Toward an Optimal Multilingual Natural Language Generator: Deep Source Analysis and Shallow Target Analysis

Tod Allman
Graduate Institute of Applied Linguistics
7500 W. Camp Wisdom Rd.
Dallas, TX 75236
+1 214-587-0721
tod_allman@gial.edu

Stephen Beale
University of Maryland, Baltimore County
1000 Hilltop Circle
Baltimore, MD 21250
+1 410-455-3372
sbeale@csee.umbc.edu

Richard Denton
Dartmouth College
6127 Wilder Lab
Hanover, NH 03755
+1 603-646-2732
richard.e.denton@dartmouth.edu

ABSTRACT

Linguist's Assistant (LA) is a large scale multilingual 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 with English, Korean, Kewa (Papua New Guinea), and Jula (Cote d'Ivoire), and proof of concept lexicons and grammars have been developed for a variety of other languages. The system has generated initial draft translations of texts in each of the test languages, and when experienced mother-tongue translators edit those drafts into publishable texts, their productivity is typically quadrupled when compared with manual translation.

An optimal NLG will be able to generate high quality texts in a wide variety of languages with minimal knowledge of the target language grammars. In order to increase the quality of the drafts generated by LA, deep source analysis techniques have been adopted. And in order to minimize the target language knowledge that is required to generate the drafts, a new approach to grammar development has been designed into LA's synthesizing grammar. This paper will: 1) summarize the major components of the generation system, 2) describe several of the source analysis techniques that have been adopted during the development of LA's semantic representations, and 3) present an example of the new type of synthesizing rule that was added to LA's grammar. The adoption of deep source analysis techniques combined with shallow target analysis has proven to be a very efficient model.

Keywords

Natural language generation, ontology, interlingua, semantic representation

1. INTRODUCTION

Linguist's Assistant is a software system which enables linguists to document a language and simultaneously generate translations of numerous texts for the speakers of that language. Linguists are able to build a lexicon and grammar for a language in LA. Then the system applies that lexicon and grammar to the many semantic representations which have been developed, and produces initial draft translations of those texts. A model of LA is shown below in Figure 1.

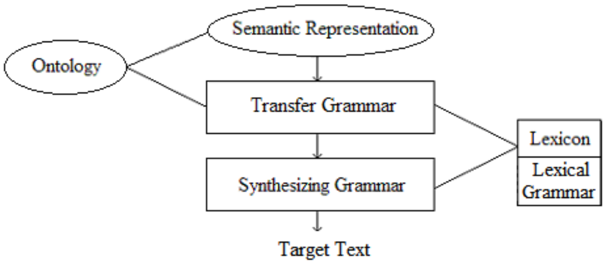


Figure 1. Model of Linguist's Assistant

As seen in the figure, there are five primary components in LA: 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 language text. This system has been thoroughly described in [1] and [2].

1.1 LA's Ontology

LA's ontology was developed using the foundational principles of Natural Semantic Metalanguage theory (NSM) [4]. NSM theorists propose that there is a small set of innate concepts which are present in every language. This set consists of approximately 65 semantically simple concepts such as *I*, *you*, *thing*, *be*, *do*, *think*, *feel*, *want*, *see*, *hear*, *good*, *bad*, *big*, *small*, etc. NSM theorists call these innate concepts "semantic primitives," and they claim that every word in every language may be explicated using these primitives. In order to simplify the process of explicating thousands of words, they have also identified concepts which are semantically more complex than the primitives, but are still semantically simple. They call these concepts "semantic molecules," and these concepts are used repeatedly when explicating words. Their semantic molecules include body parts (e.g., *head*, *hand*, etc.), actions (e.g., *make*, *drink*, *eat*, *hold*, etc.), manners (e.g., *quickly*, *slowly*, etc.), etc. LA's ontology contains concepts which have been organized into five categories according to their semantic complexity: 1) the NSM semantic primitives, 2) our semantic molecules[1], 3) semantically complex concepts which may be inserted into the semantic representations

[1] For our semantic molecules, we use the Defining Vocabulary for Longman's Contemporary English Dictionary [12].

by a rule if the target language has a lexical equivalent, 4) semantically complex concepts that don't yet have an insertion rule, and 5) concepts which are inexplicable (e.g., proper names, numbers, etc.). According to NSM theory, semantically simple words are more likely to have lexical equivalents in other languages than are semantically complex words. We've found that the use of semantically simple concepts in our semantic representations has significantly reduced the problem of lexical mismatch when working with languages that are unrelated to English. However, texts that consist of only semantically simple words are unwieldy, drawn out, and the message becomes distorted. Therefore we've developed a technique to insert semantically complex words into the texts when a target language has a lexical equivalent. For example, the word *shepherd* is semantically complex. Whenever the word *shepherd* occurs in a source document, it is replaced with the phrase *man that takes care of sheep* in the semantic representation. Speakers of languages that have a word for *shepherd* don't want to read texts that contain *man that takes care of sheep*; instead they want to read texts that contain the word *shepherd*. Therefore a complex concept insertion rule will search for all occurrences of *man that takes care of sheep* in the semantic representations, and replace that phrase with the word *shepherd* if the linguist activates the associated complex concept insertion rule.

