다. 그러나 이니꼬가 집에 달려가기 전에 어머니가 이니꼬를 빨리 들어서 말하였다. "아버지께서는 이미 그 비밀에 대해서 알고 계셔. 우리는 아버지께서 주무시도록 해야만 해."
Iniko frowned. She loved to climb on Father, but lately she had hardly been allowed to go near him. For a	1:7 Iniko frowned. She liked climbing on her father. But these days mother didn't allow Iniko to go near her	1:7 이니꼬는 인상을 썼고 아버지 위로 올라가는 것을 좋아하였다. 그러나 요즈음에 어머니가 이니꼬가 아버지 가까이에 가는 것을 허락하지

Table A-1 - continued

long time he had been lying down, never working, and he looked weak. The whole family worried about him.	father. Kande's father had slept for many days. And these days he didn't work. He always wanted to rest because he was weak. Mother and all the children worried about father.	않았다. 칸디의 아버지가 여러 날 동안 잤고 요즈음에 일하지 않았고 약하였기 때문에 항상 쉬고 싶어하였다. 어머니와 모든 아이들은 아버지에 대해서 걱정하였다.
One day Kande was at the market. She was telling her friends that her mother was having a baby. A boy walked by and jeered, "That baby is going to have AIDS just like your father!"	1:8 One day Kande was at the market. She told some friends that her mother will give birth to a baby soon. A boy heard Kande talking to her friends. He said to Kande, "The baby will have AIDS like your father has AIDS!"	1:8 어느 날 칸디는 시장에 있었다. 칸디는 친구들에게 어머니가 곧 아기를 낳을 것이다라고 말하였다. 소년이 칸디가 친구들과 말하는 것을 들었고 칸디에게 "네 아버지께서 에이즈가 있는 것처럼 아기는 에이즈가 있을 거야!" 라고 말하였다.
Kande did not know what he meant. Father didn't have AIDS, did he? "Don't listen to him," her friends said.	1:9 Kande didn't understand the things that that boy said. She said, "My father doesn't have AIDS." Kande's friends said to her, "Don't listen to that boy."	1:9 칸디는 그 소년이 말한 것을 이해할 수 없었고 "우리 아버지께서는 에이즈가 없어" 라고 말하였다. 칸디의 친구들은 칸디에게 "저 소년의 말을 듣지 마" 라고 말하였다.
"Does Father have AIDS?" Kande asked her mother that night. "I'm old enough to know."	1:10 Then Kande went to her house. That evening she asked her mother, "Does father have AIDS? Please tell the truth to me."	1:10 그리고서 칸디는 자기 집으로 가서 그 날 저녁에 어머니에게 물었다. "아버지께서는 에이즈가 있습니까? 저에게 진실을 말씀하여 주세요."
Kande's mother looked away. Kande could see that she was crying. She answered, "yes, I am sorry that you heard it as a rumor first."	1:11 Kande's mother turned around quickly. But Kande saw her mother crying. Kande's mother answered, "Yes. Your father has AIDS. I wanted to tell this bad news to you. But I didn't want you to worry about your father. I'm sorry	1:11 칸디의 어머니는 빨리 뒤돌아 섰다. 그러나 칸디는 어머니가 울고 있는 것을 봤다. 칸디의 어머니는 대답하였다. "그래. 아버지께서는 에이즈가 있어. 나는 너에게 그 나쁜 소식을 말하고 싶었어. 그러나 나는 네가 아버지에 대해서

Table A-1 - continued

	because you heard this news while you were at the market."	걱정하기를 원하지 않았어. 네가 시장에 있는 동안 그 소식을 들어서 미얀해."
What will we do if Father dies?" Kande asked. "How will we survive?" "God will help us," Mama said. They cried together for a while.	1:12 Kande asked her mother, "If father dies, how will we buy food? How will we live?" Kande's mother answered, "God will protect us." Then Kande and her mother cried for a short time.	1:12 칸디는 어머니에게 물었다. "만일 아버지께서 돌아가시면 저희들은 어떻게 음식을 살 것입니까? 저희들은 어떻게 살 것입니까?" 칸디의 어머니는 "하나님께서는 우리를 보호하실 거야" 라고 대답하였다. 그리고서 칸디와 어머니는 잠시 동안 울었다.
Kande's father died just before the rainy season. Friends and relatives came and mourned for him.	1:13 Before the rainy season began, Kande's father died. Friends and relatives came to Kande's house to mourn for her father.	1:13 우기가 시작되기 전에 칸디의 아버지께서는 돌아가셨다. 친구들과 친척들이 칸디의 아버지의 죽음을 애도하러 칸디의 집에 왔다.
"Why didn't they visit when he was so sick and so lonely?" Kande thought. The pastor had been the only visitor before Father died.	1:14 Kande wondered, "Why didn't these people visit my father while he was sick?" Before Kande's father died, only the pastor visited her family.	1:14 칸디는 "아버지께서 아프시는 동안 이 사람들은 왜 우리 아버지를 방문하지 않았어?" 라고 궁금해하였다. 칸디의 아버지께서 돌아가시기 전에 오직 목사님만 칸디의 가족을 방문하셨다.
Some time later Kande and her mother were out gathering firewood. Mama was breathing hard, and she looked very weak.	1:15 A few weeks later Kande and her mother were gathering wood. Kande's mother wanted to cook some food for the children. But she was very tired and very weak.	1:15 몇 주 후에 칸디와 어머니는 나무를 모으고 있었다. 칸디의 어머니는 아이들을 위해 음식을 요리하고 싶어하였다. 그러나 칸디의 어머니는 매우 피곤하고 매우 약하였다.
Kande took her hand, and they sat down to rest. Mama said, "Sometimes I feel like I don't have the strength to do anything	1:16 So Kande held her mother's hand. Then Kande and her mother sat down to rest. Kande's mother said, "Sometimes I feel very weak and	1:16 그래서 칸디는 어머니의 손을 잡았다. 그리고서 칸디와 어머니는 쉬려고 앉았다. 칸디의 어머니는 "나는 가끔씩 약하게 느껴지고 가끔씩 일할

Table A-1 - continued

anymore."	sometimes am not able to work."	수 없어" 라고 말하였다.
Chapter 2: More Trouble for Kande's Family	2:1 Title: Kande's family has more problems.	2:1 제목: 칸디의 가족은 문제가 더 있어요.
Kande and her mother sat under the tree. Mama didn't look well, and she was so tired, Kande had to help her up.	2:2 One day Kande and her mother were sitting near a tree. Kande's mother was sick and very tired. So Kande helped her mother stand up.	2:2 어느 날 칸디와 어머니는 나무 가까이에 앉아 있었다. 칸디의 어머니는 아프며 매우 피곤하였다. 그래서 칸디는 어머니가 일어서는 것을 도와주었다.
Kande's father had died, and with her mother pregnant and feeling weak, the children had to work much harder.	2:3 Kande's father had died. Kande and the other children had to work very hard because Kande's mother was pregnant.	2:3 칸디의 아버지께서는 돌아가셨다. 칸디의 어머니가 임신하였기 때문에 칸디와 다른 아이들은 매우 열심히 일해야만 하였다.
Kande fussed at them when she thought they weren't working hard enough. Mama scolded her for that. She said, "I'm still the mother in this family."	2:4 When Kande saw the other children not working hard, she scolded them. But then mother scolded Kande. Mother said, "I still am the mother for these children."	2:4 칸디는 다른 아이들이 열심히 일하고 있지 않는 것을 봤을 때 아이들을 꾸짖었다. 그러나 그 때 어머니가 칸디를 꾸짖었고 "나는 이 아이들을 위해 여전히 어머니이다" 라고 말하였다.
Then two women from the local church came to visit. One was a health worker; the other was a great story teller.	2:5 One day two women who attended the church that was near Kande's house came to visit her mother. One woman was a nurse. And the other woman told stories very well.	2:5 어느 날 칸디의 집 가까이에 있는 교회를 다니는 여자 두 명이 칸디의 어머니를 방문하러 왔다. 여자 한 명이 간호사였다. 그리고 다른 여자는 이야기를 매우 잘 말하였다.
They did a lot of mama's work. They brought food. They started coming over often. They joked and told stories. Kande was happy to hear mother laughing more often	2:6 These two women did a lot of work for Kande's mother and also brought food for the family. These women came to Kande's house often. These women told jokes and interesting	2:6 이 여자 두 명은 칸디의 어머니를 위해 일을 많이 하였고 또한 가족을 위해 음식을 가져왔고 자주 칸디의 집에 왔다. 그리고 이 여자들은 칸디의 어머니에게 농담과 재미있는 이야기를 말하였다.

Table A-1 - continued

now.	stories to Kande's mother. When Kande heard her mother laughing, she was happy.	칸디는 어머니가 웃는 것을 들었을 때 행복하였다.
Kande overheard the health worker talking to Mama. She did not understand everything they said, but she learned that her father had been unfaithful to Mama.	2:7 But one day Kande heard the nurse talking to her mother. Kande didn't understand all the things that the nurse said. But she learned that her father had slept with another woman.	2:7 그러나 어느 날 칸디는 간호사가 어머니와 말하는 것을 들었고 간호사가 말한 모든 것을 이해할 수 없었다. 그러나 칸디는 아버지가 다른 여자와 잤다는 것을 알았다.
He must have got HIV from another woman. At first Father didn't know he had HIV so he didn't wear a condom and he didn't protect Mama.	2:8 The other woman that Kande's father slept with had HIV. So when Kande's father slept with that woman, he caught HIV from that woman. But he didn't know that he had HIV. So when Kande's father slept with her mother, he didn't use a condom. Condoms protect people from catching HIV.	2:8 칸디의 아버지가 잔 다른 여자는 인체 면역 결핍 바이러스가 있었다. 그래서 칸디의 아버지는 이 여자와 잤을 때 이 여자에게서 인체 면역 결핍 바이러스를 옮았다. 그러나 칸디의 아버지는 자기가 인체 면역 결핍 바이러스가 있는지를 몰랐다. 그래서 칸디의 아버지는 칸디의 어머니와 잤을 때 콘돔을 쓰지 않았다. 콘돔이 항상 사람들이 인체 면역 결핍 바이러스에 걸리는 것을 막는다.
Mama might have caught it from him, and now the baby could get it from Mama. "You must come to the clinic to get tested," the health worker told Mama.	2:9 So perhaps Kande's mother caught HIV from father. And the baby might catch HIV from mother. The nurse said to Kande's mother, "You have to come to the clinic so that we could examine your blood."	2:9 그래서 칸디의 어머니는 아마도 아버지에게서 인체 면역 결핍 바이러스를 옮았을 것이다. 그리고 아기는 아마도 어머니에게서 인체 면역 결핍 바이러스를 옮을 것이다. 간호사는 칸디의 어머니에게 "우리가 아주머니의 피를 검사할 수 있도록 아주머니는 진료소에 와야만 해요" 라고 말하였다.
Mama went to the	2:10 So Kande's mother	2:10 그래서 칸디의 어머니는

Table A-1 - continued

clinic. The health worker drew blood from her arm. It did not hurt, but Mama would have to wait for the results.	went to the clinic. Then the nurse drew blood from mother's arm. But she didn't hurt mother. Then mother returned to her house. The nurses had to examine the blood for two weeks. So mother waited for two weeks.	진료소로 갔다. 그리고서 간호사는 어머니의 팔에서 피를 뽑았다. 그러나 간호사는 어머니를 해치지 않았다. 그리고서 어머니가 자기 집으로 돌아왔다. 간호사들은 이 주 동안 피를 검사해야만 하였다. 그래서 어머니가 이 주 동안 기다렸다.
Two weeks later, Mama told Kande the terrible news. Mama was infected with HIV. The baby might be also.	2:11 Two weeks later a nurse told mother that she had HIV. Then mother told the terrible news to Kande. Mother said, "I have HIV. The baby also might have HIV."	2:11 이 주 후에 간호사가 어머니에게 어머니가 인체 면역 결핍 바이러스가 있다라고 말하였다. 그리고서 어머니가 칸디에게 끔찍한 소식을 말하였다. "나는 인체 면역 결핍 바이러스가 있어. 또한 아기는 아마도 인체 면역 결핍 바이러스가 있을 거야."
Mama began to get sicker. The HIV infection turned into AIDS. She got sores on her skin.	2:12 A few months later mother became very sick. HIV became AIDS. Then sores appeared on mother's skin.	2:12 몇 달 후에 어머니가 매우 아프게 되었다. 인체 면역 결핍 바이러스는 에이즈가 되었다. 그리고서 염증이 어머니의 피부에 났다.
"Will I get AIDS from touching Mama?" Kande asked the women from the church. "Not if you are careful," said the health worker.	2:13 One day Kande asked the nurse, "If I touch mother, will I catch AIDS?" The nurse answered, "If you're careful, you won't catch AIDS from your mother."	2:13 어느 날 칸디는 간호사에게 "만일 제가 어머니를 만지면 저는 에이즈에 걸릴 것입니까?" 라고 물었다. 간호사는 "만일 네가 조심하면 너는 어머니에게서 에이즈를 옮지 않을 거야" 라고 대답하였다.
She showed Kande the safest ways to take care of Mama and taught her the best foods to give her. Kande was reassured.	2:14 The nurse showed Kande how to take care of her mother safely. And the nurse taught Kande how to cook food for her mother. Kande thanked the nurse and took care of her mother.	2:14 간호사는 칸디에게 어떻게 안전하게 어머니를 돌보는 지를 보여주었고 칸디에게 어떻게 어머니를 위해 음식을 요리하는 지를 가르쳤다. 칸디는 간호사에게 감사하였고 어머니를 돌봤다.

