Applications of Trees in Syntax

Abiyyu Avicena Ismunandar/13517083
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
13517083@std.stei.itb.ac.id

***Abstract*—Linguistics is a scientific study of languages and its structures. It studies how every part of languages from how words are formed to sentences structure. This paper will focus on a branch of linguistics called Syntax which covers sentences structures. Syntax and discrete mathematics intersect in their usage of trees. The paper will discuss the basics of syntax, how trees model sentences structure, and a brief discussion of generalizing the model.**

***Keywords*— Linguistics, Syntax, Trees, X-Bar Theory**

I. INTRODUCTION

Human languages have evolved for centuries, yet we have not completely figured out how we understand languages. There are still ongoing researches on how the sounds we make are understood by other people who also make sounds that we understand. The field of study that researches this phenomenon is called Linguistics. It covers every part of the languages we speak. There is a branch that studies how different sets of sounds if put together becomes something that has meaning for us called Semantics. Also, a branch that studies how sentences are structured so that it becomes meaningful and comprehendible called Syntax.

It may be surprising that a field that studies languages intersects with discrete mathematics. Their intersection happens in the study of Syntax where Trees are used to parse through sentences and group them into phrases. Parse trees are commonly used in the study of Syntax as they help researchers discerns how sentences are structured. Using these trees, they can hypothesize grammars that are understood by the speakers of the natural language.

II. BASIC THEORY

A. Syntax

In order to understand how and why we apply our Tree knowledge in Syntax, we need to know what is Syntax. The word itself is derived from the Greek word *syntaxis*, which means arrangement [1]. In Linguistics, the Syntactic branch studies phrases and sentence formation of a language.

First, we need to define what is a sentence. In linguistics, there are many definitions for a sentence, however, we will use the syntactic definition that defines a sentence as "largest unit to which syntactic rules can apply" [1]. Syntactic rules refer to the arrangements of words that makes a sentence. This will be further elaborated on the next sub-chapter. The syntactic rule that usually govern the English language is the subject and predicate combination. Most English sentences can be broken down into a subject and a predicate. However, some sentences do not have this syntactic structure such as emotive sentences, imperatives, elliptic, and small talk phrases [1].

Sentences containing subject and predicate can then be broken down into three types of sentences: simple sentences, compound sentences, and complex sentences. Simple sentences are sentences that contains at least one subject and one predicate (e.g. "Haifa read *Dilan"*). Compound sentences are two or more simple sentences combined with conjunctions such as *and* or *or* (e.g. "Ainun read *1984* and Haifa read *Dilan*"). Complex sentences are sentences that we usually see in everyday English such as "Ronald made Haifa read *Dilan*". Since, complex sentences are sentences in which a syntactic type is embedded into another sentence. In previous example the simple sentence "Haifa read *Dilan*" becomes the object for the verb *made*. However, complex sentences cannot be broken down into two simple sentences since "Ronald made" is not a syntactically correct sentence which will be explained later [1].

Lastly, in syntax there is a term that explains whether a combination of words and phrases form a sentence or not. The term grammatical means the combination forms a sentence, while ungrammatical means the combination does not form a sentence [2]. Consider these two sentences:

1) Ronny likes Chelsea
2) Likes Ronny Chelsea*

Unlike operators in mathematics that has the form of infix, postfix, and prefix, the English language does not have that. The first sentence is what considered to be a grammatical sentence since the string of words are what is considered to be an English sentence. On the other hand, the second sentence is ungrammatical since it does not form a sentence known in the English language. We can also say that the second sentence is syntactically ill-formed [2].

Before moving on to syntactic rules, it is important to mention that there is a difference in the understanding of grammar in Linguistics and everyday knowledge. Usually, people associate grammaticality with how English sentences must be formed such as the formation of present tenses, past tenses, etc. However, in Linguistics a sentence does not have to follow these set of prescriptive rules. For example, the sentence "We is family" may be considered an ungrammatical sentence to the general public whereas by linguist's standard that sentence is grammatical since the string of words form an English sentence that can be understood by the native speaker [5]. Therefore,

when reading this manuscript try to forget the prescriptive grammar that has been taught in English class.

