A Natural Language Generator for Minority Languages

Tod Allman, Stephen Beale

Linguistics Department
University of Texas at Arlington
todallman@sbcglobal.net

Computational Linguistics Department
University of Maryland at Baltimore
sbeale@cs.umbc.edu

Abstract

The Bible Translator's Assistant (TBTA) is a natural language generator (NLG) designed specifically for field linguists doing translation work in minority languages. In particular, TBTA is intended to generate drafts of the narrative portions of the Bible as well as numerous community development articles in a very wide range of 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 specifically for minority languages. The grammar in TBTA has two sections: a restructuring grammar and a synthesizing grammar. The restructuring 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 triples the productivity of professional mother tongue translators without any loss of quality. A model of TBTA is shown below in Figure 1.

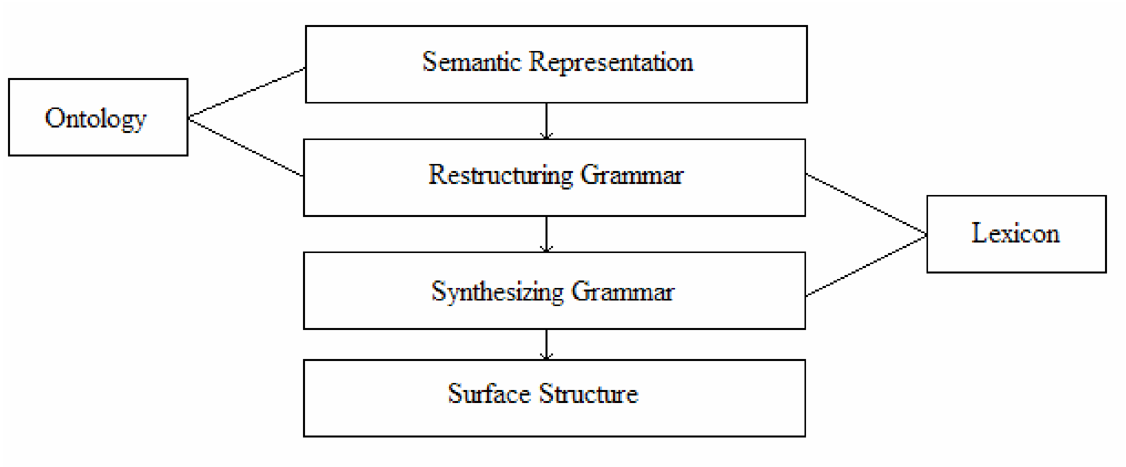


Figure 1. Underlying model of The Bible Translator's Assistant

1. The Semantic Representations

The development of an adequate method of meaning representation for TBTA's source texts proved to be a challenge. Formal semantics (Cann, 1993; Rosner, 1992), conceptual semantics (Jackendoff, 2002) and generative semantics (Lakoff, 1975) 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 (Wierzbicka, 1996; Goddard, 1998). 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 expansion rule that automatically expands the complex concept in terms of the semantic molecules for those target languages that don't have a lexicalized 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 4.

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

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

Semantic Role	Participant, Patient, State, Source, Destination, Instrument, Addressee, Beneficiary, Not Applicable

Table 3. Partial listing of the Features for Thing Phrases (NPs)

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
Notional Structure Schema (Longacre, 1996)	Narrative-Exposition, Narrative-Inciting Incident, Narrative-Developing Conflict, Narrative-Climax, Narrative-Denouement, Narrative-Final Suspense, Narrative-Conclusion, Hortatory-Authority Establishment, Hortatory-Problem or Situation, etc.
Salience Band (Longacre, 1996)	Pivotal Storyline, Primary Storyline, Secondary Storyline, Script Predictable Actions, Backgrounded Actions, Flashback, Setting, Irrealis, Evaluation, Cohesive Material, Not Applicable
Direct Quote	Man to Woman, Woman to Man, Man to Man, Woman to Woman, Father to Child, Child to Father, Mother to Child, Child to Mother, Husband to Wife, Wife to Husband, Employer to hired Worker, Hired Worker to Employer, Teacher to Student, Student to Teacher, King to Man, Man to King, King to Woman, Woman to King, Queen to Man, Man to Queen, Queen to Woman, Woman to Queen, etc.

