Linguistics 21N - Linguistic Diversity and Universals: The Principles of Language Structure

Ben Newman

March 1, 2018

1 What are we studying in this course?

This course is about *syntax*, which is the subfield of linguistics that deals with how words and phrases can be combined to form correct larger forms (usually referred to as sentences). We're not particularly interested in the structure of words (morphemes), sounds (phonetics), or writing systems, but instead on the rules underlying how words and phrases can be combined across different languages. These rules are what make up the formal grammar of a language. Formal grammar is similar to what you learn in middle and high school English classes, but is a lot more, well, formal. Instead of classifying words based on meaning or what they "do" in a sentence, formal grammars depend a lot more on where words are in the sentence.

For example, in English class you might say an adjective is "a word that modifies a noun", such as *red* in the phrase the *red ball*. A more formal definition of an adjective might be "a word that precedes a noun" or "the first word in an adjective phrase" where the adjective phrase is *red ball*. Describing a formal grammar involves writing down a lot of rules for a language.

2 I-Language and E-Language

Before we get into the nitty-gritty grammar stuff, I want to take a look at two ways language has traditionally been described by linguists. One of these descriptions centers around the rules that a person has in his/her mind for constructing sentences. This is referred to as the I-language. Another describes what constructions are typically observed in a corpus (body) of textual data. This is called the E-language.

- **I-language**: The I-language consists of the set of rules that a native speaker uses to construct grammatically correct, comprehensible sentences. The "I" stands for "Innate" as many proponents of the I-language view

of language believe that humans possess some innate quality that allows them to learn how to use language. The argument is that there is so much language that children are not exposed to when they are babies, that they cannot possibly be as good at forming sentences as they are without some inborn language-understanding capabilities. The I-language itself is composed of the rules we normally think of as grammar rules. These will be the subject of the rest of this summary.

- **E-language**: The E-language refers to all of the language that we observe from reading and listening on a day-to-day basis. The "E" stands for "Empirical" as this is the language that we actually interact with. It is sometimes useful to think of the E-language as the result of the I-language; the I-language has the rules, and the E-language is the product of those rules. Many proponents of E-language claim that there are no fixed rules for language, as there are edge cases that can be ambiguous (like "near" and "kind/sort of")[1], so you can only know what a word means or what is grammatically correct by seeing if it is used commonly enough or matches commonly-used constructions. All of the current field of statistical natural language processing depends on the E-language for training language models.

In reality, these two language descriptions exist on a continuum. Neither can exist without the other. If you just have the I-language you can?t actually say anything, and if you only have the E-language, what you say won?t make any sense (as there are no rules that language follows).

3 Word Order

When looking at syntactical differences between languages, one of the most glaring differences is in the relative position of the subject, the verb and the object. We should first define what these concepts are (for the purpose of this section—these definitions might change a little bit as we go on). These initial definitions are going to be a little fuzzy because we haven?t developed the framework for dealing them in a formal way yet.

- Subject: The "thing" that is "doing" the action or is in a state of being
- Verb: What the subject is doing
- Object: Everything that?s not part of the subject or verb. Also what the subject is doing the verb action to/on.

These are really amorphous, circular definitions, and don't really make sense unless you already know what a subject, verb, and object are, but they're what we have for now.

[1] See the introduction of Chris Manning and Hinrich Schutze's ("Foundations of Statistical Natural Language Processing") for more discussion of these words? ambiguities.

When describing word order we call the subject S, the verb V, and the object O. So English is an SVO language because the subject precedes the verb which precedes the object:

I ate the pie

I is the subject, ate is the verb, and the pie is the object. Other languages might be SOV (like Japanese), VOS, VSO, OSV, or OVS. These word orders are found with varying frequencies in human languages, with SOV and SVO being the most common, VOS and VSO following, and OSV and OVS being almost exinct. The reason for this unequal distribution might be clear soon.

4 A Formal Grammar of English

A good few weeks of the class was spent solidifying a basic formal grammar of English. It is important to note that many different formal grammars exist, and the one presented here is just one example, and goes over the process behind constructing a formal grammar.

