Tree-Based TM for Computer-Assisted Bible Translation

Andi Wu

Global Bible Initiative (GBI)

**Abstract**

TM (translation memory) is a database that stores the translation units that have been previously translated. It can be used in computer-assisted Bible translation to provide suggestions when new texts are translated, thus improving efficiency and consistency in translation. This paper presents an automatic way of incrementally creating a TM in real time as a translation project goes on. What is required in this approach is (1) an automatic word aligner, and (2) syntactic treebanks of the original Hebrew and Greek texts. After each verse is translated, the auto-aligner is used to align the translation words to their corresponding Hebrew/Greek words which are the leaf nodes in a syntactic tree. Since each node in the tree (a subtree) represents a word, phrase, or clause, phrase/clause alignment can also be automatically created by mapping the sequence of leaf nodes in the subtree to the translation words aligned to these nodes, resulting in a TM that contains linguistically valid translation units of any textual size. As such a TM grows, Bible translators get increasingly better suggestions. This approach differs from traditional methods of TM creation where translation units beyond the word level are arbitrary word sequences which are hard to identify automatically, and which are not always legitimate linguistic units. We are able to avoid these problems due to existence of syntactic treebanks of Biblical texts. Such resources are seldom available in other domains.

**Introduction**

A translation memory (TM) is a database that stores the translation units (TUs) in previously translated content. The TUs can be words, phrases, clauses or sentences. Each TU consists of a “source”, a “target”, and possibly a “count” which keeps track of how many times the source has been translated into the target. In computer-assisted translation, TMs are used to suggest translation candidates for text segments that are identical or similar to the ones stored in the TMs. The translator can accept a candidate, modify it, or produce a fresh translation.

In computer-assisted Bible translation, TMs have been used to improve the *efficiency* of translation by minimizing duplicated efforts which “reinvent the wheel”, so to speak. TMs also make the translators aware of what they have done previously on identical or similar texts, which results in better *consistency*. In cases where the translation is done by a team, TMs also keep the team members informed of each other’s work, thus encouraging *sharing and mutual checking*.

For a TM to be effective, it needs to contain TUs that are likely to recur. Individual words are certainly common enough to recur, but they usually have multiple meanings (thus multiple possible translations) and often cannot be translated correctly in isolation because of the ambiguity. At the other extreme, whole sentences can be translated without much context, but they rarely recur. The most valuable TUs are phrases and clauses. They tend to be contextualized enough for disambiguation and are likely to recur. Therefore, harvesting phrase-level TUs from existing translations is crucial to the success of a TM-based system. In this paper, we present a special way of automatically acquiring phrase-level TUs which leverages a unique resource available in the Bible domain: the syntactic treebanks of the original Hebrew and Greek texts.

**Statistical Phrase Alignment**

Phrase alignment has been an important part of statistical machine translation (SMT). In SMT, both the source text and target text are unannotated. They are just word sequences without any indication of phrase boundaries. Any segments of two or more words can be a phrase. Take Mark 1:1 for example:

Ἀρχὴ τοῦ εὐαγγελίου Ἰησοῦ Χριστοῦ υἱοῦ θεοῦ

Any N-gram in this verse (where N > 1) can be a potential phrase.

Bigrams: Ἀρχὴ τοῦ ; τοῦ εὐαγγελίου ; εὐαγγελίου Ἰησο ; Ἰησοῦ Χριστοῦ ; Χριστοῦ υἱοῦ ; υἱοῦ θεοῦ

Trigrams: Ἀρχὴ τοῦ εὐαγγελίου ; τοῦ εὐαγγελίου Ἰησοῦ ; εὐαγγελίου Ἰησοῦ Χριστοῦ ; …

