Documentation for Asia Bible Society’s SBL Greek New Testament Treebank (1.0)

Principal Authors: Randall Tan, Andi Wu

**1 Introduction**

Since a detailed syntactic analysis of biblical Greek and biblical Hebrew is of interest to Bible translation as well as linguistic and biblical studies, the Asia Bible Society launched projects to build treebanks of both the Greek New Testament and the Hebrew Bible. This document describes the SBL Greek New Testament version of Asia Bible Society’s Greek Treebank.

The Asia Bible Society’s Greek Treebank is automatically generated initially by a parser running a context-free grammar and then manually checked and edited. Because Greek appears to be configurational at the phrase level and non-configurational at the clause level, the Asia Bible Society’s Greek trees are designed to be hybrid trees: phrase structure trees for phrases and dependency trees for clauses. However, the dependency trees for clauses are also built from phrase-structure rules. Specifically, the grammar used belongs in the head-driven phrase structure grammar family. To achieve the desired goals of making the treebank sufficiently theory neutral as to be adaptable to different linguistic frameworks and ever-changing syntactic theories and adequately general to allow the scheme to be extended to additional Hellenistic and Classical Greek texts, a compromise was made between linguistic correctness and engineering convenience. As a result, our trees are not as complex as full-blown head-driven phrase structure trees that may be found in current linguistic literature, nor can we claim that we can supply the “correct” analysis for every sentence in the corpus.

**2 Overview**

The SBL Greek New Testament trees are built on the SBL Greek New Testament digital text. As the official website notes, “the SBLGNT is edited by Michael W. Holmes, who utilized a wide range of printed editions, all the major critical apparatuses, and the latest technical resources and manuscript discoveries as he established the text. The result is a critically edited text that differs from the Nestle-Aland/United Bible Societies text in more than 540 variation units” (<http://sblgnt.com/>).

Instead of building a new morphological parser, the SBL Greek New Testament trees rely on the morphological disambiguation provided by MorphGNT. MorphGNT, which stands for morphological tagging of the SBL Greek New Testament (<http://morphgnt.org/>), supplies 8 standard CCAT-based (Center for Computer Analysis of Texts at University of Pennsylvania) morphological attributes (person, tense, voice, mood, case, number, gender, degree) as well as part of speech and lemma information for each Greek word in the SBLGNT text.

2.1 Nodes and Attributes

Each node in a tree has zero or more child nodes. Terminal nodes are nodes that have no child nodes. All nodes in a tree that have child nodes are non-terminal nodes. The following table describes attributes in all terminal and/or non-terminal nodes of the trees besides the 8 CCAT-based standard morphological attributes mentioned above:

|  |  |  |
| --- | --- | --- |
| Attribute | Applies To |  |
| Surface (shows actual Greek) | Terminals  Non-terminals | The surface string of a terminal, or the surface string covered by a non-terminal. |
| Cat | Terminals  Non-terminals | Syntactic category |
| Start | Terminals  Non-terminal | The start position of a sub-tree (zero-based) |
| End | Terminals  Non-terminal | The end position of a sub-tree (zero-based)" |
| Rule | Non-terminals | Label of rule used to derive the top node of the tree |
| Head | Non-terminals | The zero-based index of the “Head” term for this tree node |
| UnicodeLemma | Terminals | Lemma of the surface Greek form in Unicode |
| Unicode | Terminals | Surface Greek form in Unicode |
| Type | Terminals | Currently distinguishes different types of nouns and pronouns |
| ClType | Non-terminals | Explicitly marks Verbless Clauses, Verb Elided Clauses, and Minor Clauses |
| HasDet | Non-terminals | HasDet value is True when a sub-tree is modified by a determiner |
| morphId | Terminals | An 11-digit unique id in the format  BBCCCVVVWWW  where  BB => zero-padded book, NT starts at 40  CCC => zero-padded chapter  VVV => zero-padded verse  WWW => zero-padded word index (for the word instance within the verse) |
| nodeId | Terminals  Non-terminals | A 15-digit unique id in the format BBCCCVVVWWWSSSL  where  BB => zero-padded book, NT starts at 40  CCC => zero-padded chapter  VVV => zero-padded verse  WWW represents the beginning position (the Nth word) of a node/sub-tree  SSS represents the SPAN of a node (how many words it covers)  L (Level) is used to distinguish nodes which have the same span (in cases of non-branching nodes) |

A terminal node corresponds to the base level of the syntactic analysis, usually corresponding to the type of speech analysis for each word as provided by MorphGNT. Eleven categories of parts of speech are distinguished, using the following abbreviations: adj, adv, conj, det, intj, noun, num, prep, ptcl, pron, verb.