B. Syntactic Rules

To start the discussion of Syntactic rules we need to introduce the syntactic categories that are used in the study of Syntax. Since we are only discussing syntactic rules that are used in the English language, the syntactic properties will also be the ones used in the English language. The syntactic properties are Sentence (S), Noun (N), Determiner (Det), Adjective (Adj), Noun Phrases (NP), Verb Phrase (VP), Preposition (P), Preposition Phrase (PP), Transitive Verb (TV), Ditransitive Verb (DTV), Sentential Verb (SV), and Adverb (Adv) [2]. These syntactic categories will have their order of occurrences according to a specific set of rules each phrase has, which will be discussed next.

Noun phrases (NP) are constituent of a sentence that is built around a noun [3]. A noun phrase consists of a noun and other elements such as determiners and adjectives. Combining the NP elements will then derive NP rules. For example, from the sentence "the big dog" we can derive an NP rule NP →Det Adj N which means a combination of a determiner, adjective and a noun will create a noun phrase. Notice how the rule gives clarity on how a grammatical sentence should be formed since a sentence that has the determiner after the noun or adjective will result into something ungrammatical.

Next is Prepositional phrases (PP). It is a constituent that can consists of a preposition and its NP complement [2]. Prepositional phrases have a simple set of rule, but is crucial to the fundamental properties of a language. Consider this sentence "the dog in the green yard" where "in" is preposition and "the green yard" is the NP as we have defined previously. Therefore, PP has the rule PP→P NP. Simple yet this reveals that a finite set of rule can create infinitely many type of sentences since the PP rule allows recursion–which is the fundamental property of languages.

The last phrase that will be discussed is Verb phrases (VP). The rule for verb phrases will depend on whether the verb is a Transitive verb (TV), Ditransitive verb (DTV), or a Sentential verb (SV) [3]. What sets them apart is the quantity of objects that need to follow the verb. Transitive needs to be followed by an object, Ditransitive needs to be followed by two, and Sentential does not need to be followed by an object. Therefore, a VP rule can be as followed VP → V for SV, VP → V NP for TV, VP → V NP NP for DTV.

Putting all these rules together we can then create a sentence that is grammatical. The rule for a sentence is S → NP VP. Therefore, a sentence like "The big dog under the table ran" is a grammatical sentence since it follows the rules we have previously mentioned. It should be noted that the rules above are just ones of the many other rules that the English language has. Since there are also rules that allows recursion of sentences inside Prepositional Phrases which allows infinitely long string of sentences.

C. Tree Definition

In Discrete Mathematics, a tree is defined a graph that has no simple circuits [4]. Which means a tree is a graph that has a definite destination. An example of a tree can be seen at figure A, which shows for each family member there is a distinct path to them from the root. The example also shows the more generally used tree which is a rooted tree. A rooted tree is a tree that has one of its vertex specified as the root and every edge is directed away from the root [4].

In mathematical notation, a Tree is defined as T (V, E) where T is a graph T with V vertexes and E edges. The amount of edges in a tree follows $E = V - 1$ which means the number of edges of a tree is its vertex minus by one.

Trees can then be further broken down to different types of trees based on their n-ary and whether they are balanced or not [4] . N-ary refers to the amount of edges that can come out of a vertex in a tree. For example, a binary/2-ary tree is a tree that has all of its vertex have only two edges coming out of it. A balanced tree means a tree that has the least amount of parity between its levels which means its height is relatively the same for each side of the tree [4]. Level refers to the length a vertex to its root. While height is the maximum of the levels in the tree.