Table A-1 - continued

But Mama worried about her baby. None of the local clinics had the right medicine to help prevent the baby from getting AIDS.	2:15 But mother worried about her baby. The clinic didn't have the medicine that prevents babies from catching AIDS.	2:15 그러나 어머니가 아기에 대해서 걱정하였다. 진료소는 항상 아기들이 에이즈에 걸리는 것을 막는 약이 없었다.
The baby was born. Mama was very weak. She held the new baby and cried. "Yatima," she said. "Orphan."	2:16 A few weeks later mother gave birth to the baby. But she was very weak. She held the new baby and cried. She said, "Yatima." Yatima means, "This child doesn't have parents."	2:16 몇 주 후에 어머니가 아기를 낳았다. 그러나 어머니가 매우 약하였다. 어머니가 새로 태어난 아기를 안았고 울었다. 그리고 어머니가 "야티마야" 라고 말하였다. 야티마는 "이 아이는 부모님이 없어" 라는 뜻이다.
Mama died a few days later, and Kande named the baby Yatima.	2:17 A few days later mother died. Then Kande named the baby Yatima.	2:17 며칠 후에 어머니께서는 돌아가셨다. 그러자 칸디는 아기를 야티마라고 이름지었다.
Kande took the baby under the tree and held her. "I won't let you be an orphan," she said. "You are my baby now."	2:18 Kande carried the baby to the tree and held him. She said, "I'll be your mother now. I'll take care of you. You're my baby now."	2:18 칸디는 나무로 아기를 들고 갔고 그를 안았다. 그리고 칸디는 말하였다. "나는 이제 네 어머닐 거야. 나는 너를 돌볼 거야. 너는 이제 내 아기야."
Chapter 3: Dangers for Kande's Family	3:1 Title: Kande's family has some difficult problems.	3:1 제목: 칸디의 가족은 어려운 문제가 있어요.
Kande sat under the tree feeding her baby sister, Yatima. It would have been best if someone could have breastfed the baby, but since Mama had died of AIDS, people were afraid they would get it from the baby.	3:2 One day while Kande was sitting near a tree, she was feeding Yatima. If Kande's mother had given breast-milk to the baby, he would have been very healthy. But Kande's mother caught AIDS from her father. So she died. Other people didn't want to help the baby.	3:2 어느 날 칸디는 나무 가까이에 앉아 있는 동안 야티마를 먹이고 있었다. 만일 칸디의 어머니가 아기에게 젖을 주었더라면 아기는 매우 건강하였을 것이다. 그러나 칸디의 어머니는 칸디의 아버지에게서 에이즈를 옮았다. 그래서 칸디의 어머니께서는 돌아가셨다. 다른 사람들이 아기를

Table A-1 - continued

	The other people were afraid to touch the baby. The other people thought that they might catch AIDS from the baby.	도와주고 싶지 않았다. 다른 사람들은 아기를 만지는 것을 두려워하였다. 다른 사람들은 자기들이 아마도 아기에게서 에이즈를 옮을 것이다라고 생각하였다.
The church helped Kande get baby milk powder and clean water. Kande was so glad that Yatima didn't get AIDS from Mama.	3:3 The people who attended the church helped Kande buy milk and clean water. One day the nurse told Kande that Yatima didn't catch AIDS from Kande's mother. When Kande knew that the baby was healthy, she was very happy.	3:3 교회를 다니는 사람들은 칸디가 우유와 깨끗한 물을 사는 것을 도와주었다. 어느 날 간호사는 칸디에게 야티마가 칸디의 어머니에게서 에이즈를 옮지 않았다라고 말하였다. 칸디는 아기가 건강한다는 것을 알았을 때 매우 행복하였다.
Kande's younger brother, Jumoke came to talk to her. "I want to quit school just like you and Teshi did."	3:4 One day while Kande was feeding the baby, her younger brother named Jumoke came to talk to her. Jumoke said, "I want to stop attending school like you and Teshi stopped attending school."	3:4 어느 날 칸디가 아기를 먹이는 동안 주목이라는 자기 남동생은 칸디와 말하러 왔다. 주목은 "누나와 태쉬가 학교를 다니는 것을 그만뒀던 것처럼 나는 학교를 다니는 것을 그만두고 싶어요" 라고 말하였다.
"No, first you need to finish school," Kande said. "Then you can help the rest of us and Teshi can go back to school. And you must be careful not to make friends with boys in your school who use intravenous drugs.	3:5 Kande said, "No. You have to finish attending school. Then you will be able to work. If you work, you will be able to buy food for our family. When you work, Teshi will start attending school again. So you have to continue attending school. When you're at school, be very careful. Some boys who attend school use drugs. Although those boys use	3:5 칸디는 말하였다. "아니다. 너는 학교를 다니는 것을 끝내야만 하고 일할 수 있을 거야. 만일 네가 일하면 너는 우리의 가족을 위해 음식을 살 수 있을 거야. 네가 일할 때 태쉬는 다시 학교를 다니기 시작할 거야. 그래서 너는 계속 학교를 다녀야만 해. 학교에 있을 때 매우 조심해. 학교를 다니는 소년들이 마약을 사용해. 그 소년들이 마약을 사용하지만 너는 마약을

Table A-1 - continued

	drugs, you must not use drugs.	사용하지 말아야만 해.
Drugs are very bad for you and you can get HIV from sharing dirty needles."	3:6 Drugs are very dangerous. If you use drugs, you will become sick. If you use drugs with those boys, you might catch HIV."	3:6 마약이 매우 위험해. 만일 네가 마약을 사용하면 너는 아프게 될 거야. 만일 네가 그 소년들과 함께 마약을 사용하면 너는 아마도 인체 면역 결핍 바이러스에 걸릴 거야."
Jumoke said he would study hard. He promised he would not take drugs.	3:7 Jumoke said, "I'll continue attending school and will study hard." He also promised Kande that he won't use drugs.	3:7 주목은 "나는 계속 학교를 다닐 것이고 열심히 공부할 것이요" 라고 말하였다. 그리고 또한 주목은 칸디에게 마약을 사용하지 않을 것을 약속하였다.
One day Kande's relatives visited. "By tradition, this land belongs to me because your father died," he told her. "But we have nowhere to go," said Kande.	3:8 One day Kande's uncle came to her house to visit her. He said to her, "I own this house and this land now because your father died." Then Kande asked her uncle, "Where will my family live?"	3:8 어느 날 칸디의 삼촌은 칸디를 방문하러 칸디의 집에 와서 칸디에게 "네 아버지께서 돌아가셨기 때문에 나는 이제 이 집과 이 땅을 소유해" 라고 말하였다. 그러자 칸디는 자기 삼촌에게 "제 가족은 어디에서 살 것입니까?" 라고 물었다.
"Your parent's death has been hard on our whole village. We all try to help you as much as we can, and you can stay here for now," he said.	3:9 Kande's uncle said, "All the people who live in this village have to do more work because your parents died. But I'll help you buy food. You and your family may continue living in this house for a short time. But you have to give me half of the crops that you harvest."	3:9 칸디의 삼촌은 말하였다. "네 부모님께서 돌아가셨기 때문에 이 마을에 사는 모든 사람들은 일을 더 해야만 해. 그러나 나는 네가 음식을 사는 것을 도와줄 거야. 너와 네 가족은 잠시 동안 이 집에서 계속 살아도 돼. 그러나 너는 나에게 네가 추수하는 곡식의 반을 줘야만 해."
"Will we have to move away?" Falala asked Kande. "No, our cousin said we could stay here. But we do need to share	3:10 After Kande's uncle left the house, Falala asked Kande, "Did our uncle say that we have to move to another house?"	3:10 칸디의 삼촌이 집을 떠난 후에 팔라라는 칸디에게 "삼촌께서는 우리가 다른 집으로 이사해야만 한다고

Table A-1 - continued

half of our crop with him."	Kande said, "No. Our uncle said that we may continue living in this house. But we have to give our uncle half of the crops that we harvest."	말씀하셨어요?" 라고 물었다. 칸디는 말하였다. "아니다. 삼촌께서는 우리가 이 집에서 계속 살아도 된다고 말씀하셨어. 그러나 우리는 삼촌에게 우리가 추수하는 곡식의 반을 줘야만 해."
"That doesn't leave enough for us," said Falala. "We will have to do something else to earn a living."	3:11 Falala said, "But if we give our uncle half of the crops that we harvest, we won't have enough food. We have to start working to earn money."	3:11 팔라라는 말하였다. "그러나 만일 우리가 삼촌에게 우리가 추수하는 곡식의 반을 주면 우리는 음식이 충분하지 않을 것이요. 우리는 돈을 벌기 위해 일하기 시작해야만 해요."
One morning, Kande and Teshi were carrying the baby to the health clinic for a checkup. Teshi pointed to a man near the market. "He is the man who gave me this bracelet. Maybe he can help us make a living," she said.	3:12 One morning Kande and Teshi took the baby to the clinic so that the nurse could see him. While Kande and Teshi were walking to the clinic, Teshi pointed at a man who was standing near the market. She said to Kande, "That man gave this necklace to me. That man likes me. He might help us earn some money."	3:12 어느 날 아침에 간호사가 아기를 볼 수 있도록 칸디와 태쉬는 진료소로 아기를 데리고 갔다. 칸디와 태쉬가 진료소로 걸어가는 동안 태쉬는 시장 가까이에 서 있는 남자를 가리켰고 칸디에게 말하였다. "저 남자는 나에게 이 목걸이를 주었어요. 저 남자는 나를 좋아하고 아마도 우리가 돈을 버는 것을 도와줄 것이요."
The healthcare worker said that the baby was doing fine. But she also talked to Kande and Teshi about important issues for girls and boys their age.	3:13 After the nurse looked at the baby, she said to Kande and Teshi, "The baby is very healthy." Then the nurse talked to Kande and Teshi about things that children have to know.	3:13 간호사는 아기를 바라본 후에 칸디와 태쉬에게 "아기는 매우 건강해" 라고 말하였다. 그리고서 간호사는 칸디와 태쉬에게 아이들이 알고 있어야만 하는 것에 대해서 말하였다.
"Because you are orphans, there are some men who might try to give you food and	3:14 The nurse said, "Some men might give food and gifts to you because your parents	3:14 간호사는 말하였다. "너희들의 부모님께서 돌아가셨기 때문에 남자들이

Table A-1 - continued

presents to persuade you to have sex with them. Do not let them trick you. There is too much risk of getting pregnant, or of getting HIV and other diseases that come through sex."	died. Those men want to sleep with you. So you have to be very careful. If you sleep with those men, you might become pregnant. If you sleep with those men, you might catch HIV and other diseases from them."	아마도 너희들에게 음식과 선물을 줄 거야. 이 남자들은 너희들과 자고 싶어해. 그래서 너희들은 매우 조심해야만 해. 만일 너희들이 이 남자들과 자면 너희들은 아마도 임신될 거야. 만일 너희들이 이 남자들과 자면 너희들은 아마도 이 남자들에게서 인체 면역 결핍 바이러스와 다른 병을 옮을 거야."
Teshi wondered if the man who gave her the bracelet might be trying to get her to have sex with him. Kande, Teshi, and Falala promised each other they would each wait for sex until they were married.	3:15 After Kande and Teshi left the clinic, Teshi asked Kande, "Does the man who gave this necklace to me want to sleep with me?" Kande, Teshi, and Falala promised each other that they won't sleep with a man until they marry him.	3:15 칸디와 태쉬가 진료소를 떠난 후에 태쉬는 칸디에게 "나에게 이 목걸이를 주었던 남자는 나와 자고 싶어해요?" 라고 물었다. 칸디와 태쉬와 팔라라는 서로에게 남자와 결혼할 때까지 남자와 자지 않을 것을 약속하였다.
Chapter 4: Kande finds hope	4:1 Title: Kande has hope.	4:1 제목: 칸디는 희망이 있어요.
Kande and her family lived through some hard times. Her parents had died of AIDS, and taking care of her younger siblings was difficult for her. Some days they went hungry, but Kande always worked hard for them, and she tried to be like a mother to her baby sister.	4:2 Kande and her family had many difficult problems. After Kande's parents caught AIDS, they died. So Kande took care of her younger brothers and her younger sisters. Sometimes the children were not able to buy food. But Kande always worked hard to buy food for her family. And she tried to be like a mother for the baby.	4:2 칸디와 그녀의 가족은 많은 어려운 문제가 있었다. 칸디의 부모님께서는 에이즈에 걸리신 후에 돌아가셨다. 그래서 칸디는 자기 남동생들과 자기 여동생들을 돌봤다. 아이들은 가끔씩 음식을 살 수 없었다. 그러나 칸디는 자기 가족을 위해 음식을 사기 위해 항상 열심히 일하였고 아기에게 어머니와 비슷하려고 노력하였다.
A boy from their village, Ajani,	4:3 A boy named Ajani lived in Kande's village.	4:3 아자니라는 소년이 칸디의 마을에 살았다. 아자니는

Table A-1 - continued

sometimes visited Kande. He brought his baby brother with him. "Come to church with us," Ajani would say. "Not this time," Kande always said. "There's too much work to do."	Sometimes Ajani visited Kande and sometimes brought his younger brother. One day Ajani said to Kande, "Go to church with me." Kande said, "I'm not able today to go to church. I have to do a lot of work."	가끔씩 칸디를 방문하였고 가끔씩 자기 남동생을 데려왔다. 그리고 어느 날 아자니는 칸디에게 "나와 함께 교회로 가" 라고 말하였다. 칸디는 말하였다. "나는 오늘 교회로 갈 수 없어요. 나는 일을 많이 해야만 해요."
But one day Kande's sister Falala said, "I'll go. Maybe I'll learn something." Teshi, another sister, said, "I'll go too. Maybe I'll make some new friends."	4:4 But one day Kande's younger sister named Falala said to Ajani, "I'll go to church with you. I might learn some things." Then Teshi said, "I also will go to church. I might meet some new friends."	4:4 그러나 어느 날 팔라라라는 칸디의 여동생은 아자니에게 말하였다. "나는 오빠와 함께 교회로 가겠어요. 나는 아마도 무엇인가를 배울 것이요." 그리고서 태쉬는 말하였다. "나도 교회로 가겠어요. 나는 아마도 새 친구들을 만날 것이요."
"Take Iniko and Yatima with you, then," said Kande. "Jumoke and I will stay here and get some work done."	4:5 Then Kande said to Falala, "Take Iniko and Yatima to church with you. Jumoke and I will stay here. We will work."	4:5 그러자 칸디는 팔라라에게 말하였다. "너와 함께 교회로 이니꼬와 야티마를 데리고 가. 주목과 나는 여기에 머무를 거야. 우리는 일할 거야."
Later, when her sisters returned and told her about the church's community garden, Kande did get involved. The church let them work a large plot, and they were allowed to keep all the food to eat or trade at the market.	4:6 While Falala and Teshi were at church, some people told Falala about a garden that was at the church. Then Falala went to her house. She told Kande about the church's garden. Then Kande and Falala went to the church. The pastor told Kande that she may grow vegetables in the church's garden. Kande and her family may eat all the vegetables that she harvests. And Kande	4:6 팔라라와 태쉬가 교회에 있는 동안 사람들이 팔라라에게 교회에 있는 정원에 대해서 말하였다. 그리고서 팔라라는 자기 집으로 가서 칸디에게 교회 정원에 대해서 말하였다. 그리고서 칸디와 팔라라는 교회로 갔다. 목사님께서는 칸디에게 그녀가 교회 정원에서 채소를 재배해도 된다라고 말씀하셨다. 칸디와 그녀의 가족은 칸디가 거두는 모든 채소를 먹어도 된다. 그리고 칸디는 시장에서 다른

