Linguist's Assistant: A Multi-Lingual Natural Language Generator based on Linguistic Universals, Typologies, and Primitives

Tod Allman
Graduate Institute of Applied Linguistics
7500 W. Camp Wisdom Rd.
Dallas, TX 75236
tod_allman@gial.edu

Stephen Beale
University of Maryland, Baltimore County
1000 Hilltop Circle
Baltimore, MD 21250
sbeale@csee.umbc.edu

Richard Denton
Dartmouth College
6127 Wilder Lab
Hanover, NH 03755
richard.e.denton@dartmouth.edu

Abstract

Linguist's Assistant (LA) is a large scale semantic analyzer and multi-lingual natural language generator 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), Jula (Cote d'Ivoure), and North Tanna (Vanuatu), and proof-of-concept lexicons and grammars have been developed for Spanish, Urdu, Tagalog, Chinantec (Mexico), and Angas (Nigeria). This paper will summarize the major components of the NLG system, and then present the results of experiments that were performed to determine the quality of the generated texts. The experiments indicate that when experienced mother-tongue translators use the drafts generated by LA, their productivity is typically quadrupled without any loss of quality.

1 Introduction

The fundamental goal underlying LA was to develop a system capable of generating high quality texts in a wide variety of languages, particularly minority and endangered languages. Drafts produced by LA are always easily understandable, grammatically correct, semantically equivalent to the source documents, and at approximately a sixth grade reading level. Because the system is based on linguistic research, LA is expected to work well for typologically diverse languages; it works equally well for languages that are coranking or clause chaining, highly isolating or highly polysynthetic, fusional or agglutinative, etc. A natural language generator of this type is practical only when translating large quantities of texts into many different languages. Therefore semantic representations for a large variety of texts are being developed for LA. This system is a tool which enables linguists to document a language and simultaneously generate numerous texts for the speakers of that language. A model of LA is shown in Figure 1.

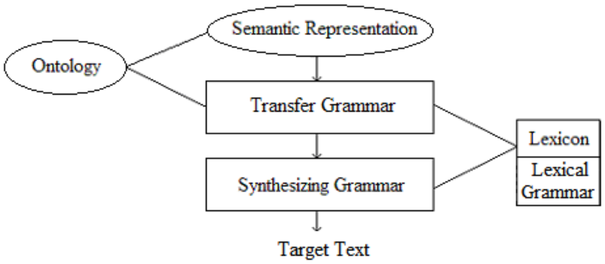


Figure 1. Model of Linguist's Assistant

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

INLG 2012 Proceedings of the 7th International Natural Language Generation Conference, pages 59–66,
Utica, May 2012. ©2012 Association for Computational Linguistics

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.

2 The Ontology

One of the foundational principles of Natural Semantic Metalanguage theory (Goddard & Wierzbicka, 1994; Wierzbicka, 1996) proposes that there is a small set of innate concepts which are present in every language. These innate concepts can be used to explicate every word in every language. If semantic representations were developed using only these innate primitives, the problem of lexical mismatch between languages would be eliminated. However, building semantic representations using only the innate concepts is unwieldy, so semantically simple molecules were identified in a principled manner. For our semantic molecules, we elected to use the defining vocabulary in Longman's Contemporary English Dictionary (2003). By using these semantically simple concepts, the problem of lexical mismatch between source and target languages is significantly reduced. There are certainly still instances of lexical mismatch, and we have an approach for dealing with them which will be described below. LA also permits the automatic insertion of semantically complex concepts into the semantic representations, but only if the linguist indicates that the target language has a lexical equivalent.

3 LA's Semantic Representational System

The development of an adequate method of meaning representation for LA's source texts proved to be a challenge. Formal semantics (Cann, 1993; Rosner, 1992), conceptual semantics (Jackendoff, 1990) and generative semantics (Lakoff, 1987) 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 earlier. 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.

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, 1996)	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

Table 1. Several Features and their Values

The semantic representation for "Paulus started walking from the market to a village named Terpen" is shown in Figure 2.

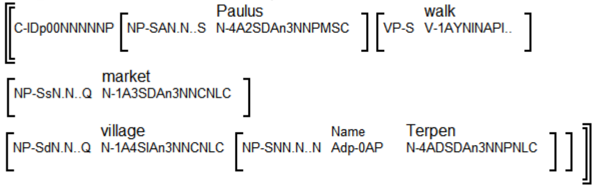


Figure 2. LA's Semantic Representational System

As seen in Figure 2, 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.