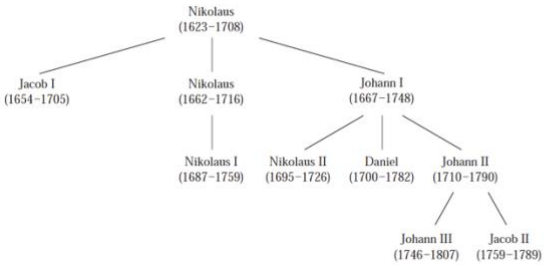

Figure A An Example of a tree

D. Important Terminologies

1. Parent: Is a vertex that is unique to another vertex so that there is an edge that connects the parent to the other vertex. For example, in fig B vertex a is a parent to vertex b since there is unique edge that connects them both. Also note how vertex a is also the parent for vertex c and d.
2. Child: When a vertex has a parent then that vertex is a child of the parent vertex. In fig B, vertex b, c, and d are children of vertex a since they all are adjacent to the parent a.
3. Siblings: Vertices that have the same parent. We can say that vertex f and g are siblings since they both have the same parent, vertex b, shown in figure B.
4. Ancestors: Vertices that are in path the from the root to the destined vertex. For instance, in fig B it can be seen that vertex e's ancestors are c and a.
5. Descendants: When a vertex has ancestors then that vertex is descendant of the ancestors' vertices. In fig B, vertex f is the descendant of vertex b and a.
6. Leaf: A vertex that has no children. Vertices j, g, e, and d are leaves since they do not have any children as shown in figure B.
7. Internal vertex: Vertices that has children. However, if a root is the only vertex in the tree then the root is a leaf.
8. Level: Length of the path of a unique vertex in a rooted tree from its root.

9. Height: The maximum level of the vertices.
10. Balanced: If a rooted n-ary tree has all leaves are at levels equal to the height or height minus 1.

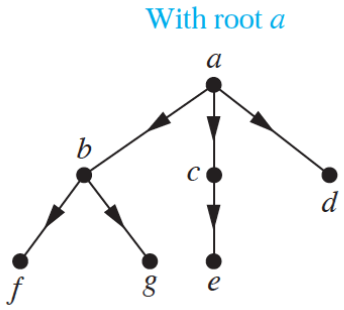

Figure B A Rooted Tree

III. APPLICATION OF TREE IN SYNTAX

In this chapter, the collaboration of discrete mathematics and the study of sentence structure will be seen through the usage of trees. Previously, the syntactic rules have been explained and is given a few examples. Those rules will then become our basic knowledge for creating trees that shows a sentence structure.

A. Identifying Constituents

In order to create syntax trees, constituents of the sentence have to be identified. It has been mentioned that phrases be it noun, verb, or preposition are constituents of the sentence. There are many ways to test if a set of words can be group as a constituent or not. This manuscript will cover three constituent tests which are clefting, answers to a question, and pro-form substitution [2].

Clefting is a test where the tested constituent is displaced to the left of the sentence. A sentence in the form of "it was X that Y" where X is the constituent being tested and Y as the remainder of the sentence. If after being displaced the sentence remains grammatical then the group of words are a constituent. An example of clefting:

Given the sentence: "The cat was sleeping on the desk"

Tested constituent: *on the desk*
Cleft form: it was *on the desk* that *the cat was sleeping.*
Conclusion: *on the desk* is a constituent

Tested constituent: *on the*
Cleft form: it was *on the* that *the cat was sleeping desk**
Conclusion: *on the* is not a constituent

From the above examples, it can be seen how one can identify if a group of words are constituents or not. As long as the resulting cleft form is grammatical then the group of words are a constituent.

Another constituency tester is Answer to Question. Similar to clefting, a group of words need to be chosen then it will be tested. After the tested constituents have been chosen, construct a question using the remaining string of words. If the tested constituent answers the question formed then it passes the test and can be declared as a constituent. If it cannot answer the question then it is not a constituent. When forming the question or answer try to use the original string of words as much as possible. Example of Answers to Questions constituency test:

Given the sentence: "The cat was sleeping on the desk."

