

**Natural Language Processing**  
**Prof. Pawan Goyal**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 21**  
**Syntax introduction**

Welcome to the 5th week of this course. So, we have so in this week, we are going, we will be moving up, the hierarchy of solving this language processing tasks. So, we had started with discussing about how do we handle individual words? How do we use, how do we make use of the word or information using language models and then how do we assign various grammatical categories; a part of speech categories to the words that is what we are doing till now.


Now, we will go to next level where we will try to, see can we arrange these words in certain groups. So, this is what we study in this topic of syntax. So, how are the words being arranged together and what can we; how we can automatically find out this arrangement given a new sentence? This will be the topic of parsing that we will discuss in this week.

What is syntax? So, in general syntax refers to the way the words are arranged together and also you will see, what is the relationship between various words and the word groups that is what we will talk about in syntax, so, when we were talking about language models, we had discussed what is the importance of modeling word order. So, which words occur after what are the words? How we can make use of that in assigning probability for a sentence or finding out the next word in completion task or in spelling correction task, similarly when we were talking about part of speech task, we saw what are the grammatical categories, part of speech categories, so this defines in one sort of equivalence class for words that is all these words behave like they are verbs behave in some equivalent manner, similarly all these words are nouns, they behaving in some equivalent manner, similarly all these are adjectives. So, you are defining some equivalent classes.

(Refer Slide Time: 02:25)

*What is Syntax?*

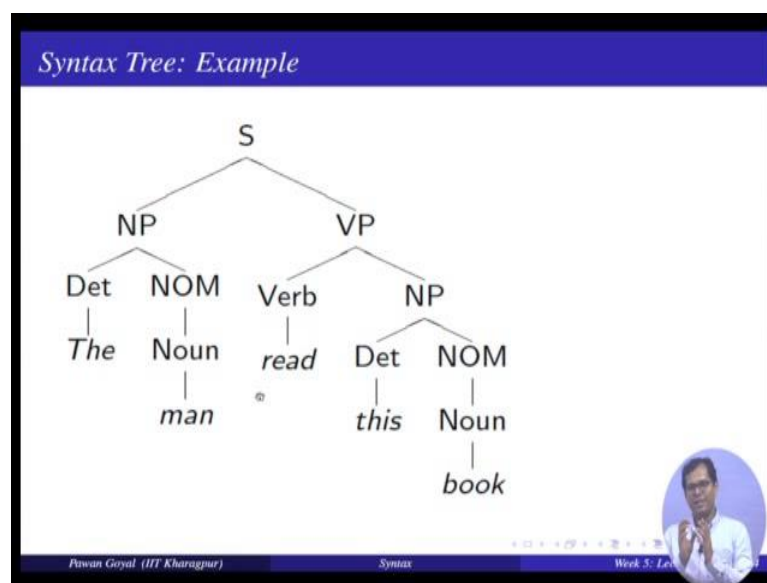
- Refers to the way words are arranged together, and the relationship between them.
- **Language Models:** Importance of modeling word order
- **POS categories:** An equivalence class for words
- More complex notions: constituency, grammatical relations, subcategorization etc.



Pawan Goyal (IIT Kharagpur) Syntax Week 5: L

Now, in syntax we will find out some more complex notions like word is constituency, what are different grammatical relations among the words? What groups or what words are grouped together in constituency and sub categorization etcetera also included in this in this topic and yeah just to make you understand, what is this notion of syntax? So, you can relate to that if you see this particular meme. So, I hope you remember this particular character from Star Wars, this syntax the only part of the language it is not, but important is it is right. So, you can see the way words are arranged normally in the way we speak and how they arranged in this particular sentence that was the special characteristics of this character.

(Refer Slide Time: 03:17)



Now so what are we are going to study in syntax. So, let me give you a simple example, So, I have the sentence, the man read this book. So, by syntax, we are trying we are trying to find out what are the various groups of word that are coming together for example, in part of speech text we found out that the word the as part of speech of determiner man as the part of speech of noun read as verb and. So, on now we are going one level up. So, now, we are saying this is determiner this is noun this is kind of nominal, but they both together make a phrase noun phrase.

