Linguist's Assistant: Gleaning a Tagalog Lexicon and Grammar from a Small, Lightly Annotated Corpus

Tod Allman
National University
551 M.F. Jhocson St.
Sampaloc, Manila, 1008
todallman@yahoo.com

Philippe John F. Sipacio
National University
551 M.F. Jhocson St.
Sampaloc, Manila, 1008
pjfsipacio@national-u.edu.ph

Abstract

Linguist's Assistant (LA) is a large scale semantic analyzer and multi-lingual natural language generator (NLG) designed and developed entirely from a linguist's perspective. The system incorporates extensive typological, semantic, syntactic, and discourse research into its semantic representational system and its transfer and synthesizing grammars. LA has been tested extensively with English, Korean, Kewa (Papua New Guinea), and Jula (Cote d'Ivoure), and proof of concept lexicons and grammars have been developed for Spanish, Urdu, North Tanna (Vanuatu), Angas (Nigeria), and Chinantec (Mexico). LA is presently being used to build a lexicon and grammar for Tagalog. This paper will (i) summarize the major components of the NLG system, (ii) discuss how a significant portion of the Tagalog grammar was gleaned and constructed automatically from a specially designed corpus, and (iii) present the results of experiments that were performed to determine the quality of the generated Tagalog texts. The experiments indicate that when experienced mother-tongue translators use the drafts generated by LA, their productivity is more than tripled without any loss of quality.

1. Introduction

LA was designed and developed specifically for the purpose of generating high quality translations in a wide variety of languages, particularly minority and endangered languages. Translations produced by LA are always easily understandable, grammatically correct, semantically equivalent to the source documents, and at approximately a sixth grade reading level. A model of LA is shown in Figure 1.

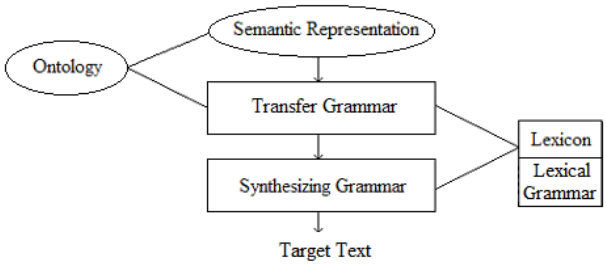

Figure 1. Model of Linguist's Assistant

As seen in the figure, there are five primary components: 1) the ontology, 2) the semantic representations, 3) the lexicon, 4) the transfer grammar, and 5) the synthesizing grammar. The two components in ovals are static knowledge which is supplied with LA, and the three items in rectangles are user-supplied target language knowledge. The final product of LA is target text. LA uses the rich interlingua approach, and employs linguistically based grammar rules rather than stochastic techniques in order to generate its texts. Therefore LA does not require large bilingual corpora for training purposes, and the system works very well even for languages which have very little or no literature.

A sample of the English and Tagalog texts generated by LA are shown below in Figure 2. These samples are from the first half of a short story by World Vision that describes how to prevent eye infections. The texts shown in the figure have not been edited; they're included here to provide a sense of the quality and content of LA's texts.

1. Title: Melissa's eyes are sore.	1. Pamagat: Makirot ang mga mata ni Melissa.
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?"	2. Isang araw, ang babaeng nagngangalang Melissa ay nakaupo sa labas ng kanyang bahay. Ngunit hindi masaya si Melissa dahil napakakirot nang kanyang mga mata. Inisip ni Melissa na mayroong buhangin sa kanyang mga mata. Kaya't tumawag si Melissa sa kaibigang nagngangalang Janet. Saka sinabi ni Melissa kay Janet, "Paki tingnan mo ang aking mga mata. Mayroon bang buhangin sa aking mga mata?"
3. Janet said to Melissa, "Nothing is in your eyes. But your eyes are very red."	3. Sinabi ni Janet kay Melissa, "Walang anumang bagay sa iyong mga mata. Ngunit napakapula nang iyong mga mata."
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!	4. Tapos ay sinabi ni Janet kay Melissa, "Paki tingnan mo ang aking mga mata dahil napakakirot din nang aking mga mata." Kaya't tiningnan ni Melissa ang mga mata ni Janet. Napakapula nang mga mata ni Janet!
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.	5. Tapos ay pumasok si Melissa sa kanyang bahay para magpahinga. Natulog si Melissa sandali. Tapos ay gumising si Melissa. Habang si Melissa ay nasa kanyang bahay, narinig niya si Janet na kinakausap ang kaibigang nagngangalang Alex.
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!"	6. Tinawag nang malakas ni Melissa si Alex. Sumigaw si Melissa, "Alex, pumarito ka sa loob ng aking bahay. Hinahadlangan ako nang isang bagay sa pagdilat nang aking mga mata! Hindi ako makakita!"
7. Then Alex entered Melissa's house quickly. There were many flies inside Melissa's 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."	7. Kaya't pumasok nang mabilis si Alex sa bahay ni Melissa. Maraming langaw sa loob ng bahay ni Melissa. At marami ding langaw malapit sa mga mata ni Melissa. Alam ni Alex na napakalubha nang mga mata ni Melissa. Sinabi ni Alex kay Melissa, "Nababalot nang dilaw na nana ang iyong mga mata. Hinahadlangan ka ng nanang ito sa pagdilat nang iyong mga mata."
8. Then Alex said to Melissa, "I'll try to clean your eyes. But I don't have a clean cloth to clean your eyes." There was a towel in Melissa's house. But that towel was dirty. And Alex's hands also were dirty.	8. Tapos ay sinabi ni Alex kay Melissa, "Susubukan kong linisin ang iyong mga mata. Ngunit wala akong malinis na tela para linisin ang iyong mga mata." Mayroong tuwalya sa bahay ni Melissa. Ngunit marumi ang tuwalyang iyon. At marumi din ang mga kamay ni Alex.
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."	9. Kaya't sinabi ni Alex kay Melissa, "Tatawagin ko si Netty upang tingnan ni Netty ang iyong mga mata. Tutulungan ka ni Netty upang gumaling ang iyong mga mata."
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 entered your eyes. We have to wash your eyes. And we have to clean your eyes thoroughly. We have to clean your eyes each day until they become healthy again.	10. Kaya't pumunta si Netty sa bahay ni Melissa. Matapos tingnan ni Netty ang mga mata ni Melissa, sinabi ni Netty kay Melissa, "Napakalubha nang iyong mga mata. Pumasok ang mga mikrobyo sa iyong mga mata. Dapat nating hugasan ang iyong mga mata at dapat linising mabuti ang iyong mga mata. Dapat nating linisin ang iyong mga mata sa bawat araw hanggang gumaling muli ang iyong mga mata."
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 that cloth.	11. Tapos ay hinugasang mabuti ni Netty ang kanyang mga kamay nang malinis na tubig at naglagay nang malinis na tubig sa kutsarita. Saka inilagay ni Netty ang asin sa tubig na iyon. Saka nagtubog si Netty nang maliit na piraso ng tela sa tubig na nasa kutsarita at hinugasan ang kaliwang mata ni Melissa ng telang iyon.
12. After Netty washed Melissa's left eye, she burned that cloth and washed her hands thoroughly again. Then she told Alex to clean Melissa's other eye.	12. Matapos hugasan ni Netty ang kaliwang mata ni Melissa, sinunog ni Netty ang telang iyon. Saka muling hinugasang mabuti ni Netty ang kanyang mga kamay. Saka sinabihan ni Netty si Alex na linisin ang isa pang mata ni Melissa.
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 spread."	13. Sinabi ni Netty kay Alex, "Bago mo linisin ang mata ni Melissa, dapat mong hugasan muna ang iyong mga kamay. At dapat kang gumamit nang malinis na tela. Kaya't hindi kakalat ang mga mikrobyo."
14. After Netty and Alex finished washing Melissa's eyes, she said, "I'm able to see things now!" Then Netty said to Alex, "You have to burn that cloth. And you have to wash your hands thoroughly."	14. Matapos hugasan nina Netty at Alex ang mga mata ni Melissa, sinabi ni Melissa, "Nakakakita na ako!" Tapos ay sinabi ni Netty kay Alex, "Dapat mong sunugin ang telang iyan. At dapat mong hugasang mabuti ang iyong mga kamay."