Tested Constituent: *sleeping on the desk*
Question: What was *the cat* doing?
Answer: *sleeping on the desk.*
Conclusion: *sleeping on the desk* is constituent

Tested Constituent: *on the*
Question: Was *cat sleeping desk*?*
Conclusion: *on the* is not a constituent

Using different ways of testing, still yield the same result. It can be seen that for both tests *on the* is not a constituent since it creates an ungrammatical sentence. Therefore, if the tested constituent creates an ungrammatical question or does not answer a form question then the string of words is not a constituent.

The last form of constituency testing is pro-form substitution. After choosing a string of words to be tested if they are a constituent, it will then be substituted with one word in the form of pro-forms. If the resulting sentence still creates a grammatical sentence then the grouped words is a constituent. In pro-form substitution, the tested constituent will be replaced with pro-nouns such as he/him, she/her, it, they/them, one, that, or they can be replaced with pro-verbs such as *do*, *be*, *have*, *there*, *then*, and *such.*

Given the sentence: "The cat was sleeping on the desk."

Tested Constituent: *the cat*
Pro-form: <u>she</u> *was sleeping on the desk*
Conclusion: *The cat* is a constituent

Tested Constituent: *on the*
Pro-form: *the cat was sleeping* <u>{there, then, such}</u> *desk**
Conclusion: *on the* is not a constituent

Using three different tests, it can be securely concluded that combination of *on the* is not a constituent. Since in the pro-form substitution there are pro-forms that can replace the *on the*, therefore, *on the* is not a constituent. This information will be important in the next discussion of forming trees.

It is advisable to use more than one constituency test since sometimes a constituent that passes one test does not pass another test. Therefore, use as many constituency tests before determining a constituent in a sentence.

B. Identifying Parts of Speech

After figuring out the constituents of the sentence, next comes the part where each word in the sentence are labeled into their parts of speech. Parts of speech includes nouns, verbs, adjectives, preposition, adverbs, determiners, auxiliaries, degree words, and complementizers [6].

In this chapter, the goal is to create a tree for this sentence: "Adel works at a car retailer to pay for her education". Therefore, to ease the process of making the tree the parts of speech will be identified in this sub-chapter while also serving as an example of identifying parts of speech.

When identifying parts of speech, it is best to start with the easier ones to identify such as nouns, verbs, and preposition then move on to the ones that are harder to identify such as determiners, auxiliaries, and degree words. From the sentence above it can be identified that the words can be grouped into:

Nouns: Adel, car, retailer, education, her
Verb: works, pay
Preposition: to, at, for
Determiners: a

Turns out the sentence does not contain any other parts of speech and, therefore, have labeled correctly into their separate parts of speech group. It is possible to first label the tree then identify the constituents or vice-versa.

C. Drawing A Tree

After identifying the parts of speech and the sentence's constituents, it is time to draw the tree. Our goal sentence will not be drawn here, instead the sentence "the cat was sleeping on the desk" will be used for this sub-chapter since some of the constituents have been identified. Knowledge of syntactic rules will also be used during the tree drawing. Such rules include: S → NP VP, NP → Det Adj N, VP → V NP, and PP → P NP. Usually in an English language syntax tree, the tree will usually skew towards the right, therefore, try to give space towards the right for the tree to grow.

Firstly, when drawing a syntax tree start from the top instead of the bottom. It can be easily deduced that the sentence will become the leaves for the tree and from there we can start. However, doing so could cause an asymmetrical tree. Therefore, start from the top with S as the root of the tree. S will have NP and VP as its branch. At first, the tree should look like figure C. For most of the time, an English sentence will start with this structure [6].

Next, put the identified constituents under each syntactic category of the tree. For instance, put the subject of the sentence in the first NP branch. So, for our example, it has been identified that the subject for the sentence is *the cat* and it is an NP since it follows the syntactic rule NP → Det N with *the* as the Det and *cat* as the N. The syntax tree should now look like figure D. After identifying the subject of the sentence, next is figuring out the VP. Since, *sleeping on the desk* has been identified as a constituent and is a VP since it follows the VP syntactic rule of VP → V PP, PP → P NP, NP → Det N. Where *sleeping* is the V, *on* is the P, *the* is the determiner, and *desk* is the N. Consequently, *was* is left as a lone constituent which is a VP connected to another VP as *was* is an intransitive verb. It has been previously mentioned that an intransitive verb does not have an object attached on it instead the second verb *sleeping* acts as the object of *was*. Therefore, for an intransitive verb the VP will have the rule VP → VP. We have identified where all the constituents will be placed and, therefore, have all of the words in the sentence in the tree. The complete tree should look like figure E.