4 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.

5 LA's Transfer Grammar

The purpose of LA's transfer grammar[1] is to restructure the English influenced semantic representations in order to produce a new underlying representation that is appropriate for the target language. This new underlying representation consists of the target language's words, structures, and features. For example, many languages have rules that are based on grammatical relations, but the object 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 languages in the world 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 in Figure 3. The transfer grammar consists of nine different types of rules, several of which will be briefly described below.

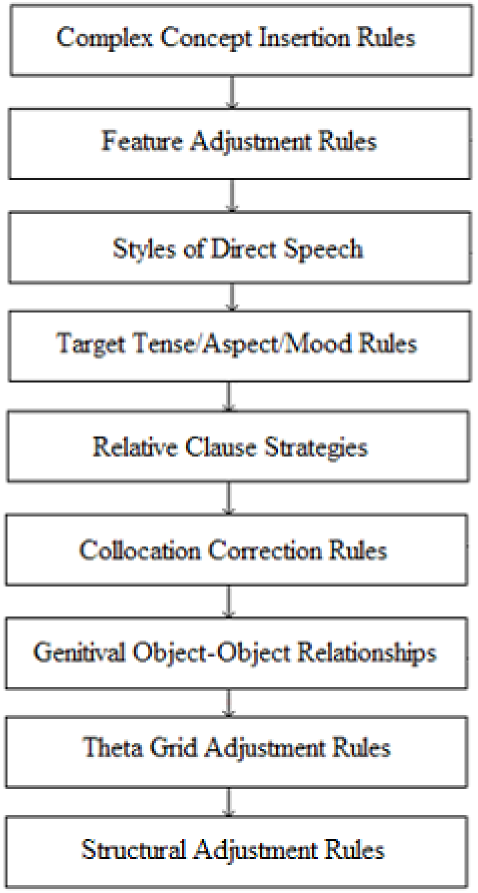

Complex Concept Insertion Rules
Feature Adjustment Rules
Styles of Direct Speech
Target Tense/Aspect/Mood Rules
Relative Clause Strategies
Collocation Correction Rules
Genitival Object-Object Relationships
Theta Grid Adjustment Rules
Structural Adjustment Rules

Figure 3. Model of LA's Transfer Grammar

Complex Concept Insertion Rules: These rules are prebuilt for specific complex concepts and may be activated by the user if his target language has a lexical equivalent for a particular complex concept. For example, the concept *blind* is semantically complex and is not permitted in the semantic representations. Whenever the adjective "blind" is used attributively in a source document, it is replaced in the semantic representations with the relative clause *who is not able to see*. But if the target language has a lexical equivalent for *blind*, the user can activate the complex concept insertion rule which will replace all occurrences of *who is not able to see* in the semantic representations with *blind.*

Styles of Direct Speech: Many languages employ techniques for indicating relative status when two people talk to one another. Therefore in the semantic representations all propositions that are di-

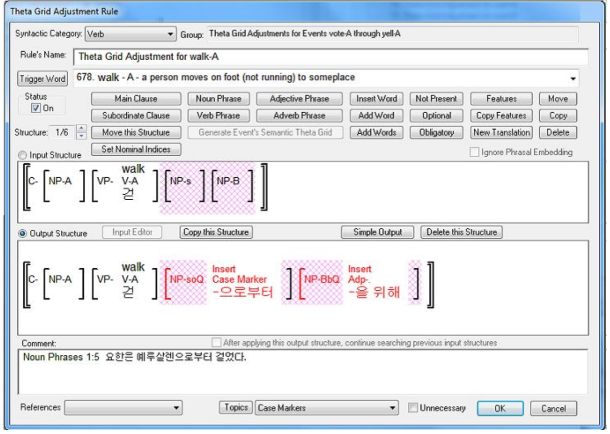
[1] The translation process is often divided into three fundamental steps: 1) analysis: analyze the source document to determine its meaning, 2) transfer: reconstruct that meaning using the target language's lexemes, structures, and world view, and 3) synthesis: synthesize the final surface forms. The term "Transfer Grammar" here refers to the grammar in LA that performs the second step of the translation process.

rect speech are marked with five features indicating: 1) the general category of the speaker (e.g., father, mother, child, political leader, religious leader, employer, employee, etc.), 2) the general category of the listener, 3) the speaker's attitude, 4) the speaker's approximate age, and 5) the age of the speaker relative to the listener. Linguists are able to define the styles of direct speech that are pertinent to the target language, and then use these features and rules to set the style appropriately. Subsequent rules then insert the appropriate pronouns or honorific morphology to indicate the relative status of the speaker to the listener.