Figure 2. Examples of LA's English and Tagalog Texts

2. LA's Semantic Representational System

As was mentioned above, LA uses the rich interlingua approach. This rich interlingua consists of a controlled English influenced metalanguage augmented by a feature system that was designed to accommodate a wide variety of languages. Fundamentally the semantic representations consist of (i) concepts, (ii) features, and (iii) structures. Each of these constituents will be briefly described in the following sections.

2.1 The Concepts in LA's Semantic Representations

Natural Semantic Metalanguage (NSM) theorists have proposed that there is a small set of innate concepts which are indefinable and are found in every language. They claim that every word in every language can be explicated using this small set of innate concepts. They also claim that, while translating a document from one language to another, the problem of cross linguistic lexical mismatch can be significantly reduced by using semantically simple concepts in the source documents. NSM theorists have proposed a systematic method for identifying semantically simple concepts: they explicate numerous words, and the words which appear most frequently in the explications are the semantically simple concepts. They used this method to identify concepts which they call "semantic molecules." The semantic molecules are semantically more complex than the primitives, but they are still semantically simple and occur very frequently in the explications of many words. LA uses the defining vocabulary in Longman's Dictionary of Contemporary English as its semantic molecules. Only the semantic primitives and the words in Longman's defining vocabulary are permitted in LA's semantic representations. Additionally, each word in LA's ontology is very precisely defined, and is used in consistent environments throughout all the semantic representations.

2.2 The Features in LA's Semantic Representations

The features incorporated in LA's semantic representational system have been gleaned from a wide variety of languages. Examples of several features and their possible values are shown below in Table 1.

Table 1. Several Features and their Values

Semantic Category	Feature Name	Feature Values
Object	Number	Singular, Dual, Trial, Quadrial, Plural, Paucal
Object	Participant Tracking	First Mention, Routine, Interrogative, Frame Inferable, Exiting, Restaging, Generic
Object	Proximity	Near Speaker and Listener, Near Speaker, Near Listener, Remote within Sight, Remote out of Sight, Temporally Near, Temporally Remote, Contextually Near with Focus
Event	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, …, Immediate Future, Later Today, Tomorrow, 2 to 3 days from now, …
Proposition	Illocutionary Force	Declarative, Imperative, Content Interrogative, Yes-No Interrogative
Proposition	Salience Band (Longacre, 2012)	Pivotal Storyline, Script Predictable Actions, Backgrounded Actions, Flashback, Setting, Irrealis, Evaluation, Cohesive Material
Object Phrase	Semantic Role	Agent, Patient, State, Source, Destination, Instrument, Beneficiary, Addressee

2.3 The Structures in LA's Semantic Representations

The structures that are permitted in LA's semantic representational system are simple English sentence structures. Propositions always consist of a single event with its arguments and oblique phrases; optional event modifying propositions may also be embedded in a proposition. The argument phrases may include object attributes (i.e., adjectives), modifying phrases (i.e., modifying NPs), and modifying propositions (i.e., relative clauses).

2.4 Example of LA's Semantic Representational System

Every proposition, phrase, and concept in LA's semantic representations has numerous features associated with them. Linguists using LA are able to write grammatical rules that generate the appropriate structures and morphology based on the feature values. The semantic representation of "*I should finish reading these books*" is shown below in Figure 3. As seen in the popup dialog under "read," the verb's Time is 'Future,' its Aspect is 'Completive,' and its Mood is 'should.' These values produce in English "*should finish reading*." The feature values in the popup below "John" indicate that Number is 'Singular' and Person is 'First.' This produces the pronoun "*I*." In the popup below "book," Number is 'Plural,' and Proximity is 'Near Speaker.' These values produce "*these books*."

