TBTA Tutorial Lesson 1:
Introduction to The Bible Translator's Assistant

(video: TBTA-1B Introduction to TBTA)

1.0 Introduction

The Bible Translator's Assistant (TBTA) is a natural language generator (NLG) designed specifically for field linguists doing translation work in minority languages. A natural language generator is a computer program that takes a semantic representation of a text, and a lexicon and grammar of a language, and then generates text in that language that has the same meaning as the semantic representation. If you learn how to use this generation system well, it will save you many years of work. TBTA will guide you through the process of developing a lexicon and grammar for your language. After your lexicon and grammar are sufficiently developed, TBTA will generate drafts of all the analyzed source materials. Then you and your mother-tongue assistants will be able to edit the generated drafts into publishable texts.

A high level model of TBTA is shown in Figure 1-1.

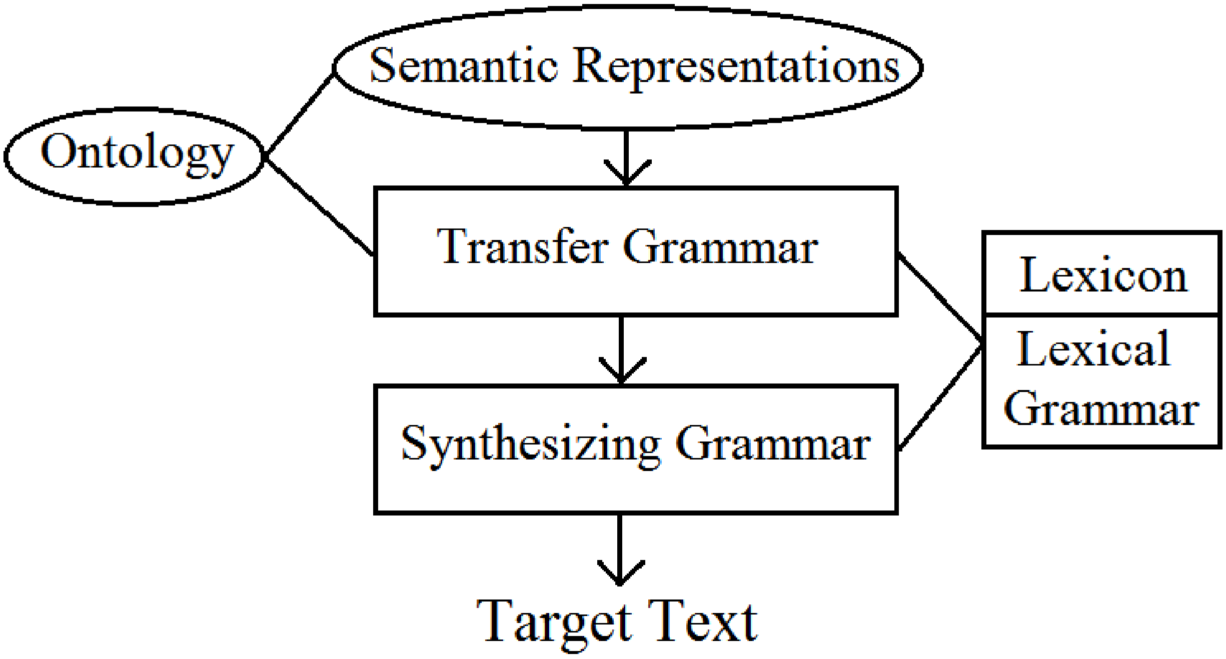

Figure 1-1. A High Level Model of TBTA

As seen in the figure above, there are five fundamental components in this generation system: 1) the semantic representations, 2) the ontology, 3) the target lexicon, 4) the transfer grammar, and 5) the synthesizing grammar. The model indicates that TBTA begins with a semantic representation of a proposition or text, and then executes all the rules in the target language's transfer grammar. The output of the transfer grammar is a deep structure representation of that proposition or text, and it consists of the target language's lexemes and structures. That deep structure representation is then supplied to the synthesizing grammar where all the target language's synthesizing rules are executed. The output of the synthesizing grammar is target language surface structure text. The pertinent concepts in the ontology are included in the semantic representations, and they are also available to the rules in the transfer grammar. Similarly the words in the target lexicon are available to both the transfer and synthesizing grammars. The two components in ovals represent static knowledge that is supplied with TBTA. The three components in rectangles represent target language knowledge that is supplied by a linguist. The final output of TBTA is target language text.

