# The Bible Translator's Assistant:
# A Multilingual Natural Language Generator Based on Linguistic Universals and Typologies

Tod Allman
Linguistics Department
University of Texas at Arlington
todallman@yahoo.com

## 1. Introduction

The Bible Translator's Assistant (TBTA) is a natural language generator (NLG) designed specifically for linguists doing translation work in a very wide variety of languages. In particular, TBTA is intended to generate drafts of the entire Bible and numerous community development articles in the world's 3000+ minority languages. TBTA uses the rich interlingua approach. The semantic representations developed for TBTA consist of a controlled English based metalanguage augmented by a feature system designed to accommodate a very wide range of languages. The grammar in TBTA consists of two parts: a transfer grammar and a synthesizing grammar. The transfer grammar restructures the semantic representations in order to produce a new underlying representation that is appropriate for a particular target language. Then the synthesizing grammar synthesizes the final surface forms. To date TBTA has been tested with four languages: English, Korean, Jula (Cote d'Ivoire), and Kewa (Papua New Guinea). Experiments with the Jula text indicate that TBTA's rough drafts triple the productivity of professional mother tongue translators without any loss of quality, and experiments with the Korean text indicate that TBTA's drafts quadruple the productivity of experienced mother tongue translators. A model of TBTA is shown below in Figure 1.

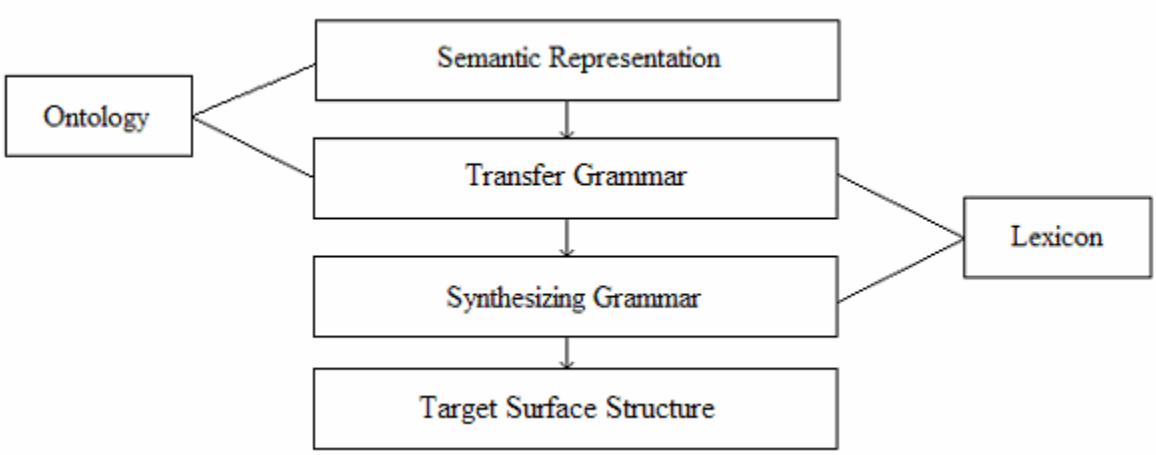


Figure 1. Underlying model of The Bible Translator's Assistant

## 2. The Semantic Representations

The development of an adequate method of meaning representation for TBTA's source texts proved to be a challenge. Both formal semantics and conceptual semantics were each considered but found inadequate. Using the foundational principles of Natural Semantic Metalanguage theory, a set of semantically simple English molecules was identified in a principled manner. These semantic molecules serve as the primary lexemes in TBTA's ontology. The ontology also includes semantically complex lexemes, but each of those lexemes

has an associated insertion rule that automatically inserts the complex concept only if the target language has a lexical semantic equivalent.

The feature set developed for TBTA encodes semantic, syntactic and discourse information. Each feature is an exhaustive etic list of the values pertinent to the world's languages. For example, each nominal is marked for Number, and the possible values are Singular, Dual, Trial, Quadrial and Plural. Each of these values is necessary because some languages morphologically distinguish all five of these categories. Examples of some of the features and their values are listed below in Tables 1 through 5.