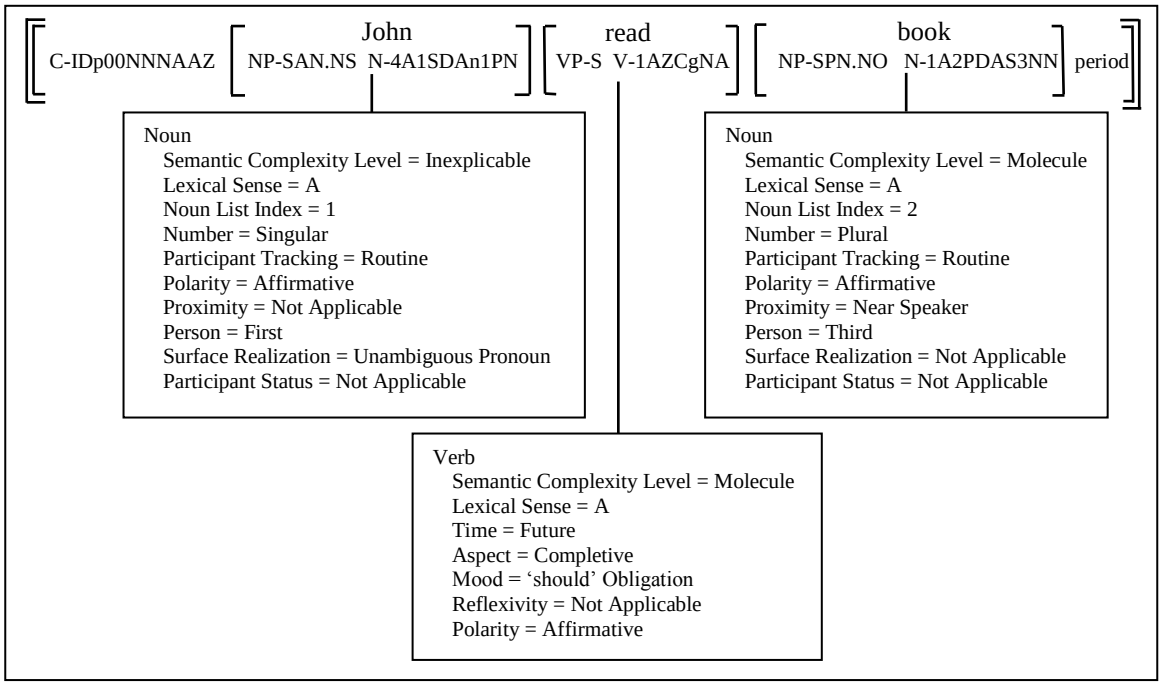

Figure 3. LA's Semantic Representation of "*I(John) should finish reading these books.*"

3. LA's Transfer Grammar

Linguists have known for several decades that it's impossible to develop a language neutral underlying representation that accommodates every language. Therefore the task of LA's transfer grammar is to restructure the semantic representations into new underlying representations that are appropriate for each particular target language. These new underlying representations consist of the target language's words, structures, and features. For example, many languages have rules that are based on grammatical relations, but the noun phrases in the semantic representations are marked with semantic roles rather than grammatical relations. Therefore a rule in the transfer grammar must generate grammatical relations from the semantic roles. For another example, many of the world's

languages are clause chaining rather than coranking, so a rule in the transfer grammar must build appropriate clause chains from the coranking propositions in the semantic representations.

A model of LA's transfer grammar is shown below in Figure 4. The transfer grammar consists of nine different types of rules, each rule type performing a particular task in the process of converting the semantic representations into appropriate underlying representations for the target language.

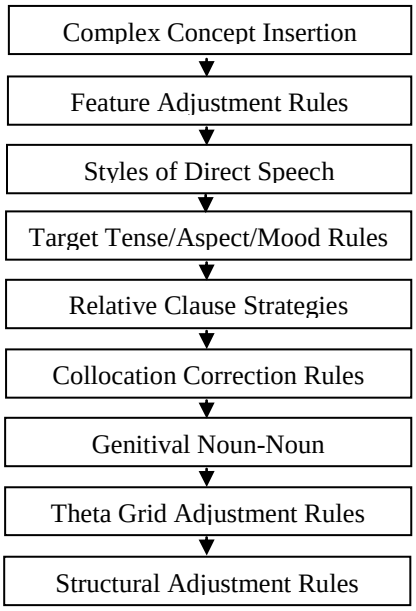

Figure 4. Model of LA's Transfer Grammar

4. LA's Synthesizing Grammar

After the transfer grammar has produced an underlying representation that consists of the target language's words, structures, and features, the synthesizing grammar is responsible for synthesizing the final surface forms. The synthesizing grammar consists of eight different types of rules as shown below in Figure 5.

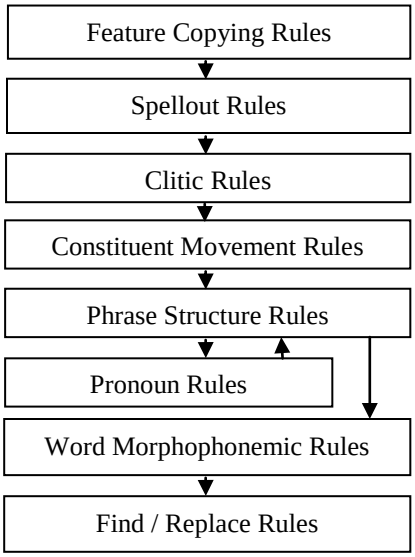

Figure 5. Model of LA's Synthesizing Grammar

5. Gleaning the Tagalog Grammar from a Lightly Annotated Corpus

As was mentioned in the abstract, lexicons and grammars have been developed for a wide variety of languages using LA. Each language typically required somewhere between 500 and 1,000 rules. Developing those rules manually was a time consuming and difficult task that required a skilled computational linguist. After examining the grammars for each of the test languages, it became clear that the majority of the rules for each language were Spellout rules in the synthesizing grammar, and Theta Grid Adjustment rules in the transfer grammar. In order to reduce the time and skill required to develop these grammars, a new technique was developed specifically to build Spellout rules and Theta Grid Adjustment rules. This new approach consists of using a specially designed corpus. The corpus contains numerous tables with sample sentences in English, and a mother-tongue translator is asked to translate those sentences into the target language, and then lightly annotate the translations. This corpus consists of two sections: (i) a section for gleaning and building Phrase Builder Spellout rules, and (ii) a section for gleaning and building Theta Grid Adjustment rules.