1.2 LA's Semantic Representations

Many natural language generators and machine translation projects use the rich interlingua approach. For example, the Knowledge Based Accurate Natural Language Translation project (KANT) [9] developed at Carnegie Mellon during the '90s and early 2000s used an interlingua formatted like the one shown below in Figure 2. This interlingua representation is for the sentence "*The truck must be parked on a level surface*."

```
(*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. Example of KANT's Interlingua

LA initially used an interlingua representation similar to the one shown above. Interlinguas such as these work well when the target languages are closely related to English. However, since LA is intended to generate texts in a wide variety of languages, a much richer representation was required. Formal semantics [3], conceptual semantics [5], generative semantics [6], and ontological semantics [11] were each considered but found unsuitable because they didn't include sufficient information for minority languages. Therefore a new format was developed specifically for LA's semantic representational system. LA's semantic representations are comprised of a controlled, English influenced metalanguage augmented by a feature system which was designed to accommodate a wide variety of languages. Fundamentally these semantic representations consist of concepts, structures, and features. The concepts that are permitted in the semantic representations are all semantically simple, as was described in the previous section. The structures permitted in the semantic representations are a small restricted set of English-like sentence structures. The feature system developed for LA includes semantic, syntactic, and discourse information. The feature values have been gleaned from a wide variety of diverse languages. Table 1 shows a few examples of these features and their values.

Table 1. Several of LA's Features and their Values

Feature	Possible Values
Noun Number	Singular, Dual, Trial, Quadrial, Plural, Paucal
Noun Participant Tracking	First Mention, Routine, Interrogative, Frame Inferable, Exiting, Restaging, Generic, …
Noun 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 without 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, …
Proposition Illocutionary Force	Declarative, Imperative, Content Interrogative, Yes-No Interrogative
Proposition Salience Band	Pivotal Storyline, Script Predictable Actions, Backgrounded Actions, Flashback, Setting, Irrealis, Evaluation, Cohesive Material
Noun Phrase Semantic Role	Agent, Patient, State, Source, Destination, Instrument, Beneficiary, Addressee

As seen in Table 1 above, every noun in the semantic representations is marked for Number, and the possible values are Singular, Dual, Trial, Quadrial, Plural, and Paucal. All of these values are necessary because some languages morphologically distinguish each of them. LA's semantic representation for the sentence "*Paulus started walking from the market to a village named Terpen*" is shown below in Figure 3.

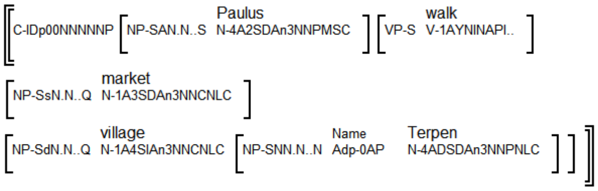


Figure 3. Example of LA's Semantic Representation

As seen in Figure 3, every concept, phrase, and proposition has numerous features associated with it; the letters and numbers below the concepts and beside the phrase and proposition boundaries represent specific feature values. For example, the phrase containing *Paulus* has its Semantic Role set to Agent, the phrase containing *market* has its Semantic Role set to Source, the phrase containing *village* has its Semantic Role set to Destination, the event *walk* has its Time set to Discourse and its Aspect set to Inceptive, the proposition's Illocutionary Force is set to Declarative and its Salience Band is set to Pivotal Storyline, etc.

1.3 LA's Lexicon

The target lexicon serves as a repository for all of the target language's words and their associated features and forms. Within the lexicon a linguist defines the features that are pertinent to each syntactic category for his particular target language. For example, each noun can be assigned a gender value, an honorific value, a class value, etc. Similarly the required forms are defined in the target lexicon (e.g., English verbs have a stem plus a past tense form, a perfect participle form, a gerund form, and a third singular present form). Then lexical spellout rules are used to generate the various forms of each target word. All instances of suppletion are entered into the target lexicon manually.

1.4 LA's Transfer Grammar

Linguists have known for several decades that it's impossible to build 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 a 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.

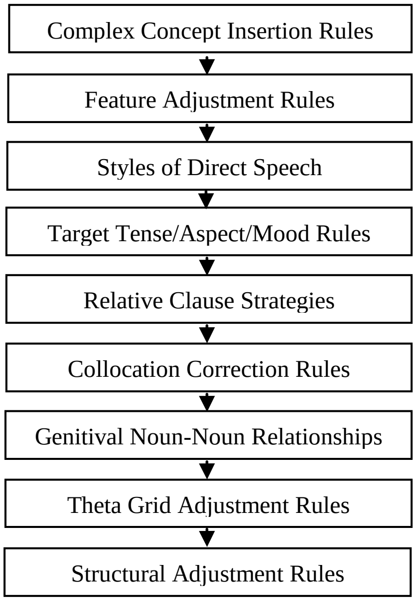
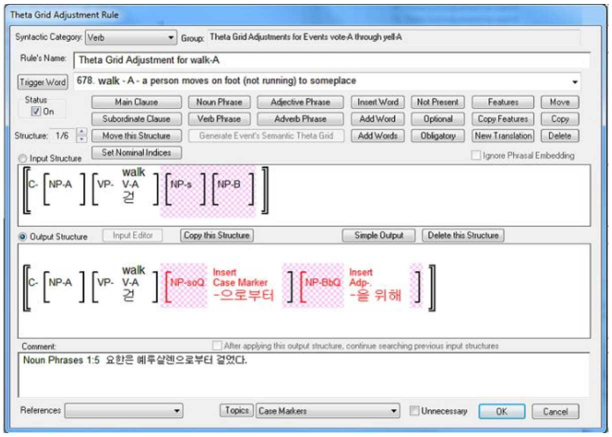

Complex Concept Insertion Rules
Feature Adjustment Rules
Styles of Direct Speech
Target Tense/Aspect/Mood Rules
Relative Clause Strategies
Collocation Correction Rules
Genitival Noun-Noun Relationships
Theta Grid Adjustment Rules
Structural Adjustment Rules

Figure 4. Model of LA's Transfer Grammar

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.