Table 1. Partial listing of the Features for Things (Nominals)

| | |
|---|---|
| Number | Singular, Dual, Trial, Quadrial, Plural |
| Participant Tracking | First Mention, Integration, Routine, Exiting, Offstage, Restaging, Generic, Interrogative, Frame Inferable |
| Polarity | Affirmative, Negative |
| Proximity | Near Speaker and Listener, Near Speaker, Near Listener, Remote within sight, Remote out of sight, Temporally Near, Temporally Remote, Contextually Near, Contextually Remote, Not Applicable |
| Person | First, Second, Third, First & Second, First & Third, Second & Third, First & Second & Third |
| Participant Status | Protagonist, Antagonist, Major Participant, Minor Participant, Major Prop, Minor Prop, Significant Location, Insignificant Location, Significant Time, Not Applicable |

Table 2. Partial listing of the Features for Events (Verbs)

| | |
|---|---|
| Time | Discourse, Present, Immediate Past, Earlier Today, Yesterday, 2 days ago, 3 days ago, a week ago, a month ago, a year ago, During Speaker's lifetime, Historic Past, Eternity Past, Unknown Past, Immediate Future, Later Today, Tomorrow, 2 days from now, 3 days from now, a week from now, a month from now, a year from now, Unknown Future, Timeless |
| Aspect | Discourse, Habitual, Imperfective, Progressive, Completive, Inceptive, Cessative, Continuative, Gnomic |
| Mood | Indicative, Definite Potential, Probable Potential, 'might' Potential, Unlikely Potential, Impossible Potential, 'must' Obligation, 'should' Obligation, 'should not' Obligation, Forbidden Obligation, 'may' (permissive) |
| Reflexivity | Not Applicable, Reflexive, Reciprocal |
| Polarity | Affirmative, Negative, Emphatic Affirmative, Emphatic Negative |

Table 3. Partial listing of the Features for Attributes (Adjectives)

| | |
|---|---|
| Degree | Comparative, Superlative, Intensified, 'too' or 'overly', 'less', 'least', Not Applicable |

Table 4. Partial listing of the Features for Thing Phrases (NPs)

| | |
|---|---|
| Type | Simple, Coordinate, First Coordinate, Last Coordinate |
| Semantic Role | Participant, Patient, State, Source, Destination, Instrument, Addressee, Beneficiary, Not Applicable |

Table 5. Partial listing of the Features for Propositions

| | |
|---|---|
| Type | Independent, Coordinate Independent, Restrictive Thing Modifier, Descriptive Thing Modifier, Event Modifier, Participant, Patient, Attributive Patient |
| Illocutionary Force | Declarative, Imperative, Content Interrogative, Yes-No Interrogative |
| Topic NP | Participant, Patient, State, Source, Destination, Instrument, Beneficiary |
| Discourse Genre | Narrative, Expository, Hortatory, Procedural, Expressive, Descriptive, Epistolary, Dramatic Narrative, Dialog |
| Salience Band | Pivotal Storyline, Primary Storyline, Secondary Storyline, Script Predictable Actions, Backgrounded Actions, Flashback, Setting, Irrealis, Evaluation, Cohesive Material, Not Applicable |
| Direct Quote Speaker | Adult Daughter, Adult Son, Angel, Animal, Boy, Brother, Crowd, Daughter, Demon, Disciple, Employee, Employer, Father, Girl, God, Government Leader, Government Official, Group of Friends, Holy Spirit, Husband, Jesus, King, Man, Military Leader, Mother, Prophet, Queen, Religious Leader, Satan, Servant, Sister, Slave, Slave Owner, Soldier, Son, Wife, Woman, Written Material (letter,law,etc.) |

| | |
|---|---|
| Direct Quote Listener | Adult Daughter, Adult Son, Angel, Animal, Boy, Brother, Crowd, Daughter, Demon, Disciple, Employee, Employer, Father, Girl, God, Government Leader, Government Official, Group of Friends, Holy Spirit, Husband, Jesus, King, Man, Military Leader, Mother, Prophet, Queen, Religious Leader, Satan, Servant, Sister, Slave, Slave Owner, Soldier, Son, Wife, Woman |
| Speaker's Attitude | Neutral, Familiar, Endearing, Honorable, Derogatory, Friendly, Antagonistic, Complimentary, Anger, Rebuke |
| Speaker/Listener Age | Older, Same, Younger |