2.2 Syntactic Categories at Word Level: Part of Speech Abbreviations Spelled Out and Defined

Adjective (adj): A word that belongs to a class whose members modify nouns. An adjective specifies the properties or attributes of a noun it modifies.

Adverb (adv): Narrowly defined, a word that belongs to a class of words whose members modify verbs for such categories as time, place, manner, direction, etc. Broadly defined, includes words that modify any constituent class of words other than nouns, such as verbs, adjectives, adverbs, phrases, clauses, or sentences. Some examples of this broader class are degree words and negatives.

Conjunction (conj): A word that syntactically links words or larger constituents and expresses a semantic relationship between them.

Determiner (det): A word that belongs to a class whose members modify nouns, expressing their reference. For the Greek article, this typically indicates definite reference or given status in the discourse. When used with words other than nouns, often functions to make the resulting phrase a nominal phrase.

Interjection (intj): A syllable or word that expresses emotion or directly addresses the audience. It is most often an exclamation.

Noun: A word that belongs to a class whose members may act as the subjects, objects or indirect objects of the verb or the object of a preposition. Includes words that typically refer to concrete or abstract entities like people, places, things and concepts.

Numeral (num): A word that expresses a number or a relation to a number in terms of frequency, quantity, or sequence. It is typically functioning adjectivally or in place of a noun. Only indeclinable words are marked as numerals. Declinable numerals (e.g., the Greek numerals one and three) have been categorized as adjectives.

Preposition (prep): A word that occurs before a nominal phrase, forming a single unit with it (a prepositional phrase) to express the prepositional phrase’s semantic relation to another unit within the clause (typically indicating when, where, how or why).

Particle (ptcl): A catchall term for words that do not belong to the main classes of words. It is typically invariable in form and partially overlaps with the broader definition of adverbs that modify at the clause or sentence level.

Pronoun (pron): A word substituting for a noun or nominal phrase whose referent is recoverable from the linguistic or extralinguistic context.

Verb: A word that belongs to a class whose members typically indicate events or actions. It governs the number and types of other constituents that may occur in a clause.

2.3 Syntactic Categories at Phrase Level Defined and Abbreviations Spelled Out

The phrase level is the intermediate level between word level and clause level. Phrase level nodes are either non-terminal nodes that are the immediate parent nodes of the part-of-speech terminal nodes or parent nodes of other phrase level non-terminal nodes that together form multi-word phrases. From the perspective of the clause, single words or combinations of words form phrases, which are the minimal constituents with a specific function at the clause level. Only six categories of phrases are distinguished, using the following abbreviations: adjp, advp, np, nump, pp, vp.

Adjectival Phrase (adjp): A phrase with an adjective as its head. Adjectival phrases typically function as dependent modifiers within a nominal phrase. In many cases, an adjectival phrase functions in place of an elided noun and thus functions as a nominal phrase (this is the reason why adjectival phrases are often promoted to nominal phrases in the tree). Some adjectives function as the Predicate in an attributive relational clause (whether with a Verbal Copula or in a Verbless clause). Some adjectives have an Adverbial function at the clause level.

Adverbial Phrase: A phrase with an adverb as its head. Adverbial phrases typically have Adverbial function at the clause level, indicating the “when”, “where”, “why” and “how” of the verb or predicate. Some adverbial phrases also modify nouns or adjectives.

Nominal Phrase: A phrase with a noun (or adjective or pronoun functioning as a noun) as its head. Nominal phrases typically have Subject, Object, or Indirect Object function at the clause level. Nominal phrases are also often modified by a preposition, forming a prepositional phrase.

Numeral Phrase: A phrase with an indeclinable numeral as its head (including all single word numerals, which are automatically promoted to numeral phrase).