4-grams: Ἀρχὴ τοῦ εὐαγγελίου Ἰησοῦ ; τοῦ εὐαγγελίου Ἰησοῦ Χριστοῦ ; …

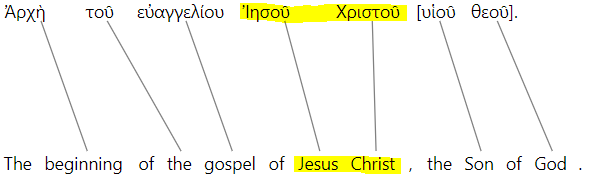
5-grams: Ἀρχὴ τοῦ εὐαγγελίου Ἰησοῦ Χριστοῦ ; τοῦ εὐαγγελίου Ἰησοῦ Χριστοῦ υἱοῦ ; …

…

Humans who read Greek can tell that most of them are not phrases, but they all look the same to the machine in this single verse. Things are different when we have a large corpus, because real phrases tend to recur. The machine can then recognize Ἰησοῦ Χριστοῦ (Jesus Christ) as a phrase, because these two words often go together.

In phrase alignment, we need to identify phrases on both the source side and the target side. This actually makes the job a bit easier, as the phrases on the two sides provide mutual evidence for each other’s phrase-hood. If we see “ Ἰησοῦ Χριστοῦ” on the source side whenever “Jesus Christ” appears on the target side, we can assume that both “Ἰησοῦ Χριστοῦ” and “Jesus Christ” are phrases. Furthermore, they are probably translations of each other.

This is the basic intuition behind statistical phrase alignment. In practice, we often do word alignment first, and then do phrase alignment on top of that. (Some approaches also do phrase alignment directly.) Given a word alignment between Mark 1:1 and its English translation:



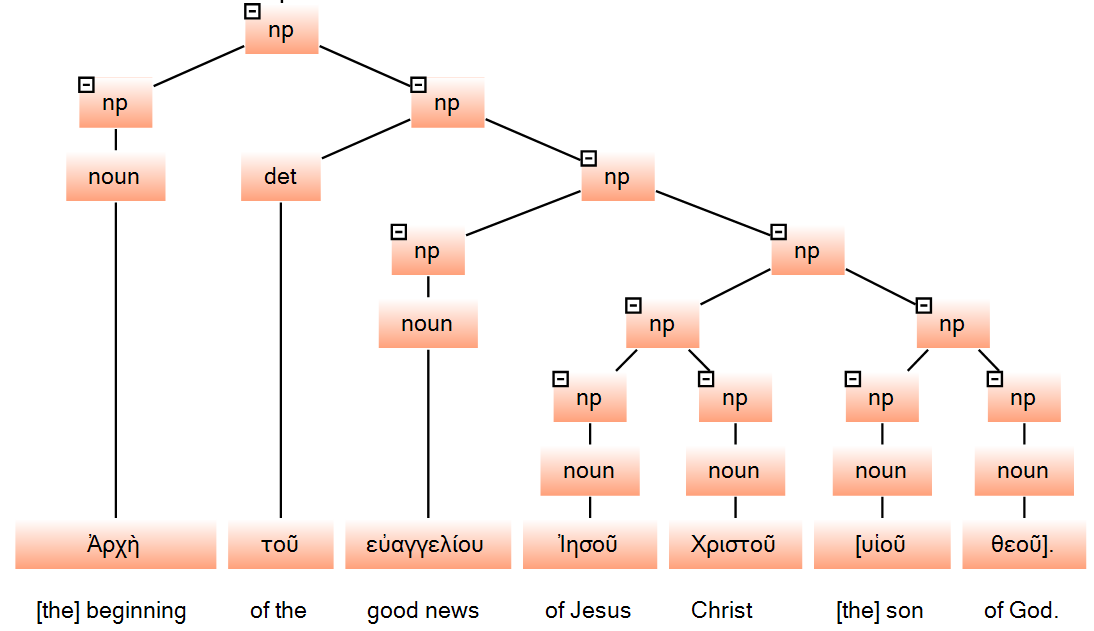
“Ἰησοῦ Χριστοῦ 🡪 Jesus Chris” can be identified as a phrasal translation unit if this mapping also occurs in other verses.