Table 4. Partial listing of the Features for Propositions

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.

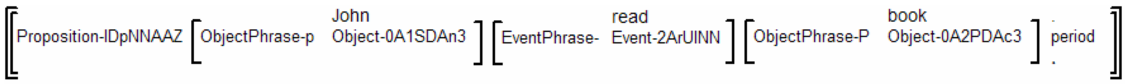


Figure 2. Semantic Representation of *John did not read those books.*

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

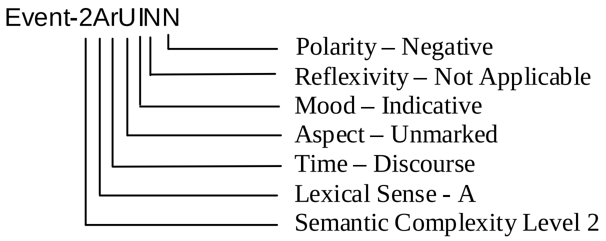


Figure 3. Expansion of Features associated with *read* shown in Figure 2

2. The Generator's Grammar

As was mentioned above, users of TBTA build a restructuring grammar and a synthesizing grammar for their target languages. The restructuring 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 synthesizing grammar and the sequence in which they're executed.

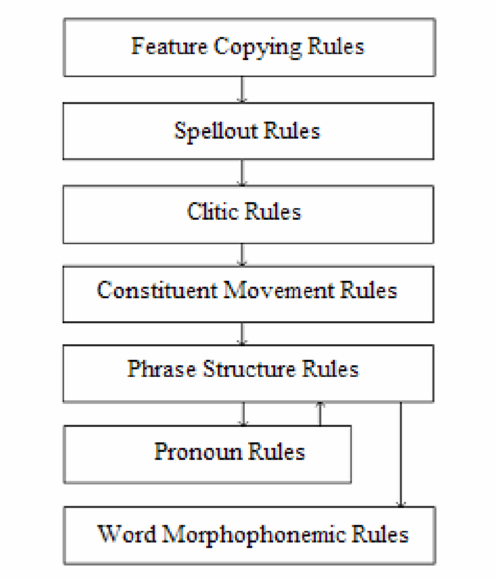
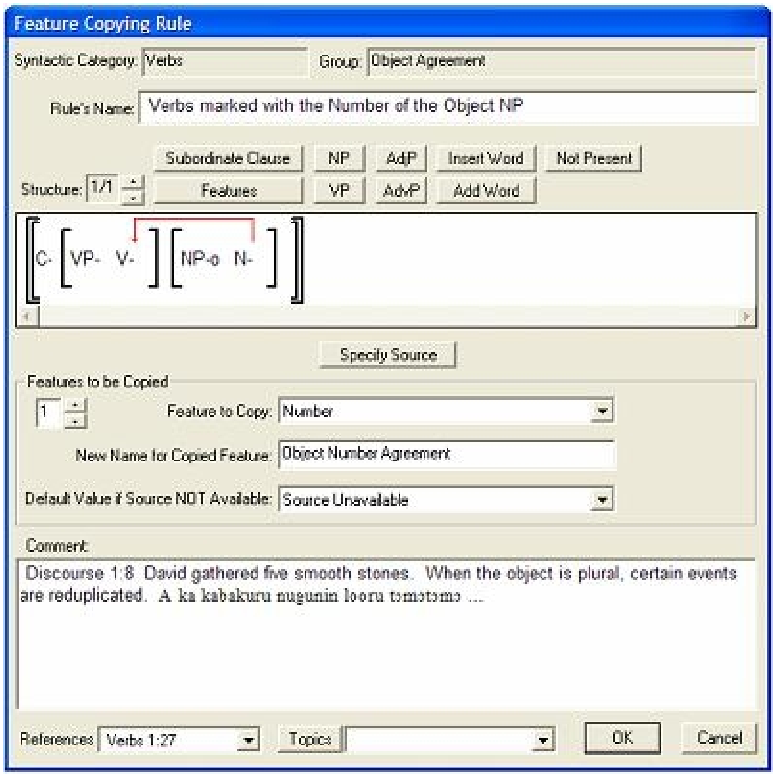