There are two main units of a grammar: the word and the phrase. The word is the atomic syntactic unit, i.e. it cannot be broken up into simpler units. Each word has its own syntactic label outlining where it can go in a sentence or phrase. Ideally words of the same syntactic label should be interchangeable. For example, any noun should be able to take the place of any noun and any verb should be able to take the place of every verb. (We know that in practice this can't be the case, but we will discuss this a little later) The phrase is the larger syntactic unit, made up of words and phrases. Phrases are labeled very similarly to how words are labeled: A phrase should be able to stand in for a phrase with the same name. Phrases are labeled after their "heads", which are the words that determine the name of the phrase. (This is very circular, but we'll circle back to it later. It has to do with which word is controlling what the other words in the phrase can be).

Ok, so now let's start constructing a formal grammar of English. As we said before, a formal grammar is just a list of rules. These rules take the form of:

<Phrase> → (optional phrase/word) required phrase/word <head> required phrase/word (optional phrase/words)

Let's start with simple nouns (**N**). In English class these are often defined as "people, places, or things." This is a useless definition from a structural standpoint. Let's circularly define a noun as the head of a noun phrase (**NP**). (We have to start somewhere, so might as well start here).

$$\boldsymbol{NP \rightarrow N}$$

Nouns can be preceded by articles (*a*, *the*) and other words that belong to the syntactic class called determiners (**D**). This group also includes possessives like *my*, *her*, *its*, as well as demonstratives like *this*, *that*. In English these precede the nouns (*like the milk*) and are optional components for the noun phrase (*happiness*)

$$\boldsymbol{NP \rightarrow (D)N}$$

Adjectives can also be placed into these rules, falling between the determiner and the noun (*the white milk*). These are optional, however:

$$\boldsymbol{NP \rightarrow (D)(A)N}$$

Nouns can also optionally be followed by prepositional phrases (**PP**) and complementizer phrases (**CP**). We will talk about these more later, but for now:

$$\boldsymbol{NP \rightarrow D(A)N(PP)(CP)}$$

This rule will DEFINITELY change before we finish, but it's a good place to start.

Next let's move onto adpositions: *of*, *for*, *on*, *in*, etc. In English these are at the beginning of adpositional phrases (they **pre**cede the rest of the phrase), so they are called **pre**positions (**P**). The constructions that prepositions precede are often referred to as the "object of the preposition." This object can be a noun phrase (*in the bathtub*) or a complementizer phrase (**CP**) (which, I assure you, we will get to later)

$$\boldsymbol{PP \rightarrow P(NP)(CP)}$$

Ok, let's move on to verbs and verb-like words now. We're going to define a verb very similarly to how we defined a noun. A verb is the head of a verb phrase. The rest of the verb phrase consists of the verb's "object", which can be a noun phrase (**NP**) (*I ate the pie*), a complementizer phrase (**CP**), an adjective (**A**) (*I am happy*), or even a prepositional phrase (**PP**) (*I live in California*).

$$\boldsymbol{VP \rightarrow V(NP)(CP)(A)(PP)}$$

Next let's take a look at auxiliaries (**Aux**). These are similar to verbs, but always precede them, such as *has eaten*, *must eat*, is eating. These can be bundled into auxiliary phrases (**AuxP**), which contain an auxiliary and a verb phrase (**VP**).

$$\boldsymbol{AuxP \rightarrow AuxVP}$$

Now let's combine verbs with their subjects. Subjects are usually noun phrases and verbs are either verb phrases or auxiliary phrases. Combining a subject and a verb gives a sentence (**S**). These are usually referred to as "clauses" in English classes.

$$\boldsymbol{S \rightarrow NPAuxP}$$

$$\boldsymbol{S \rightarrow NPVP}$$

Now let's talk about complementizers (**C**). Complementizers are words that connect sentences to larger phrases and include subordinating conjunctions and adverbs like *that*, *which*, *who*, *because*, *if*. Complementizer phrases (**CP**) consist of the complementizer and the sentence that follows it. These are usually referred to as subordinate clauses in English classes.

$$\boldsymbol{CP \rightarrow CS}$$

Complementizer phrases can also be subjects of sentences: *That you can be happy makes me so glad!*

$$\boldsymbol{S \rightarrow CPAuxP}$$

$$\boldsymbol{S \rightarrow CPVP}$$

There are going to be changes made to these rules and a final list (with some extras) can be found in the Appendix.