Some people have applied statistical phrase alignment in Bible translation for the purpose of building translation memories. However, they have only been successful in acquiring small segments, usually no more than two or three words long. The segments are not necessarily phrases in the linguistic sense, but this is good enough for statistical machine translation where a translation can be stitched together from any kinds of text segments. This is not sufficient for building translation memories that suggest linguistically valid translation units of any textual size, such as this paper presents.

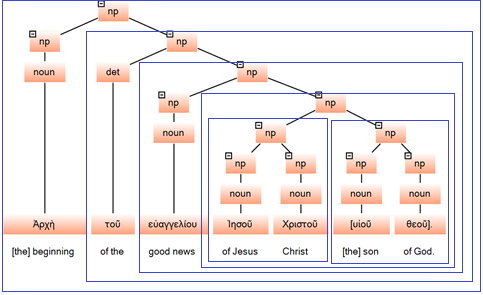
**Tree-based phrase alignment**

Statistical phrase alignment is the best option available if the texts are not annotated. It is the standard way of doing phrase alignment because we cannot expect every text to be linguistically analyzed before it is aligned. In the Bible domain, however, the source texts in Hebrew and Greek, which generally stay the same, have been richly annotated, both morphologically and syntactically. Phrase structure trees have been built for both the Hebrew Old Testament and the Greek New Testament. This puts us in a unique situation where we can use alignment methods that leverage those rich resources, though they are not available in the general domain.

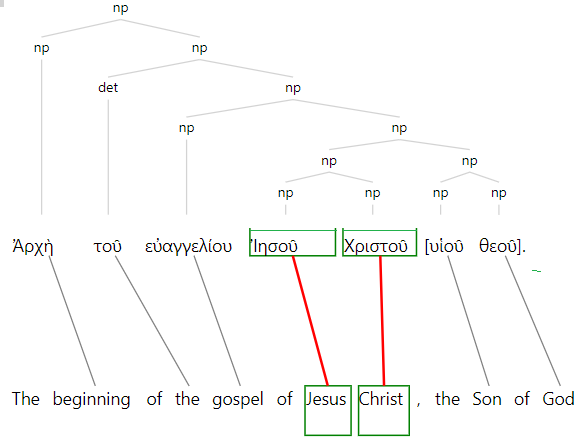
In the syntactic treebank developed by Global Bible Initiative (GBI), every verse has been syntactically analyzed with its structure represented in tree diagram. Take Mark 1:1 again as an example (where “np” stands for “noun phrase”):



6 noun phrases are found in this tree:



Now that we know what the phrases are on the source side, it is much easier to identify the corresponding phrases on the target side. If we have good word alignment already, the phrases on the source side can be projected onto the target side through the word links:



Among other things, “Ἰησοῦ Χριστοῦ” as a phrase is projected onto “Jesus Christ” through the word links between Ἰησοῦ and “Jesus” and between Χριστοῦ and “Christ”. Likewise, “υἱοῦ θεοῦ” can be projected onto “Son of God”, “Ἰησοῦ Χριστοῦ υἱοῦ θεοῦ” onto “Jesus Christ the Son of God”, etc. Here are all the phrase alignments we get for this verse. They go into the translation memory in addition to the single word alignments:

Ἀρχὴ τοῦ εὐαγγελίου Ἰησοῦ Χριστοῦ υἱοῦ θεοῦ -- the beginning of the gospel of Jesus Christ the Son of God

τοῦ εὐαγγελίου Ἰησοῦ Χριστοῦ υἱοῦ θεοῦ -- the gospel of Jesus Christ the Son of God

