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.

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 Meta-language 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.”

![Figure 2. Example of KANT’s Interlingua](image)

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>
<thead>
<tr>
<th>Feature</th>
<th>Possible Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun Number</td>
<td>Singular, Dual, Trial, Quadrial, Plural, Paucal</td>
</tr>
<tr>
<td>Noun Participant Tracking</td>
<td>First Mention, Routine, Interrogative, Frame Inferable, Exiting, Restaging, Generic,…</td>
</tr>
<tr>
<td>Noun Proximity</td>
<td>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</td>
</tr>
<tr>
<td>Event Time</td>
<td>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, …</td>
</tr>
<tr>
<td>Proposition Illocutionary Force</td>
<td>Declarative, Imperative, Content Interrogative, Yes-No Interrogative</td>
</tr>
<tr>
<td>Proposition Salience Band</td>
<td>Pivotal Storyline, Script Predictable Actions, Backgrounded Actions, Flashback, Setting, Irrealis, Evaluation, Cohesive Material</td>
</tr>
<tr>
<td>Noun Phrase Semantic Role</td>
<td>Agent, Patient, State, Source, Destination, Instrument, Beneficiary, Addressee</td>
</tr>
</tbody>
</table>

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](image)
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.

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.

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.
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, 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.

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.

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.

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 text languages: Jula, Kewa, and Korean.

Table 2. Summary of Productivity Experiments

<table>
<thead>
<tr>
<th>Language</th>
<th>Ratio of Edited Words to Manually Translated Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jula</td>
<td>4.3</td>
</tr>
<tr>
<td>Kewa</td>
<td>6.7</td>
</tr>
<tr>
<td>Korean</td>
<td>4.6</td>
</tr>
</tbody>
</table>

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 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

<table>
<thead>
<tr>
<th>Language</th>
<th>LA’s Texts</th>
<th>Manual Texts</th>
<th>Equal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jula</td>
<td>12</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td>Korean</td>
<td>88</td>
<td>71</td>
<td>33</td>
</tr>
</tbody>
</table>

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-Indo-European 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

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

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

<table>
<thead>
<tr>
<th>Salience Band</th>
<th>English Grammatical Encoding Mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Storyline</td>
<td>simple past tense verbs</td>
</tr>
<tr>
<td>2. Background</td>
<td>past progressive -ing verbs</td>
</tr>
<tr>
<td>Activities</td>
<td></td>
</tr>
</tbody>
</table>

---

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.
There are one, and de the necessary information, and h Par
s
participant Tracking.  Table.  But the
ast, and the verb's
ground, and that
The
eaker and Listener, Near Speaker,
sh encodes by using the past
story used a past tense
thor of this sentence used a
variety of languages.
ner events, and that
mechanism of the target language to encode events that happened
Flashback Salience Band to the appropriat
linguists building their grammars in LA are able to link the
Salience Band value of Background Activity to the
background, so linguists using LA are able to write rules that will
have
some grammatical mechanism for putting events into 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. 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 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.

| 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
run 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.

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.

| 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>
<thead>
<tr>
<th>Participant Tracking Value</th>
<th>English Article</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Mention</td>
<td>a, some</td>
</tr>
<tr>
<td>Routine</td>
<td>the</td>
</tr>
<tr>
<td>Interrogative</td>
<td>which</td>
</tr>
<tr>
<td>Frame Inferable</td>
<td>the</td>
</tr>
<tr>
<td>Exiting</td>
<td>the</td>
</tr>
<tr>
<td>Restaging</td>
<td>the</td>
</tr>
<tr>
<td>Generic</td>
<td>(no article)</td>
</tr>
</tbody>
</table>

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.

| 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>
<thead>
<tr>
<th>Proximity Value</th>
<th>English Demonstrative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Near Speaker and Listener</td>
<td>this, these</td>
</tr>
</tbody>
</table>
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>
<thead>
<tr>
<th>Speaker’s Attitude</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government Official, Religious Official, Father, Husband, Mother, Wife, man, woman, boy, girl, son, daughter, …</td>
<td></td>
</tr>
<tr>
<td>Neutral, Familiar, Endearing, Honorable, Derogatory, Antagonistic, Angry, Rebuke, …</td>
<td></td>
</tr>
<tr>
<td>Child (0-17), Young Adult (18-24), Adult (25-49), Elder (50+)</td>
<td></td>
</tr>
<tr>
<td>Older - different generation, Older - same generation, Essentially the same age, Younger - different generation, Younger - same generation</td>
<td></td>
</tr>
</tbody>
</table>

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 stopped 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.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government Official, Religious Official, Father, Husband, Mother, Wife, man, woman, boy, girl, son, daughter, …</td>
<td></td>
</tr>
<tr>
<td>Neutral, Familiar, Endearing, Honorable, Derogatory, Antagonistic, Angry, Rebuke, …</td>
<td></td>
</tr>
<tr>
<td>Child (0-17), Young Adult (18-24), Adult (25-49), Elder (50+)</td>
<td></td>
</tr>
<tr>
<td>Older - different generation, Older - same generation, Essentially the same age, Younger - different generation, Younger - same generation</td>
<td></td>
</tr>
</tbody>
</table>

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,” 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).

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.

![Figure 12. Part of a VP Phrase Builder Rule for Tagalog](image)

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.

![Figure 13. Part of a Bisakol NP Phrase Builder Rule](image)

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.

![Figure 14. Part of a Tagalog AdvP Phrase Builder Rule](image)

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.

![Figure 15. English VP Phrase Structure Rule before Phrase Builder Rules](image)

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.

![Figure 16. English VP Phrase Structure Rule after Phrase Builder](image)

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