A note on replaceability: The way that I defined a word or phrase included the idea that any word/phrase of the same syntactical label could be replaced by a different word/phrase of the same syntactical label. This is obviously not always the case. You can say *I ate on the plane*, but you can not say *I devoured on the plane*, because you need to specify what you devoured. (This is the whole idea of transitive versus intransitive verbs). All words can only be surrounded by specific kinds of syntactic categories (like transitive verbs need to be followed by noun or complementizer phrases), or even particular words. The specifics of what words/phrases need to be around a particular word are called the "arguments" of that particular word. Arguments can be a little messy, and the class did not really go into them that deeply. Arguments do however provide an alternative way to define the head of a phrase: the head of the phrase is the word (or words) whose arguments specify the other syntactic components in the phrase. So in the verb phrase *devoured the pie on the plane*, the head of the phrase (the verb *devoured*) has the arguments that specify the inclusion of the noun phrase (*the pie*) and the prepositional phrase (*on the plane*).

5 Parameters

Parameters are general characteristics of languages that lead to linguistic variation. We will see concrete examples of parameters in the next sections, but the important thing to not for now is that a single parameter can often be responsible for seemingly distinct linguistic variation. For the parameters that we're going to look at, there are two distinct settings, each one leads to a different effect, but the same setting tends to lead to the same effect cross-linguistically. In other words, if language A exhibits setting 1 of a parameter, if language B exhibits the same setting of the setting of the same parameter, languages A and B will share certain characteristics.

Now let's get into some specific parameters so this gobbly-gook makes sense.

6 Head Directionality Parameter

As was defined in the "Formal Grammar of English" section, the head of a phrase is the word that determines how the phrase is named. The head directionality parameter has two settings: head initial or head final. In head initial languages (like English), the phrase head is found at the front of the phrase. So, for example, the adpositional phrases are **pre**positions: *at home*, *to the beach*, *about the author*. In head final languages (like Japanese), the phrase head is found at the end of the phrase. So, for example, the adpositional phrases are called **post**positions.

In an overwhelming number of cases, this carries through for all phrase types in a given language. If you look at the grammar rules we came up with above, an overwhelming number of the phrases begin with their heads. Some languages are of mixed headedness (i.e. some phrases are head-initial while others are head-final), but these tend to quickly align themselves to one headedness or the other.

You might notice that the rule for noun phrases we constructed before is not head initial:

$$\boldsymbol{NP \rightarrow (D)(A)N(PP)(CP)}$$

We can change this by adding another type of phrase, called an adjective phrase(**AP**) that is head initial and consists of an adjective and another type of phrase: (*proud parent*, *proud of me*, *proud that you came here tonight*).

$$\boldsymbol{AP \rightarrow A(NP)(PP)(CP)}$$

Now the noun phrase rule might look something like:

$$\boldsymbol{NP \rightarrow N(PP)(CP)}$$

And what about the determiner? Well maybe we can create a determiner phrase (**DP**) with rules like:

$$\boldsymbol{DP \rightarrow DNP}$$
$$\boldsymbol{DP \rightarrow DAP}$$

Now everything looks nice and head initial. Except it doesn't really work out that simply. For one, it's sort of weird to have two rules for determiner phrases. One of these rules associates a determiner phrase with an adjective phrase even though determiners have traditionally been thought of as relating to nouns. Also, adjectives can be modified, can't they? There are multiple types of words that modify adjectives (most are lumped into the the "adverb" category in English class). One of these types is a group known as degree words (**Deg**): *very*, *quite*, *enormously*, etc. These words all in front of the adjectives they describe—*He is very disappointed in you*—and suggest a rule as follows:

$$\boldsymbol{AP \rightarrow (Deg)A(NP)(PP)(CP)}$$

This rule is clearly not head-initial. So, while English is very often head initial, there definitely are some more unclear cases.

This minor discrepancy underlies some of the fuzziness underlying how these parameters appear to operate. They point out overwhelming trends, but are not necessarily unbreakable laws. There is also some linguist discretion when defining which phrases exist in a language. As long as there is evidence to support a certain phrase structure, there is no reason why it cannot exist.