εὐαγγελίου Ἰησοῦ Χριστοῦ υἱοῦ θεοῦ -- gospel of Jesus Christ the Son of God

Ἰησοῦ Χριστοῦ υἱοῦ θεοῦ -- Jesus Christ the Son of God

Ἰησοῦ Χριστοῦ -- Jesus Christ

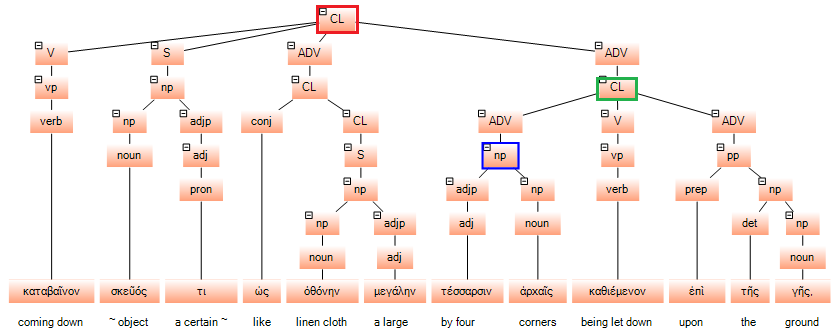
υἱοῦ θεοῦ -- Son of God

**Tree-Based Suggestions**

Once a phrase-based translation memory is built, it can be used to make suggestions during the translation process. Since the translation units are based on the trees, it is only natural to use the trees in making suggestions.

Potentially any node in the tree may cover texts that have been translated before and therefore have translations in the TM, as we have seen above in Mark 1:1. Which of them should we show as suggestions? Single words tend to be ambiguous without context. Therefore, we want to show phrases whenever possible rather than the words they contain. Likewise, given two phrases P1 and P2, with P2 properly contained in P1, we want to show P1 instead of P2. Moreover, phrases, especially longer phrases, are more useful as building blocks of the translation.

To achieve this effect, we perform a top-down traversal of the tree, visiting the nodes one by one. As soon as we hit a node whose text is in the TM, we use its translations as suggestions and stop the traversal there, not going down further to its child nodes. For illustration, we look at a subtree in Acts 11:5.



The noun phrase “τέσσαρες ἀρχή” (the node marked blue) also appears in Acts 10:11. If this book is being translated sequentially, the translation unit “τέσσαρες ἀρχή – by its four corners” should already be in the TM. Also in the TM will be “ἀρχή -- beginning”, “ἀρχή -- rulers”, “ἀρχή -- first”, as “ἀρχή” has multiple meanings. If we show suggestions for the single words, we will have a list of translations to choose from. However, we will not reach the single word here because “τέσσαρες ἀρχή – by its four corners” is in the TM. Now let’s start from the beginning. If you look at the tree here, the tree traversal will start at the red node at the top. Not finding this phrase in the TM, it will go down and reach the green node in time. Since the green node is not in TM, either, we continue until we reach the blue node which is in the TM. We will show “by its four corners” as the suggestion and stop there, not going down to its child nodes. Once we have the translation of the whole phrase, we do not need the translations of its component words. This results in the disambiguation of ἀρχή, as its other meanings are ruled out by context.

The translation memory is built incrementally during the translation process. After each verse is completed, the verse text is aligned, first at the word level and then at the phrase level, with the translation units thus obtained added to the TM. If we also record the locations of where the phrases occur, a phrase-based concordance can be built incrementally at the same time. This can be very useful for the final checking of the translation, as pointed out by Patrick Rosendahl, a translation consultant who works in Nigeria.

**Issues in Tree-Based Phrase Alignment**

The tree-based approach presented above seems straightforward enough. Despite its simplicity, however, there are some problems which must be solved in order for this approach to be successful. First, this method assumes the use of an automatic word aligner. Given a good word alignment, phrase alignment is easy. But the alignment is not guaranteed to be 100% correct. What happens when the alignment is incomplete? Secondly, the number of identical phrases is limited, but the TM approach seems to work for identical phrases only. How can we increase the coverage of the data and make suggestions available in more places? Finally, since the phrases in this approach are defined by the trees and the trees are based on Hebrew and Greek, what happens to the linguistic units in translations that do not correspond to any constituent in the tree? We will address these issues one by one below.