Table A-1 - continued

	may trade some vegetables for other things at the market.	것과 채소를 교환해도 된다.
Now they worked very hard, but they were making a better living than before	4:7 So Kande and her family worked very hard. They were able to grow many vegetables at the church's garden.	4:7 그래서 칸디와 그녀의 가족은 매우 열심히 일하였다. 칸디와 그녀의 가족은 교회 정원에 채소를 많이 재배할 수 있었다.
One day Kande's relative sent word that it was time now for Kande and her siblings to leave their father's farm. Kande was very sad.	4:8 One day Kande's uncle visited her again. He said to her, "You and your family have to leave this house now. I want to live at this house now." So Kande was very sad.	4:8 어느 날 칸디의 삼촌은 다시 칸디를 방문하였고 칸디에게 말하였다. "너와 네 가족은 이제 이 집을 떠나야만 해. 나는 이제 이 집에 살고 싶어." 그래서 칸디는 매우 슬펐다.
A woman from the church invited the children to come and live with her. She had helped them when their mother was sick. She lived next to the church and the garden.	4:9 A few days later a woman who attended the church invited the children to live at her house. When Kande's mother was sick, this woman had helped Kande take care of her. This woman lived in a house that was near the church's garden.	4:9 며칠 후에 교회를 다니는 여자가 아이들이 자기 집에 살도록 초대하였다. 칸디의 어머니가 아팠을 때 이 여자는 칸디가 어머니를 돌보는 것을 도와주었다. 이 여자는 교회 정원 가까이에 있는 집에서 살았다.
The children moved in with her, and their cousins took back the old home and garden.	4:10 So the children moved to this woman's house. Then Kande's uncle moved to her house. He also grew vegetables at her house.	4:10 그래서 아이들은 이 여자의 집으로 이사하였다. 그리고서 칸디의 삼촌은 칸디의 집으로 이사하였고 또한 칸디의 집에서 채소를 재배하였다.
Kande and the whole family, even Iniko, worked in the garden. Teshi and Falala also learned to sew, and the church helped them to buy sewing machines. Jumoke learned	4:11 So Kande and her family grew vegetables in the garden that was near the church. Teshi and Falala also learned how to sew clothes. The pastor helped Teshi buy machines that sewed	4:11 그래서 칸디와 그녀의 가족은 교회 가까이에 있는 정원에서 채소를 재배하였다. 또한 태쉬와 팔라라는 어떻게 옷을 바느질하는지 배웠다. 목사님께서는 태쉬와 팔라라가 옷을 바느질하는 기계를 사는

Table A-1 - continued

carpentry in the church's work shop.	clothes. And Jumoke learned how to build things with wood.	것을 도와주셨다. 그리고 주목은 어떻게 나무로 물건을 만드는지 배웠다.
Kande was glad that Ajani often helped her in the garden while the babies played together.	4:12 Ajani helped Kande often work in the church's garden. Kande liked Ajani. Ajani always was very kind to Kande.	4:12 아자니는 자주 칸디가 교회 정원에서 일하는 것을 도와주었다. 칸디는 아자니를 좋아하였다. 아자니는 항상 칸디에게 매우 친절하였다.
Kande told Ajani, "When my parents died, I thought our whole family would die too. Life is still hard, but now we have hope."	4:13 One day Kande said to Ajani, "After my parents died, I thought that my family also will die. I and my family have to work very hard now. But we have hope now."	4:13 어느 날 칸디는 아자니에게 말하였다. "내 부모님께서 돌아가신 후에 나는 내 가족도 죽을 것이다라고 생각하였어요. 나와 내 가족은 이제 매우 열심히 일해야만 해요. 그러나 우리는 이제 희망이 있어요."
Chapter 5: Kande's Community learns about AIDS	5:1 Title: Kande's friends and her neighbors learn about AIDS.	5:1 제목: 칸디의 친구들과 이웃사람들은 에이즈에 대해서 배워요.
After church one day, Kande talked to Ajani. "The church has helped us so much! They have let us grow crops on their land. They have taught us to make a good living, and they have been friends to us in so many ways. How can we ever repay the good they have done for us?"	5:2 One day Kande and Ajani went to church. After the pastor finished preaching, Kande said to Ajani, "The people who attend this church have helped my family much! The pastor said that we may grow vegetables at the church's garden. People who attend this church have taught my younger sisters how to sew clothes. And these people have become our friends. We want to thank these people."	5:2 어느 날 칸디와 아자니는 교회로 갔다. 목사님께서 설교하는 것을 끝내신 후에 칸디는 아자니에게 말하였다. "이 교회를 다니는 사람들은 내 가족을 많이 도와주었어요! 목사님께서는 우리가 교회 정원에 채소를 재배해도 된다고 말씀하셨어요. 이 교회를 다니는 사람들이 내 여동생들에게 어떻게 옷을 바느질하는 지를 가르쳤고 우리의 친구들이 되었어요. 우리는 이 사람들에게 감사하고 싶어요."
The church soon hosted a conference on AIDS prevention. Trainers	5:3 A few weeks later the pastor invited many people to come to a big	5:3 몇 주 후에 목사님께서는 많은 사람들이 큰 회의에 오도록 초대하셨다. 그래서

Table A-1 - continued

and learners came from all around. Kande, Teshi, and Falala attended. They brought their brother Jumoke. Ajani came too.	meeting. So many people came to this meeting. Kande, Teshi, Falala, and Jumoke attended this meeting. Ajani also came to this meeting. The pastor taught these people about AIDS. Then the people who attended this meeting talked to each other about AIDS.	많은 사람들이 그 회의에 왔다. 칸디와 태쉬와 팔라라와 주목은 그 회의에 참석하였다. 아자니도 그 회의에 왔다. 목사님께서는 이 사람들에게 에이즈에 대해서 가르치셨다. 그리고서 그 회의에 참석한 사람들은 서로 에이즈에 대해서 말하였다.
One of the conference leaders approached Kande. "We want you and your sisters to become AIDS prevention trainers. Who knows the need to prevent AIDS better than you? And you all read well. People know that you know the facts about HIV and AIDS." They accepted happily.	5:4 One teacher said to Kande, "We want you and your younger sisters to become teachers. We want you to teach other people about AIDS. You know that AIDS is dangerous because your parents died. You know about AIDS and read books well. People know that you know about HIV." Then Kande and her younger sisters were very happy.	5:4 선생님 한 분이 칸디에게 말하였다. "우리는 너와 네 여동생들이 선생님들이 되기를 원해. 우리는 네가 다른 사람들에게 에이즈에 대해서 가르치기를 원해. 네 부모님께서 돌아가셨기 때문에 너는 에이즈가 위험하다는 것을 알고 있고 에이즈에 대해서 알고 있어. 그리고 너는 책을 잘 읽었어. 사람들이 네가 인체 면역 결핍 바이러스에 대해서 알고 있다는 것을 알고 있어." 그러자 칸디와 그녀의 여동생들은 매우 행복하였다.
Teshi started right away, helping with other workshops in the region. With her energy and easy laugh, she made people listen to the hard facts about AIDS. She soon became a leading trainer in the region, and lots of people went to her seminars.	5:5 Teshi immediately started teaching other people about AIDS. She helped the teachers plan more meetings. Teshi and the other teachers wanted to teach people who lived in other villages about AIDS. People listened to Teshi because she was very kind. So Teshi taught the people about AIDS. She taught the people about	5:5 태쉬는 다른 사람들에게 에이즈에 대해서 즉시 가르치기 시작하였고 선생님들이 회의를 더 계획하는 것을 도와주었다. 태쉬와 다른 선생님들은 다른 마을에 사는 사람들에게 에이즈에 대해서 가르치고 싶어하였다. 태쉬가 매우 친절하였기 때문에 사람들이 태쉬의 말을 들었다. 그래서 태쉬는 사람들에게 에이즈에

Table A-1 - continued

	AIDS very well. When Teshi talked to people about AIDS, many of them listened to her.	대해서 가르쳤고 사람들에게 에이즈에 대해서 매우 잘 가르쳤다. 태쉬가 사람들에게 에이즈에 대해서 말하였을 때 많은 사람들이 태쉬의 말을 들었다.
Falala began drawing pictures and writing lessons to use in the seminars. She made booklets in her own language that explained how to avoid HIV and how to care for people sick with AIDS.	5:6 Falala started writing short books that explained AIDS. The teachers used at the meetings the books that Falala wrote. These books explained how people caught HIV. The books also explained how a person should take care of other people who have AIDS.	5:6 팔라라는 에이즈를 설명하는 짧은 책을 쓰기 시작하였다. 선생님들은 회의에 팔라라가 쓴 책을 사용하였다. 그 책은 사람들이 어떻게 인체 면역 결핍 바이러스에 걸리는지 설명하였고 또한 어떤 사람이 어떻게 에이즈가 있는 다른 사람들을 돌보아야 하는지 설명하였다.
Kande and Ajani helped by inviting people to the courses and making sure the leaders had the supplies they needed. They made a special effort to invite teenage boys.	5:7 Kande and Ajani invited many people to come to these meetings. Kande gave the books that Falala wrote to the people who attended these meetings. Kande and Ajani invited many young men to come to these meetings. Kande knew that young men have to learn about AIDS.	5:7 칸디와 아자니는 많은 사람들이 그 회의에 오도록 초대하였다. 칸디는 그 회의에 참석한 사람들에게 팔라라가 쓴 책을 주었다. 칸디와 아자니는 많은 젊은 남자들이 그 회의에 오도록 초대하였다. 칸디는 젊은 남자들이 에이즈에 대해서 배워야만 한다는 것을 알았다.
Sometimes boys think that to become real men, they must have sex. Ajani let the boys know that he and Kande had promised not to have sex until they got married. That way they would be sure not to get HIV.	5:8 Some young men think that a man who sleeps with a woman is a strong man. So Ajani said to the young men, "I won't sleep with Kande until I marry her and won't sleep with other women. Therefore I know that I won't catch	5:8 젊은 남자들이 여자와 잔 남자가 강한 남자이다라고 생각한다. 그래서 아자니는 젊은 남자들에게 말하였다. "저는 칸디와 결혼할 때까지 칸디와 자지 않을 것이고 다른 여자들과 자지 않을 것입니다. 따라서 저는 제가 인체 면역 결핍 바이러스에 걸리지 않을

Table A-1 - continued

	HIV. And I know that Kande won't catch HIV."	것이다는 것을 알고 있고 칸디가 인체 면역 결핍 바이러스에 걸리지 않을 것이다는 것을 알고 있습니다."
Kande and Ajani got married, and later they had a baby. One day their large family gathered under Kande's favorite tree. "I used to sit here and talk with Mama," she said.	5:9 A few months later Kande married Ajani. Then she gave birth to a baby. One day Kande's family went to the tree that she liked. Kande said, "When our mother was living, she and I sat here often. While we were sitting near this tree, we talked to each other about many things."	5:9 몇 달 후에 칸디는 아자니와 결혼하였고 아기를 낳았다. 어느 날 칸디의 가족은 칸디가 좋아하는 나무로 갔다. 칸디는 말하였다. "어머니께서 살아 계셨을 때 어머니와 나는 자주 여기에 앉았어. 우리는 이 나무 가까이에 앉아 있는 동안 서로 많은 것에 대해서 말하였어."
Iniko, who had been very young when her parents died, said, "I miss Mama and Father, but I think they would be proud of us now."	5:10 Then Iniko said, "I miss our mother and our father. But I think that our parents are proud of us now."	5:10 그리고서 이니꼬는 말하였다. "나는 어머니와 아버지를 그리워해요. 그러나 나는 우리의 부모님께서 이제 우리를 자랑스러워하신다라고 생각해요."
	5:11 Footnote: An organization named Shell Publishing owns this story.	5:11 각주: 셸 프블리싱이라는 기관이 이 이야기를 소유하고 있습니다.

Table A-2. TTA's English and Korean Drafts of "Melissa's Eyes are Sore"

1:1 Title: Melissa's eyes are sore.	1:1 제목: 멜리사는 눈이 아파요.
1:2 One day a girl named Melissa was sitting outside her house. But Melissa was not happy because her eyes were very sore. She thought that some sand was in her eyes. So she called a friend named Janet and said to her, "Please look at my eyes. Is some sand in my eyes?"	1:2 어느 날 멜리사라는 소녀가 자기 집 바깥에 앉아 있었다. 그러나 멜리사는 눈이 매우 아팠기 때문에 행복하지 않았다. 멜리사는 자기 눈 안에 모래가 있다라고 생각하였다. 그래서 멜리사는 재닛이라는 친구를 불러서 말하였다. "내 눈을 봐. 내 눈 안에 모래가 있어?"
1:3 Janet said to Melissa, "Nothing is	1:3 재닛은 멜리사에게 말하였다. "네 눈