The previously mentioned fact how a usual English sentence syntax tree skews more to the right since it is seldom that the subject of the sentence will contain a lot of branches—unless for a test question where the professor wants to give a hard time to their students. It can also be noted that the leaves are the words in the sentence and if read from left to right will yield the sentence we started with.

Sometimes there may be ambiguous sentences where if the constituents are group differently could cause a whole new meaning to the sentence. Therefore, there can be numerous ways of drawing a syntax tree. As long as the procedure of drawing the syntax tree is followed closely, the procedure is identifying parts of speech and constituents then drawing the tree, then the resulting tree shouldn't stray away to something incorrect.

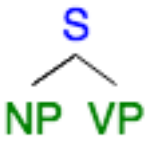

Figure C S --> NP VP

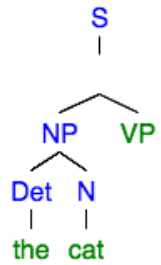

Figure D Added Subject

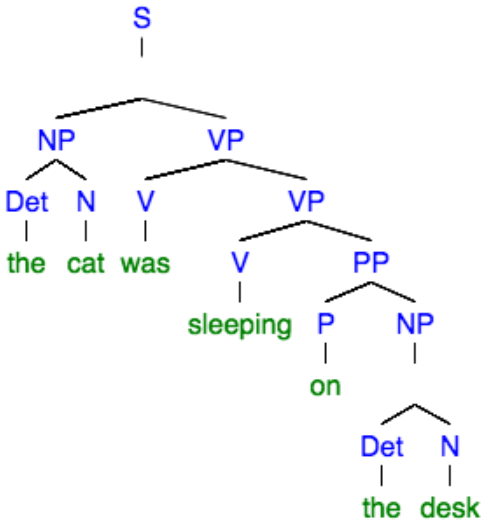

Figure E Complete Tree For "the cat was sleeping on the desk"

D. Generalizing a Tree

There is a way to generalize a tree, that is by using the X-bar theorem [7]. The main idea of using X-Bar theory is to generalize the syntactic rules of NP, VP, PP, etc, by explaining their similarity between the rules.

Generalizing the syntactic rules does not change anything in the surface, however, the theorem shows a general rule that every phrase usually has. The X in the X-Bar theory can be replaced with any variable over categories. When XP is used the

X can then be replaced with N, V, P, etc [7].

The X-bar theorem states that most phrases have the rules of the following form [7]:

XP: ZP X'
X': (YP) X'
X': X' (YP)
X': X (WP)

Notice by using the X-bar form rules the tree created will have a binary tree form since most rules generate into two branches.

The X-bar theorem can be applied to any Phrase structure rules, it will now be applied to the Noun Phrase syntactic rules. Given this NP structure rule: NP → (Det) (AdjP) N (PP), it will then be broken down into a more generalized rule. In figure F, there is a tree diagram of the NP for "this big book of poems with the blue cover". To generalize the rule, a technique quite similar to the pro-form substitution will be utilized. By substituting *book of poems with the blue cover* with *one* yields the sentence *the big one*. Meaning the group of words are a constituent but it is not shown in the tree structure. More embedded constituents can be found by utilizing the pro-form substitution on many other parts of the sentence such as substituting *book of poems* with *one*, again, which yields the sentence *this big one with the blue cover*. After finding each embedded constituent in the NP tree structure a more grouped tree structure like the one in figure G can be seen. From the final NP tree diagram the generalized NP rule can be yield as follows:

NP: (D) N'
N': AdjP N'
N': N' PP
N': N (PP)

N' means it is the intermediate nodes for NP (N-bar). Notice how the resulting NP rule have similar form with XP rules previously mentioned. The NP rule also branches out in a binary way; therefore, an X-bar tree will result in a binary tree.