Because it's impossible to represent meaning in a completely language neutral way, it was decided that a subset of English sentence structures would be used. Taking all of the above into consideration, the semantic representation for the very simple sentence *John did not read those books* is shown below in Figure 2.

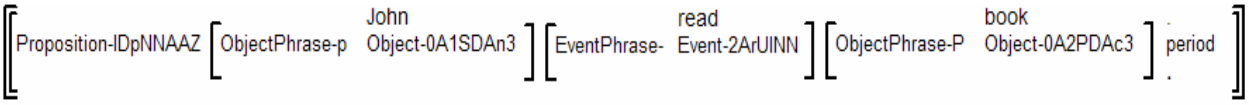


Figure 2. Semantic Representation of *John did not read those books.*

As seen in Figure 2, each lexeme has a set of features associated with it represented by the numerals and letters immediately below it, each Object Phrase (NP) is marked for its semantic role, and the proposition has a set of features characterizing it. The features associated with the event *read* in Figure 2 are expanded below in Figure 3.

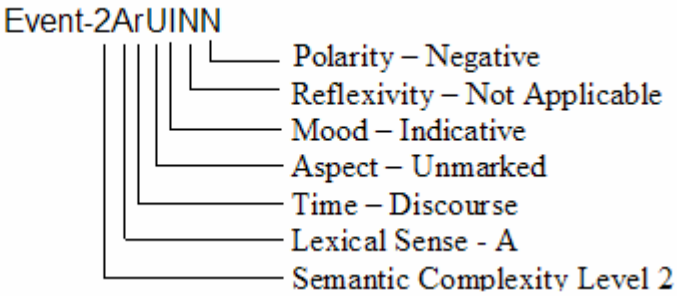


Figure 3. Expansion of Features associated with *read* shown in Figure 2

3. The Generator's Grammar

As was mentioned above, users of TBTA build a transfer grammar and a synthesizing grammar for their target languages. The transfer grammar restructures the semantic representations so that they contain the target language's structures, lexemes and features. The synthesizing grammar then synthesizes the final surface forms. The synthesizing grammar in TBTA has been designed to look as much as possible like the descriptive grammars that linguists routinely write. Therefore the synthesizing grammar includes phrase structure rules, constituent movement rules, clitic rules, spellout rules, morphophonemic rules, and feature copying rules. Figure 4 shows all of the types of rules in the transfer grammar and the sequence in which they're executed, and Figure 5 shows all the rules in the synthesizing grammar and their sequence of execution.