Table A-2 - continued

in your eyes. But your eyes are very red."	안에 아무것도 없어. 그러나 네 눈은 매우 빨개."
1:4 Then Janet said to Melissa, "Please look at my eyes because they also are very sore." So Melissa looked at Janet's eyes. Janet's eyes also were very red!	1:4 그리고서 재닛은 멜리사에게 "내 눈도 매우 아프기 때문에 내 눈을 봐" 라고 말하였다. 그래서 멜리사는 재닛의 눈을 봤다. 재닛의 눈도 매우 빨갰다!
1:5 Then Melissa entered her house to rest. She slept for a short time. Then she woke up. While Melissa was in her house, she heard Janet talking to a friend named Alex.	1:5 그리고서 멜리사는 쉬기 위해 자기 집으로 들어가서 잠시 동안 잤다. 그리고서 멜리사는 깨었고 집에 있는 동안 재닛이 알렉스라는 친구와 말하는 것을 들었다.
1:6 Melissa called Alex loudly. She shouted, "Alex, come into my house. Something is preventing me from opening my eyes! I'm not able to see things!"	1:6 멜리사는 알렉스를 큰 소리로 불러서 외쳤다. "알렉스야, 우리 집에 들어와. 나는 무언가 때문에 눈을 뜰 수 없어! 나는 볼 수 없어!"
1:7 Then Alex entered Melissa's house quickly. There were many flies inside the house. There were many flies near Melissa's eyes also. Alex knew that Melissa's eyes were very sick. He said to Melissa, "Yellow pus is covering your eyes. This pus is preventing you from opening your eyes."	1:7 그러자 알렉스는 멜리사의 집으로 빨리 들어갔다. 집 안에 파리들이 많았다. 멜리사의 눈 가까이에도 파리들이 많았다. 알렉스는 멜리사의 눈이 매우 아프다는 것을 알았고 멜리사에게 말하였다. "노란 고름이 네 눈을 덮고 있어. 너는 그 고름 때문에 눈을 뜰 수 없어."
1:8 Alex said to Melissa, "I'll try to clean your eyes. But I don't have a clean cloth to clean your eyes." A towel that was hanging on a rope was dirty. And Alex's hands also were dirty.	1:8 알렉스는 멜리사에게 말하였다. "나는 네 눈을 깨끗하게 하려고 노력할께. 그러나 나는 네 눈을 깨끗하게 하기 위해 깨끗한 천이 없어." 밧줄 위에 걸려 있는 수건이 더러웠다. 그리고 알렉스의 손도 더러웠다.
1:9 So Alex said to Melissa, "I'll call Netty so that she could look at your eyes. Netty will help your eyes become well."	1:9 그래서 알렉스는 멜리사에게 말하였다. "네티가 네 눈을 볼 수 있도록 나는 네티를 부를께. 네티는 네 눈이 건강해지는 것을 도와줄 거야."
1:10 Then Netty came to Melissa's house. After Netty looked at Melissa's eyes, she said to Melissa, "Your eyes are very sick. Some germs have entered your eyes. We have to wash	1:10 그리고서 네티는 멜리사의 집에 와서 멜리사의 눈을 본 후에 멜리사에게 말하였다. "네 눈은 매우 아파. 세균이 네 눈에 들어갔어. 우리는 네 눈을 씻어야만 하고 그것들을 철저히 깨끗하게 해야만

Table A-2 - continued

your eyes. We have to clean your eyes thoroughly. And we have to clean your eyes each day until they become healthy again."	해. 그리고 네 눈이 다시 건강해질 때까지 우리는 매일 네 눈을 깨끗하게 해야만 해."
1:11 Then Netty washed her hands with clean water thoroughly and put clean water in a teaspoon. Then she put some salt in that water and dipped a small piece of cloth in it. Then she washed Melissa's left eye with this cloth.	1:11 그리고서 네티는 깨끗한 물로 손을 철저히 씻었고 찻숟가락 안에 깨끗한 물을 부었고 그 물 속에 소금을 넣었다. 그리고서 네티는 찻숟가락에 있는 물에 작은 천을 살짝 담가서 그 천으로 멜리사의 왼쪽 눈을 씻었다.
1:12 After Netty washed Melissa's left eye, she burned the cloth and washed her hands thoroughly again. Then she told Alex to clean Melissa's other eye.	1:12 네티는 멜리사의 왼쪽 눈을 씻은 후에 천을 태웠고 다시 손을 철저히 씻었다. 그리고서 네티는 알렉스에게 멜리사의 다른 눈을 깨끗하게 하라고 말하였다.
1:13 Netty said to Alex, "Before you clean Melissa's eye, you have to wash your hands first. And you have to use a clean cloth. Then the germs won't be able to spread."	1:13 네티는 알렉스에게 말하였다. "먼저 너는 멜리사의 눈을 깨끗하게 하기 전에 손을 씻어야만 하고 깨끗한 천을 쓰야만 해. 그러면 세균은 퍼질 수 없을 거야."
1:14 After Netty and Alex finished washing Melissa's eyes, she said, "I'm able to see things now!" Then Netty said to Alex, "Burn the cloth with fire. And you have to wash your hands thoroughly."	1:14 네티와 알렉스가 멜리사의 눈을 씻는 것을 끝낸 후에 멜리사는 "저는 이제 볼 수 있습니다!" 라고 말하였다. 그리고서 네티는 알렉스에게 말하였다. "천을 불에 태워. 그리고 너는 손을 철저히 씻어야만 해."
1:15 But Janet's eyes still were sore. So Janet asked Melissa to clean her eyes. Then Melissa said to Janet, "I'll clean your eyes." But Alex said to Melissa, "You have to wash your hands first. And you have to use a clean cloth."	1:15 그러나 재닛의 눈은 여전히 아팠다. 그래서 재닛은 멜리사가 자기 눈을 깨끗하게 하여 주기를 부탁하였다. 그러자 멜리사는 재닛에게 "나는 네 눈을 깨끗하게 할께" 라고 말하였다. 그러나 알렉스는 멜리사에게 "먼저 너는 손을 씻어야만 하고 깨끗한 천을 쓰야만 해" 라고 말하였다.
1:16 So Melissa did all the things that Alex said. She cleaned Janet's left eye thoroughly and burned the cloth. Then she washed her hands and washed Janet's right eye with another cloth.	1:16 그래서 멜리사는 알렉스가 말하였던 모든 일들을 하였고 재닛의 왼쪽 눈을 철저히 깨끗하게 하였고 천을 태웠다. 그리고서 멜리사는 손을 씻었고 다른

Table A-2 - continued

Then she burned that cloth and washed her hands again.	천으로 재닛의 오른쪽 눈을 씻었고 그 천을 태웠다. 그리고서 멜리사는 다시 손을 씻었다.
1:17 Melissa burned the cloth so that the germs could not spread and washed her hands so that they could not spread. Then Alex asked Netty, "Why did Melissa's eyes and Janet's eyes become sick? Why did Melissa's eyes and Janet's eyes become red?"	1:17 세균이 퍼질 수 없도록 멜리사는 천을 태웠다. 그리고서 세균이 퍼질 수 없도록 멜리사는 손을 씻었다. 그리고서 알렉스는 네티에게 물었다. "멜리사의 눈과 재닛의 눈은 왜 아프게 되었습니까? 멜리사의 눈과 재닛의 눈은 왜 빨개졌습니까?"
1:18 Netty said, "If you touch your eyes with a dirty towel, germs will enter them. When you touch your eyes with dirty hands, germs also enter them. If you wash your face with dirty water, germs will enter your eyes."	1:18 네티는 말하였다. "만일 네가 더러운 수건으로 눈을 만지면 세균이 네 눈에 들어갈 거야. 또한 네가 더러운 손으로 눈을 만질 때 세균이 네 눈에 들어가. 만일 네가 더러운 물로 네 얼굴을 씻으면 세균이 네 눈에 들어갈 거야."
1:19 Melissa said to Netty, "We have to clean our eyes carefully!" Then Netty agreed with Melissa. Netty said to Melissa, "When you wash your eyes, you have to wash them with clean water. You also have to wash your face and your eyes with clean water each day.	1:19 멜리사는 네티에게 "저희들은 눈을 조심스럽게 깨끗하게 해야만 합니다!" 라고 말하였다. 그러자 네티는 멜리사의 의견에 동의하였고 멜리사에게 말하였다. "너는 눈을 씻을 때 깨끗한 물로 눈을 씻어야만 해. 또한 너는 매일 깨끗한 물로 얼굴과 눈을 씻어야만 해.
1:20 You have to chase all the flies away from your house. And you have to take all the garbage away from your house. You should eat three kinds of food each day. Fresh food is the best food. And before you eat the food, you have to wash your hands. Then your eyes will become healthy."	1:20 너는 집 밖으로 모든 파리들을 쫓아 버려야만 하고 네 집에 있는 모든 쓰레기를 내다버려야만 해. 그리고 너는 매일 세 종류의 음식을 먹어야 해. 신선한 음식이 가장 좋은 음식이야. 그리고 너는 음식을 먹기 전에 손을 씻어야만 해. 그러면 네 눈이 건강해질 거야."
1:21 Footnote: An organization named World Vision owns this story.	1:21 각주: 선명회라는 기관이 이 이야기를 소유하고 있습니다.

Table A-3. TTA's English and Korean Drafts of "Avian Influenza"

1:1 One day a doctor named Paulus returned from the market to his village named Terpen. While Paulus had been at the market, some people had told him	1:1 어느 날 팔러스라는 의사가 시장에서 터펜이라는 자기 마을로 돌아왔다. 팔러스가 시장에 있는 동안 사람들이 팔러스에게 어떤 병에 대해서 말하였다.

Table A-3 - continued

about a certain disease. So when Paulus returned to his village, he said to Isak, who was the village chief, and the other people who lived in Terpen, "A new disease named Avian Influenza has killed most of the birds that are at the market. This disease has killed many chickens and many ducks.	그래서 팔러스는 자기 마을로 돌아왔을 때 마을 이장인 아이작과 터펜에 사는 다른 사람들에게 말하였다. "조류 인플루엔자라는 새 병이 시장에 있는 대부분 새들을 죽였습니다. 이 병은 닭들과 오리들을 많이 죽였습니다.
1:2 Many people who own chickens and ducks are very sick. Those people are at the clinic. This disease is in many countries and has killed many people.	1:2 닭들과 오리들을 소유하고 있는 많은 사람들이 매우 아픕니다. 이 사람들은 진료소에 있습니다. 이 병은 많은 나라에 있고 사람들을 많이 죽였습니다.
1:3 Guards are watching the farms that have many chickens. Veterinarians who know about this disease go to the market each day. These veterinarians examine the birds. If a bird catches this disease, these veterinarians put it in a special cage. Then these veterinarians give a vaccine to that bird. These veterinarians burn birds that die because of this disease.	1:3 보초병들이 많은 닭들이 있는 농장을 감시하고 있습니다. 이 병에 대해서 알고 있는 수의사들이 매일 시장으로 가서 새들을 검사합니다. 만일 새가 이 병에 걸리면 그 수의사들은 특별한 새장 안에 새를 넣고 그 새에게 백신을 놓아 줍니다. 그리고 그 수의사들은 이 병 때문에 죽은 새들을 태웁니다.
1:4 People stopped watching roosters fight other roosters because some roosters have died. When roosters fight other roosters, germs move from the sick rooster to the other rooster."	1:4 수탉들이 죽었기 때문에 사람들이 수탉들이 다른 수탉들과 싸우는 것을 구경하는 것을 멈추었습니다. 수탉들이 다른 수탉들과 싸울 때 세균이 아픈 수탉에서 다른 수탉으로 옮겨갑니다."
1:5 Chief Isak asked Paulus, "How does this disease spread?" Paulus answered, "When people touch birds that have this disease, they catch it. And when animals touch birds that have this disease, they catch it.	1:5 아이작 이장은 팔러스에게 "이 병은 어떻게 퍼집니까?" 라고 물었다. 팔러스는 대답하였다. "사람들이 이 병이 있는 새들을 만질 때 이 병에 걸립니다. 그리고 동물들이 이 병이 있는 새들에 접촉할 때 이 병에 걸립니다.
1:6 People, animals, and birds carry this disease from one place to another place. The germs travel through the air. This disease causes people, animals, and birds to become very sick. When people catch this disease, they think that they have a cold. But this disease kills people quickly."	1:6 사람들과 동물들과 새들이 한 곳에서 다른 곳으로 이 병을 옮깁니다. 세균은 공기를 통해 퍼집니다. 이 병은 사람들과 동물들과 새들을 매우 아프게 되게 합니다. 사람들은 이 병에 걸릴 때 자기들이 감기가 있다라고 생각합니다. 그러나 이 병은 사람들을 빨리

Table A-3 - continued

	죽입니다."
1:7 The people who lived in Terpen started worrying about their families and their chickens. So Chief Isak asked all the people to come to a meeting. Then the people started talking to each other about this disease. The people wanted to prevent the disease from spreading. The people didn't want their families or their chickens to catch the disease. So the people decided to protect the village from the disease. The people decided to do these things.	1:7 터펜에 사는 사람들은 자기들 가족과 자기들 닭들에 대해서 걱정하기 시작하였다. 그래서 아이작 이장은 모든 사람들이 회의에 와 주기를 부탁하였다. 그리고서 사람들은 서로 이 병에 대해서 말하기 시작하였고 병이 퍼지는 것을 막고 싶어하였다. 그리고 사람들은 자기들 가족과 사람들의 닭들이 병에 걸리기를 원하지 않았다. 그래서 사람들은 병으로부터 마을을 보호하기로 결정하였고 이 일들을 하기로 결정하였다.
1:8 1) The people will build fences around their chickens. The people will give clean food and clean water to their chickens. The people will also clean the chickens' cages. 2) The people will carry their chickens to the market in bamboo cages. After the people sell the chickens at the market, they will burn the cages at the market. 3) Before the people return to the village, they will wash themselves with soap in the river.	1:8 첫번째, 사람들은 자기들 닭들 주변에 울타리를 만들 것이고 자기들 닭들에게 깨끗한 음식과 깨끗한 물을 줄 것이고 또한 닭장을 청소할 것이다. 두번째, 사람들은 대나무로 만들어진 새장 안에 자기들 닭들을 넣어서 시장으로 갈 것이다. 사람들은 시장에서 닭들을 팔은 후에 시장에서 새장을 태울 것이다. 세번째, 사람들은 마을로 돌아오기 전에 강에서 비누로 몸을 깨끗이 할 것이다.
1:9 4) The people will tell their children about this disease. The people will also tell their children to chase away other birds that are close to their chickens. 5) People who live in the village won't buy more chickens until Avian Influenza leaves this region.	1:9 네번째, 사람들은 자기들 아이들에게 이 병에 대해서 말할 것이고 또한 자기들 아이들에게 사람들의 닭들 가까이에 있는 다른 새들을 쫓아내라고 말할 것이다. 다섯번째, 조류 인플루엔자가 그 지역을 떠날 때까지 마을에 사는 사람들이 닭들을 더 사지 않을 것이다.
1:10 6) If a person sees a chicken that has this disease, he has to tell Paulus or Isak about it. 7) People who live in other villages must not bring chickens, eggs, or manure to Terpen.	1:10 여섯번째, 만일 누군가가 이 병이 있는 닭을 보면 이 사람은 팔러스 또는 아이작에게 그 닭에 대해서 말해야만 한다. 일곱번째, 다른 마을에 사는 사람들이 터펜으로 닭들 또는 달걀 또는 똥을 가져오지 말아야만 한다.
1:11 8) When people who work for the	1:11 여덟번째, 정부기관에서 일하는