7 Null Subject Parameter

The next parameter I'll be overviewing is the null subject parameter. At the most basic level, the setting of this parameter indicates **whether or not a language requires an explicit subject in every sentence**. To English speakers, this might seem like a bit of a no-brainer. We?re taught in English classes that all sentences have a subject and a verb, and one particular setting of this parameter appears to contradict that. One of the clearest examples of this parameter at work is in differences between English and Spanish:

I run to school every day (English)

Corro a la escuela cada día (Spanish)

In the Spanish sentence, the word for I, *Yo*, never appears. Even without a subject, the Spanish sentence is perfectly comprehensible. At this point, there is a clear distinction to make: the Spanish sentence definitely has a **thematic** subject. It is clear that I am talking about myself running in the Spanish sentence. This context comes from the verb conjugation and maybe even the context of the larger conversation. Despite the existence of the **thematic** subject, there is no **structural** subject—here is no word in that sentence that you can point to and say "that is the subject."

There are some other manifestations of the null subject parameter in a language:

1. **The structural subject is relatively free to move around in the sentence. It can either precede or follow the verb.**

   So in Spanish these are both acceptable:

   Yo corro a la escuela

   Corro yo a la escuela

   But in English you can say

   I run to the school

   But not

   **Run I to the school*

2. **There are no expletive pronouns.**

   Expletive pronouns are present in non-null subject parameter languages' weather expressions and impersonal expression. These pronouns have no

actual meaning, and are only present because the syntax of a language requires them to be present for a grammatically correct sentence. In English these pronouns are *it* and *there*. You can see them in sentences like:

It is necessary that he goes to the doctor. and

There are three little pigs in the house.

The words It and there add no additional meaning to the sentence. This is also apparent in weather-related expressions like *It is raining*. What is raining?? What is *it*?? You wouldn't say *The sky is raining* or *The weather is raining*. You'd just say that the thing that is raining is *IT*.

In null subject parameter languages, there are no expletive pronouns:

Es necesario que él vaya al médico

Hay tres cerditos en la casa

Even weather expressions have no expletive pronouns:

Está lloviendo

3. **Questioning embedded subjects requires no change in complementizer**

   This one is a little complicated, but applies to sentences of the form:

   (a) *I told him that Suzy would go to the store?*
   In this sentence, the embedded subject is the subject of the embedded (or subordinate) clause: *Suzy*. If you for some reason didn?t hear her name correctly, and wanted to ask about who was going to the store you would ask:

   (b) *Who did you tell him would go to the store?*

   Ok, so this is all good, but did you notice that there was a change in the complementizer between a) and b)? In a) the complementizer is *that*, and in b) the complementizer seems to have disappeared. This is actually the result of English having a "null" complementizer—in certain instances, it's acceptable to omit a spoken complementizer, and the sentence is still correct (*He said he would go to the movies with me* is as acceptable as *He said* ***that*** *he would go to the movies with me*).

   In non-null subject parameter languages, this change in complementizer is necessary when questioning these embedded subjects. You can try putting the *that* back into b), but it just doesn?t work:

   (*Who did you tell him* ***that*** *would go to the store?* — it just sounds wrong) In French this difference is even clearer, because there is no null complementizer?you actually have to change the word.

   In null subject parameter languages, there is no change:

   (a) *Yo le dije* ***que*** *María iría a la tienda.*

(b) *¿Quién le dije* ***que*** *iría a la tienda?*

And that wraps up another parameter! To summarize, the null-subject parameter controls whether:

1. A structural subject is necessary in all sentences
2. The structural subject can be inverted in simple sentences
3. Expletive pronouns exist in the language
4. Questioning embedded subjects requires a change of complementizer

8 Verb Raising Parameter

The verb raising parameter has a lot to do with tense and auxiliary verbs, but before we get into it, we have to discuss a common way of representing sentences used by linguistics: tree diagrams.