Relative Clause Strategies: Extensive typological research has been done regarding relative clauses (Comrie 1989:138, Givón 1990:645), and linguists have found that languages apply a limited number of strategies to a limited number of grammatical relations in what is commonly called the NP Accessibility Hierarchy (Keenan & Comrie 1977, Comrie 1989:156). Cross-linguistically relative clauses may be classified as either embedded or adjoined. If a language uses embedded relative clauses, they may be pre-nominal, post-nominal, or circum-nominal. If a language uses adjoined relative clauses, they are either sentence initial or sentence final. There are generally three strategies for encoding the coreferential noun in a relative clause: the gap strategy, the pronoun retention strategy, and the relative pronoun strategy. The relative clause rules in LA enable a linguist to describe what types of relative clauses are employed in his target language, and which strategies are used at the various positions in the Accessibility Hierarchy.

Collocation Correction Rules: Collocation deals with how certain words go together, and how words and phrases co-occur with certain grammatical choices. Every word in every language has its own collocational range and restrictions. Therefore collocation correction rules are used to change one target word to another target word in a particular environment. For example, in English a king *wears* his crown, but in Korean, a king 쓰다 [sseu da] *uses* his crown. So a collocation correction rule will change the Korean verb 입다 [ip da] 'to wear' to 쓰다 [sseu da] 'to use' whenever the agent is a *king* or *queen* and the patient is a *crown.*

Theta Grid Adjustment Rules: Every verb in every language has an associated theta grid which describes its 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 easily 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 4. That rule inserts the appropriate Korean postpositions into the source and destination 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
Comment: Noun Phrases 1:5 요한은 예루살렘으로부터 걸었다.
☐ After applying this output structure, continue searching previous input structures
☐ Ignore Phrasal Embedding
References:
Topics: Case Markers
☐ Unnecessary

Figure 4. Korean Theta Grid Adjustment Rule

Structural Adjustment Rules: The structural adjustment rules are used to restructure the semantic representations in any way that's necessary in order to construct an appropriate underlying representation for the target language. These rules may be used to handle lexical mismatch, convert predicate adjective constructions to verbal constructions, build clause chains from coranking propositions, make adjustments for various views of time, etc. The structural adjustment rules look identical to the theta grid adjustment rule shown in Figure 4, but they are grouped separately because they perform a variety of tasks.

The final product of the transfer grammar is a new underlying representation that is appropriate for the target language. This underlying representation consists of the target language's words, structures, and features. This underlying representation serves as the input to the synthesizing grammar.

6 LA's Synthesizing Grammar

LA's synthesizing grammar is responsible for synthesizing the final surface forms of the target text.

LA's 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 observe the capabilities that are required to synthesize surface text. A model of the final result is shown in Figure 5, and several of these rule types will be briefly described below.

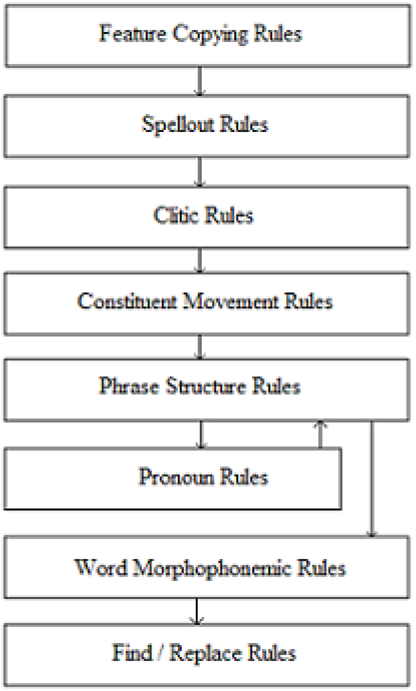
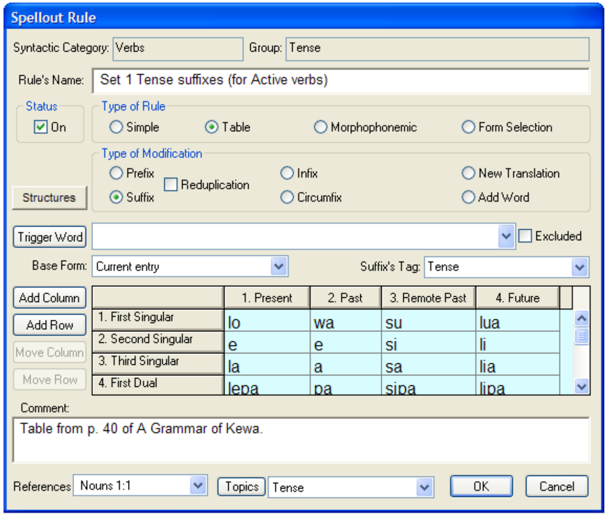