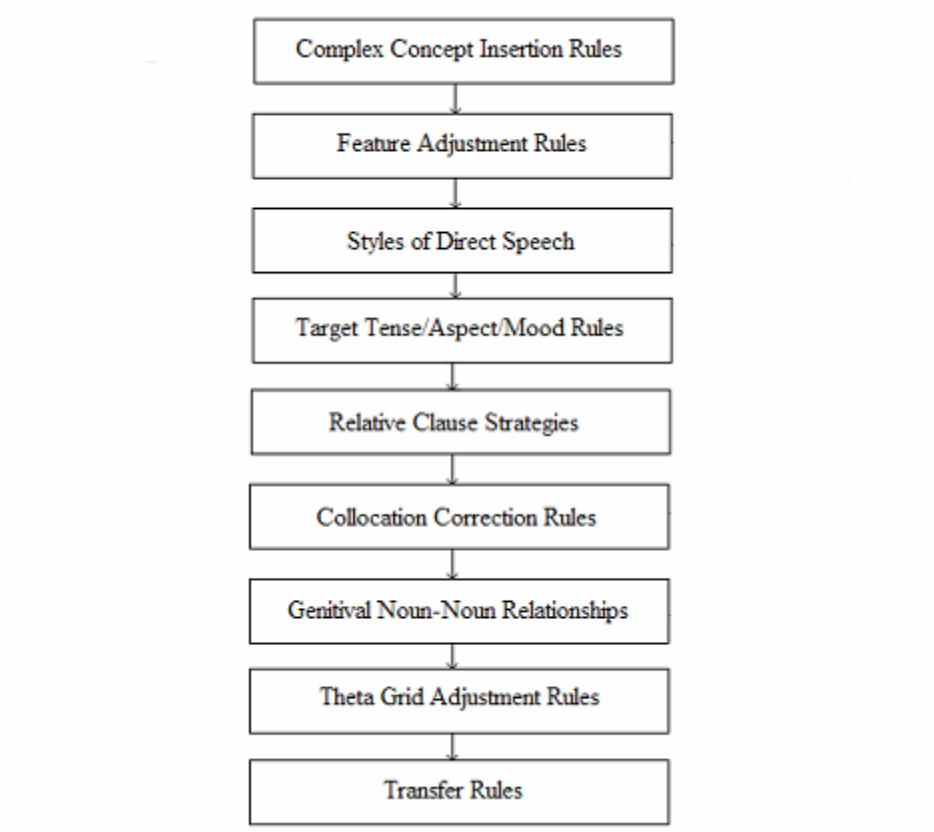


Figure 4. Overview of the Transfer Grammar in TBTA

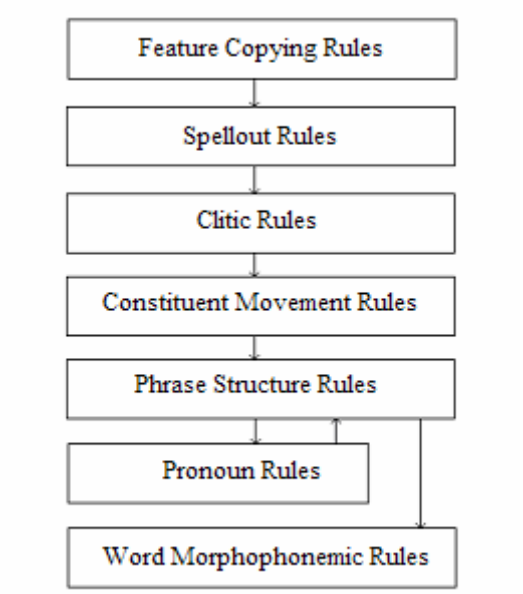


Figure 5. Overview of the Synthesizing Grammar in TBTA

Samples of some of the synthesizing rules are shown below in Figures 6 through 8. Figure 6 shows a Feature Copying rule for Jula. Certain verbs in Jula are reduplicated when their objects are plural. Therefore a Feature Copying rule copies the number of the object nominals to the verb. If there are multiple object nominals, the system finds all

of them and sums their number values (e.g., singular + singular = dual, singular + dual = trial, dual + trial = plural, etc.)