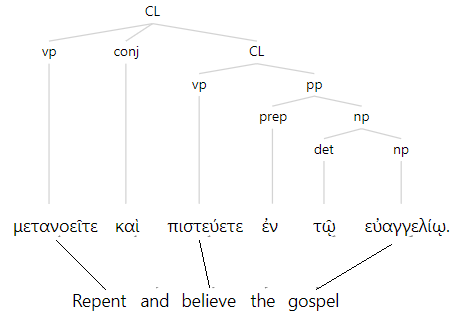
**The Word Alignment Issue**

As we have seen, phrase alignment depends on word alignment in this approach. In an ideal situation, the alignment automatically produced by the machine is manually checked to remove any errors. But this is not always practical. We cannot expect every translator to review the machine alignment and make corrections. As a result, there can be wrong links and missing links in the alignment. They can result in wrong phrase alignment or missing phrase alignment. There is no bullet-proof way of preventing it from happening, but there are a few things we can do to minimize the damage.

To minimize the number of wrong links, we need to improve the accuracy of the word aligner. In our experiment with the word aligner, it is found that the accuracy can improve significantly if we align the content words only, i.e. the nouns, verbs and adjectives . Since the Hebrew and Greek trees also contain morphological information with each word tagged a part of speech, it is easy to identify the content words. The experiments show that on average (across different languages) there is a 10 percentage points increase in accuracy when the alignment is limited to content words. This reduces the error rate and the need for manual editing.

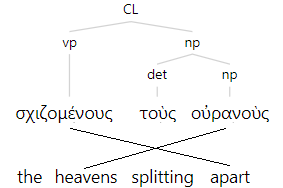
The undesirable consequence of aligning content words only, however, is that the function words (pronouns, prepositions, conjunctions, articles, etc.) are not linked, which aggravates the missing link problem. Therefore we need to be able to do phrase alignment in spite of the missing links. It turns out that not all the words in a phrase need to be linked in order for that phrase to be linked.

In phrase alignment, we assume that a phrase in Greek will be translated into a phrase in the target language. Ideally, if the words in that phrase are contiguous in Greek, the words in the translation phrase will also be contiguous, though it is not always the case. If we can identify the edges, i.e. the beginning and end of a phrase, we can get the whole phrase. In other words, as long as the first word and last word of a phrase are linked in the alignment, we can align that phrase even if some words inside that phrase are not linked. Consider the following example:



In this case, the 3 content words are linked but not the 3 function words, but this does not prevent the phrase from being aligned. Given the fact that the first word and the last word are aligned and one of the words in the middle is also linked, we can safely conclude that the chunk of text that starts at “repent” and ends with “gospel” is a phrase aligned to this subtree. The translation we get for this phrase will be “Repent and believe the gospel” which also includes the function words that are not linked, being “wrapped” by the content words.

In the example we just saw, the first word in Greek happens to link to the first word in the translation, and the last word to the last word in the translation. This is not usually the case, due to word order differences between the source language and the target language. In the following case, for instance:



The first word in Greek is linked to the last word in English, and the last word in Greek linked to the second word in English. This is not a problem, though. We just need a more general mapping procedure:

1. For each Greek word in the phrase, find the translation word it is linked to, if any.
2. Sort all the linked translation words according to their positions in the translation text.
3. Take the text that spans from the first linked word to the last linked word. This will be the translation phrase corresponding to the Greek phrase in this subtree.

In the example given here, we find that, of the 3 Greek words, two of them are linked: “σχιζομένους” to “apart” and “οὐρανοὺς” to “heavens”. Sorting the two linked words according to their positions in the translation text, we get “heavens - apart”. Then we take the text spanning from “heavens” and “apart” and get “heavens splitting apart”. What goes into the TM is then:

σχιζομένους τοὺς οὐρανοὺς – heavens splitting apart

Notice that “splitting” is included in the phrase although it is not linked. Also, the unlinked “τοὺς” is included on the source side. What is left out is “the” on the target side which should be the first word in the translation text. It is not linked and it precedes the first linked word. But the loss is minimal, as “heavens splitting apart” is still a valid phrase.

In general, phrase alignment can be successful if the first content word and the last content word in the Hebrew or Greek phrase are linked.