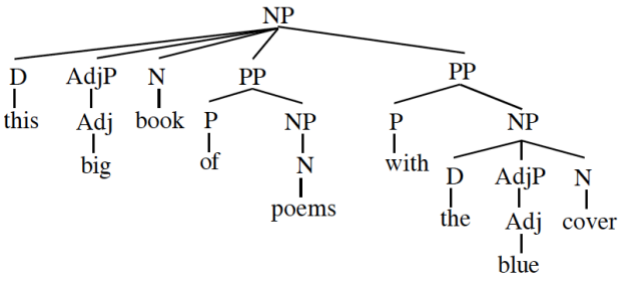

Figure F Tree Diagram of an NP [7]

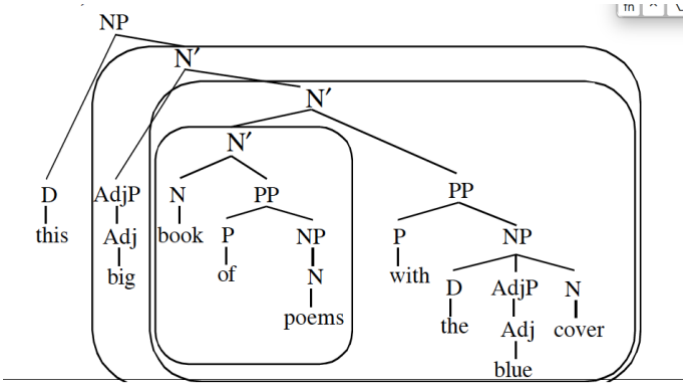

Figure G Tree Diagram with All Embedded Constituents Grouped [7]

E. *Putting It to the Test*

Previously it has been mentioned that the tree for the sentence "Adel works at a car retailer to pay for her education" will be made. The parts of speech for the sentence have been identified too so now the constituents of the sentence need to be identified.

1) Identifying Constituents
   a) Testing *Adel*

      Using Pro-form substitution to test *Adel*
      Pro-form: she *works at a car retailer to pay for her education*

      Using Answers to Questions to test *Adel*
      Question: Who *works at car retailer to pay for her education?*
      Answer: *Adel*

      Conclusion: *Adel* is a constituent
   b) Testing *works at a car retailer to pay for her education*

      Using Answers to Questions
      Question: What does *Adel* do?
      Answers: *works at a car retailer to pay for her education.*

      Using Pro-form substitution:
      Pro-form: *Adel* do so

      Conclusion: *works at a car retailer to pay for her education* is a constituent
   c) Testing *at a car retailer*

      Using Clefting:
      Cleft form: It was *at a car retailer* that *Adel works to pay for her education*

      Using Pro-form substitution:
      Pro-Form: Adel works there to pay for her education.

      Conclusion: *at a car retailer* is a constituent.

   d) Testing *to pay*

Using Clefing:
Cleft form: It was *to pay* that *Adel works at a car retailer for her education*

Using Answers to Questions:
Question: Why does *Adel works at a car retailer for her education?*
Answer: *to pay*

Conclusion: *to pay* is a constituent

e) Testing *for her education*
Using pro-form substitution:
Pro-form: *Adel works at a car retailer to pay that*

Using Clefting:
Cleft form: It was *for her education* that *Adel works at a car retailer to pay*

Conclusion: *for her education* is a constituent

f) Testing *a car retailer*

Using Answers to Questions:
Question: Where does *Adel work to pay for her education?*
Answer: *A car retailer*

Using Pro-form substitution:
Pro-form: *Adel works* there *to pay for her education*

Conclusion: *a car retailer* is a constituent

g) Testing *pay*

Using Clefting:
Cleft form: It was *pay* that *Adel works at a car retailer for her education*