Figure 5. Model of LA's Synthesizing Grammar

Feature Copying Rules: The feature copying rules copy features from one constituent to another constituent so that the spellout rules can add the necessary morphology to indicate appropriate agreement. For example, certain Jula nouns agree in person and number with their object nouns, so a feature copying rule copies the person and number of the object noun to the verb. Then a spellout rule adds the appropriate morphology to the verb.

Spellout Rules: The spellout rules add contextual morphology in order to synthesize the final form of each target word. There are four basic types of spellout rules: (i) simple spellout rules which add a prefix, suffix, infix, circumfix, 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 in Figure 6.



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

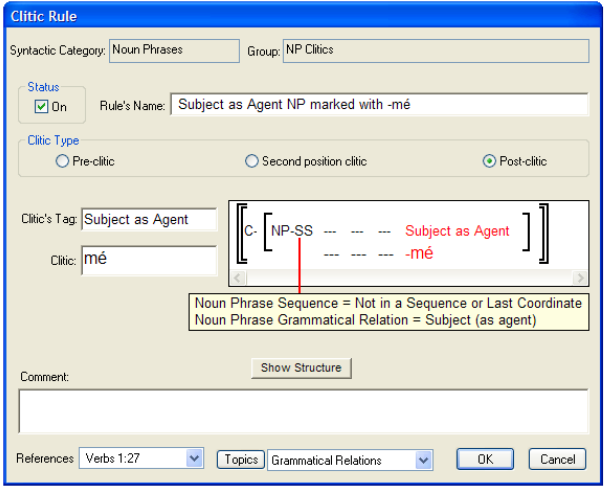
	1. Present	2. Past	3. Remote Past	4. Future
1. First Singular	lo	wa	su	lua
2. Second Singular	e	e	si	li
3. Third Singular	la	a	sa	lia
4. First Dual	lepa	pa	sipa	lipa

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

References: Nouns 1:1 Topics: Tense

Figure 6. Spellout Rule that adds Kewa Tense Affixes

Clitic Rules: Linguists have found that languages employ three different types of clitics (Payne, 1997): (i) pre-clitics which attach to the beginning of the first word in a phrase, (ii) second position clitics which attach to the end of the first word in a phrase, and (iii) post-clitics which attach to the end of the last word in a phrase. A clitic rule that adds the post-clitic *–me* to Kewa subjects is shown in Figure 7.



Clitic Rule

Syntactic Category: Noun Phrases Group: NP Clitics

Status: ☑ On

Rule's Name: Subject as Agent NP marked with -mé

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

Clitic's Tag: Subject as Agent

Clitic: mé

[[C- [NP-SS --- --- --- Subject as Agent
--- --- --- -mé]]]

Noun Phrase Sequence = Not in a Sequence or Last Coordinate
Noun Phrase Grammatical Relation = Subject (as agent)

Comment:

References: Verbs 1:27 Topics: Grammatical Relations

Figure 7. A Clitic Rule for Kewa

Pronoun Rules: There are no pronouns in the semantic representations because each language has its own rules for determining when and where pronouns are appropriate. Therefore, after the phrase structure rules have moved each constituent into its final position, the pronoun rules identify the nominals that should be realized with pronouns, and then supply the appropriate surface forms.

Word Morphophonemic Rules: The word morphophonemic rules are similar to the morphophonemic rules described in the spellout rule section above, but these morphophonemic rules operate across word boundaries rather than morpheme boundaries. For example, the English indefinite article *a* changes to *an* whenever the next word begins with a vowel.