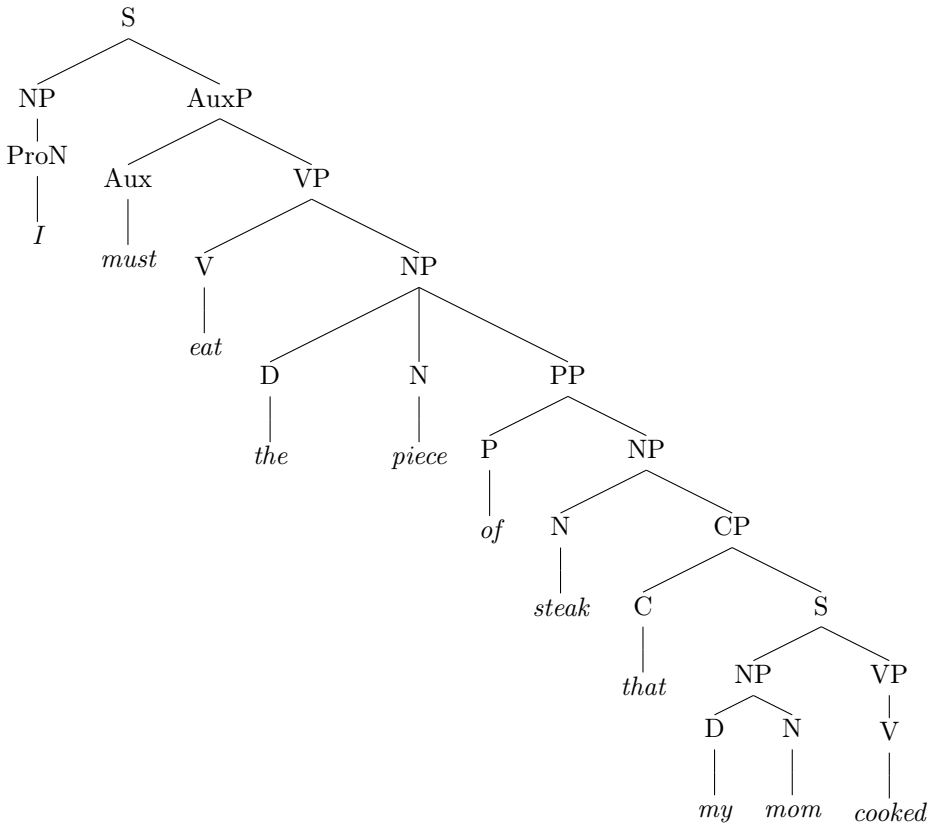
Let's take the English sentence: *I must eat the piece of steak that my mom cooked*

If we wanted to discuss the structure of the sentences, we have to determine which category each word falls into. A good start might be labeling each of the words in this sentences with its category:

I must eat the piece of steak that my mom cooked

PRO (N) AUX V D N P N C D N V

The problem with this is that it doesn't really show us anything about the structure of the sentence. We also don't have any evidence of phrases if we label the sentence in this way. What we want to do is associated certain elements with each other. For example, *must* should be closer to *I* than *I* is to say *mom*. How should these elements be associated? These associations tend to be hierarchical, which means that we can create a sentence tree. This is similar to how sentences are diagrammed:



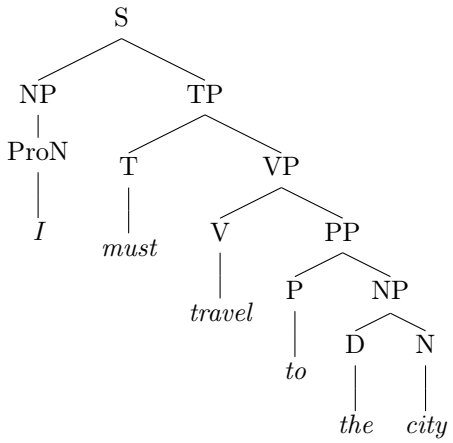
While this may be overwhelming to look at, it illustrates the structure of the sentence very clearly. Each word fits into a category and a phrase. (you can make these trees online on a website such as ("http://mshang.ca/syntree/")

Now we're going to talk about the parameter at hand.