Table A-3 - continued

government examine the people's chickens, they will give vaccine to them.	사람들이 사람들의 닭들을 검사할 때 사람들은 아픈 닭들에게 백신을 놓아 줄 것이다.
1:12 9) When people sell their chickens and eggs, they have to save one tenth of the money that they receive to buy vaccine and more chickens.	1:12 아홉번째, 사람들은 자기들 닭들과 달걀을 팔 때 백신과 닭들을 더 사기 위해 자기들이 받은 돈의 1/10 을 저축해야만 한다.
2:1 That evening a man named Nano saw that his rooster was sick. The rooster's comb was big and blue. Pus was coming from the rooster's eyes and his beak. And the rooster's foot was red. The rooster also was not able to crow.	2:1 그 날 저녁에 나노라는 남자가 자기 집 수탉이 아픈 것을 알았다. 수탉의 볏은 크고 파랬다. 수탉의 눈과 부리에서 고름이 나오고 있었다. 그리고 수탉의 발은 빨갰다. 또한 수탉은 울 수 없었다.
2:2 Nano went to Paulus' house. Then he said to Paulus, "Come to my pen to look at my rooster. My rooster is sick!" So Paulus put a cloth on his face so that he would not catch the disease. And he also put gloves on his hands so that he would not catch the disease. Then Nano also put a cloth on his face so that he would not catch the disease. And he also put gloves on his hands so that he would not catch the disease.	2:2 나노는 팔러스의 집으로 가서 팔러스에게 말하였다. "우리 집 수탉을 보러 내 우리에 오라. 우리 집 수탉은 아프다!" 그래서 팔러스는 병에 걸리지 않도록 천으로 자기 얼굴을 가렸고 또한 병에 걸리지 않도록 손에 장갑을 꼈다. 그러자 나노도 병에 걸리지 않도록 천으로 자기 얼굴을 가렸고 또한 병에 걸리지 않도록 손에 장갑을 꼈다.
2:3 After Paulus examined Nano's rooster, he said, "I think that this rooster has Avian Influenza." Then he examined all of Nano's chickens. After Paulus examined a chicken, he washed his hands with soap so that he would not carry the disease to other chickens.	2:3 팔러스는 나노의 집 수탉을 검사한 후에 "나는 이 수탉이 조류 인플루엔자가 있다라고 생각한다" 라고 말하였다. 그리고서 팔러스는 나노의 모든 닭들을 검사하였고 닭을 검사한 후에 다른 닭들에게 병을 옮기지 않도록 비누로 손을 씻었다.
2:4 Paulus put in a special cage the rooster that was sick. And Paulus also put three chickens in a basket because he thought that they had Avian Influenza. Then Nano and Paulus put all the healthy chickens in the pen.	2:4 팔러스는 특별한 새장 안에 아픈 수탉을 넣었고 또한 닭 3 마리가 조류 인플루엔자가 있다라고 생각하였기 때문에 바구니 안에 닭 3 마리를 넣었다. 그리고서 나노와 팔러스는 우리 안에 모든 건강한 닭들을 넣었다.
2:5 Then Paulus and Nano went to the stream. Paulus washed his hands with soap thoroughly. And Nano also	2:5 그리고서 팔러스와 나노는 시냇가로 갔다. 팔러스는 비누로 손을 철저히

Table A-3 - continued

washed his hands with soap thoroughly. Then Paulus went to his house. And Nano also went to his house.	씻었다. 그리고 나노도 비누로 손을 철저히 씻었다. 그리고서 팔러스는 자기 집으로 갔다. 그리고 나노도 자기 집으로 갔다.
2:6 The next morning Nano's rooster was dead. So Paulus and Nano put the rooster in a plastic bag. Then Nano burned the cage where the rooster had been. The three chickens that were sick still were living. But those chickens were very sick.	2:6 그 다음 날 아침에 나노의 집 수탉은 죽었다. 그래서 팔러스와 나노는 플라스틱 가방 안에 수탉을 넣었다. 그리고서 나노는 수탉이 있었던 새장을 태웠다. 아픈 닭 3 마리는 여전히 살아 있었다. 그러나 그 닭들은 매우 아팠다.
2:7 The other people who were living in the village heard about Nano's rooster and his chickens. So those people put their chickens in pens. Those people didn't want their chickens to be close to the three chickens that were sick.	2:7 마을에 살고 있는 다른 사람들은 나노의 집 수탉과 닭들에 대해서 들었다. 그래서 이 사람들은 우리 안에 자기들 닭들을 넣었고 자기들 닭들이 아픈 닭 3 마리 가까이에 있기를 원하지 않았다.
2:8 Paulus and Nano took to town the rooster that died. A veterinarian named Agus worked at the clinic. Paulus and Nano showed that rooster to Agus. After Agus examined that rooster, he said, "I think that this rooster died because of Avian Influenza. I'll burn this rooster."	2:8 팔러스와 나노는 도시로 죽은 수탉을 가지고 갔다. 애거스라는 수의사가 진료소에서 일하였다. 팔러스와 나노는 애거스에게 그 수탉을 보여주었다. 애거스는 그 수탉을 검사한 후에 말하였다. "나는 이 수탉이 조류 인플루엔자 때문에 죽었다라고 생각해요. 나는 이 수탉을 태울 것이요."
2:9 Then Agus said, "You have to do these things. Take these gloves and this jar of bleach to your village. Clean all of your pens with this bleach. Tomorrow I'll go to your village and will examine your pens. And I will give vaccine to all the chickens that are healthy."	2:9 그리고서 애거스는 말하였다. "아저씨는 이 일들을 해야만 해요. 아저씨의 마을로 이 장갑과 이 표백제 주전자를 가지고 가고 이 표백제로 아저씨의 모든 우리를 청소하세요. 나는 내일 아저씨의 마을로 가서 아저씨의 우리를 검사할 것이고 건강한 모든 닭들에게 백신을 놓아 줄 것이요."
2:10 Then Paulus and Nano returned to the village. Before Paulus and Nano entered the village, they went to the stream. Paulus washed his hands with soap thoroughly. And Nano also washed his hands with soap thoroughly. Paulus and Nano didn't want to carry	2:10 그리고서 팔러스와 나노는 마을로 돌아왔다. 팔러스와 나노는 마을로 들어오기 전에 시냇가로 갔다. 팔러스는 비누로 손을 철저히 씻었다. 그리고 나노도 비누로 손을 철저히 씻었다. 팔러스와 나노는 마을로 세균을 옮기고

Table A-3 - continued

germs to the village.	싶지 않았다.
3:1 Paulus called all the people who lived in Terpen. So all the people came to Nano's house. Then Paulus and Nano told the people about the things that Agus had said. Then half of the people worked with Paulus. And the other people worked with Nano.	3:1 팔러스는 터펜에 사는 모든 사람들을 불렀다. 그래서 모든 사람들은 나노의 집에 왔다. 그리고서 팔러스와 나노는 사람들에게 애거스가 말하였던 것에 대해서 말하였다. 그리고서 사람들의 반은 팔러스와 함께 일하였다. 그리고 다른 사람들은 나노와 함께 일하였다.
3:2 Nano, Chief Isak, and the other people dug a hole that was one and a half meters deep. The hole was about twenty meters from the pen where Nano's chickens lived. Then Nano made a fire in the hole.	3:2 나노와 아이작 이장과 다른 사람들은 1.5 미터 깊이의 구덩이를 팠다. 구덩이는 나노의 닭들이 사는 우리에서 20 미터 쯤에 있었다. 그리고서 나노는 구덩이 속에 불을 피웠다.
3:3 The people put cloths on their faces and put gloves on their hands. Then the people killed all of Nano's chickens and put them in the hole. Then the people burned the chickens. The people also burned all of Nano's cages and buckets. The people burned all the things that Nano's chickens had touched.	3:3 사람들은 천으로 자기들 얼굴을 가렸고 손에 장갑을 꼈고 나노의 모든 닭들을 죽였다. 그리고 사람들은 구덩이 속에 닭들을 넣어서 그것들을 태웠고 또한 나노의 모든 새장과 양동이를 태웠다. 그리고 사람들은 나노의 닭들이 접촉하였던 모든 것을 태웠다.
3:4 The people also burned all the food that Nano had bought for the chickens. And the people burned all the baskets that Nano had carried chickens in. The people also burned the chickens' manure. The people also caught all the chickens that had walked near Nano's house. Then the people killed those chickens and also burned them.	3:4 또한 사람들은 나노가 닭들을 위해 샀던 모든 사료를 태웠고 나노가 닭들을 들고 갔던 모든 바구니를 태웠다. 그리고 또한 사람들은 닭들의 똥을 태웠고 또한 나노의 집 가까이에 갔던 모든 닭들을 잡았다. 그리고서 사람들은 그 닭들을 죽였고 또한 그것들을 태웠다.
3:5 Then the people put some bleach and some water in a bucket. The people cleaned all the things that the chickens had touched. After the people burned all the chickens, they put dirt on the ashes that were in the hole. Then the people filled the hole with dirt.	3:5 그리고서 사람들은 양동이 안에 표백제와 물을 부었고 닭들이 접촉하였던 모든 것을 청소하였다. 그리고 사람들은 모든 닭들을 태운 후에 구덩이 속에 있는 재 위에 흙을 뿌렸고 흙으로 구덩이를 채웠다.
3:6 Nano said, "If more chickens become sick, we have to kill all of them so that this disease could not spread.	3:6 나노는 말하였다. "만일 더 많은 닭들이 아프게 되면 이 병이 퍼질 수 없도록 저희들은 저희들의 모든 닭들을

Table A-3 - continued

Tomorrow Agus will come to examine our chickens."	죽여야만 합니다. 애거스는 저희들의 닭들을 검사하러 내일 올 것입니다."
3:7 Paulus and the people who worked with him put gloves on their hands. And the people put cloths on their faces so that they would not catch the disease. Then the people went to each house that was in the village. Those people caught all the chickens that lived in the village. Then those people examined those chickens.	3:7 팔러스와 그와 함께 일하는 사람들은 손에 장갑을 꼈다. 그리고 사람들은 병에 걸리지 않도록 천으로 자기들 얼굴을 가렸고 마을에 있는 각각 집으로 갔다. 그리고 이 사람들은 마을에 사는 모든 닭들을 잡았고 그것들을 검사하였다.
3:8 After a person examined a chicken, he washed his hands to prevent the disease from spreading. The people finished examining chickens at Nano's house. Paulus and the people who worked with him met Nano and the people who worked with him at his house.	3:8 사람은 닭을 검사한 후에 병이 퍼지는 것을 막기 위해 손을 씻었다. 사람들은 나노의 집에서 닭들을 검사하는 것을 끝냈다. 팔러스와 그와 함께 일하였던 사람들은 나노의 집에서 나노와 그와 함께 일하였던 사람들을 만났다.
3:9 The people saw that the other chickens were not sick. So the people were happy.	3:9 사람들은 다른 닭들이 아프지 않는 것을 알았다. 그래서 사람들은 행복하였다.
3:10 After the people finished examining all the chickens, the people who owned them put bleach and water into buckets. Then those people cleaned their chicken pens. Then all the people went to the stream. The people washed their clothes with soap thoroughly and also washed themselves with soap. The people were very tired. But the people were happy because they had worked hard.	3:10 사람들이 모든 닭들을 검사하는 것을 끝낸 후에 닭들을 소유하고 있는 사람들은 양동이에 표백제와 물을 부었고 자기들 닭 우리를 청소하였다. 그리고서 모든 사람들은 시냇가로 가서 비누로 자기들 옷을 철저히 씻었고 또한 비누로 몸을 깨끗이 하였다. 사람들은 매우 피곤하였다. 그러나 사람들은 열심히 일하였기 때문에 행복하였다.
3:11 The next morning Agus came to the village. He washed his hands with soap. Then he put on special clothes. He put a cloth on his face and put gloves on his hands. Then he started giving vaccine to the chickens.	3:11 그 다음 날 아침에 애거스는 마을에 와서 비누로 손을 씻었고 특별한 옷을 입었다. 그리고 애거스는 천으로 자기 얼굴을 가렸고 손에 장갑을 꼈고 닭들에게 백신을 놓아 주기 시작하였다.
3:12 After Agus gave vaccine to all the chickens, he said to the people, "Yesterday you cleaned your pens well.	3:12 애거스는 모든 닭들에게 백신을 놓아 준 후에 사람들에게 말하였다.

Table A-3 - continued

Your chickens are very healthy. But your chickens have to stay in the pens. Please examine your chickens each day. If you see a chicken that is sick, put it in a special cage. Then please immediately tell me about that chicken. If you see a person who is sick, please tell me about that person. If you see an animal that is sick, please tell me about that animal.	"여러분은 어제 여러분의 우리를 잘 청소하셨습니다. 여러분의 닭들은 매우 건강합니다. 그러나 여러분의 닭들은 우리 안에 있어야만 합니다. 매일 여러분의 닭들을 검사하여 주세요. 만일 여러분이 아픈 닭을 보시면 특별한 새장 안에 그 닭을 넣으시고 저에게 그 닭에 대해서 즉시 말하여 주세요. 만일 여러분이 아픈 사람을 보시면 저에게 이 사람에 대해서 말하여 주세요. 만일 여러분이 아픈 동물을 보시면 저에게 이 동물에 대해서 말하여 주세요.
3:13 Don't allow other people to bring to your village chickens that are sick. And don't allow other people to bring to your village animals that are sick."	3:13 다른 사람들이 여러분의 마을로 아픈 닭들을 가져오도록 하지 말고 다른 사람들이 여러분의 마을로 아픈 동물들을 가져오도록 하지 마세요."
4:1 Two days later Agus returned to the village. Before Agus started examining the chickens, he put on special clothes. And he put gloves on his hands. Then he drew blood from each chicken. He also examined each chicken's beak. Then he took the blood to the clinic so that other people could examine it.	4:1 이 틀 후에 애거스는 마을로 돌아왔고 닭들을 검사하기 시작하기 전에 특별한 옷을 입었다. 그리고 애거스는 손에 장갑을 꼈고 각각 닭에서 피를 뽑았고 또한 각각 닭의 부리를 검사하였다. 그리고서 다른 사람들이 피를 검사할 수 있도록 애거스는 진료소로 피를 가지고 갔다.
4:2 After the people finished examining the blood, Agus wrote a letter to Chief Isak. Agus wrote, "Nano's rooster had Avian Influenza. Nano's rooster died because of Avian Influenza." Then Agus sent the letter to Chief Isak.	4:2 사람들이 피를 검사하는 것을 끝낸 후에 애거스는 아이작 이장에게 편지를 썼다. "나노의 집 수탉은 조류 인플루엔자가 있었다. 나노의 집 수탉은 조류 인플루엔자 때문에 죽었다." 그리고서 애거스는 아이작 이장에게 편지를 보냈다.
4:3 A few days later Nano became sick. He thought that he had a cold. So Paulus immediately examined Nano. Nano didn't go to the clinic because he didn't have a fever. But he stayed in a house alone so that other people would not catch the disease. Paulus examined Nano each day.	4:3 며칠 후에 나노는 아프게 되었고 자기가 감기가 있다라고 생각하였다. 그래서 팔러스는 나노를 즉시 검사하였다. 나노는 열이 나지 않았기 때문에 진료소로 가지 않았다. 그러나 다른 사람들이 병에 걸리지 않도록 나노는 집에 혼자서 머물렀다. 팔러스는