The semantic representations serve as the source documents from which TBTA will generate target language text. The ontology and semantic representations will be provided for you, but you need to become familiar with their formats. The content of the ontology and the format of the semantic representations will be the topic of the next tutorial. The remaining tutorials will then describe the lexicon, the transfer grammar, and the synthesizing grammar.

1.1 The Translation Process

Translating a document from one language to another is a very complex, labor intensive, highly skilled task. Producing a natural translation in a language that is unrelated to the source language requires a thorough knowledge of both the source and receptor languages, the two cultures, and the translation's intended audience. During the translation process, a myriad of linguistic and sociolinguistic factors must be taken into consideration. Although translation is a very complex process, it is usually divided into three fundamental steps:

- *analysis* – analyze the source document and determine its meaning,
- *transfer* – reconstruct that meaning using the target language's lexemes, structures, and world view, and
- *synthesis* – synthesize the final surface forms.

Human translators must perform all three of these steps when translating a document from one language to another. Machine translation projects must also perform all three of these steps. However, making a computer perform analysis has proven extraordinarily difficult. Computational linguists have found that there are three particular issues that arise when a computer attempts to perform source analysis:

- part-of-speech disambiguation,
- word sense disambiguation, and
- structural disambiguation.

To circumvent these three difficulties, computational linguists have begun developing natural language generators which perform only two steps of the translation process: transfer and synthesis. NLGs avoid the difficulties associated with automatic source analysis because they use manually developed semantic representations. During the manual development of the semantic representations, all part-of-speech ambiguity is eliminated, all word sense ambiguity is resolved, and all structural ambiguity is made explicit. Therefore NLGs are able to produce much better translations than are machine translation projects. However, developing good semantic representations is a very time consuming, labor intensive task. Developing a semantic representation of a document and using an NLG to produce drafts of translations is only practical when many large documents need to be translated into many languages. The Bible is a very large document that needs to be translated into thousands of languages, so it is an ideal candidate for a NLG.

1.2 Limitations of NLGs

All NLGs have their limitations. NLGs of TBTA's type have the following limitations:

1) They produce simple, basic sentences using simple vocabulary. No NLG is able to produce high quality texts with sophisticated sentences and complex vocabulary.
2) The texts produced by NLGs always require human editing before being published.
3) NLGs work well for certain discourse genres, but very poorly for other genres. For example, NLGs typically work well for narrative and procedural texts, but very poorly for poetry.
4) NLGs are never aware of sociolinguistic or cultural issues. For example, consider translating the question "If a son asks his father for a fish, will the father give him a snake instead?" If in a particular culture, snake meat is considered a delicacy, then a direct translation completely distorts the intended message. But NLGs are never aware of issues such as snake meat being a delicacy, and even if they were aware of these issues, they wouldn't be able to resolve them.
5) Identifying which nominals may be replaced with pronouns is always difficult for NLGs. The approach usually adopted is to under-generate pronouns. This means that pronouns are generated in only the very clear cases, and the remaining nominals are left unchanged so that the human editor can make the final decision.

1.3 Installing TBTA

Before beginning the next tutorial, you should download and install TBTA. You can always (even after you've completed this course) download the latest version of TBTA by going to www.TheBibleTranslatorsAssistant.org. At the Downloads tab you'll find an installation program and the latest executable. If you haven't yet installed TBTA, you should download the installation program and go through the installation process. Whenever there are new versions of TBTA, they will be made available at the web site, and you can simply download the latest version without going through the installation process again. So please go to the web site and download and install TBTA before starting the next tutorial. The installation program includes an English database with a well developed lexicon and grammar. During these tutorials, if you need help building a particular rule, you can always switch to that database, look at the rule, and then switch back to your own database and build the rule.

The remaining tutorials will deal primarily with English data. For example, we’ll enter English nouns and verbs into the lexicon, and define their pertinent features and forms. Later we’ll write spellout rules for English words, and develop phrase structure rules for English phrases and clauses. As you’re going through these processes, you should be thinking about how to handle similar situations in your particular target language.