Figure 4. Overview of the Synthesizing Grammar in TBTA

Samples of some of these rules are shown below in Figures 5 through 7. Figure 5 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, etc.).



Syntactic Category: Verbs

Group: Object Agreement

Rule's Name: Verbs marked with the Number of the Object NP

Structure: 1/1

Features to be Copied

1

Feature to Copy: Number

New Name for Copied Feature: Object Number Agreement

Default Value if Source NOT Available: Source Unavailable

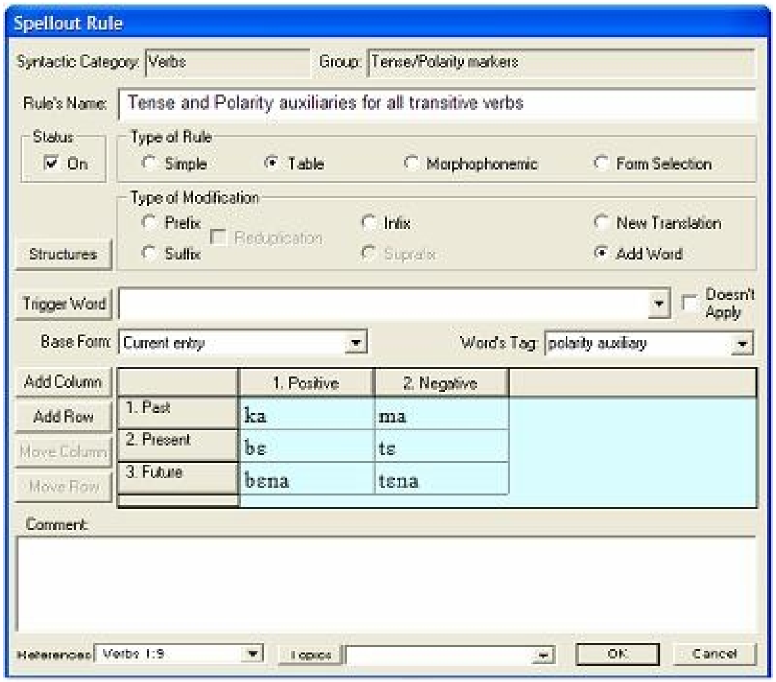
Comment:

Discourse 1:8 David gathered five smooth stones. When the object is plural, certain events are reduplicated. A ka kabakuru nugunin looru tɔmɔtɔmɔ ...

References: Verbs 1:27

Figure 5. Feature Copying rule for Jula

Figure 6 below shows a table spellout rule for Jula. All transitive verbs in Jula are marked with an auxiliary that indicates both tense and polarity. The table in this rule shows the six auxiliary verbs and their environments.



Syntactic Category: Verbs

Group: Tense/Polarity markers

Rule's Name: Tense and Polarity auxiliaries for all transitive verbs

Status: ☑ On

Type of Rule: ○ Simple ● Table ○ Morphophonemic ○ Form Selection

Type of Modification: ○ Prefix ○ Suffix □ Reduplication ○ Infix ○ Suprafix ○ New Translation ● Add Word