**The Coverage Issue**

The usefulness of a TM depends on how many phrases are repeated in the text. If we consider identical phrases only, the coverage is limited. So far, we have assumed that the phrases are identical in their surface forms. This assumption holds in the general domain where we do not expect the text to be morphologically analyzed. In the Bible domain, however, the source texts in Hebrew and Greek have been analyzed and we know the underlying/root form (the lemma) of each word. Since more words are likely to have the same lemma than surface forms, we can improve the coverage by using lemmas instead of surface forms in the TM. For example, ἠκολούθησαν (followed) occurs only 18 times in the New Testament, but its lemma ἀκολουθέω occurs 90 times. By using the lemma, what we have learned from ἠκολούθησαν can also be applied to ἀκολούθει, ἠκολούθει, ἀκολουθεῖν, etc. This significantly increases the number of “equivalent” phrases.

Some phrases can also be “equivalent” despite the fact that they don’t have the same number of words. For instance, the following segments of text can all be translated as “I send my messenger before you to prepare the way”:

ἐγὼ ἀποστέλλω τὸν ἄγγελόν μου πρὸ προσώπου σου ὃς κατασκευάσει τὴν ὁδόν σου ἔμπροσθέν σου

(from Matthew 11:10)

ἀποστέλλω τὸν ἄγγελόν μου πρὸ προσώπου σου ὃς κατασκευάσει τὴν ὁδόν σου

(from Mark 1:2)

ἀποστέλλω τὸν ἄγγελόν μου πρὸ προσώπου σου ὃς κατασκευάσει τὴν ὁδόν σου ἔμπροσθέν σου

(from Luke 7:27)

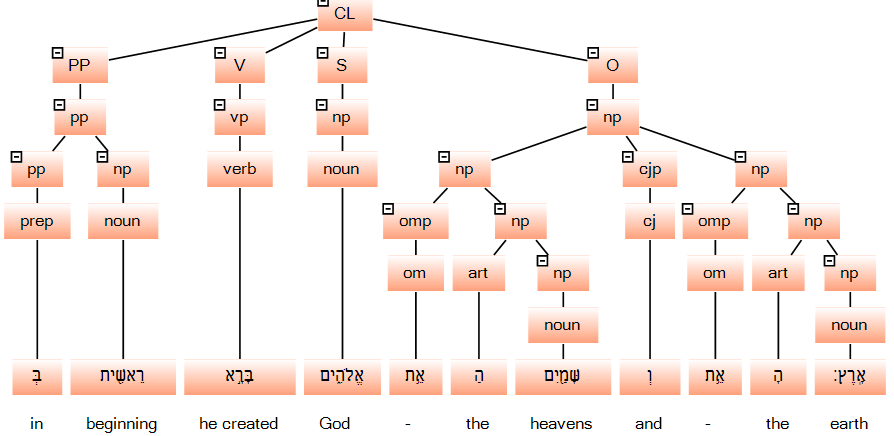
They differ only in (1) whether there is an overt subject (ἐγὼ) and (2) whether there is an extra ἔμπροσθέν σου (in front of you). If the segment from Matthew is in the translation memory, it should be able to help translate the segments from Mark and Luke.

To improve the coverage of the TM, we have compared all the phrases in the Hebrew and Greek texts and grouped them into clusters of similar phrases. When making suggestions, if no identical phrase is found in the TM, we will suggest the translations of phrases in the same cluster.

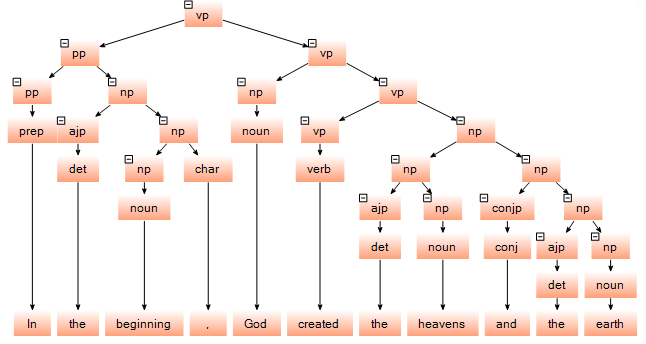
Another way to improve the re-usability of the stored translation units is to use synonyms. For example, the Hebrew words אבד, שׁמד, שׁחת, and נכה can all be translated as “destroy”. After one of these words have been translated, the translation can be provided as a suggestion for all the other words. We have built a database of Hebrew and Greek synonyms and it is to be used to improve the coverage of the translation memory.