Table A-3 - continued

	매일 나노를 검사하였다.
4:4 Nano had a daughter named Nina. Nina was two years old. One morning she became sick. She had a fever and was breathing quickly. And Nina's nose was running. Nano knew that Nina had touched the three chickens that were sick. So Nano, his wife, and Paulus took Nina to the clinic because they thought that she might have Avian Influenza.	4:4 나노는 니나라는 딸이 있었다. 니나는 두 살이었다. 어느 날 아침에 니나는 아프게 되었고 열이 있었고 숨을 빨리 쉬고 있었다. 그리고 니나는 콧물을 흘리고 있었다. 나노는 니나가 아픈 닭 3 마리를 만졌다는 것을 알았다. 그래서 나노와 그의 아내와 팔러스는 니나가 아마도 조류 인플루엔자가 있을 것이다라고 생각하였기 때문에 진료소로 니나를 데리고 갔다.
4:5 A nurse who worked at the clinic said that Nina has to stay there for ten days. Nina's mother also stayed at the clinic to take care of her. So Paulus and Nano returned to the village. Then Paulus said to Nano, "You and your children have to stay at your house for ten days. If a person becomes sick, immediately call me."	4:5 진료소에서 일하는 간호사가 니나가 10 일 동안 진료소에 머물러야만 한다고 말하였다. 또한 니나의 어머니는 니나를 돌보기 위해 진료소에 머물렀다. 그래서 팔러스와 나노는 마을로 돌아왔다. 그리고서 팔러스는 나노에게 말하였다. "너와 아이들은 10 일 동안 집에 머물러야만 한다. 만일 누군가가 아프게 되면 나를 즉시 불러라."
4:6 The nurses who worked at the clinic put Nina in a special room because they wanted to take care of her. Nina became very sick. But the nurses helped Nina become well. Then Nina's mother brought her to the village.	4:6 진료소에서 일하는 간호사들은 니나를 돌보고 싶어하였기 때문에 특별한 방에 니나를 넣었다. 니나는 매우 아프게 되었다. 그러나 간호사들은 니나가 건강해지는 것을 도와주었다. 그리고서 니나의 어머니는 마을로 니나를 데려왔다.
4:7 The other people who lived in Terpen didn't become sick. And the other chickens that were in Terpen didn't become sick. One month later the people who lived in the village helped Nano clean his chicken pen. Then the people gave one rooster and some chickens to Nano. But the people continued examining all the chickens that were in the village. And the people continued chasing away the birds that ate food with the chickens. And the	4:7 터펜에 사는 다른 사람들은 아프게 되지 않았다. 그리고 터펜에 있는 다른 닭들은 아프게 되지 않았다. 한 달 후에 마을에 사는 사람들은 나노가 자기 닭 우리를 청소하는 것을 도와주었고 나노에게 수탉 한 마리와 닭들을 주었다. 그러나 사람들은 계속 마을에 있는 모든 닭들을 검사하였고 계속 닭들과 함께 사료를 먹는 새들을 쫓아냈다. 그리고 사람들은 다른 사람들이 마을로 아픈 닭들을 가져오는 것을 허락하지 않았다.

Table A-3 - continued

people didn't allow other people to bring to the village chickens that were sick.	
5:1 You must protect your chickens and your animals from this disease. You must work with the other people who live in your village. You and the other people who live in your village must learn about this disease. If you prevent this disease from spreading, your animals will be healthy.	5:1 여러분은 이 병으로부터 여러분의 닭들과 동물들을 보호해야만 하고 여러분의 마을에 사는 다른 사람들과 함께 일해야만 합니다. 여러분과 여러분의 마을에 사는 다른 사람들은 이 병에 대해서 배워야만 합니다. 만일 여러분이 이 병이 퍼지는 것을 막으면 여러분의 동물들은 건강할 것입니다.
5:2 You must do these things to prevent Avian Influenza from killing you and your animals.	5:2 여러분은 조류 인플루엔자가 여러분과 여러분의 동물들을 죽이는 것을 막기 위해 이 일들을 해야만 합니다.
5:3 1) When you buy chickens and ducks at the market, you must be very careful. Chickens and ducks have Avian Influenza often. When you cut the meat, use a special board. You must put only raw meat on that board. You must not put on that board meat that you cooked. After you cook the meat, wash your hands with soap thoroughly.	5:3 첫번째, 여러분은 시장에서 닭들과 오리들을 살 때 매우 조심해야만 합니다. 닭들과 오리들이 자주 조류 인플루엔자가 있습니다. 고기를 자를 때 특별한 도마를 쓰세요. 여러분은 그 도마 위에 오직 날고기만 놓아야만 하고 그 도마 위에 여러분이 요리하였던 고기를 놓지 말아야만 합니다. 고기를 요리한 후에 비누로 손을 철저히 씻으세요.
5:4 2) When you buy eggs at the market, you also must be careful. Before you boil the eggs, wash them thoroughly. After you touch the eggs, wash your hands with soap.	5:4 두번째, 또한 여러분은 시장에서 달걀을 살 때 조심해야만 합니다. 달걀을 끓이기 전에 달걀을 철저히 닦고 달걀을 만진 후에 비누로 손을 씻으세요.
5:5 3) You must cook meat very well. Before you eat the meat, examine it. If you see blood, you must cook the meat more. If you eat blood, you might become sick. 4) After you cook the meat, put it on a clean plate. You must wash the plate that the raw meat touches. Then you must dry the plate.	5:5 세번째, 여러분은 고기를 매우 잘 요리해야만 합니다. 고기를 먹기 전에 고기를 검사하세요. 만일 여러분이 피를 보면 여러분은 고기를 더 요리해야만 합니다. 만일 여러분이 피를 먹으면 여러분은 아마도 아프게 될 것입니다. 네번째, 고기를 요리한 후에 깨끗한 접시 위에 고기를 놓으세요. 여러분은 날고기가 닿은 접시를 닦어야만 하고 그것을 말려야만 합니다.

Table A-3 - continued

5:6 5) You must wash all the things that touch the raw meat. You must wash these things with soap. You must wash the knives, plates, and pots with soap. And before you use these things again, you must dry them.	5:6 다섯번째, 여러분은 날고기에 닿은 모든 것을 닦어야만 하고 비누로 그것들을 닦어야만 하고 비누로 칼과 접시와 냄비를 닦어야만 합니다. 그리고 여러분은 다시 이 것을 쓰기 전에 이 것을 말려야만 합니다.
5:7 6) If your chickens are sick, you must not sell them at the market. If your chickens are sick, you must not sell the eggs at the market. If you see a person selling chickens that are sick, don't buy them.	5:7 여섯번째, 만일 여러분의 닭들이 아프면 여러분은 시장에서 닭들을 팔지 말아야만 합니다. 만일 여러분의 닭들이 아프면 여러분은 시장에서 달걀을 팔지 말아야만 합니다. 만일 여러분이 어떤 사람이 아픈 닭들을 팔고 있는 것을 보면 그 닭들을 사지 마세요.
5:8 7) If Avian Influenza is in your village, don't go to meetings. If people who are sick go to those meetings, you might catch this disease from them.	5:8 일곱번째, 만일 조류 인플루엔자가 여러분의 마을에 있으면 회의에 가지 마세요. 만일 아픈 사람들이 그 회의에 가면 여러분은 아마도 이 사람들에게서 이 병을 옮을 것입니다.
5:9 8) If you think that you have a cold, you immediately must go to the clinic. 9) Whenever you cough, and whenever you sneeze, you should cover your mouth with your hand. 10) If a person has a cold, his family must stay at his house for ten days to prevent the disease from spreading.	5:9 어덟번째, 만일 여러분이 감기가 있다라고 생각하면 여러분은 진료소로 즉시 가야만 합니다. 아홉번째, 여러분은 기침을 할 때마다 그리고 재채기할 때마다 손으로 여러분의 입을 가려야 합니다. 열번째, 만일 누군가가 감기가 있으면 자기 가족은 병이 퍼지는 것을 막기 위해 10 일 동안 자기 집에 머물러야만 합니다.
5:10 11) Whenever you help a person who has Avian Influenza, you should put gloves on your hands. After you finish helping that person, you should wash your hands with soap thoroughly.	5:10 열한번째, 여러분은 조류 인플루엔자가 있는 사람을 도와줄 때마다 손에 장갑을 끼어야 하고 이 사람을 도와주는 것을 끝낸 후에 비누로 손을 철저히 씻어야 합니다.
5:11 Footnote: An organization named The Summer Institute of Linguistics owns this story.	5:11 각주: 서머 인스티트 오브 링귀스틱이라는 기관이 이 이야기를 소유하고 있습니다.

The following text is taken from page 50 of a sixth grade Korean textbook entitled 국어 읽기 6-2 ("Language Reading Sixth Grade, Second Semester"). The sentences from this text were typed into Google-Translate and Yahoo's Babel Fish. Their translations of this text are shown below, along with a Korean speaker's translation. This short sample illustrates that both the Google and Yahoo translations are incomprehensible.

---------- Title --

백 번째 손님

Human: The 100th Guest

Google: 100th Customer

Babel Fish: Hundredth caller

---------- Sentence 1 ---

국밥집 주인 강씨 아저씨는 손님을 기다리며 신문을 뒤적였다.

Human: Mr. Kang, who is the owner of a rice soup restaurant, is waiting for guests skimming through a newspaper.

Google: Waiting for Mr. Kang, the owner gukbapjip guests dwijeokyeotda newspaper.

Babel Fish: The rice served in soup house master river uncle waited the customer and the newspaper was after.

---------- Sentence 2 ---

점심을 먹으러 온 직장인 손님들이 한차례 지나간 뒤였다.

Human: It is the time after the first round of businessmen who came for lunch had left.

Google: Guest workers come for lunch was well past the rainstorm.

Babel Fish: Have the lunch the worker callers one time after passing by, were.

---------- Sentence 3 ---

그 때, 문이 열리며 머리카락이 허연 할머니가 들어섰다.

Human: At that time the door opened and a grandmother with gray hair entered.

Google: Then, the door opened and she entered heoyeon hair.

Babel Fish: That time, the door was opened and the grandmother where the hair is very white entered.

---------- Sentence 4 ---

그 뒤에 열 살도 채 안 뒤어 보이는 소년이 마치 꼬리를 잡듯 할머니의 한 손을 꼭 잡고 들어왔다.

Human: After her, a boy, who didn't look like he's even ten years old, entered, holding his grandmother's hand tightly like one holds an animal's tail.

Google: After that, less than 10 years old dwieo Give her tail, she looks like the boy a hand in the inside.

Babel Fish: The boy where also heat ray is visible [e] after inside not yet caught the tail as after that and to certainly catch one hand of the grandmother came in.

---------- Sentence 5 ---

옷차림이 남루하고 궁기가 흘렀다.

Human: Their clothes looked shabby and showed that they were poor.

Google: Gunggi the ragged clothes and a passed.

Babel Fish: The attire was tattered and the wretchedness flowed.

---------- Sentence 6 ---

“이 쪽으로 앉으세요.”

Human: “Come this way and sit down please.”

Google: "This way, sit down."

Babel Fish: “With sits side.”

---------- Sentence 7 ---

강씨 아저씨는 가운데 자리에 있는 식탁 의자를 빼며 자리를 권하였다.

Human: Mr. Kang recommended a place while he pulled out a chair that was in the middle of the restaurant.

Google: Mr. Kang, in the middle of the table where they did ppaemyeo chair.

Babel Fish: The river uncle the place drew out the dining table chair which is to the middle place and recommended.

---------- Sentence 8 ---

“저, 국밥 한 그릇에 얼마나 하는지?”

Human: “Um, how much is one bowl of rice soup?”

Google: "Well, you know how much of a bowl of soup?"

Babel Fish: “Well, how many do in the rice served in soup one vessel?”

---------- Sentence 9 ---

“사천 원 받습니다.” 강씨 아저씨는 사람 좋은 웃음을 얼굴 가득 담으며 말하였다.

Human: “We charge four thousand won,” Mr. Kang said, with a smile on his face that showed he’s a good person.

Google: "Receive 4,000 won." Mr. Kang contain affecting a person full of good laughs and said his face.

Babel Fish: “Szuchuan circle receives.” The river uncle the person the face put in a good laugh full and said.

---------- Sentence 10 --

할머니는 조금 몸을 돌려 허리춤에서 주머니를 꺼내, 그 안에 든 동전까지 조몰락거리며 헤아렸다.

Human: The grandmother turned her body and took out her pouch that was in her waist pocket, and she counted all the coins inside.

Google: She turned a little out of pocket in his belt, holding a coin in it until the heahryeotda jomolrakgeorimyeo.

Babel Fish: The grandmother turned the body a little and pulled out until the coin which inside that holds trillion ruin distances counted the pocket from the waist-tops.

---------- Sentence 11 --
그러고 나서야 그 자리에 소년을 앉혔다. “한 그릇만 주세요.”
Human: After that, she put her boy in that chair and said, “Please give us only one bowl.”
Google: Then, and only sat in place for the boy. "A BATCH, please."
Babel Fish: The [le] came out and seated the boy in that place. “Only one vessel give.”
---------- Sentence 12 --
“네?”
Human: “What?”
Google: "Yes?"
Babel Fish: “Four?”
---------- Sentence 13 --
“난 점심을 이미 먹었다요.”
Human: “I have already eaten.”
Google: "I already ate lunch, drink it."
Babel Fish: “Already has the difficulty lunch.”
---------- Sentence 14 --
“아, 네. 맛있게 말아 드리겠습니다.”
Human: “Ah, yes. I will give you one delicious bowl.”
Google: "Oh, yes. Enjoy, I'll do. "
Babel Fish: “Oh, four. Rolls up tastefully.”

APPENDIX B

EXAMPLE OF A KOREAN QUESTIONNAIRE

아래에 영어에서 한글로 번역된 두 편의 짧은 글이 있습니다. 두 글을 읽고 아래 질문에 답하여 주세요.

보아스가 일꾼들을 감독하는 자기 종에게 물었습니다. “저 여자는 어느 집 여자인가?”

그 종이 대답했습니다. “저 여자는 나오미와 함께 모압 지방에서 온 모압 여자입니다.