□ Doesn't Apply

Base Form: Current entry

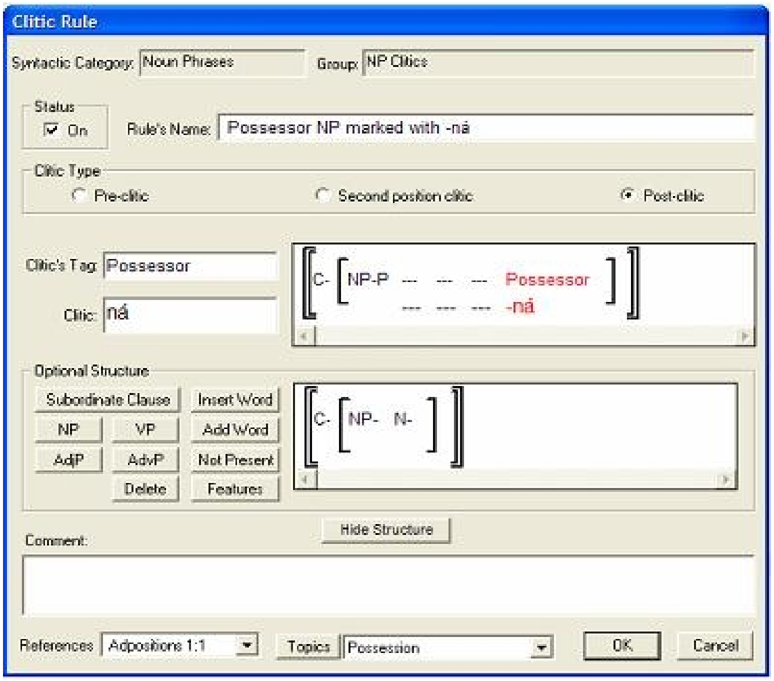
Word's Tag: polarity auxiliary

	1. Positive	2. Negative
1. Past	ka	ma
2. Present	bɛ	tɛ
3. Future	bɛna	tɛna

Comment:

References: Verbs 1:9

Figure 6. Spellout Rule for Jula



Clitic Rule

Syntactic Category: Noun Phrases

Group: NP Clitics

Status ☑ On

Rule's Name: Possessor NP marked with -ná

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

Clitic's Tag: Possessor

Clitic: ná

C- NP-P --- --- --- Possessor
--- --- --- -ná

Optional Structure

C- NP- N-

Comment:

References: Adpositions 1:1

Topics: Possession

Figure 7. Clitic Rule for Kewa

Kewa marks many of its NPs with post-clitics which signal a variety of relationships. Figure 7 above shows a Clitic Rule for Kewa that inserts the post-clitic *–ná* which indicates possession.

3. 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 8 and 9 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.