Prepositional Phrase: A phrase with a preposition and the nominal phrase it governs, forming a single unit with the nominal phrase to express semantic relation to another unit within the clause (typically indicating when, where, how or why). Prepositional phrases typically have an Adverbial function at the clause level. Some prepositional phrases modify a noun or adjective (indicating a relationship of where, when, how, why for the object of the preposition to the head noun or adjective).

Verbal Phrase: A phrase with a verb as its head.

2.4 Syntactic Categories at Clause Level Defined and Abbreviations Spelled Out

The clause level differs from the phrase level by using a dependency structure. The terminology used to describe the functions of clause level constituents is purposely conservative for ease of understanding (closer to traditional grammatical terminology in biblical Greek grammar) and to preserve a clearer link between clause level terminology and phrase and word level terminology. In general, different parts of speech have the following clause-level functions: Verbs have a Verbal Function, except for relational "to be" verbs, which function as Verbal Copula. Nominals (nouns, pronouns, and adjectives) can have three basic functions in relation to the verb: Subject, Object, and Indirect Object. Nominals can also function as a Predicate to either the Subject (Predicate in Verbal Clauses with a Verbal Copula or Predicate in Verbless Clauses) or the Object (Object Complement, considered a Second Object). Adverbs and prepositional phrases at the clause constituent level function adverbially in relation to the verb. Eight categories of clause level function are distinguished, using the following abbreviations: ADV, IO, O, O2, P, S, V, VC.

Adverbial Function (ADV): A constituent that represents when, where, how, or why of a proposition.

Indirect Object Function (IO): A constituent that represents the intended recipient of the action of a proposition.

Object Function (O): A constituent that represents the patient or goal of the action of a proposition.

Second Object Function (O2): Some verbs take two objects. There are two main types. The first type involves two accusative objects in the Greek, an object of person (the first object) and an object of thing (the second object). For example, he will teach you (object of person) all things (object of thing). The second type also involves two accusative objects in the Greek, but the first is the direct object and the second is an object complement. The object complement predicates a description of the direct object (e.g., "king" is the object complement in "God appointed David as king").

Subject Function (S): A constituent that represents the agent of typically transitive verbs and the single argument of intransitive verbs.

Predicate Function (P): A constituent that makes an attribution or identification about the Subject of a non-verbal clause.

Verbal Function (V): A constituent that represents the action/event of a proposition. In verbal clauses, this is the head of the clause, on which all other clause constituents depend.

Verbal Copula Function (VC): A constituent that facilitates an attribution or identification about the Subject. Verbal copulas are commonly known as copula verbs or linking verbs. This typically corresponds to a “to be” verb in English, which indicates a relation between the Subject and the Predicate rather than any action/event.

2.5 Sentence and Clause Type Terminology

This preliminary version of the SBL Greek New Testament trees focuses on sentence and clause structures, which will serve as the building blocks of higher levels of analysis. Clear groupings of clauses had been connected together into sentences, especially when subordination or ellipsis is involved. However, the identification of sentences has not undergone the same vigorous process of testing as the identification of structures at the clause level and below. The label "Sentence" itself has been used for the lack of a clearly preferable alternative term. At the clause level, four basic types of clauses have been distinguished: Verbal Clause, Verbless Clause, Verb Elided Clause, and Minor Clause. By default only Verbless Clauses, Verb Elided Clauses, and Minor Clauses are explicitly marked with a ClType label. All clauses without an explicit ClType label can be assumed to be Verbal Clauses.

Sentence: A sentence is composed of one or more clauses. It is the highest grammatical unit in the current version of the trees.

Verbal Clause: The clause represents a grammatical unit that expresses a proposition. Verbal clauses are the most typical form of the clause, with an explicit verbal element as the head of the clause (i.e., the constituent on which all other clause constituents are dependent).

Verbless Clause: Verbless clauses are relational clauses of identification or attribution without an explicit copula verb. Verbless clauses typically consist of a core of a Subject and a Predicate, with additional Adverbial clause constituents possible.

Verb Elided Clause: Verb elided clauses are clauses that imply the carrying over of Verbal Function (usually from the previous clause). They are clauses without Verbal Function that are not clauses of identification or attribution.

Minor Clause: Minor clauses are clauses without any predication — no Verbal Function or Predicate Function — that contain no assertions or propositions. They function interpersonally (vocative direct address or interjection to gain attention/alert) or textually (left-dislocated focus noun phrases, e.g. as for the game, I did not get to watch it.).