The Theta Grid Adjustment rules do a significant amount of the restructuring, so they will be briefly described here. Every verb in every language has an associated theta grid which describes the verb's argument structure. The theta grids for the events in the semantic representations are very similar to the theta grids for the equivalent English verbs. However, the verbs in other languages have different argument structures, so the theta grid adjustment rules enable a linguist to restructure an event's arguments according to the theta grid of the target language's equivalent verb. The Korean theta grid adjustment rule for the concept *walk* is shown in Figure 5. That rule inserts the appropriate Korean postpositions into the source and destination noun phrases.


Theta Grid Adjustment Rule

Syntactic Category: Verb

Group: Theta Grid Adjustments for Events vote-A through yell-A

Rule's Name: Theta Grid Adjustment for walk-A

Trigger Word: 678. walk - A - a person moves on foot (not running) to someplace

Status ☑ On

Structure: 1/6

○ Input Structure

● Output Structure

□ Ignore Phrasal Embedding

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

Noun Phrases 1:5 요한은 예루살렘으로부터 걸었다.

References:

Topics: Case Markers

□ Unnecessary

Figure 5. The Korean Theta Grid Adjustment Rule for *walk*

1.5 LA's Synthesizing Grammar

LA's synthesizing grammar is responsible for synthesizing the final surface forms of the target text. The synthesizing grammar was designed to resemble as closely as possible the descriptive grammars that field linguists routinely write. Before developing this grammar, dozens of descriptive grammars written by field linguists were examined in order to determine the capabilities that are required to synthesize surface text. A model of the final result is shown below in Figure 6.

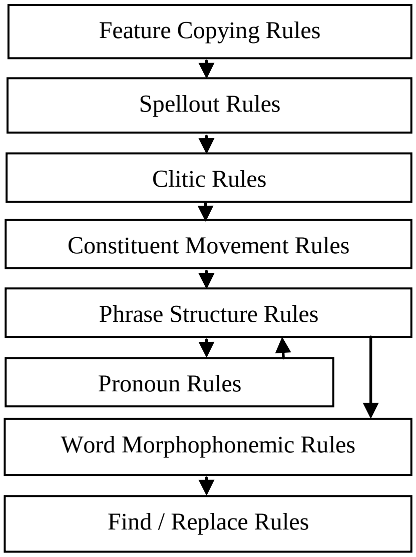
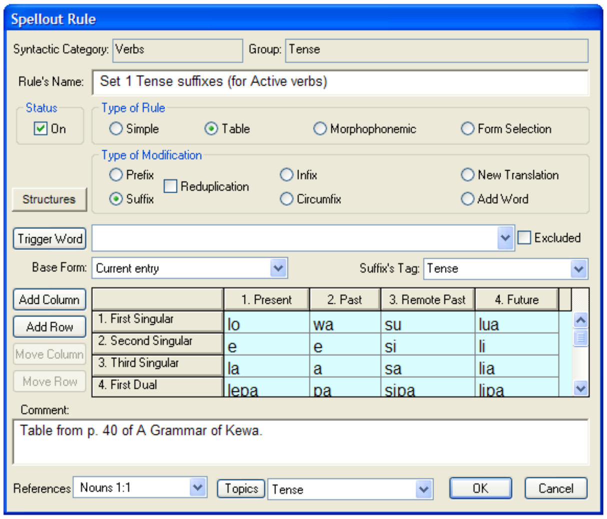


Figure 6. Model of LA's Synthesizing Grammar

As seen in the figure above, there are eight different types of rules in the synthesizing grammar. Spellout Rules are responsible for synthesizing the final forms of the target words, so they will be briefly described here. Initially LA included four basic types of spellout rules: (i) simple spellout rules which add prefixes, suffixes, infixes, circumfixes, or a new word to an existing word, or they provide a new translation of a particular target word in a given context; (ii) form selection rules which select a form of a target word from the target lexicon; (iii) morphophonemic rules which perform morphophonemic operations on the affixes that were added to the stem; and (iv) table spellout rules which group a common set of affixes together into a single rule. After these spellout rules have been executed, each target word is in its final surface form. A table spellout rule that adds tense suffixes to Kewa verbs is shown below in Figure 7.



Spellout Rule

Syntactic Category: Verbs Group: Tense

Rule's Name: Set 1 Tense suffixes (for Active verbs)

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: Current entry Suffix's Tag: Tense

	1. Present	2. Past	3. Remote Past	4. Future
1. First Singular	lo	wa	su	lua
2. Second Singular	e	e	si	li
3. Third Singular	la	a	sa	lia
4. First Dual	lepa	pa	sipa	lipa

Comment: Table from p. 40 of A Grammar of Kewa.

References: Nouns 1:1 Topics: Tense

Figure 7. Spellout Rule that adds Kewa Tense Suffixes

After the synthesizing grammar has been executed, the system displays the final form of the target language text. Then mother-tongue speakers edit the text to improve the naturalness and information flow. Samples of English and Korean texts generated by LA are shown below in Figure 8. The texts in that figure have not been edited; they are the actual texts that were generated by LA. These texts occur at the beginning of a story that describes how to prevent the spread of Avian Influenza.

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 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.	어느 날 팔러스라는 의사가 시장에서 터펜이라는 자기 마을로 돌아왔다. 팔러스가 시장에 있는 동안 사람들이 팔러스에게 어떤 병에 대해서 말하였다. 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.

Figure 8. Examples of LA's English and Korean Texts

1.6 LA's Results

When beginning a new language project, we always start by working through a series of simple sentences which we call the Grammar Introduction. The sentences in the Grammar Introduction illustrate various tenses, moods, aspects, illocutionary forces, etc., and they include various types of relative clauses, object complements, and adverbial clauses. After the grammar has been sufficiently developed to generate all the sentences in the Grammar Introduction, we begin working through actual texts. In every test language a clear trend has developed: after working through the Grammar Introduction, the number of new grammatical rules required for each subsequent chapter of text dramatically decreased. Figure 9 below shows the number of new grammatical rules required to generate each chapter of Kewa text. The number of new transfer rules required for each chapter is shown in blue, and the number of new synthesizing rules is shown in red.

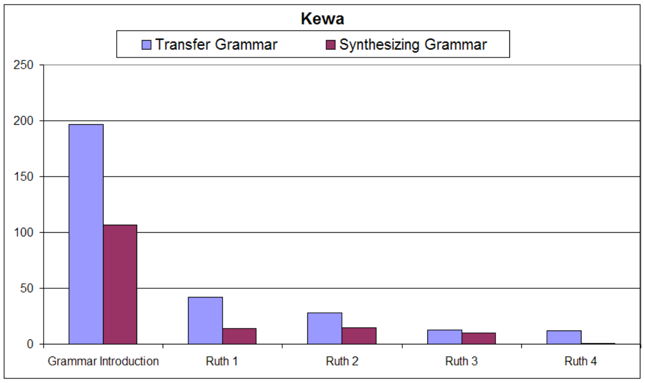


Figure 9. Graph Showing the Number of New Kewa Rules