Using Pro-form substitution:
Pro-form *Adel works at a car retailer to* do so *for her education*

Conclusion: *pay* is a constituent

h) Testing *her education*

Using Answers to Question:
Question: Why does *Adel works at a car retailer?*
Answers: For *her education*

Using Clefting:
Cleft form: It was *her education* that *Adel works at a car retailer to pay for*

Conclusion: *her education* is a constituent

From the constituents above notice that pay becomes an object rather than a verb. However, for the VP the rule will become VP → V (PP+) meaning the VP can have more than one PP as its branch. From the constituents, it can also be inferred that *Adel* is an NP and is the subject of the sentence. Therefore, a syntax tree diagram can now be drawn. Remembering again that the phrase rules are as follows: S → NP VP, NP → (Det) N, VP → V (PP+), PP → (P) NP. It is important to mention that syntactic properties that are bracketed means that they do not have to appear for the phrase to grammatical. Therefore, from all the above constituents and facts about the syntactic rules a tree diagram as seen in figure H can be drawn.

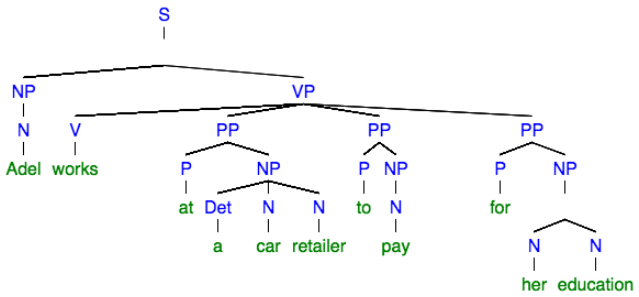

Figure H Tree for Adel works at a car retailer to pay for her education

IV. CONCLUSION

Hopefully this manuscript shows how two quite different subjects can collaborate. Discrete mathematics may sound like a foreign subject for students who are studying syntax and vice-verse is also true. It may only seem that the intersection is only in the application of trees. The potential of collaboration does not stop at the application but it may continue even to how Computer Scientist design compilers. Since this is exactly how different programming languages understand codes that are written then translated into bytecode. Therefore, the question comes back to the reader, what will you do with the knowledge that you have gained?

VI. ACKNOWLEDGMENT

I would like to thank Allah SWT first and foremost as Allah has continued on giving me the strength to finish this manuscript even during the most hectic week of my life. I would also like to thank all the Discrete Mathematics faculty for giving me the opportunity to share some of my knowledge in Linguistics that I have acquired overseas. It may not all be correct; however, I hope it may have some correctness so that the people who reads this may acquire new knowledge.

REFERENCES

[1] Vadja, Edward J., "Syntax", Western Washington University, Bellingham, WA, pandora.cii.wwu.edu/vajda/ling201/test1materials/syntax.htm.
[2] Department of Linguistic, *Language Files: Materials for an Introduction to Language and Linguistics* (E-book), 12th edition, 2016, pp. 202-228.
[3] Duncan, Bonnie. "Phrase Structure Rules" (Website), sites.millersville.edu/bduncan/221/anderson8/9.html.
[4] Rosen, Kenneth H. *Discrete Mathematics and Its Applications* (E-Book), 7th Edition, *McGraw-Hill*, New York, NY, 2012, pp. 745 – 769.
[5] Lippi, Rosalina. *English with an Accent: Language, Ideology and Discrimination in the United States* (E-book), *Psychology Press*, New York, NY, 1997, pp. 55 – 65.
[6] McCulloch, Gretchen. "How to Draw a Syntax Tree, Part 8: A step by step treedrawing guide, with gifs" (Blog), *All Things Linguistics*,

allthingslinguistic.com/post/102131750573/how-to-draw-a-syntax-tree-part-8-a-step-by-step.
[7] Roberts, Ian. *Comparative Syntax*,1997, London Arnold.

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 10 Desember 2018

Abiyyu Avicena Ismunandar