5.1. Gleaning Phrase Builder Spellout Rules

The first section of the corpus was designed to quickly glean and build Phrase Builder Spellout rules. Two examples of the tables in the corpus for Phrase Builder rules are shown below. Table 2 contains the examples for building NPs that include a Proximity value, and Table 3 contains the examples for building VPs that include an Aspect value of Inceptive. Linguists may add their own rows to the tables in the corpus, and they may also add additional tables to the corpus to account for feature values or target language constructions that are not represented in the corpus template.

Table 2. The Table in LA's Corpus for Tagalog Noun Proximity

Phrase Builder Nouns: Proximity	Target Language Translation	Features	Noun Form
[This (man)] walked.	Naglakad ang [(lalaki)ng ito].	Near Speaker and Listener	Stem
[These (men)] walked.	Naglakad ang [mga (lalaki)ng ito].	Plural, Near Speaker and Listener	Stem
[This (man)] walked.	Naglakad ang [(lalaki)ng ito].	Near Speaker	Stem
[These (men)] walked.	Naglakad ang [mga (lalaki)ng ito].	Plural, Near Speaker	Stem
[That (man)] walked.	Naglakad ang [(lalaki)ng iyan].	Near Listener	Stem
[Those (men)] walked.	Naglakad ang [mga (lalaki)ng iyan].	Plural, Near Listener	Stem
[That (man)] walked.	Naglakad ang [(lalaki)ng iyon].	Remote within Sight	Stem
[Those (men)] walked.	Naglakad ang [mga (lalaki)ng iyon].	Plural, Remote within Sight	Stem
[That (man)] walked.	Naglakad ang [(lalaki)ng iyon].	Remote out of Sight	Stem
[Those (men)] walked.	Naglakad ang [mga (lalaki)ng iyon].	Plural, Remote out of Sight	Stem
A person walked [this (year)].	Naglakad ang tao sa [(taon)ng ito].	Temporally Near	Stem
A person walked [that (year)].	Naglakad ang tao sa [(taon)ng iyon.]	Temporally Remote	Stem
[This (person)] walked.	Naglakad ang [(tao)ng ito].	Contextually	Stem

		Near with Focus	
[These (people)] walked.	Naglakad ang [mga (tao)ng ito].	Plural, Contextually Near with Focus	Stem
[This (person)] walked.	Naglakad ang [(tao)ng iyon].	Contextually Near	Stem
[These (people)] walked.	Naglakad ang [mga (tao)ng iyon].	Plural, Contextually Near	Stem

After the mother-tongue translator has translated each English sentence in the table above, he must annotate the translations. The annotation consists of putting square brackets around the NP, and putting parentheses around the head noun. Note that if a language employs case markers as Tagalog does, the case markers must not be included in the square brackets. Also note that morphophonemic operations must not be included in the translations. For example, in the row containing the English sentence "*A person walked this year*," the Tagalog translation must have "(taon)ng" even though the correct text is "taong." A subsequent morphophonemic rule will change "taon-ng" to "taong." After the translator has finished annotating the Tagalog translations in the table above, LA will scan through the table and automatically build the rule shown below in Figure 6.