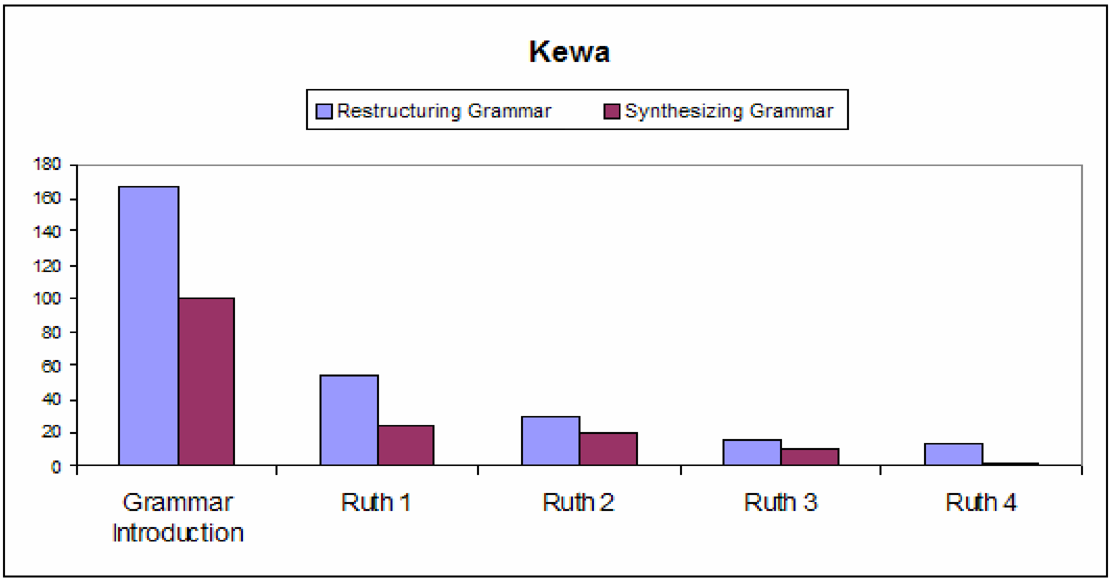


Figure 8. Number of new grammatical rules required for each chapter of Kewa text

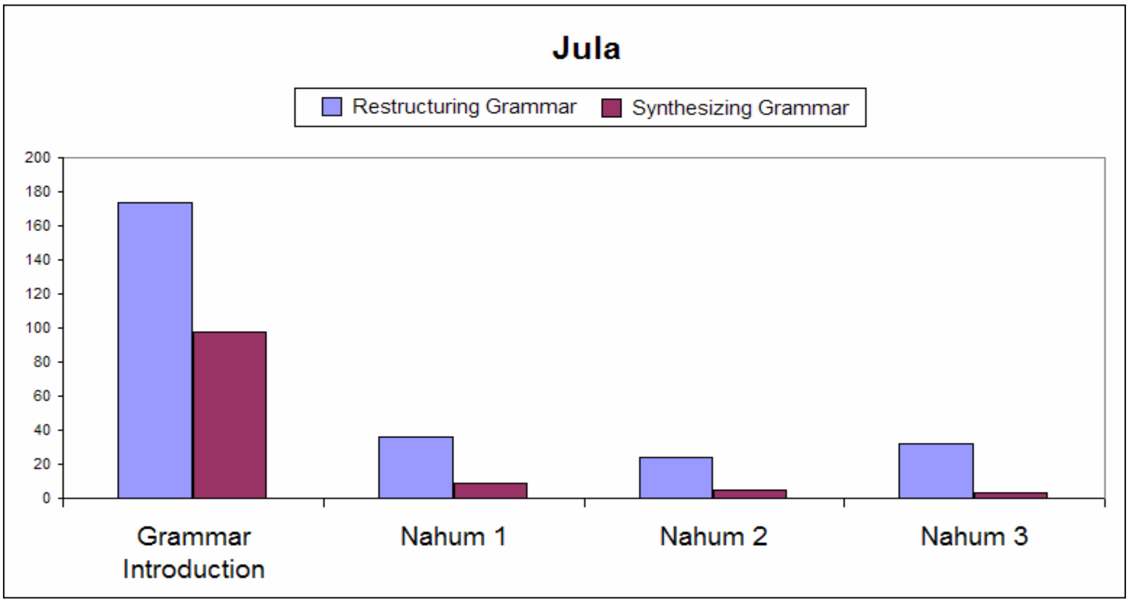


Figure 9. Number of new grammatical rules required for each chapter of Jula text

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 perfect 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 TBTA triples the productivity of professional mother tongue translators without any loss of quality. Those experiments will be described in Section 4.

4. Evaluating the Generated Text

In order to determine whether or not the quality of the text generated by TBTA is sufficient so that it actually improves the productivity of a translator, several experiments were performed with the generated Jula text. As was shown above in Figure 9, a lexicon and grammar were developed for Jula so that TBTA could generate a draft of the biblical book of Nahum. Then eight professional mother tongue Jula translators in Mali were asked to participate in an experiment that was designed to determine the quality of the generated text. In particular, four of the translators were asked to edit the first half of the generated text and make it a presentable first draft. Then they were asked to manually translate the second half of Nahum from the French *La Bible en Français Courant*, again producing a presentable first draft. The other four translators were asked to perform the same two tasks, but they manually translated the first half of Nahum and then edited the second half of the generated text. All of the translators were told that they'd be timed during each of the two tasks. Table 5 below shows the results of this experiment. On average these eight professional mother tongue translators spent three times as much time translating as they did editing. These results were encouraging, but another experiment was considered necessary to determine whether or not the translators had actually done a thorough job of editing the generated text.