**The Discontinuous Phrase Issue**

Hebrew and Greek are so called “non-configurational” languages where the word order is free and phrases can be syntactically discontinuous. Take Genesis 1:1 as an example (it is displayed from left to right for easy readability):

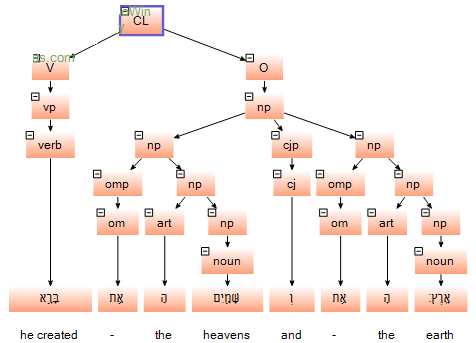
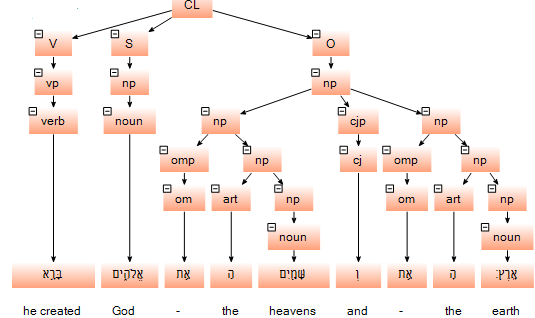


When translated into English, a configurational language, we have:



As we can see, “created the heavens and the earth” is a phrase in English, forming a single subtree. In Hebrew, however, “created” and “the heavens and the earth” are not able to form a phrase. They are discontinuous, being separated by the word for “God”. Since the phrases in the TM are based on the phrases/subtrees in Hebrew and Greek, “created the heavens and the earth” will not be able to get into the TM although it is translated and is obviously a re-usable unit.

To solve this problem, we create derived trees where the phrases can be continuous. This is made possible by the functional tags in the GBI trees: S for subject, V for Verb, O for object, and so on. Given the original tree of Genesis 1:1, for example, we can derive the following trees, among others:

With these trees, we can identify “created the heavens and the earth” and “God created the heavens and the earth” as phrases and store their translations in the TM. The derivation can also produce subtrees that correspond to “create the heavens”, “created the earth”, “God created”, “created in the beginning”, etc. In short, the derivation can generate a complete set of phrases in a verse, not only the ones in the original tree, but all the phrases formed by grammatical relations. This enriches the translation memory and maximizes the use of existing translations.

**Implementation**

The tree-based phrase alignment and TM-creation presented in this paper is being implemented in YouTranslate.Bible (YTB), a web-based Bible translation software that GBI is developing. As one of its features, this software provides automatic suggestion powered by automatic word alignment and phrase alignment. This is part of YTB’s efforts to make Bible translation accessible to everyone, everywhere. A minimally viable project of YTB will be available by the end of this year, ready to take on new Bible translation projects.

**Summary**

In this paper, we have seen how the richly annotated original Hebrew and Greek texts can help create and enrich the translation memory used in Bible translation. The morphological and syntactic analysis of the Biblical texts are fully leveraged to improve the accuracy and coverage of the translation memory. As a result, the translation process can be more efficient and more consistent, as existing content is put to maximal use to make suggestions for subsequent translation. The approaches taken here are only applicable in areas where the source text has been analyzed morphologically and syntactically. The Bible domain happens to be such an area, where a huge amount of linguistic work has been done for the purpose of understanding God’s word.