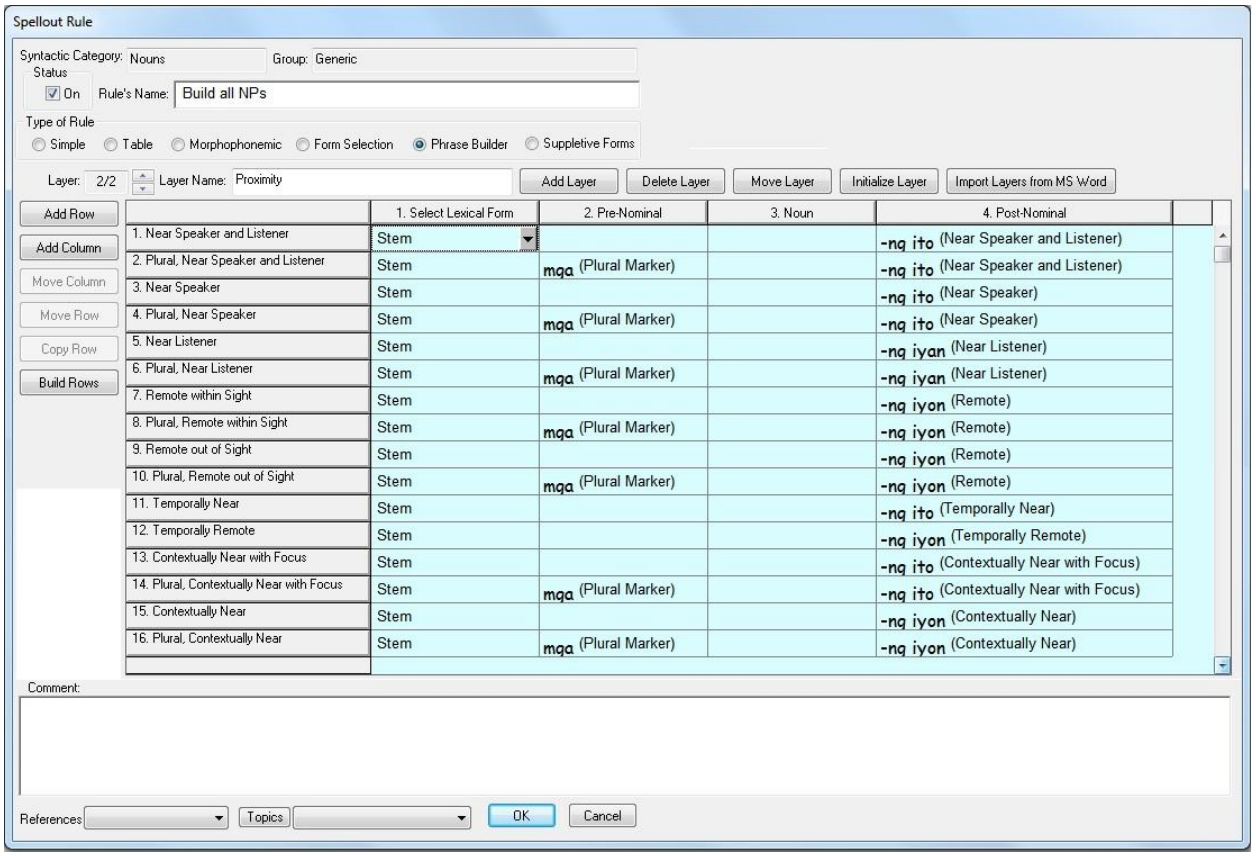

Figure 6. An NP Phrase Builder Rule that was Gleaned from LA's Annotated Corpus

As seen in the figure above, each row in Table 2 corresponds to a row in the rule. When building these rules, LA names each row using the features in the table. Then it scans through the annotated Tagalog text and finds the contents within the square brackets. All the text in the brackets that precedes the

head noun will be put in the column labeled “Pre-Nominal,” and all the text in the brackets that follows the head noun will be put in the column labeled “Post-Nominal.” LA also builds the trigger structure associated with each row, and sets the features according to the Features column in the table. Note that the text in the rule that is enclosed in parentheses was added manually (e.g., “(Plural Marker)” in row 2 of the rule).

A table for building one layer in a VP Phrase Builder rule is shown below in Table 3. In this table, the annotated English sentences are provided, and the features for each row are also provided. The mother-tongue translator translates the English sentences into Tagalog, and then puts square brackets around the VP, and parentheses around the main verb. The translator must also specify the form of the verb to be used in each sentence.

Table 3. Tagalog Inceptive Aspect, Tenses, and Polarities

Phrase Builder Verbs: Inceptive Aspect, Tenses, and Polarities	Target Language Translation	Features	Verb Form
John [started (walking)].	[Nagsimulang (maglakad)] si Juan.	Past, Inceptive	Infinitive
John [did not start (walking)].	[Hindi nagsimulang (maglakad)] si Juan.	Past, Negative, Inceptive	Infinitive
John [certainly did not start (walking)].	[Siguradong hindi nagsimulang (maglakad)] si Juan.	Past, Emphatic Negative, Inceptive	Infinitive
John [certainly started (walking)].	[Siguradong nagsimulang (maglakad)] si Juan.	Past, Emphatic Affirmative, Inceptive	Infinitive
John [will start (walking)].	[Magsisimulang (maglakad)] si Juan.	Future, Inceptive	Infinitive
John [will not start (walking)].	[Hindi magsisimulang (maglakad)] si Juan.	Future, Negative, Inceptive	Infinitive
John [will certainly not start (walking)].	[Siguradong hindi magsisimulang (maglakad)] si Juan.	Future, Emphatic Negative, Inceptive	Infinitive
John [will certainly start (walking)].	[Siguradong magsisimulang (maglakad)] si Juan.	Future, Emphatic Affirmative, Inceptive	Infinitive

After the mother-tongue translator has finished translating and annotating the sentences in the table above, LA will scan through the table and build the rule shown below in Figure 7.

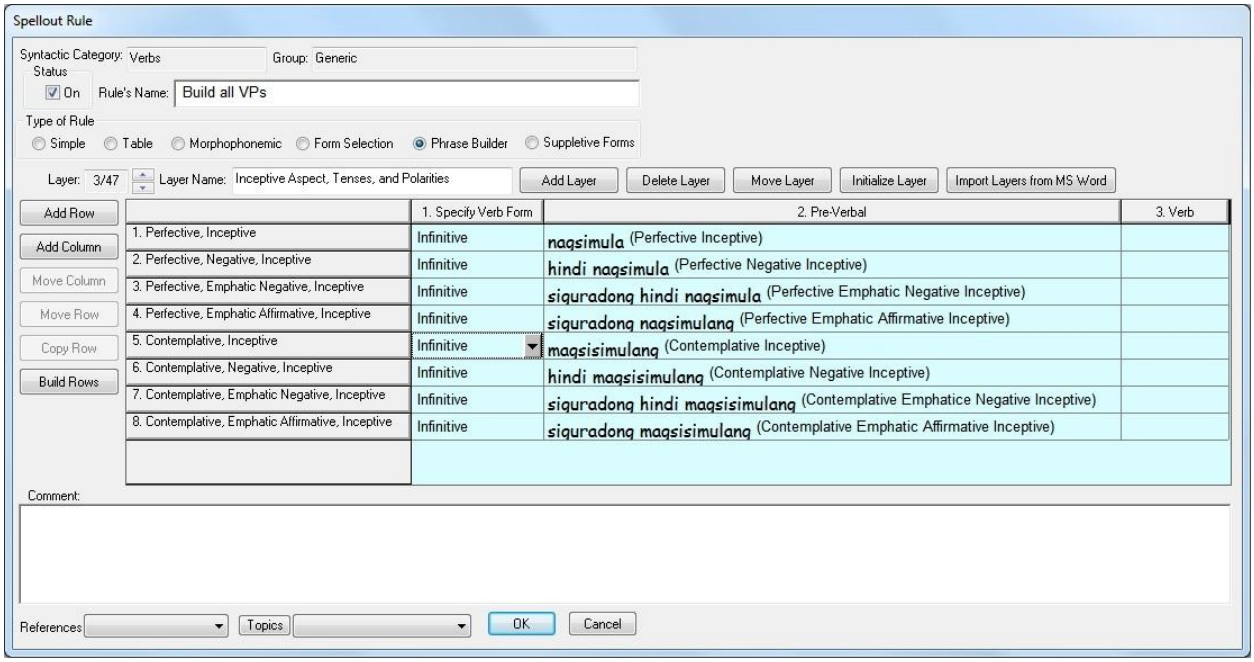

Figure 7. A VP Phrase Builder Rule that was Gleaned from LA's Annotated Corpus

Similar to Figure 6, each row in Table 3 corresponds to a row in this rule. When building the rule, LA uses the row's features for the row's name, and all text in the square brackets that precedes the main verb is put in the column labeled Pre-Verbal. In this example, none of the Tagalog VPs contain text after the main verb, so there isn't a column labeled Post-Verbal.