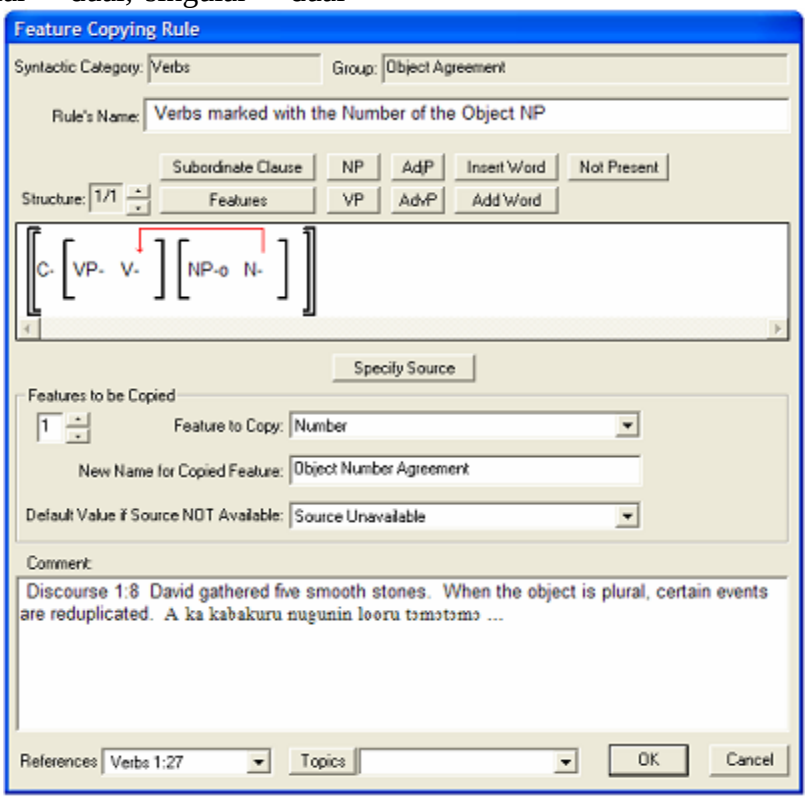


Figure 6. Feature Copying rule for Jula

Figure 7 below shows a table spellout rule for Jula. All transitive verbs in Jula are marked with an auxiliary that indicates both tense and polarity.