Figure 10 demonstrates the same trend during the development of the Korean grammar.

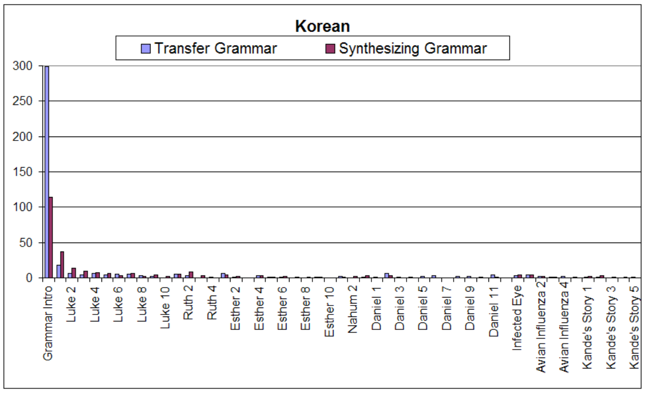


Figure 10. Graph Showing the Number of New Korean Rules

After texts have been generated in a particular target language, experiments are performed to determine the quality of the texts. The experiments have varied for each test language for a variety of reasons[2], but typically several experienced mother-tongue

[2] There are very few Jula speakers who are able to read, so the Jula translators read their texts, and the recordings were played to the people who were doing the evaluations. There are very few Kewa translators, so only one translator participated in the Kewa experiment.

translators are asked to spend a period of time (e.g., 30 minutes) editing a draft produced by LA, and then they are asked to spend the same amount of time manually translating a similar text. Then the ratio of the number of words in the edited LA draft and the number of words in the manually translated text is calculated. Typically the mother-tongue translators are able to edit more than four times as much text in the given time period as they are able to manually translate. Table 2 below summarizes the ratios for three of the test languages: Jula, Kewa, and Korean.

Table 2. Summary of Productivity Experiments

Language	Ratio of Edited Words to Manually Translated Words
Jula	4.3
Kewa	6.7
Korean	4.6

Then additional experiments were performed to compare the quality of the edited LA drafts with the quality of the manually translated texts. Short samples of the edited LA draft and the manually translated text were presented to other mother-tongue speakers who were unaware of how the two texts had been produced. Those people were asked to evaluate the two samples, and choose one of the following three options: (i) the first text is better[3] than the second text, (ii) the second text is better than the first text, or (iii) the two texts are essentially equal in quality. The results of these evaluations for Jula and Korean are shown below in Table 3.

Table 3. Summary of Evaluation Experiments

Language	LA's Texts	Manual Texts	Equal
Jula	12	11	17
Korean	88	71	33

In Table 3 the column labeled "LA's Texts" indicates the number of evaluators who indicated that the edited LA draft was better than the manually translated text, the column labeled "Manual Texts" indicates the number of evaluators who chose the manually translated text as being better, and the column labeled "Equal" indicates the number of evaluators who said the two texts were equal in quality. These evaluation experiments demonstrated that the edited LA drafts and the manually translated texts are statistically of equal quality.

Additional tests were performed to determine whether or not the edited LA drafts are semantically equivalent to the source documents. Mother-tongue speakers were asked to read the edited LA drafts, and then answer comprehension questions and produce back-translations. In every case the comprehension questions were answered correctly, and the back-translations proved that the edited LA drafts are communicating the same message as the original source documents. Therefore it was concluded that LA's drafts typically quadruple the productivity of experienced mother-tongue translators without any loss of quality.

2. Deep Source Analysis Techniques

As was mentioned above, LA initially employed a semantic representational system similar to the one shown above in Figure 2. Nouns were marked as "+Definite" or "-Definite," sentences were marked as "+Passive," etc. That system worked well for LA's first two test languages which were English and Spanish. However, when non-Indoeuropean languages were tested, it quickly became apparent that the semantic representational system had to include much more information. Additionally, we realized that deeper source analysis techniques had to be developed. Marking nouns simply as "+Definite" was inadequate when working with languages which have a much richer article system than English. Therefore we began searching for the universal underlying features that are common across a wide variety of languages, and found that discourse linguists have documented much of this information. We incorporated their findings into LA's semantic representational system, and this section will present some of the techniques that have been adopted and the additions that have been made to LA's feature system.

2.1 Salience Bands

When building the semantic representation of a particular source document, we begin by examining the VP of each proposition. For example, consider the sentence *Kande and Teshi* *ran* *to their house.* Most analysts look at that VP and mark the verb's tense as past, and the verb's aspect as perfective. That analysis is correct, but we take the analysis one step further. We always ask: Why did the author use a past perfective verb in this particular sentence? The answer lies in what linguists call "Salience Bands." Linguists have found that the grammatical form of the VP is often dictated by the sentence's Salience Band. Longacre [8] has identified seven salience bands that he proposes are present in every language; a list of these salience bands is shown below in Table 4.