5.2. Gleaning Theta Grid Adjustment Rules

The second section of the corpus was designed to quickly glean and build Theta Grid Adjustment rules. Each event in LA's ontology has an argument structure that corresponds very closely to the English perspective of that event. For example, the event "walk" has an agent, an optional source, and an optional destination. The Theta Grid Adjustment rules are responsible for restructuring the argument structure associated with each event in order to produce a new argument structure that's appropriate for the target language. Every language requires several hundred Theta Grid Adjustment rules, so gleaning and building these rules automatically was imperative.

The events in LA's ontology may be divided into two broad categories: those that take nominal arguments, and those that take object complement clauses. A sample of the table for gleaning Theta Grid Adjustment rules for events that take nominal arguments is shown below in Table 4.

Table 4. Table for Gleaning Tagalog Theta Grid Adjustment Rules for Events with Nominal Arguments

Theta Grid Adjustment Rules - Verbs with Noun Phrase Arguments	English	Target	Annotated English	Annotated Target
eat-A to intentionally consume food	John ate that fish.	Kinain ni Juan ang isdang.	[A X] ate [P Y]. Mary = X that fish = Y	Kinain [S X] [O Y].
give-A to transfer ownership of something to someone	John gave this book to Mary.	Ibinigay ni Juan ang librong ito kay Maria.	[A X] gave [P Y] [D to Z]. John = X that book = Y Mary = Z	Ibinigay [S X] [O Y] [1 Z].
know-D to be aware of something	John knows about that problem.	May alam si Juan tungkol sa problemang iyon.	[A X] knows [P about Y]. John = X that problem = Y	May alam [S X] [O tungkol Y]
move-A to change one's residence from one place to another place	John moved from this house to that house.	Lumipat si Juan mula sa bahay na iyon patungo sa bahay na ito.	[A X] moved [S from Y] [D to Z]. John = X this town = Y that town = Z	Lumipat [S X] [1 mula Y] [2 patungo Z].

As seen in the table above, the annotation of the Tagalog text consists of three tasks: (i) place square brackets around each NP, (ii) insert the appropriate preposition into each NP (case markers are not to be included in the annotated text), and (iii) label each NP with the appropriate grammatical relation. When LA scans through this table, it will build the Theta Grid Adjustment rule for each verb in the table. While building the rule, it will assign the appropriate grammatical relation to each NP, and it will insert the appropriate preposition into each NP. The Theta Grid Adjustment rule for "move-A" that was built automatically after scanning through this table is shown below in Figure 8.

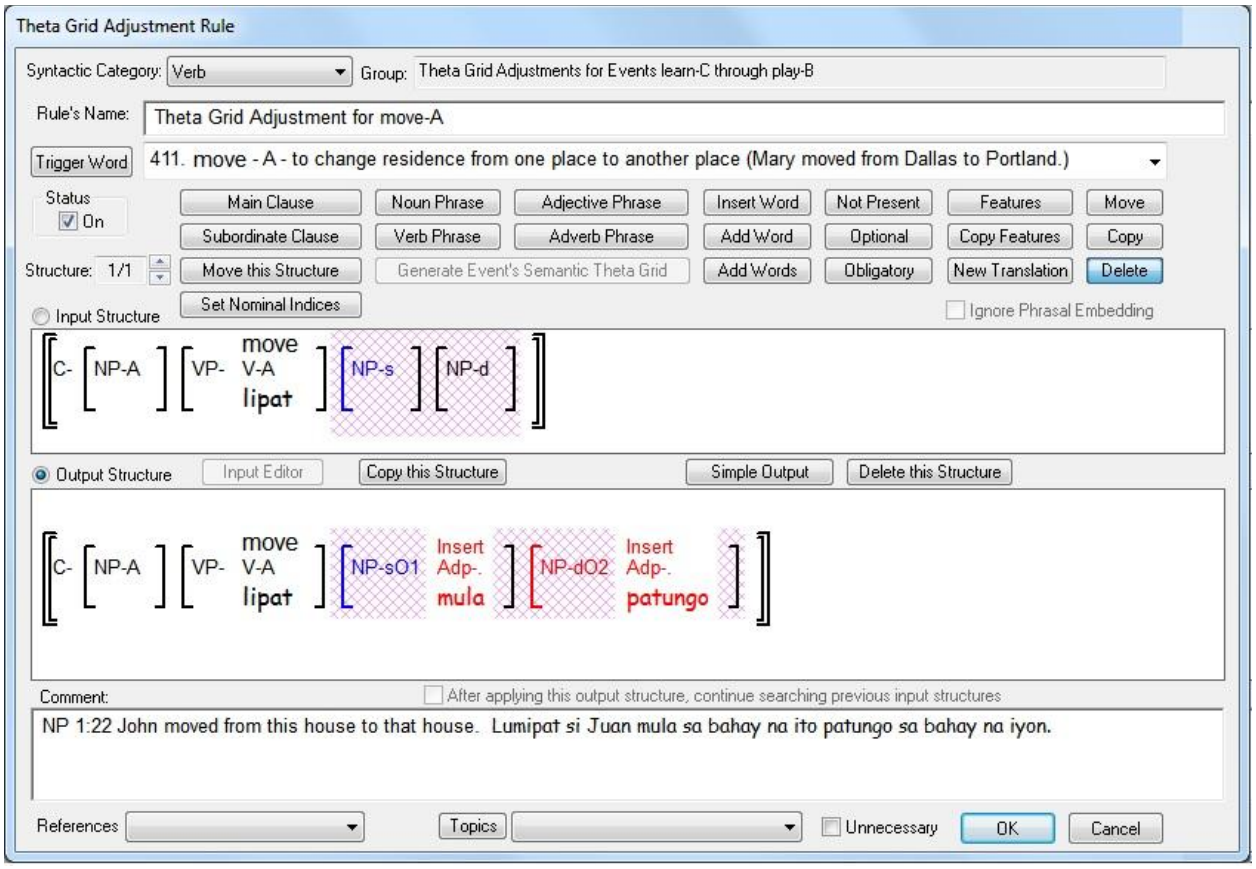

Figure 8. Theta Grid Adjustment Rule for "move-A"