일꾼들 뒤를 따라다니며 땅에 떨어진 이삭을 줍도록 해 달라고 했습니다.

그녀는 잠시 오두막에서 쉰 것 말고는 아침부터 지금까지 계속 이삭을 줍고 있습니다.”

보아스가 룻에게 말했습니다. “여인이여, 나의 말을 잘 들으시오. 이삭을 줍기 위해 다른 밭으로 가지 말고 여기에서 주우시오. 내 일꾼들 뒤만 따라다니시오. 그들이 가는 밭을 잘 보고 그 뒤를 따라가시오. 나의 일꾼들에게 당신을 건드리지 말라고 일러두었소. 목이 마르거든 물항아리 있는 곳으로 가서 일꾼들이 길어 온 물을 마시도록 하시오.”

보아스가 룻을 봤을 때 하인들의 감독에게 "저 여자는 누구냐?" 라고 말하였다. 감독은 보아스에게 대답하였다. "저 여자는 나오미와 함께 모압에서 왔습니다. 저 여자는 저에게 "제가 일꾼들 뒤에서 보리를 주어도 됩니까?" 라고 말하였습니다. 그래서 저는 저 여자에게 "예" 라고 말하였습니다. 그러자 저 여자는 아침에 열심히 일하였고 잠시 동안 천막 안에서 쉬었습니다."

그리고서 보아스는 룻에게 걸어갔고 룻에게 말하였다. "내 딸이여, 다른 남자의 밭에서 보리를 줍지 마세요. 만일 당신이 다른 남자의 밭에서 보리를 주으면 그 남자는 아마도 당신에게 친절하지 않을 것이요. 내 밭에 머무르고 이 밭에서 일하고 있는 여자들 뒤에서 보리를 주으세요.

이 여자들을 따라가세요. 나는 젊은 남자들에게 당신을 괴롭히지 않도록 명령하였어요. 당신의 목이 마를 때 당신은 저 물통 안에 있는 물을 마셔도 됩니다. 하인들이 우물 안에 있었던 물로 저 물통을 채웠어요."

위의 두 글을 읽고 다음 중 하나에 동그라미하세요.

A. 이 이야기를 잘 알지 못하는 6 학년 학생이 이 두 글을 읽을 때 첫번째 글이 두번째 글보다 훨씬 더 낫습니다.

B. 이 이야기를 잘 알지 못하는 6 학년 학생이 이 두 글을 읽을 때 두번째 글이 첫번째 글보다 훨씬 더 낫습니다.

C. 이 이야기를 잘 알지 못하는 6 학년 학생이 이 두 글을 읽을 때 첫번째 글과 두번째 글이 거의 동일합니다.

12-17-7A
2:5-9 EB TTA

APPENDIX C

THE FEATURE SYSTEM DEVELOPED FOR THE TRANSLATOR’S ASSISTANT

Table C-1. Object Features

Number	Singular, Dual, Trial, Quadrial, Plural, Paucal
Participant Tracking	First Mention, Integration, Routine, Exiting, Offstage, Restaging, Generic, Interrogative, Frame Inferable
Polarity	Affirmative, Negative
Proximity	Near Speaker and Listener, Near Speaker, Near Listener, Remote within sight, Remote out of sight, Temporally Near, Temporally Remote, Contextually Near with Focus, Contextually Near, Not Applicable
Person	First, Second, Third, First & Second, First & Third, Second & Third, First & Second & Third
Surface Realization	Always a Noun, Unambiguous Pronoun, Not Applicable
Participant Status	Protagonist, Antagonist, Major Participant, Minor Participant, Major Prop, Minor Prop, Significant Location, Insignificant Location, Significant Time, Not Applicable

Table C-2. Event Features

Time	Discourse, Present, Immediate Past, Earlier Today, Yesterday, 2 to 3 days ago, 4 to 6 days ago, 1 to 4 weeks ago, 1 to 5 months ago, 6 to 12 months ago, 1 to 9 years ago, 10 to 20 years ago, During Speaker's lifetime, Historic Past, Eternity Past, Unknown Past, Immediate Future, Later Today, Tomorrow, 2 to 3 days from now, 4 to 6 days from now, 1 to 4 weeks from now, 1 to 5 months from now, 6 to 12 months from now, 1 to 9 years from now, 10 to 20 years from now, during speaker's lifetime, Historic future, Eternity future, Unknown Future, Timeless
Aspect	Unmarked, Completive, Inceptive, Cessative, Continuative, Habitual, Gnomic, Imperfective
Mood	Indicative, Definite Potential, Probable Potential, 'might' Potential, Unlikely Potential, Impossible Potential, 'must' Obligation, 'should' Obligation, 'should not' Obligation, Forbidden Obligation, 'may' (permissive)
Reflexivity	Not Applicable, Reflexive, Reciprocal
Polarity	Affirmative, Negative, Emphatic Affirmative, Emphatic Negative

Table C-3. Object Attribute Features

Degree	Comparative, Superlative, Intensified, Extremely Intensified, 'too' or 'overly', 'less', 'least', Not Applicable

Table C-4. Event Attribute Features

Degree	Comparative, Superlative, Intensified, 'too' or 'overly', 'less', 'least', Not Applicable

Table C-5. Object Phrase Features

Sequence	Not in a Sequence, Coordinate, First Coordinate, Last Coordinate
Semantic Role	Most Agent-like, Most Patient-like, State, Source, Destination, Instrument, Beneficiary, Addressee, Not Applicable

Table C-6. Event Phrase Features

Sequence	Not in a Sequence, Coordinate, First Coordinate, Last Coordinate

Table C-7. Object Attribute Phrase Features

Sequence	Not in a Sequence, Coordinate, First Coordinate, Last Coordinate
Usage	Attributive, Predicative

Table C-8. Event Attribute Phrase Features

Sequence	Not in a Sequence, Coordinate, First Coordinate, Last Coordinate

Table C-9. Proposition Features

Type	Independent, Restrictive Thing Modifier, Descriptive Thing Modifier, Event Modifier, Agent, Patient, Attributive Patient, Closing Quotation Frame
Illocutionary Force	Declarative, Imperative, Content Interrogative, Yes-No Interrogative
Topic NP	Most Agent-like, Most Patient-like
Discourse Genre	Narrative-Story, Narrative-Prophecy, Hortatory, Procedural, Expository
Notional Structure Schema	Narrative-Exposition, Narrative-Inciting Incident, Narrative-Developing Conflict, Narrative-Climax, Narrative-Denouement, Narrative-Final Suspense, Narrative-Conclusion, Hortatory-Authority Establishment, Hortatory-Problem or Situation, Hortatory-Issuing of Commands, Hortatory-Motivation, Procedural-Problem or Need, Procedural-Preparatory Procedures, Procedural-Main Procedures, Procedural-Concluding Procedures, Persuasive-Problem or Question, Persuasive-Proposed Solution or Answer, Persuasive-Supporting Argumentation, Persuasive-Appeal, Expository-Problem or Situation, Expository-Solution or Answer, Expository-Supporting Argumentation, Expository-Evaluation of Solutions, Not Applicable
Salience Band	Pivotal Storyline, Primary Storyline, Secondary Storyline, Script Predictable Actions, Backgrounded Actions, Flashback, Setting, Irrealis, Evaluation, Cohesive Material, Not Applicable
Speaker	Not Applicable, Adult Daughter, Adult Son, Angel, Animal, Boy, Brother, Crowd, Daughter, Demon, Disciple, Employee, Employer, Father, Girl, God, Government Leader, Government Official, Group of Friends, Holy Spirit, Husband, Jesus, King, Man, Military Leader, Mother, Prophet, Queen, Religious Leader, Satan, Servant, Sister, Slave, Slave Owner, Soldier, Son, Wife, Woman, Written Material to General Audience (letter, law, etc.)
Listener	Not Applicable, Adult Daughter, Adult Son, Angel, Animal, Boy, Brother, Crowd, Daughter, Demon, Disciple, Employee, Employer, Father, Girl, God, Government Leader, Government Official, Group of Friends, Holy Spirit, Husband, Jesus, King, Man, Military Leader, Mother, Prophet, Queen, Religious Leader, Satan, Servant, Sister, Slave, Slave Owner, Soldier, Son, Wife, Woman
Speaker's Attitude	Not Applicable, Neutral, Familiar, Endearing, Honorable, Derogatory, Friendly, Antagonistic, Complimentary, Anger, Rebuke
Speaker's Age	Not Applicable, Child (0-17), Young Adult (18-24), Adult (25-49), Elder (50+)
Speaker to Listener's Age	Not Applicable, Older - Different Generation, Older - Same Generation, Essentially the Same Age, Younger - Different Generation, Younger - Same Generation
Alternative Analysis	Not Applicable, Primary Analysis, First Alternative Analysis, Second Alternative Analysis, Third Alternative Analysis, Fourth Alternative Analysis, Fifth Alternative Analysis
Implicit Information	Not Applicable, Implicit Cultural Information, Implicit Situational Information, Implicit Historical Information, Implicit Background Information, Implicit Subactions
Sequence	Not in a Sequence, First Coordinate, Last Coordinate, Coordinate
Location in Paragraph	Not Applicable, First, Last, Discourse Title, Aperture, Finis, Footnote

APPENDIX D

LIST OF ABBREVIATIONS

Table D-1. List of Abbreviations

1st	First Person
2nd	Second Person
3rd	Third Person
Aff	Affirmative
Affirm	Affirmative Polarity
Ben	Benefactive Case
Col	Collective
Comp	Complementizer
Decl	Declarative
Def	Deferential
Dl	Dual
DS	Different Subject
Exist	Existential
Fut	Future Tense
Imp	Imperative
Imperf	Imperfective Aspect
IndObj	Indirect Object
Inf	Infinitive
Inst	Instrumental Case
Inter	Interrogative
Loc	Locative Case
Neg	Negative Polarity
Obj	Object
Past	Past Tense
Pl	Plural
Pos	Possessor
Pres	Present Tense
Pst	Past Tense
Rel	Relativizer
RPast	Remote Past
Sg	Singular
Sim	Simultaneous
Sing	Singular
SS	Same Subject
Subj	Subject
TopicCl	Topic Clause

REFERENCES

Allan, Keith. 1986. Linguistic Meaning, 2 vols. London: Routledge & Kegan Paul.

Appelt, Doug. 1985. Planning English Referring Expressions. Cambridge University Press, New York.

Bartsch, Carla. 1995. Is Satan the Main Character in your Translation of the Gospels? Notes on Translation, 9(4):47-50.

Bateman, John. 1997. Sentence Generation and Systemic Grammar: An Introduction. Iwanami Lecture Series: Language Sciences, Volume 8. Tokyo: Iwanami Shoten Publishers.

Bateman, John, and Elke Teich. 1995. Selective information presentation in an integrated publication system: an application of genre-driven text generation. Information Processing and Management 31(5), 753–768.

Bateman, John, Elke Teich, Geert-Jan Kruijff, Ivana Kruijff-Korbayova, Serge Sharoff, and Hana Skoumalova. 2000. Resources for multilingual text generation in three Slavic languages. Proceedings of the 18th Conference on Computational Linguistics. 1:474-480.

Bateman, John, Renate Henschel, and Fabio Rinaldi. 2005. The Generalized Upper Model 2.0. On-line publication at http://www.fb10.uni-bremen.de/anglistik/langpro/webspace/ jb/repository/pdf/gum2.pdf, accessed December 11, 2010.

Bateman, John. 2009. Web Site Title: Bateman/Zock: NLG list system entry: FoG. www.fb10.uni-bremen.de/anglistik/langpro/NLG-table/details/FoG.htm, accessed December 11, 2010.

Bateman, John. 2010a. Web Site Title: NLG Systems Wiki. www.fb10.uni-bremen.de/anglistik/langpro/NLG-table/NLG-table-root.htm, accessed December 11, 2010.

Bateman, John. 2010b. Web Site Title: KPML one-point access page. www.fb10.uni-bremen.de/anglistik/langpro/kpml/README.html, accessed December 13, 2010.

Beale, Stephen, Benoit Lavoie, Marjorie McShane, Sergei Nirenburg, and Tanya Korelsky. 2004. Question Answering Using Ontological Semantics. Proceedings of ACL-2004 Workshop on Text Meaning and Interpretation. Barcelona, Spain. 41-48.

Belz, Anja. 2007. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. Natural Language Engineering, 14(4):431-455.

Biberauer, Theresa, and Ian Roberts. 2008. Subjects, Tense and Verb Movement in Germanic and Romance. In Cambridge Occasional Papers in Linguistics 3, ed. E. Lash, Y. Lin and T. Rainsford, 24–43.

Cann, Ronnie. 1993. Formal Semantics, Cambridge: Cambridge University Press.

Carbonell, Jaime, Teruko Mitamura, and Eric Nyberg. 1992. The KANT Perspective: A Critique of Pure Transfer (and Pure Interlingua, Pure Statistics, …), in Proceedings of TMI-92, Montreal, Canada.

Cho, YoungMee, HyoSang Lee, Carol Schulz, HoMin Sohn, and SungOck Sohn. 2000. Integrated Korean. Honolulu: University of Hawai'i Press.

Chomsky, Noam. 1957. The Logical Structure of Linguistic Theory. Harvard University. New York: Plenum.

Chomsky, Noam. 1970. Remarks on Nominalization. In R. Jacobs and P. Rosenbaum, eds., Readings in Transformational Grammar. Boston: Ginn.

Coch, José, Raphaël David, and Jeannine Magnoler. 1995. Quality tests for a mail generation system. Proceedings of Linguistic Engineering '95. 435-443. Montpellier, France.

Coch, José. 1998. Interactive generation and knowledge administration in MultiMeteo. Proceedings of the Ninth International Workshop on Natural Language Generation. 300-303. Niagara-on-the-lake, Ontario, Canada.

Comrie, Bernard. 1976. Aspect. Cambridge: Cambridge University Press.

Comrie, Bernard. 1985. Tense. Cambridge: Cambridge University Press.

Comrie, Bernard. 1989. Language Universals and Linguistic Typology. 2nd edition. Chicago: The University of Chicago Press.

Dahl, Ősten. 1985. Tense and Aspect Systems. New York: Basil Blackwell.

Davey, Anthony. 1979. Discourse Production: a computer model of some aspects of a speaker. Doctoral Dissertation. Edinburgh University Press, Edinburgh, Scotland.

Dorr, Bonnie, Eduard Hovy, and Lori Levin. 2006. Natural Language Processing and Machine Translation: Interlingual Methods in Encyclopedia of Language and Linguistics, 2nd edition. Cambridge: Cambridge University Press.