Up until this point we've basically been considering any verb that goes before another verb as an "auxiliary." This includes the classic auxiliaries like a form of the verb *to have* in the phrase *to have eaten* and *to be* in the phrase *to be eating*, as well as the modals like *can (could)*, *shall (should)*, *will (would)*, *must (might)*, and *ought*, to name a few. These two groups are actually distinct categories, but for now we're going to say that they affect the **tense** of the verb in the verb phrase that makes up their other component. As a result, we're going to make a huge simplification and put all auxiliaries and modals under a single umbrella term we're going to call **T** for tense. (It was briefly mentioned in class that auxiliaries don't really fall into this category, but they behave pretty similarly to the words that do, so we include them.) So now we're going to re-write the rule for auxiliaries and auxiliary phrases as the rule for T's and T phrases (**TP**).

$$\boldsymbol{TP \rightarrow TVP}$$

Here's an example sentence: *I must travel to the city*

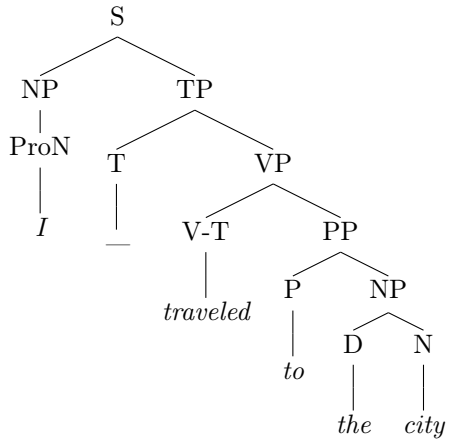


Now there's one aspect of tense that we have to address. Sometimes the tense isn't specified by a separate word. In English's present and past tense, the tense is specified by a certain ending (or lack of ending):

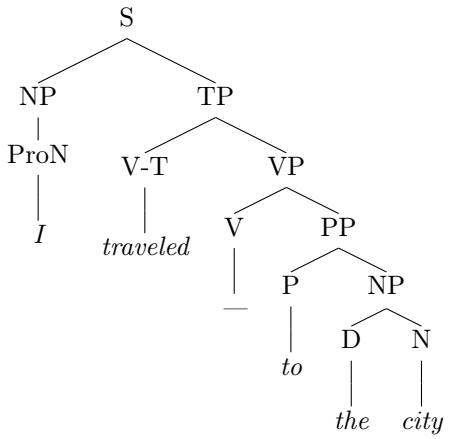
I travel to the city.

I traveled to the city.

How do we reconcile this difference with what we described as tense above? Well, one thing we can do is allow for the **merging of elements**. So the past tense *traveled* really consists of the verb *travel* and the T *ed*. When they merge, it is called V-T-fusion. Ok, so this works. Let's try to draw a syntax tree for the second sentence:



Wait... or is it:



Huh...

The fused V-T word can live in the spot of the T or the spot of V, so which is it? This question underlies our next parameter: the verb-raising parameter. There are two settings of this parameter:

1. **Verb-raising**: when the verb is fused with the T, the whole verb-t fusion blob thingy resides in the position of the T (the verb "raised" in the syntax tree to the position of the T.)

2. **T-lowering**: when the verb is fused with the T, the whole verb-t fusion word resides in the position of the V (the T "lowered" in the syntax tree to the position of the V.)

This is all good, but how do we tell which setting English follows? We just said that it was ambiguous in our example. Well, what we do is we find a word that resides between the T and the verb when they are separate words, and we use this word as a signpost—whichever side the fused V-T falls on indicates the setting of this parameter. (If the V-T falls on the left, then it?s V-raising, and if it falls on the right, then it?s T-lowering.) Now we need a signpost. Hmm so our example before was *I must travel to the city*, so what would we say if we don't ever want to travel to the city: *I must never travel to the city.* And where does *never* fall? Between the T and the verb.

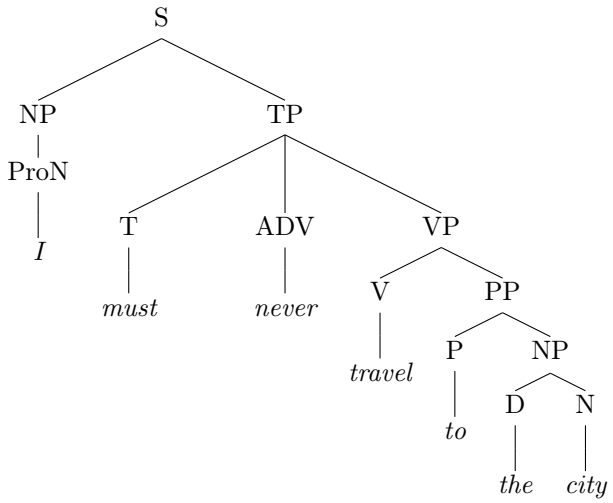
We can use this class of words (certain adverbs) as our signpost: *never*, *always*, *sometimes*, etc. We can modify our **TP** rule to include these optional adverbs:

I travel to the city. I traveled to the city.

$$\boldsymbol{TP \rightarrow T(ADV)VP}$$

(Note that this is still head-initial)

The sentence *I must never travel to the city* would have this tree:

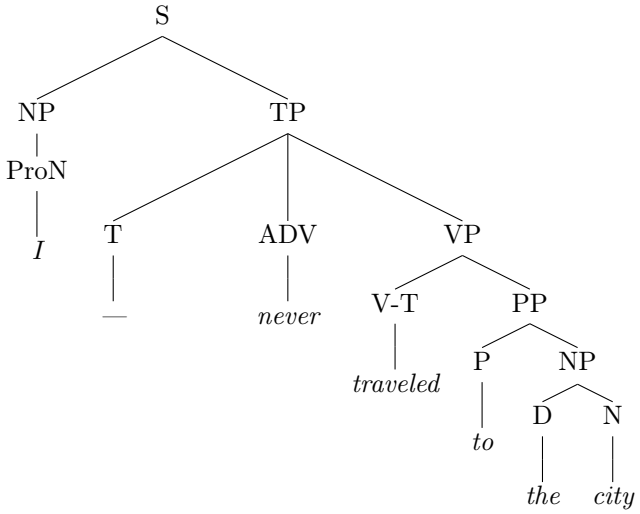


So now, when we fuse the verb and the T and include one of these adverbs, what do we get?

I never traveled to the city

** I traveled never to the city*

So the word *traveled* is on the right side of the adverb *never*, indicating that the T lowered to the position of the verb:



9 Question Movement Parameter

The final parameter we discussed in class was the Question Movement Parameter. This is a very straightforward parameter, and requires a lot less discussion than the others. The parameter has to do with where question words fall in interrogative statements. These question words are called **WH**'s in the linguistics literature and refer to words like *who*, *what*, *where*, *when*, *why*, *which*, and

how. Sentences that contain these words are called WH-phrases (**WHP**).

Let?s take a look at the question formation rules for English:

Let?s take the following sentence:

John walked the dog

Let?s say we want to know what John was walking.

1. Replace the word with the appropriate WH:

   John walked ***what***

2. Now we move the WH to the beginning of the sentence and make whatever adjustments we need:

   What did John walk?

(The whole did business is just a quirk of English.)

In English, the WH moves to the front of the sentence. Other languages stop after step one and merely replace the word to question with the appropriate WH. The parameter that decides which one a language does is known as the question movement parameter. It has two settings:

1. WH-movement: The WH moves to the front (or back) of the sentence in a question

2. WH-*In-Situ*: The WH stays where it is in the sentence in a question

English is a WH-movement language while Chinese is not. Interestingly, English is known as a special type of WH-movement language called **single WH-movement language**. This is because if more than one item is being questioned in an interrogative statement, only one of the WH's moves to the front:

I threw the ball to John $\rightarrow$ ***Who*** *did I throw* ***what*** *to?*

The WH *what* referring to the ball stays in place.

10 Appendix

English Syntax Rule List (THIS IS NOT COMPLETE):

$$\boldsymbol{PP \rightarrow P(NP)(CP)}$$

$$\boldsymbol{NP \rightarrow (D)N(PP)(CP)}$$

$$\boldsymbol{NP \rightarrow PRO}$$

$$\boldsymbol{NP \rightarrow NAME}$$

$$\boldsymbol{AP \rightarrow (Deg)A(NP)(PP)(CP)}$$

$$\boldsymbol{DP \rightarrow DAP}$$

$$\boldsymbol{DP \rightarrow DNP}$$

$$\boldsymbol{VP \rightarrow V(NP)(CP)(AP)(PP)}$$

$$\boldsymbol{AuxP \rightarrow AuxVP}$$
$$\boldsymbol{TP \rightarrow T(Adv)VP}$$

$$\boldsymbol{S \rightarrow NPT}$$
$$\boldsymbol{S \rightarrow CPT}$$

$$\boldsymbol{XP \rightarrow XCONJ.X}$$

(Basically you can chain together phrases of the same type with conjunctions like and, or, but, etc.)