In the rule shown above, the event move-A takes an obligatory agent, an optional source NP, and an optional destination NP. Optionality is indicated by the cross hatching in the input and output structures. The output structure of this rule inserts the preposition 'mula' into the source NP if one is present, and it inserts the preposition 'patungo' into the destination NP if one is present. The source NP also has its features set to a Case value of 'Oblique,' and a Grammatical Relation value of 'Oblique 1.' Similarly the destination NP has its features set to a Case value of 'Oblique,' and a Grammatical Relation value of 'Oblique 2.' A subsequent Spellout rule will insert the case marker 'sa' into every NP that has a Case value of 'Oblique,' and the phrase structure rule for clauses will order the NPs according to their grammatical relation values. Thus it can be seen how this rule is able to produce Tagalog sentences such as "*Lumipat si Juan mula sa bahay na ito patungo sa bahay na iyon*."

6. Determining the Quality of the Generated Tagalog Texts

Extensive experiments have been performed to determine the quality of the Tagalog texts generated by LA. The purpose of these experiments is to determine whether the LA drafts are (i) grammatically correct, (ii) semantically equivalent to the original documents, and (iii) easily understandable to junior high students. The experiments may be divided into four categories, and each category will be described below.

6.1. The Backtranslation Experiments

The first experiment that is performed with a text generated by LA is the backtranslation experiment. The purpose of this experiment is to determine whether or not the Tagalog texts generated

by LA are communicating the same information as the original documents. During these experiments, mother-tongue Tagalog speakers who speak English well were asked to backtranslate into English the LA Tagalog draft. Specifically, three adult mother-tongue Tagalog speakers were asked to backtranslate the LA Tagalog draft of the Infected Eye story into English. They were also told that if any text in the Tagalog translation was unclear or incomprehensible, they should underline that text. All three of these participants produced backtranslations which communicated the same information as the original document. None of these participants indicated that any part of the Tagalog text was unclear or incomprehensible.

6.2. The Comprehension Questions Experiments

In order to determine whether or not the texts generated by LA are easily understandable, mother-tongue Tagalog speakers were asked to read the LA Tagalog translation, and then answer comprehension questions related to the story. Five college students and two junior high students were asked to read the entire Tagalog draft of the story generated by LA, and then answer five comprehension questions. All five of the college students and both of the junior high students answered all of the comprehension questions correctly. Thus it is clear that the Tagalog translation produced by LA is easily understandable, even to junior high students.

6.3. The Productivity Experiments

After the backtranslation experiments and comprehension question experiments confirmed that the Tagalog draft generated by LA was easily understandable and communicating the same information as the original document, the productivity experiments were performed. The purpose of the productivity experiments is to determine whether or not the Tagalog draft produced by LA is of sufficient quality that it actually increases the productivity of experienced mother-tongue translators. More specifically, these experiments determine whether editing the LA Tagalog text is more efficient than manually translating the text. If the draft produced by LA requires more time to edit than is required to manually translate the same text, then using LA is counter-productive. But if this experiment indicates that experienced mother-tongue translators are able to edit the LA drafts in less time than is required for manual translation, then using LA improves their productivity, and should therefore be used.

For the productivity experiments, ten graduate students at National University (NU) were selected. These ten students speak English well, and each has some experience translating English into Tagalog. All ten of the students are mother-tongue Tagalog speakers. Before the experiment began, the students were told that a new technique for translating junior high texts into Tagalog was being evaluated; these students were not told that a computer had translated the story into Tagalog. The Infected Eye story was divided into two halves based on the word count in the English draft. Then five of the NU students were asked to translate LA's English draft of the first half of the story into Tagalog. The students were told that they would be timed during this task, and that they should produce a presentable first draft translation appropriate for junior high students. After those five students had completed the translation task, they were then asked to edit the second half of the LA generated Tagalog draft of the story, and they were again told that they should produce a presentable first draft appropriate for junior high students. The students were also timed during this second task. The other five NU graduate students performed the same two tasks, but those students edited the LA Tagalog draft first, and then manually translated the second half of the story from English into Tagalog. The results of these productivity experiments are shown below in Table 5. For each participant there is an Editing Time and a Translating Time recorded in minutes. Each participant's ratio of Translating Time to Editing Time is shown in the final column.

Table 5. Summary of the Productivity Experiments

Participant	Tasks	Editing Time (minutes)	Translating Time (minutes)	Ratio
#1	translated 1st half, edited 2nd half	5	45	9
#2	edited 1st half, translated 2nd half	4	19	4.8
#3	edited 1st half, translated 2nd half	8	20	2.5
#4	edited 1st half, translated 2nd half	10	36	3.6
#5	translated 1st half, edited 2nd half	17	25	1.5
#6	edited 1st half, translated 2nd half	14	27	1.9
#7	translated 1st half, edited 2nd half	13	42	3.2
#8	translated 1st half, edited 2nd half	17	25	1.5
#9	translated 1st half, edited 2nd half	16	27	1.7
#10	edited 1st half, translated 2nd half	10	27	2.7
			Average Ratio:	3.2

As seen in the last row, the average ratio is 3.2. Thus the results of this experiment indicate that on average, the students were able to edit the computer generated text more than three times as quickly as they were able to manually translate the same amount of text. However, the results of this experiment were skewed because these students worked in groups. Each student did his or her own work during this experiment, but some of the students would wait until their classmates had completed the task before they would hand in their work. For example, participants 5, 8, and 9 appeared to wait until all three had finished a particular task, and then they handed in their work together. They each spent approximately 25 minutes translating the first half of the story, and then approximately 17 minutes editing the second half of the LA draft. Before the experiment began, all the students were told that they would be timed during each task, and that they should submit their work as soon as they were finished so that the time could be recorded. But it appeared to the moderator that some of these students would complete their work, and then wait to submit their work until other students were ready to submit their work also. Therefore the results of this experiment are skewed to some degree. If this productivity experiment is repeated in the future, it will be done with just one or two participants at a time so that they will not be influenced by their peers.