7 LA's Target Text

After the synthesizing grammar has been executed and produced the final form of the target text, mother-tongue speakers edit the text to improve the naturalness and information flow. Samples of English and Korean texts generated by LA are shown 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

8 LA's Results

Extensive grammars and lexicons were developed for English, Korean, Kewa, and Jula. We began each project by working through a set of sentences called the Grammar Introduction. The Grammar Introduction consists of approximately 500 basic sentences, each illustrating a particular feature or construction of the semantic representational system. For example, the Grammar Introduction includes a series of propositions dealing with the various tenses, aspects, and moods, there's a set of propositions dealing with relative clauses, object complement clauses, and adverbial clauses, another set of propositions dealing with pronouns, etc. After completing the Grammar Introduction[2], a very thorough foundation has been developed for the lexicon and grammar, but the Grammar Introduction is intentionally restricted to a very small set of concepts. Therefore rules that deal with concept-specific issues must be dealt with while working through actual texts. While working through the semantic representations of these texts, a very clear trend developed for each of the test languages: the number of new grammatical rules required per chapter of text decreased very quickly as seen in Figures 9 through 12.

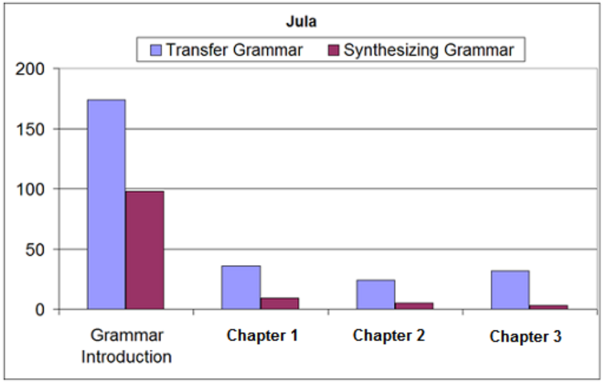


Figure 9. Graph of New Rules for Jula

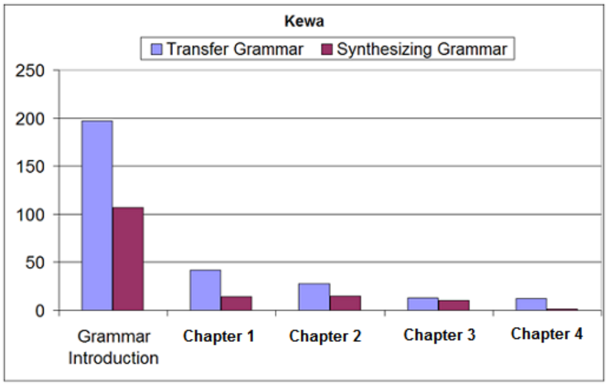


Figure 10. Graph of New Rules for Kewa

[2] For each test language, working through the Grammar Introduction took approximately 40 to 50 hours.

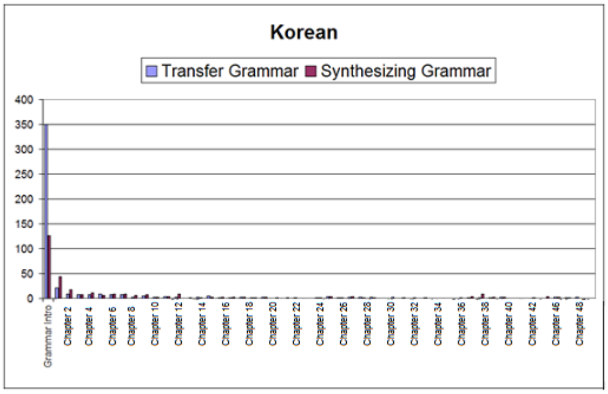


Figure 11. Graph of New Rules for Korean

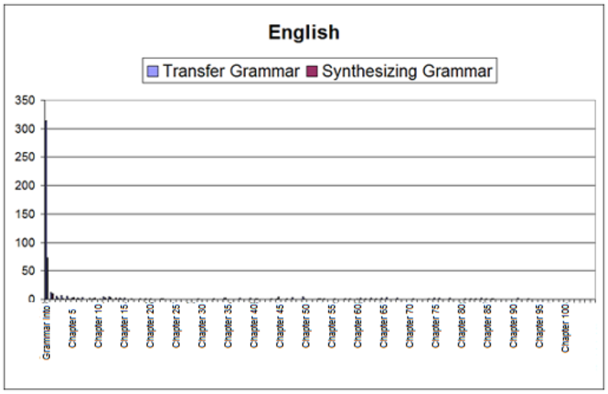


Figure 12. Graph of New Rules for English

The graphs shown above conclusively demonstrate that the grammars developed in LA are accurately capturing the significant generalizations of these four languages.