Farrar, Scott, William Lewis, and Terence Langendoen. 2002. A Common Ontology for Linguistic Concepts.

Farrar, Scott, and Terry Langendoen. 2003. A linguistic ontology for the semantic web. Glot International Vol. 7, No. 3. 97-100.

Farrar, Scott. 2007. Using 'Ontolinguistics for Language Description' in Ontolinguistics: How Ontological Status shapes the linguistic coding of concepts. Edited by Andrea Schalley and Dietmar Zaefferer. Berlin: Mouton de Gruyter. 175-192.

Fillmore, Charles. 1968. The Case for Case. In Bach and Harms (Ed.): Universals in Linguistic Theory. New York: Holt, Rinehart, and Winston, 1-88.

Fodor, Janet. 1977. Semantics: Theories of Meaning in Generative Grammar. New York: Thomas Y. Crowell.

Foley, William. 1986. The Papuan Languages of New Guinea. Cambridge: Cambridge University Press.

Frantz, Donald. 1974. Generative Semantics: An Introduction. Dallas: Summer Institute of Linguistics.

Gazdar, Gerald, Ewan Klein, Geoffrey Pullum, and Ivan Sag. 1985. Generalized Phrase Structure Grammar. Cambridge: Harvard University Press.

Givón, Talmy. 1990. Syntax: A Functional-Typological Introduction, 2 vols. Amsterdam: John Benjamins.

Goddard, Cliff, and Anna Wierzbicka. 1994. Semantic and Lexical Universals: Theory and Empirical Findings. Amsterdam: John Benjamins.

Goddard, Cliff. 1998. Semantic Analysis. Oxford: Oxford University Press.

Goddard, Cliff. 2008. Cross-linguistic Semantics. Amsterdam: John Benjamins.

Goldberg, Eli, Norbert Driedger, and Richard Kittredge. 1994. Using natural-language processing to produce weather forecasts. IEEE Expert, 9(2):45-53.

Goldman, Neil. 1975. Computer Generation of Natural Language from a Deep Conceptual Base. Doctoral Dissertation. Stanford University.

Haegeman, Liliane. 1994. Government and Binding Theory. Oxford: Blackwell Publishers.

Hana, Jirka. 2001. The AGILE System. PBNL, Praha. 39-67.

Harris, Randy. 1993. The Linguistics Wars. New York: Oxford University Press.

Healy, Phyliss. 1966. Levels and Chaining in Telefol Sentences. Pacific Linguistics, Series B, no. 5.

Humphreys, Kevin, Mike Calcagno, and David Weise. 2001. Reusing a Statistical Language Model for Generation. Proceedings of the 8th European workshop on Natural Language Generation – Volume 8. Toulouse, France. 1-6.

Hüske-Kraus, Dirk. 2008. Suregen-2: a shell system for the generation of clinical documents. Oeynhausen: Department of Hospital Informatics, Heart and Diabetes Center.

Hunter, Jim, Albert Gatt, Francois Portet, Ehud Reiter and Somayajulu Sripada. 2008. Using natural language generation technology to improve information flows in intensive care units. Proceedings of the 5th Conference on Prestigious Applications of Intelligent Systems, PAIS-08.

Hunter, Jim, Ehud Reiter, and Yaji Sripada. n.d.a. Web Site Title: BabyTalk – Generating Textual Summaries of Clinical Temporal Data. www.csd.abdn.ac.uk/research/babytalk, accessed December 11, 2010.

Hunter, Jim, Ehud Reiter, S. G. Somayajulu, and Jin Yu. n.d.b. Web Site Title: SumTime – Generating Summaries of Time Series Data. www.csd.abdn.ac.uk/research/sumtime, accessed December 11, 2010.

Hwang, Shin Ja J. 1990. The relative clause in narrative discourse. Language Research 26:373-400.

Hwang, Shin Ja J. 1996. The grammar and discourse of relative clauses. In Bates Hoffer (ed.), The twenty-second LACUS forum 1996, 144-56. Chapel Hill, NC: Linguistic Association of Canada and the United States.

Jackendoff, Ray. 1990. Semantic Structures. Cambridge, Massachusetts: The MIT Press.

Jackendoff, Ray. 2007. Language, Consciousness, Culture. Cambridge, Massachusetts: The MIT Press.

Keenan, Edward, and Bernard Comrie. 1977. Noun Phrase Accessibility and Universal Grammar. Linguistic Inquiry, Vol. 8, No. 1. pp. 3-99. The MIT Press.

Lakoff, George. 1987. Women, Fire, and Dangerous Things. Chicago: University of Chicago Press.

Langacker, Ronald. 1986. Foundations of Cognitive Grammar. Vol 1. Stanford: Stanford University Press.

Larson, Mildred. 1984. Meaning Based Translation. Oxford: University Press of America.

Lasnik, Howard. 1995. Verbal Morphology: Syntactic Structures meets the Minimalist Program in Evolution and Revolution in Linguistic Theory, edited by Héctor Campos and Paula Kempchinsky. Washington D.C.: Georgetown University Press. pp. 251-275.

Leavitt, John, Deryle Lonsdale, and Alexander Franz. 1994. A Reasoned Interlingua for Knowledge-Based Machine Translation. Proceedings of CSCSI-94, Banff, Canada.

Levi, Judith. 1978. The Syntax and Semantics of Complex Nominals. New York: Academic Press.

Levin, Beth. 1993. English Verb Classes and Alternations. Chicago: University of Chicago Press.

Levine, John, and Chris Mellish. 1994. CORECT: combining CSCW with natural language generation for collaborative requirements capture. Proceedings of the 7th. InternationalWorkshop on Natural Language generation (INLGW '94). Kennebunkport, Maine. pp. 236–239.

Lewis, Paul (ed.). 2009. Ethnologue: Languages of the World. 16th ed. Dallas: SIL International. Online version: **http://www.ethnologue.com/** accessed December 11, 2010.

Li, Tangqiu, Eric Nyberg, and Jaime Carbonell. 1996. Chinese Sentence Generation in a Knowledge-Base Machine Translation System, in 'Technical Report CMU-CMT-96-148'.

Pittsburgh, Pennsylvania.

Longacre, Robert. 1972. Hierarchy and Universality of Discourse Constituents in New Guinea Languages. Vol. 1. Georgetown University Press.

Longacre, Robert. 1995. Some interlocking concerns which govern participant reference in Discourse. Language Research, 31:697-714.

Longacre, Robert. 1996. The Grammar of Discourse. 2nd ed. New York: Plenum Press.

Longacre, Robert. 2007. Sentences as Combinations of Clauses. In Timothy Shopen (ed.), Language Typology and Syntactic Description, vol. 2: Complex Constructions. pp. 372-420. Cambridge: Cambridge University Press.

Lonsdale, Deryle, Alexander Franz, and John Leavitt. 1994. Large-Scale Machine Translation: An Interlingua Approach. Proceedings of IEAAIE-94. Ottawa, Canada.

McCarthy, Joy. 1965. Clause Chaining in Kanite. Anthropological Linguists, Vol. 7, No. 5, Part II, pp. 59-70.

McKeown, Kathleen. 1985. Text Generation. Cambridge University Press, Cambridge.

Mitamura, Teruko, Eric Nyberg, and Jaime Carbonell. 1991. An Efficient Interlingua Translation System for Multi-lingual Document Production. Proceedings of Machine Translation Summit III. Washington DC.

Mitamura, Teruko, Eric Nyberg, Kathy Baker, David Svoboda, Enrique Torrejon, and Michael Duggan. 2001. The KANTOO MT System: Controlled Language Checker and Knowledge Maintenance Tool, in ‘Proceedings of NAACL 2001’, Pittsburgh, PA.

Mitamura, Teruko, and Eric Nyberg. 2001. Automatic Rewriting for Controlled Language Translation. ‘Proceedings of the NLPRS 2002 Workshop on Automatic Paraphrasing: Theories and Applications’. Tokyo.

Montague, Richard. 1974. Formal Philosophy. Richmond Thomason ed. London. Yale University Press.

Montague, Richard. 2002. "The Proper Treatment of Quantification in Ordinary English", reprinted in Formal Semantics: The Essential Readings, by Paul Portner, Barbara H. Partee, eds. Blackwell.

Nirenburg, Sergei, and Victor Raskin. 2004. Ontological Semantics. Cambridge, Massachusetts: The MIT Press.

Noonan, Michael. 2007. Complementation. Language Typology and syntactic description: Complex constructions, Vol. 2, ed. by Timothy Shopen, 42-140. Cambridge: Cambridge University Press.

Noy, Natalya and Carole Hafner. 1997. The State of the Art in Ontology Design: A Survey and Comparative Review. AI Magazine Volume 18, Number 3. 53-74.

Nyberg, Eric, and Teruko Mitamura. 1992. The KANT System: Fast, Accurate, High-Quality

Translation in Practical Domains. Proceedings of COLING-92. Nantes, France.

Nyberg, Eric, Teruko Mitamura, and Jaime Carbonell. 1997. The KANT Machine Translation System: From R&D to Initial Deployment, in 'Proceedings of LISA Workshop in Integrating Advanced Translation Technology', Washington DC.

Nyberg, Eric, and Teruko Mitamura. 2000. The KANTOO Machine Translation Environment. Proceddings of AMTA 2000. Cuernavaca, Mexico.

Nyberg, Eric, Teruko Mitamura, Kathryn Baker, David Svoboda, Brian Peterson, and Jennifer Williams. 2002. Deriving Semantic Knowledge from Descriptive Texts using an MT System. Proceedings of AMTA 2002. Tiburon, California.

Nyberg, Eric. 2004. Web Site Title: The KANT Project Home Page. www.lti.cs.cmu.edu/Research/Kant, accessed December 13, 2010.

Paris, Cécile, Keith Vander Linden, Markus Fischer, Anthony Hartley, Lyn Pemberton, Richard Power, and Donia Scott. 1995. A support tool for writing multilingual instructions. Proceedings of International Joint Conference on Artificial Intelligence. Montreal, Canada. 1398–1404.

Partee, Barbara. 2001. Montague Grammar. In: Smelser, Neil and Paul Baltes (Ed.), International Encyclopedia of the Social and Behavioral Sciences. Oxford: Oxford University Press.

Payne, Thomas. 1997, Describing Morphosyntax. Cambridge: Cambridge University Press.

Pease, Adam, Ian Niles, and John Li. 2002. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. AAAI Technical Report WS-02-11.

Portet, François, Ehud Reiter, Jim Hunter and Somayajulu Sripada. 2007. Automatic Generation of Textual Summaries from Neonatal Intensive Care Data. In: Bellazzi, Riccardo, Ameen Abu-Hanna and Jim Hunter (Ed.), 11th Conference on Artificial Intelligence in Medicine (AIME 07). pp. 227-236.

Prince, Ellen. 1981. Toward a taxonomy of given-new information. Radical pragmatics. P. Cole (Ed.). 223-55. New York: Academic Press.

Reiter, Ehud, Chris Mellish, and John Levine. 1995. Automatic generation of technical documentation. Applied Artificial Intelligence 9.

Reiter, Ehud, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. 2005. Choosing Words in Computer-Generated Weather Forecasts. To appear in Artificial Intelligence.

Reiter, Ehud and Robert Dale. 2000. Building Natural Language Generation Systems. Cambridge: Cambridge University Press.

Reiter, Ehud and Robert Dale. 1997. Building applied natural language generation systems. Natural Language Engineering, 3(1):57-87.

Rogers, Michael, Clare You, and Kyungnyun Richards. 1992. College Korean. Berkeley: University of California Press.

Rösner, Dietmar, and Manfred Stede. 1994. Generating multilingual documents from a knowledge base: the techdocproject. Proceedings of the 15th. International Conference on Computational Linguistics (Coling 94), Vol. I, Kyoto, Japan. 339 – 346.

Rosner, Michael, and Roderick Johnson. 1992. Computational Linguistics and Formal Semantics, Cambridge: Cambridge University Press.

Sheremetyeva, Svetlana, Sergei Nirenburg, and Irene Nirenburg. 1996. Generating patent claims from interactive input. Proceedings of the 8th. International Workshop on Natural Language Generation (INLG'96). Herstmonceux, England. 61–70.

Sinclair, John. 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Springer, Stephen, Paul Buta, and Thomas Wolf. 1991. Automatic letter composition for customer service. R. Smith & C. Scott, eds. Innovative applications of artificial intelligence 3. AAAI Press. Proceedings of CAIA-1991.

Sripada, Somayajulu, Ehud Reiter, Ian Davy, and Kristian Nilssen. 2004 Lessons from Deploying NLG Technology for Marine Weather Forecast Text Generation. Proceedings of PAIS session of ECAI-2004. 760-764.

Summers, Della, et al. 2003. Longman Dictionary of Contemporary English. Edinburgh, England: Pearson Education Limited.

Summers, Ray. 1950. Essentials of New Testament Greek, Nashville: Broadman Press.

Talmy, Leonard. 1985. Lexicalization Patterns: Semantic Structure in Lexical Forms. In T. Shopen et al., eds., Language Typology and Syntactic Description, vol. 3. Cambridge: Cambridge University Press.

Tarski, Alfred. 1956. Logic, semantics, metamathematics: papers from 1923 to 1938. Oxford: Oxford University Press.

Van Valin, Robert. 2001. An Introduction to Syntax. Cambridge: Cambridge University Press.

Wierzbicka, Anna. 1992. Semantic Primitives and Semantic Fields. In A. Lehrer et al., eds., Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Wierzbicka, Anna. 1992. Semantics, Culture, and Cognition. Oxford: Oxford University Press.

Wierzbicka, Anna. 1996. Semantics: Primes and Universals. Oxford: Oxford University Press.

BIOGRAPHICAL INFORMATION

Tod Allman completed a Bachelor’s degree in Systems Engineering at Harvey Mudd College in 1981, and a Master’s degree in Systems Engineering at Claremont Graduate School in 1982. He then went to Talbot Theological Seminary at Biola University and completed a Master’s of Divinity in 1986. While attending Talbot, he decided to spend his life serving God as a Bible translator. He then enrolled in several courses taught by the Summer Institute of Linguistics, and thoroughly enjoyed them. He continued studying linguistics, and completed a Master’s of Linguistics at the University of Texas at Arlington in 1989. While studying at UTA, he began developing a computer program that helps linguists who are translating the New Testament into other languages. After graduating from UTA, he spent approximately twenty years developing that software, and it is now known as The Translator’s Assistant. This dissertation describes that project.