The changes that the students made while editing the LA draft generally consist of the following types:

1) Changing one conjunction to a different conjunction (e.g., change “tapos” to “pagkatapos,” change “kaya’t” to “kaya,” or change “saka” to “at”), or deleting the conjunction.
2) Replacing a full noun phrase with a pronoun.
3) Changing a word or expression to a different but more appropriate word or expression (e.g., change ‘barat araw’ to ‘araw araw’).

All of the changes that the students made to the LA draft were examined, and whenever two or more students made the same change, those changes were incorporated into the draft generated by LA. The description of the next experiment refers to the “edited LA draft” which is the draft generated by LA, but also incorporates the students’ edits.

6.4. The Quality Evaluation Experiments

The purpose of the quality evaluation experiments is to compare the quality of the edited LA draft with the quality of the manually translated texts that were produced in the productivity experiments. Twenty NU college students were asked to compare the quality of the manually translated texts with the quality of the edited LA text. These students were not told how either of the two translations had been produced. Twenty questionnaires were prepared; two questionnaires contained an excerpt from the text manually translated by Participant #1 in Table 5. One of those questionnaires

had the edited LA draft at the top of the page, and the manually translated text at the bottom. The other questionnaire had the manually translated text at the top, and the edited LA draft at the bottom. Similarly two questionnaires containing an excerpt from the text manually translated by Participant #2 were prepared, two questionnaires containing an excerpt from the text manually translated by Participant #3 were prepared, etc. So each questionnaire contained one paragraph that had been manually translated, and the same paragraph from the edited LA draft. The students who participated in the evaluation experiments were asked to read the two short paragraphs, and then choose one of the following three options[1]: (i) the first translation is better[2] than the second translation for sixth grade students, (ii) the second translation is better than the first translation for sixth grade students, and (iii) the two translations are equally good for sixth grade students. The results of the evaluations performed by the college students are shown below in Table 6. The numbers in the column labeled "LA" indicate the number of college students who thought that the edited LA text was better for sixth grade students than the manually translated text. The numbers in the column labeled "Manual" indicate the number of students who thought that the manually translated text was better than the LA text. And the numbers in the column labeled "Equal" indicate the number of students who thought that the two texts were equal in quality.

Table 6. Evaluations by the College Students

Participant	LA	Manual	Equal
#1	2	0	0
#2	1	1	0
#3	2	0	0
#4	0	2	0
#5	1	1	0
#6	1	1	0
#7	1	1	0
#8	0	1	1
#9	1	1	0
#10	1	1	0
Totals:	10	9	1

Since LA is intended to generate texts at a junior high reading level, the quality evaluation experiments were also performed with fifty junior high students. Fifty questionnaires identical to the ones used by the college students were prepared, but the choices at the bottom of these questionnaires were: (i) the first translation is better than the second, (ii) the second translation is better than the first, and (iii) the two translations are equally good. The results of the evaluations by the junior high students are shown below in Table 7. Similar to Table 6, the numbers in the column labeled "LA" indicate the number of junior high students who thought that the edited LA text was better than the manually translated text. The numbers in the column labeled "Manual" indicate the number of students who thought that the manually translated text was better than the LA text. And the numbers in the column labeled "Equal" indicate the number of students who thought that the two texts were equal in quality.

[1] The questionnaires were written entirely in Tagalog; the options shown here have been translated into English.

[2] The word "better" was intentionally used in these three options even though it is very vague. We did not want to ask the evaluators which text was easier to understand, which text had better information flow, which text was more natural, etc. Instead we wanted to know if either text was better in any way than the other text.

Table 7. Evaluations by the Junior High Students

Participant	LA	Manual	Equal
#1	4	2	2
#2	2	2	2
#3	1	1	4
#4	0	0	6
#5	1	0	3
#6	0	2	2
#7	1	0	3
#8	0	1	3
#9	1	0	3
#10	0	1	3
Totals:	10	9	31

The results of these evaluation experiments indicate that both the college students and the junior high students consider the edited LA draft to be equal in quality with the manually translated texts. It is interesting to note that 14 of the college student evaluators chose the second paragraph on the questionnaire as being the better text, only 5 of them chose the first paragraph in the questionnaire as being better, and 1 of them said the two texts were equal. So apparently the translation that the college students read last was overwhelmingly the preferred text, regardless of whether that text was translated by LA or manually translated. This again confirms that the edited LA draft is identical in quality with the manually translated texts.

7. Conclusions

At the present time a lexicon and grammar for Tagalog are being developed. The texts generated to date have been tested for quality by adults and junior high students. The results of the tests indicate that LA's Tagalog texts are:

- easily understandable,
- grammatically correct,
- communicate the same information as the original documents,
- at a reading level that can be understood by junior high students, and
- of sufficient quality that they triple the productivity of experienced mother-tongue translators.

Additionally, the edited LA text is of the same quality as the manually translated texts. These results are very encouraging, but more work remains to be done. The Tagalog lexicon must be expanded to include more words, and the grammar must be refined to produce a wider range of Tagalog sentence structures.

References:

Allman, Tod. 2010. *Translator's Assistant: A Multilingual Natural Language Generator based on Linguistic Universals, Typologies, and Primitives*. Arlington, TX: University of Texas dissertation.

Goddard, Cliff. 2008. *Cross-linguistic Semantics*. Amsterdam, The Netherlands: John Benjamins.

Longacre, Robert E., and Shin Ja J. Hwang. 2012. *Holistic Discourse Analysis*. Dallas, TX: SIL International.

Reiter, Ehud, and Robert Dale. 2006. *Building Natural Language Generation Systems*. Cambridge, UK: Cambridge University Press.

Wierzbicka, Anna. 1996. *Semantics: Primes and Universals*. Oxford, UK: Oxford University Press.