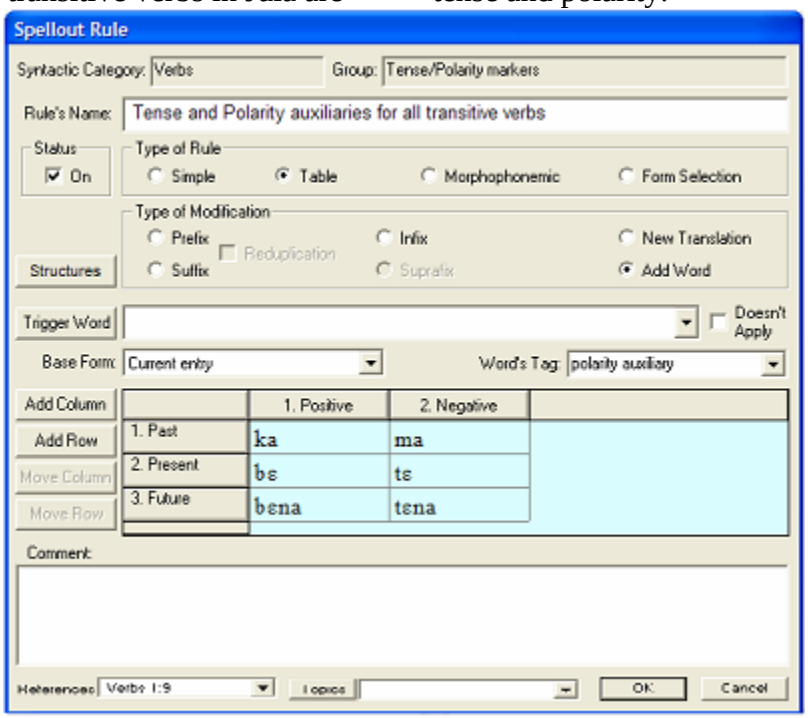


Figure 7. Spellout Rule for Jula

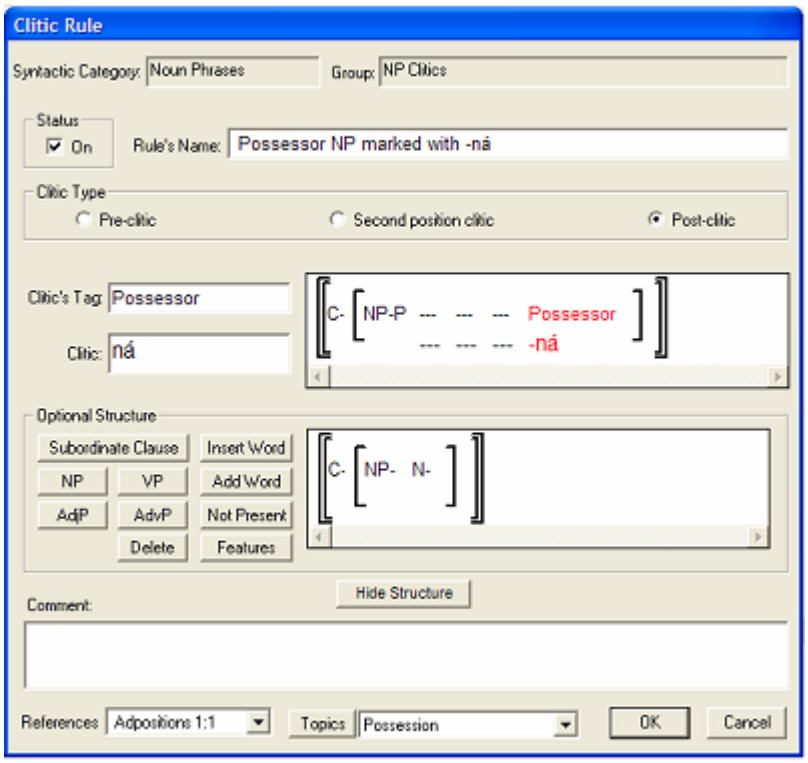


Figure 8. Clitic Rule for Kewa

Kewa marks many of its NPs with post-clitics which signal a variety of relationships. Figure 8 above shows a Clitic Rule for Kewa that inserts the post-clitic *–ná* which indicates possession.

## 4. Generating Target Text

As the linguist builds his lexicon and grammar, TBTA acquires knowledge of the target language and is able to generate target text; the more knowledge the linguist enters, the less assistance TBTA requires. Figures 9 through 11 shown below indicate that each subsequent chapter of text requires less effort by the linguist. Eventually TBTA acquires sufficient knowledge of the target language that it is able to generate drafts of all the analyzed source materials without any additional assistance from the linguist.

As was mentioned above, TBTA has been tested with four languages: English, Korean, Jula which is spoken in Cote d'Ivoire and Mali, and Kewa which is a clause chaining language with a switch reference system spoken in Papua New Guinea. In each of these four tests TBTA has produced text that is easily understandable, grammatically correct and semantically equivalent to the source texts. However, the generated texts lack naturalness and need to be post-edited in order to produce presentable first drafts. Experiments with the Jula text indicate that using TBTA's rough drafts tripled the productivity of eight professional mother tongue translators without any loss of quality. Additional experiments with the Korean text indicated that using TBTA's drafts quadrupled the productivity of six experienced mother tongue translators without any loss of quality.