In the second experiment, the eight drafts of Nahum were evaluated by forty other Jula speakers in order to compare the quality of the two halves of each text. These other speakers had no idea how the texts had been produced or where the texts had come from. Each of the evaluators was given one text that consisted of two halves – one half had been manually translated and the other half had been generated by TBTA and then edited by the same translator. The evaluators were each asked just one question: Is the quality of either half significantly better than the quality of the other half, or are the two halves essentially equal in quality? The results of this experiment are also summarized below in Table 5.

Translator	Editing Time	Translating Time	Ratio	Evaluations
Translator #1	24 minutes	65 minutes	2.7:1	C1 - M1 - E3
Translator #2	51 minutes	89 minutes	1.7:1	C1 - M2 - E2
Translator #3	56 minutes	132 minutes	2.4:1	C4 - M1
Translator #4	40 minutes	150 minutes	3.8:1	C2 - M3
Translator #5	70 minutes	145 minutes	2.1:1	C1 - E4
Translator #6	52 minutes	120 minutes	2.3:1	E5
Translator #7	62 minutes	192 minutes	3.1:1	C2 - M1 - E2
Translator #8	20 minutes	296 minutes	14.8:1	C1 - M3 - E1

Table 5. Evaluating the Quality of the generated Jula text

Average translation time: 1189/8 = 149 minutes
Average editing time: 375/8 = 47 minutes
Ratio: 3.2:1

In the Evaluations column of Table 5, the numbers prefaced with a 'C' indicate the number of evaluators that chose the computer generated half as better, the numbers prefaced with an 'M' indicate the number of evaluators that considered the manually translated half to be better, and the numbers prefaced with an 'E' indicate the number of evaluators that said the two halves of the text were equal in quality. Considering all of the evaluations together, a total of twelve evaluators thought that the edited computer generated half was better, eleven evaluators chose the manually translated half as being better, and seventeen evaluators considered the two halves to be of equal quality. Therefore twenty-nine of the forty evaluators said that the halves that had been generated by TBTA and then manually edited were as good as or better than the halves that had been professionally translated. So this second experiment confirmed that the translators had done a thorough job of editing the generated text even though they had only spent a third as much time editing as translating. Therefore, in this particular case, TBTA tripled the productivity of professional mother tongue translators without any loss of quality.

5. Conclusions

TBTA is a tool that will help field linguists who are translating texts into a variety of languages. The information encoded in the semantic representations combined with the capabilities of the restructuring and synthesizing grammars enables this project to generate target text that is easily understandable, grammatically perfect, and semantically equivalent to the source texts. The generated texts lack naturalness, but this problem may be easily corrected with post-editing. Additional experiments are currently being performed to ascertain the quality of the generated texts in other languages. It is hoped that this project will help produce translations of many different documents into the world's minority languages.

6. References

Allman, T., Beale, S. (2004). An environment for quick ramp-up multi-lingual authoring. In *International Journal of Translation*, Vol. 16, No. 1.

Beale, S., Nirenburg, S., McShane, M., and Allman, T. (2005). Document Authoring the Bible for Minority Language Translation. In *Proceedings of MT-Summit.* Phuket, Thailand.

Cann, R. (1993). *Formal Semantics*, Cambridge: Cambridge University Press.

Goddard, C. (1998). *Semantic Analysis: A Practical Introduction*, New York: Oxford University Press.

Jackendoff, R. (2002) *Foundations of Language*, New York: Oxford University Press.

Lakoff, G. (1975). *Pragmatics in Natural Logic.* In Keenan, E. (ed.) (1975) *Formal Semantics of Natural Language*, Cambridge: Cambridge University Press.

Longacre, R. (1996). *The Grammar of Discourse*, New York: Plenum Press.

Rosner, M., Johnson, R. (1992). *Computational Linguistics and Formal Semantics*, Cambridge: Cambridge University Press.

Wierzbicka, A. (1996) *Semantics: Primes and Universals*, New York: Oxford University Press.