Table 4. Longacre's Salience Scheme

Salience Band	Function
1. Storyline	These propositions carry the story.
2. Background Activities	These propositions provide extraneous details about background activities.
3. Flashback	These propositions portray events that happened prior to the story, but they are significant at a certain point in the story.
4. Setting	These propositions set the discourse stage.
5. Irrealis	These propositions describe events that might have happened, or could have happened but did not.
6. Author Intrusion	These propositions express the narrator's opinions.
7. Cohesion	These propositions serve to tie the story together.

Longacre proceeded to identify the grammatical mechanisms by which these bands are encoded in a wide variety of languages. For example, his salience scheme for English is shown below in Table 5.

Table 5. Longacre's Salience Scheme for English

Salience Band	English Grammatical Encoding Mechanism
1. Storyline	simple past tense verbs
2. Background Activities	past progressive *-ing* verbs

[3] The term "better" is intentionally very generic. We didn't want to ask the evaluators which text was more natural, or was easier to read, etc. Instead we let the evaluators choose whichever text they thought was better for any reason.

3. Flashback	*had* verbs (past perfect)
4. Setting	*be* clauses, active verbs with inanimate subjects
5. Irrealis	negatives and modals
6. Author Intrusion	
7. Cohesion	adverbial clauses and participial clauses

Taking this salience scheme into consideration, we analyze the VP *ran* as Time = Discourse, Aspect = none, and Salience Band = Storyline. The reason the author of this story used a past tense verb in this situation is because English speakers tell their narratives using past tense. However, that is not universal. Speakers of Banjar, an Austronesian language spoken in Indonesia, tell their narratives using present tense. Therefore if a past tense verb in an English text is translated with a past tense verb in Banjar, the result is unnatural. In LA's semantic representations, all of the verbs in a discourse are marked with a Time value of Discourse. Linguists are then able to link this value of Time to the appropriate tense in the target language. The reason the author of this sentence used a perfective verb is because English speakers portray storyline events with past perfective verbs. Linguists using LA are able to link the Salience Band value of Storyline to the appropriate grammatical mechanism in the target language. Every language will use a particular grammatical mechanism for indicating that an event is in the foreground rather than the background, and that grammatical mechanism may or may not be a past tense verb with perfective aspect.

For another example, consider the sentence *One day a girl named Kande* was sitting *near a tree.* Most analysts look at that VP and mark the verb's tense as past, and the verb's aspect as imperfective. But again we take the analysis one step further and ask: Why did the author use a past imperfective in this particular sentence? The reason is because the author wanted to put this event in the background rather than the foreground, and English speakers encode background events with past imperfective verbs. So we analyze this VP as Time = Discourse, Aspect = none, and Salience Band = Background Activity. Again, every language will have some grammatical mechanism for putting events into the background, so linguists using LA are able to write rules that will link the Salience Band value of Background Activity to the appropriate grammatical mechanism in the target language, and that may or may not be a past tense verb with imperfective aspect.

For a final example, consider the VP in the sentence *Kande's father* had slept *for many days.* Most analysts would mark that VP as past tense, perfect aspect. But we ask: Why did the author use a past perfect in this sentence? The answer is that this proposition is portraying an event that began or happened earlier, but is relevant at this particular point in the story. This is an example of Flashback, which English encodes by using the past perfect form of the verb. Therefore we analyze this VP as Time = Discourse, Aspect = None, Salience Band = Flashback. Then linguists building their grammars in LA are able to link the Flashback Salience Band to the appropriate grammatical mechanism of the target language to encode events that happened earlier, but are significant at a particular point in the narrative. Every language will have some grammatical mechanism for encoding flashback, but very few languages have anything resembling perfect aspect. Therefore this deep source analysis approach enables LA to generate texts that are natural in a wide variety of languages.

2.2 Participant Tracking

Many of the world's languages don't employ any articles (e.g, Korean), while other languages have a much richer article system than English. When we examined the discourse linguistic literature [8] regarding articles, we found that linguists attribute articles to a feature called "Participant Tracking." This feature was presented in Table 1 above, and is repeated here in Table 6.