9 Quality of Generated Texts

After generating texts in Korean, Kewa, and Jula, experiments were performed to determine whether or not the drafts generated by LA are of sufficient quality that they improve the productivity of experienced mother-tongue translators. Two sets of experiments were performed: the first set tested for increased productivity, and the second set tested for quality. The first set compared the quantity of text an experienced translator could translate in a given period of time with the quantity of text generated by LA that the same person could edit in the given time. Eight professional mother-tongue translators participated in the Jula experiment, one translator participated in the Kewa experiment, and eighteen translators participated in the Korean experiment. In these experiments, quantity was determined by word count. Table 2 summarizes the results of these productivity experiments.

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

Table 2. Summary of Productivity Experiments

The table shown above indicates that in each test language, the drafts generated by LA were of such high quality that they more than quadrupled the productivity of experienced mother-tongue translators. The results of these experiments were certainly encouraging, but at this point we didn't know whether or not the editors had done a thorough job of editing the generated texts. Therefore we performed another set of experiments to determine whether or not the edited texts were comparable in quality with professionally translated texts.

10 Quality of Edited Texts

The second set of experiments was performed with Jula and Korean speakers in order to determine the quality of the edited LA drafts. Speakers of these languages were asked to compare the edited LA texts with the manually translated texts. These evaluations were performed by people who did not know how either of the texts had been produced. Forty evaluations were performed by Jula speakers, and 192 evaluations were performed by Korean speakers. Although no evaluations were performed by Kewa speakers, the edited Kewa draft was ultimately published. Table 3 summarizes the Jula and Korean evaluations.

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

Table 3. Summary of the Evaluation Experiments

In Table 3, the column labeled "LA Texts" indicates the number of evaluators who said that the edited LA text was better[3] than the manually translated text, the column labeled "Manual Texts" indicates the number of evaluators who said the manually translated text was better than the edited LA text, and the column labeled "Equal" indicates the number of evaluators who said that the edited LA text was equal in quality to the manually trans-

[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.

lated text. In both languages the evaluation experiments indicate that the edited LA texts are considered as good as the manually translated texts.

11 Conclusions

LA is a tool which drastically reduces the amount of time and effort required to produce an initial draft of a translation of a text. This tool enables linguists to build large scale lexicons and grammars for a very wide variety of languages, particularly minority and endangered languages. After a lexicon and grammar have been completed, LA generates drafts of texts which are at approximately a sixth grade reading level. We hope to eventually have a large library of community development texts which will describe how to prevent the spread of various diseases such as AIDS, Avian Influenza, etc. This tool works equally well for languages that are thoroughly studied, languages that have only slightly been studied, and languages that are endangered. Similarly, this tool works equally well for languages that are typologically diverse with respect to their morphological and syntactic features. It is hoped that this tool will empower speakers of minority languages around the world by providing them with translations of vital information, which will not only enable them to live longer, healthier, and more productive lives, but will also enable them to participate in the larger world.

References

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

Allman, Tod, and Stephen Beale. 2006. A Natural Language Generator for Minority Languages. In Proceedings of Speech and Language Technology for Minority Languages (SALTMIL). Genoa, Italy.

Beale, Stephen. In print. Documenting Endangered Languages using Linguist’s Assistant. Language Documentation and Conservation. Draft available at http:/onyxcons.com/LA/

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.

Cann, Ronnie. 1993. *Formal Semantics.* Cambridge: Cambridge University Press.

Givón, Talmy. 1990. *Syntax: A Functional-Typological Introduction,* 2 vols. Amsterdam: John Benjamins.

Goddard, Cliff, and Anna Wierzbicka. 1994. *Semantic and Lexical Universals: Theory and Empirical Findings.* Amsterdam: John Benjamins.

Jackendoff, Ray. 1990. *Semantic Structures.* Cambridge, Massachusetts: The MIT Press.

Lakoff, George. 1987. *Women, Fire, and Dangerous Things.* Chicago: University of Chicago Press.

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

Payne, Thomas. 1997, *Describing Morphosyntax.* Cambridge: Cambridge University Press.

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