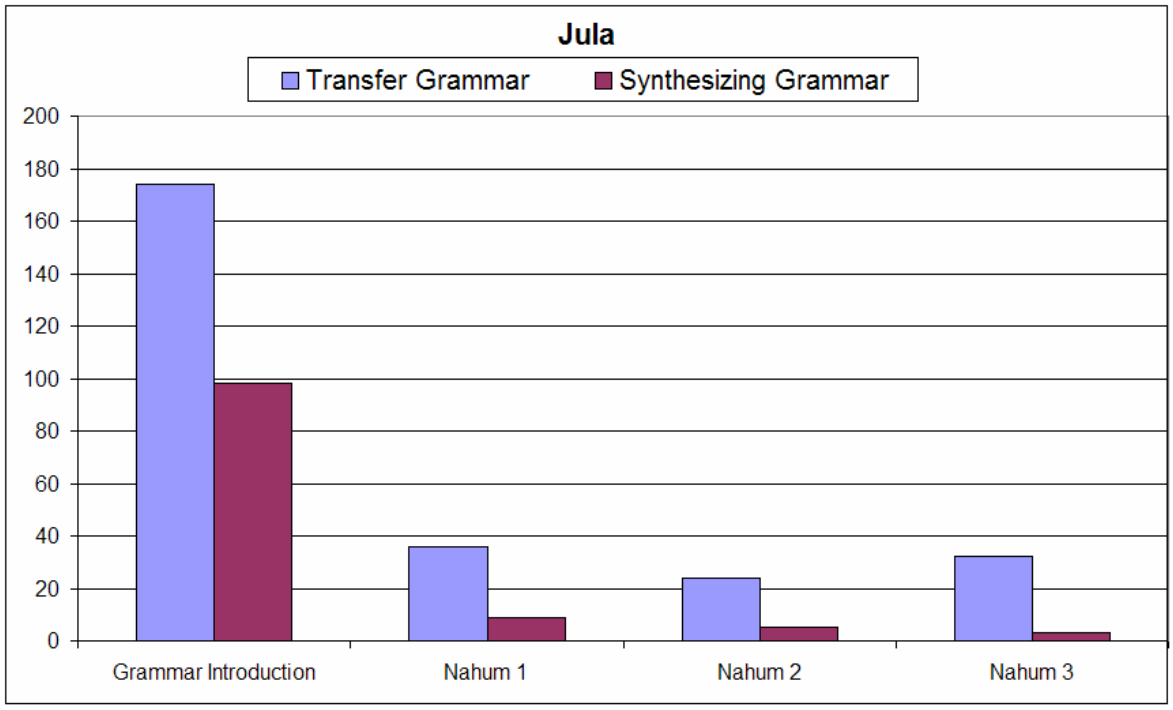

Figure 9. Number of new grammatical rules required for each chapter of Jula text

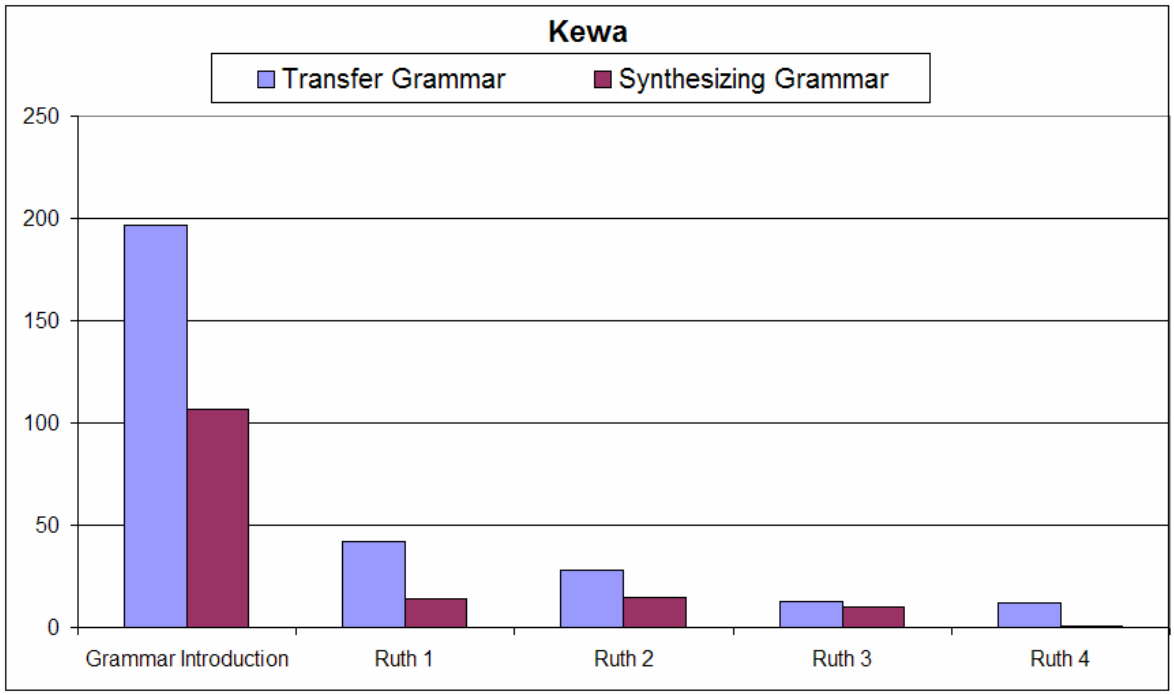

Figure 10. Number of new grammatical rules required for each chapter of Kewa text

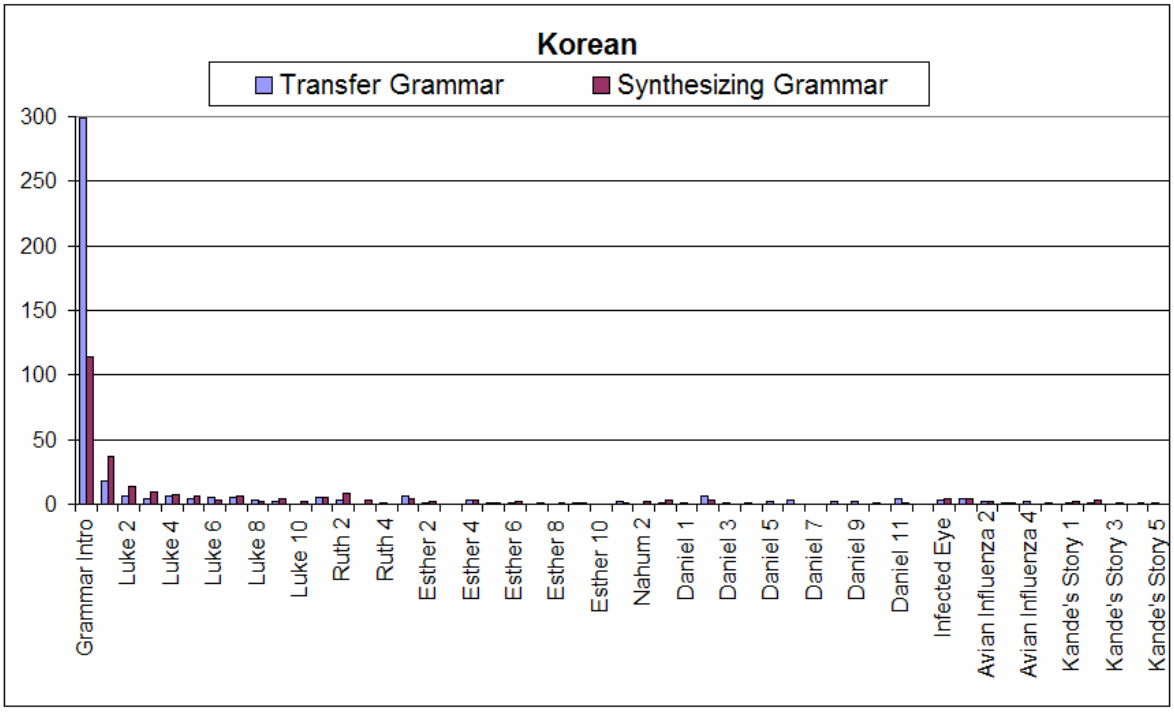


Figure 11. Number of new grammatical rules required for each chapter of Korean text

## 5. Conclusion

TBTA is a tool that will help linguists who are translating texts into a variety of languages. The information encoded in the semantic representations combined with the capabilities of the transfer and synthesizing grammars enables this project to generate target language text that is easily understandable, grammatically correct, and semantically equivalent to the source texts. The generated texts lack naturalness, but mother tongue speakers are able to edit the rough drafts and resolve these issues in a fraction of the time required to manually translate the same text. It is hoped that this project will help produce translations of many different documents into the world's many different languages.