Table 6. Participant Tracking

Noun Participant Tracking	First Mention, Routine, Interrogative, Frame Inferable, Exiting, Restaging, Generic, …

When a nominal is first mentioned in a discourse, English marks it with the indefinite article "*a.*" Subsequent references to that nominal are marked as definite with the article "*the.*" Frame inferable nouns are also marked with "*the*" in English as in the sentence "*The steering wheel on John's car needs to be replaced.*" In certain environments English doesn't mark its nominals with any article as in the sentence "*There are lions in Africa.*" These situations all correspond well to the feature values associated with Participant Tracking. Table 7 below lists the values of Participant Tracking and the associated English articles.

Table 7. English Articles and Participant Tracking

Participant Tracking Value	English Article
First Mention	*a, some*
Routine	*the*
Interrogative	*which*
Frame Inferable	*the*
Exiting	*the*
Restaging	*the*
Generic	(no article)

After incorporating the Participant Tracking feature into LA's semantic representational system, LA was able to generate the appropriate articles for a wide variety of languages.

2.3 Proximity

Most languages include deictics such as *this* and *that*. However, similar to the situation with articles, many languages have a much richer deictic system than English. Therefore LA's semantic representational system had to include the necessary information to generate the appropriate deictic markers for a wide variety of languages. The values of Proximity that are used in LA were listed above in Table 1, and are repeated here in Table 8.

Table 8. Proximity

Noun 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 without Focus

Table 9 below lists the values of Proximity and the associated English demonstratives.

Table 9. English Demonstratives and Proximity

Proximity Value	English Demonstrative
Near Speaker and Listener	*this, these*

Near Speaker	*this, these*
Near Listener	*that, those*
Remote within Sight	*that, those*
Remote out of Sight,	*that, those*
Temporally Near	*this, these*
Temporally Remote	*that, those*
Contextually Near with Focus	*this, these*
Contextually Near without Focus	*that, those*

2.4 Styles of Direct Speech

Many languages employ multiple styles of direct speech; when people talk to one another, their speech reflects their relative status. For example, Korean has six speech styles: 1) Plain, which is used most frequently, 2) Deferential, which is used when talking to an elder or an audience, 3) Polite, which is used when talking to someone you don't know, 4) Intimate, which is used when talking to someone you know well, 5) Familiar, which is used when talking to someone you know casually, and 6) Blunt, which is used when scolding a child or subordinate. Languages may indicate speech styles in a variety of ways such as adding an honorific morpheme to the verb, using honorific case markers, employing deferential pronouns, selecting honorific lexical forms, etc. The speech style is determined by the relative status of the speaker and listener, their ages, and the speaker's attitude toward the listener. English doesn't encode honorifics, so an analysis based solely on English surface structure can't possibly include this information. In order to accommodate languages that encode honorifics, five features were added to every proposition that is direct speech. These five features are summarized in Table 10 below.

Table 10. Direct Speech Features

Feature	Values
Speaker	Government Official, Religious Official, Father, Husband, Mother, Wife, man, woman, boy, girl, son, daughter, …
Listener	Government Official, Religious Official, Father, Husband, Mother, Wife, man, woman, boy, girl, son, daughter, …
Speaker's Attitude	Neutral, Familiar, Endearing, Honorable, Derogatory, Antagonistic, Angry, Rebuke, …
Speaker's Age	Child (0-17), Young Adult (18-24), Adult (25-49), Elder (50+)
Speaker-Listener Age	Older - different generation, Older - same generation, Essentially the same age, Younger - different generation, Younger - same generation

Linguists using LA are able to write rules which examine these speech features, and then set another feature called Speech Style to the appropriate value such as Plain, Deferential, Polite, Intimate, etc. Then subsequent rules look at the value of Speech Style, and add the appropriate morphology, make the appropriate lexical selections, etc.

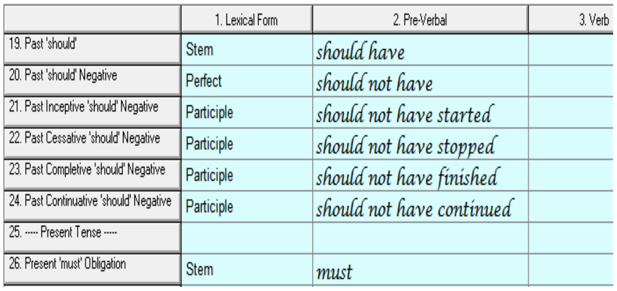
3. Shallow Target Analysis

An example of a Spellout Rule was shown above in Figure 7. The discussion of that rule mentioned that initially LA had four types of spellout rules: (i) Simple, (ii) Form Selection, (iii) Morphophonemic, and (iv) Table. With those four types of spellout rules, grammars were built for a variety of languages, but the task was complex. Numerous spellout rules were required to insert the aspectual auxiliaries, question auxiliaries, passive auxiliaries, salience auxiliaries, tense markers, mood markers, polarity markers, etc. The VP phrase structure rule was also quite complex because it had to order all of these constituents correctly. For example, consider the English sentence "*John should not have stopped running"* which is shown below in example (i):

(i)	John	should	not	have	stopp-ed	running.
		Mood	Polarity	Mood.Aux	Aspect-Tense	Verb

The VP consists of a mood marker *should*, a polarity marker *not*, a mood auxiliary *have*, an aspectual auxiliary *stop*, the tense marker *-ed*, and the semantically main verb in its participial form *running*. In earlier versions of LA, each of those constituents was inserted into the VP by a separate rule, and then the VP phrase structure rule ordered them properly. That approach worked, but it was difficult and required extensive knowledge of the target language.

Recently we discovered that we could generate the same high quality texts but with a much shallower analysis of the target language. To achieve this, we added a new type of spellout rule to LA's synthesizing grammar called "Phrase Builder." This new type of spellout rule is able to build entire phrases by simply inserting target language strings. For example, a single row in a Phrase Builder rule inserts the string *should not have stopped* into an English text whenever the verb in the semantic representation is marked as Time = Past, Aspect = Cessative, Mood = 'should', and Polarity = Negative. Part of the Phrase Builder rule that performs this insertion is shown below in Figure 11.



	1. Lexical Form	2. Pre-Verbal	3. Verb
19. Past 'should'	Stem	*should have*	
20. Past 'should' Negative	Perfect	*should not have*	
21. Past Inceptive 'should' Negative	Participle	*should not have started*	
22. Past Cessative 'should' Negative	Participle	*should not have stopped*	
23. Past Completive 'should' Negative	Participle	*should not have finished*	
24. Past Continuative 'should' Negative	Participle	*should not have continued*	
25. ----- Present Tense -----			
26. Present 'must' Obligation	Stem	*must*	

Figure 11. Part of a VP Phrase Builder Rule for English

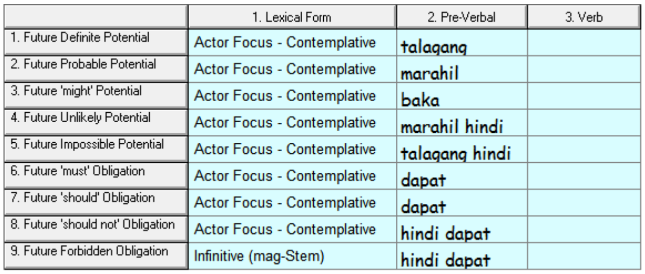
As seen in the figure above, row 22 applies to all verbs in the semantic representations marked as Past, Cessative, Negative, and 'should' mood. That row then inserts the string *should not have stopped* with the part of speech label "Pre-Verbal[4]," and selects the participle form of the verb from the lexicon. The VP phrase structure rule then simply positions the Pre-Verbal string before the verb.

An example showing part of a Phrase Builder rule for Tagalog VPs is shown below in Figure 12. There are many layers in that rule, and each layer contains multiple rows. The section shown in the figure inserts the appropriate Tagalog strings for various moods, and then selects the appropriate lexical form of the verb. The Tagalog equivalent of *John will definitely not walk* is shown in example (ii).

[4] In Phrase Builder rules, the column's name is used as the part of speech tag, and the user is able to specify the column names. We suggest keeping the column names very simple such as "Pre-Verbal," "Post-Verbal," etc. These column names then appear in the VP phrase structure rule which is shown in Figure 16.

(ii) Talagang hindi magla-lakad si John.
definitely not Future-walk Abs John
'John will definitely not walk.'

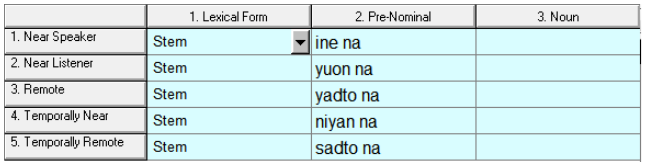
As seen in Figure 12 below, row 5 applies to all verbs marked as Future, Impossible Potential. That row selects the Actor Focus Contemplative form of the verb from the lexicon, and inserts the string *talagang hindi* into the VP with the label Pre-Verbal. Then the VP phrase structure rule positions that string before the verb.



	1. Lexical Form	2. Pre-Verbal	3. Verb
1. Future Definite Potential	Actor Focus - Contemplative	talagang	
2. Future Probable Potential	Actor Focus - Contemplative	marahil	
3. Future 'might' Potential	Actor Focus - Contemplative	baka	
4. Future Unlikely Potential	Actor Focus - Contemplative	marahil hindi	
5. Future Impossible Potential	Actor Focus - Contemplative	talagang hindi	
6. Future 'must' Obligation	Actor Focus - Contemplative	dapat	
7. Future 'should' Obligation	Actor Focus - Contemplative	dapat	
8. Future 'should not' Obligation	Actor Focus - Contemplative	hindi dapat	
9. Future Forbidden Obligation	Infinitive (mag-Stem)	hindi dapat	

Figure 12. Part of a VP Phrase Builder Rule for Tagalog

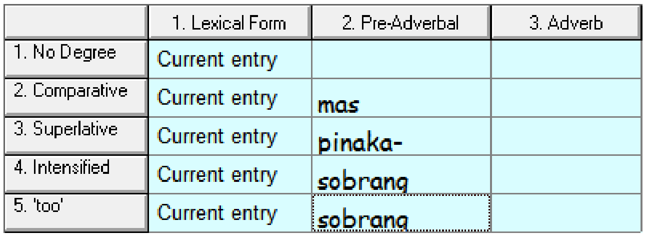
Similar Phrase Builder rules are used to build entire noun phrases with the appropriate articles and demonstratives. An example showing the Phrase Builder rule for Bisakol's demonstratives is shown below in Figure 13. As seen in that figure, Bisakol's demonstrative system is considerably richer than the English demonstrative system.



	1. Lexical Form	2. Pre-Nominal	3. Noun
1. Near Speaker	Stem	ine na	
2. Near Listener	Stem	yuon na	
3. Remote	Stem	yadto na	
4. Temporally Near	Stem	niyan na	
5. Temporally Remote	Stem	sadto na	

Figure 13. Part of a Bikasol NP Phrase Builder Rule

A final example is shown in Figure 14. That figure shows part of the Phrase Builder rule that builds Tagalog adverbial phrases. Each row in that layer inserts the appropriate degree modifier into the adverbial phrase.



	1. Lexical Form	2. Pre-Adverbal	3. Adverb
1. No Degree	Current entry		
2. Comparative	Current entry	mas	
3. Superlative	Current entry	pinaka-	
4. Intensified	Current entry	sobrang	
5. 'too'	Current entry	sobrang	

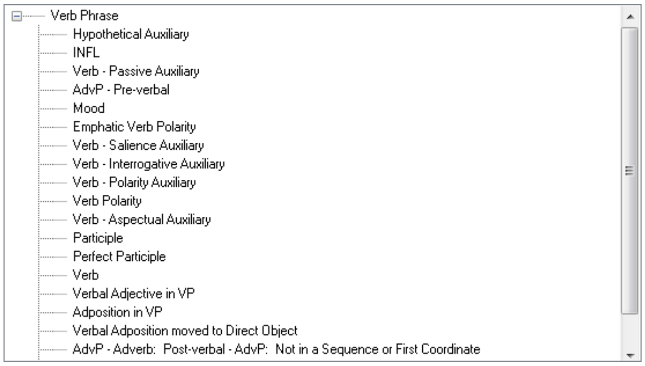
Figure 14. Part of a Tagalog AdvP Phrase Builder Rule

The Tagalog equivalent of "*John walked the most quickly*" is shown in example (iii).

(iii) Nag-lakad nang pinaka-mabilis si John.
past-walk how superlative-quickly Abs John
'John walked the most quickly.'

Row 3 in the rule shown above applies to all adverbs with a Degree value of Superlative. That row inserts the string *pinaka-* into the adverbial phrase, and the phrase structure rule for adverbial phrases positions that string before the adverb. Since that string ends with a dash ('-'), that string will attach to the word that follows it.

The use of Phrase Builder rules has significantly simplified the phrase structure rules. Figure 15 below shows the primary phrase structure rule for English VPs before Phrase Builder rules were introduced.

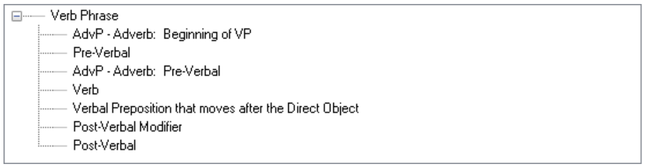


- Verb Phrase
  - Hypothetical Auxiliary
  - INFL
  - Verb - Passive Auxiliary
  - AdvP - Pre-verbal
  - Mood
  - Emphatic Verb Polarity
  - Verb - Salience Auxiliary
  - Verb - Interrogative Auxiliary
  - Verb - Polarity Auxiliary
  - Verb Polarity
  - Verb - Aspectual Auxiliary
  - Participle
  - Perfect Participle
  - Verb
  - Verbal Adjective in VP
  - Adposition in VP
  - Verbal Adposition moved to Direct Object
  - AdvP - Adverb: Post-verbal - AdvP: Not in a Sequence or First Coordinate

Figure 15. English VP Phrase Structure Rule before Phrase Builder Rules

As seen in the rule above, the English VPs contain numerous constituents which must be ordered correctly. In addition to the phrase structure rule shown above, there were nine other VP phrase structure rules that specified the constituent order for very specific verb phrases (e.g., negated imperfectives, passive flashbacks, etc.).

After Phrase Builder rules were added to LA, the English VP phrase structure rule was simplified to that shown in Figure 16. That rule is the only phrase structure rule now required for English VPs.



- Verb Phrase
  - AdvP - Adverb: Beginning of VP
  - Pre-Verbal
  - AdvP - Adverb: Pre-Verbal
  - Verb
  - Verbal Preposition that moves after the Direct Object
  - Post-Verbal Modifier
  - Post-Verbal

Figure 16. English VP Phrase Structure Rule after Phrase Builder

The use of Phrase Builder rules has significantly simplified the target language grammars, but yet the generated texts are still of very high quality. This new type of rule enables linguists to build their target grammars more quickly with a much shallower analysis of the target language.

4. Conclusions

This paper has provided a brief introduction to the multilingual NLG called Linguist's Assistant. The ontology, semantic representations, lexicon, transfer grammar, and synthesizing grammar were each briefly described. Then several of the techniques developed for building semantic representations of source documents were presented. In particular, the salience scheme developed by Robert Longacre has proven tremendously helpful for analyzing the source documents while building the semantic representations. Including the Salience Band information in LA's semantic representations has significantly increased the naturalness of the texts generated by LA. This paper also presented a new type of spellout rule that was recently added to LA’s synthesizing grammar. This rule type is called “Phrase Builder” because it enables linguists to build entire phrases with simple target language text. Prior to Phrase Builder rules, numerous spellout rules were required to insert the many constituents required in VPs, NPs, etc. Then complex phrase structure rules were required to order the constituents correctly. Phrase Builder rules have significantly simplified the grammar development process in LA.

5. REFERENCES

[1] Allman, Tod. 2010. The Translator’s Assistant: A Multilingual Natural Language Generator based on Linguistic Universals, Typologies, and Primitives. Arlington, Texas: University of Texas dissertation.

[2] Beale, Stephen, and Tod Allman. 2011. Linguist’s Assistant: a Resource for Linguists. In Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP-11), The 9th Workshop on Asian Language Resources, Chiang Mai, Thailand.

[3] Cann, Ronnie. 1993. Formal Semantics, Cambridge: Cambridge University Press.

[4] Goddard, Cliff. 2008. Cross-linguistic Semantics. Amsterdam: John Benjamins.

[5] Jackendoff, Ray. 1990. Semantic Structures. Cambridge, Massachusetts: The MIT Press.

[6] Lakoff, George. 1987. Women, Fire, and Dangerous Things. Chicago: University of Chicago Press.

[7] Langacker, Ronald. 1986. Foundations of Cognitive Grammar. Vol 1. Stanford: Stanford University Press.

[8] Longacre, Robert. 1996. The Grammar of Discourse. 2nd ed. New York: Plenum Press.

[9] 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.

[10] 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.

[11] Nirenburg, Sergei, and Victor Raskin. 2004. Ontological Semantics. Cambridge, Massachusetts: The MIT Press.

[12] Summers, Della, et al. 2003. Longman Dictionary of Contemporary English. Edinburgh, England: Pearson Education Limited.