I am saying the man is a noun phrase; similarly this book is also a noun phrase. So, there are 2 noun phrases here, the man and this book, but when the verb read or any verb comes before the noun phrase it makes a verb phrase.

All these 3 words act as a single unit of a verb phrase, there is a no such unit for man and read, there is a unit for read this book and the man and then I go up saying that a noun phrase and verb phrase are making the sentence and this gives me complete hierarchy structure of how the words arranged in the sentence the sentence is nothing, but a noun phrase and the verb phrase this noun phrase contains a determiner and noun verb phrase contains a verb and noun phrase which contains and so on. So, this is the complete hierarchy of the sentence that I come to know by this syntax tree and that is the; what is the topic of this week, how do we come up with such syntax tree for some sentences, what is the particular formulation that we will use.

(Refer Slide Time: 05:07)

*Defining the notions: Constituency*

*Constituent*  
A group of words acts as a single unit - phrases, clauses etc.

*Part of Speech - "Substitution Test"*  
The {sad, intelligent, green, fat, ...} one is in the corner.

*Constituency: Noun Phrase*

- Kermit the frog
- they
- December twenty-sixth
- the reason he is running for president

Pawan Goyal (IIT Kharagpur) Syntax Week 5: Lecture 5

Let me start by defining some basic notions like what is constituency? So, in the last example we saw that a group of words; they act they acting as some single unit and you can call them as phrases clauses etcetera. So, in part of speech, we could have done this substitution test. So, I have this sentence, there is a fill in the blank, one in the room and I can fill in any adjective, the green one, the fat one and intelligent one, sad one, all that can be filled in.

So, I can fill in any word that belongs to that particular part of speech category, now here it will be a particular constituent, it will be a particular group of words that can behave similarly like here. So, all these 4 things; Kermit, the frog, the December, 26th the reason he running for president, all these are noun phrases and they can occur in a given context. So, now, for substitution test, you can substitute any of this these 4 noun phrases and yeah we will see an example where all these can come in the same in a very very similar context.

(Refer Slide Time: 06:25)

*Constituent Phrases*

Usually named based on the word that heads the constituent:

<i>the man from Amherst</i>	is a Noun Phrase (NP) because the head man is a noun
<i>extremely clever</i>	is an Adjective Phrase (AP) because the head clever is an adjective
<i>down the river</i>	is a Prepositional Phrase (PP) because the head down is a preposition
<i>killed the rabbit</i>	is a Verb Phrase (VP) because the head killed is a verb

*Words can also act as phrases*

*Joe grew potatoes*  
Joe and potatoes are both nouns and noun phrases  
Compare with: *The man from Amherst grew beautiful russet potatoes.*  
Joe appears in a place that a larger noun phrase could have been.

Pawan Goyal (IIT Kharagpur) Syntax Week 5: LA

How do we name these constituent phrases? So, last slide I was showing some noun phrases. So, why do we call them noun phrases or something else? So, usually the names are given based on the words that are heading these constituent, what is the head and usually speaking you can find the head by the word that can substituted for the whole thing, let me take the first example the man from Amherst, this is the phrase, they are 4 words, now which of the 4 word do you think can substitute the whole thing, now that can be used in the grammatical function of the complete unit and that will be man where the man from Amherst, the word man can be used to denote the grammatical function of the whole unit. So, the head here is a noun; man. So, this will be called a noun phrase because the head man is a noun.

Similarly, extremely clever; the head here is clever; this is the adjective. So, this is called in adjective phrase down the river here; head is down preposition. So, this will called in prepositional phrase, killed the rabbit; the head is killed, the word killed which is a verb. So, this is called verb phrase. So, like that we are, we defined, what are the constituent by taking what is the head of that phrase?

Now in general, words can also act as phrases. So, a phrase need not have always multiple heads, a single word can also be a phrase. So, let us take the simple example Joe grew potatoes. So, Joe itself it is a noun phrase, potato also a noun phrase, they are nouns, but also a noun phrases, in this case now compare the sentence with the man from

Amherst grew beautiful russet potatoes. So, what do you see? So, instead of Joe I have substituted the man from Amherst, a complete 4 word unit that is again a noun phrase and beautiful russet potatoes instead of potatoes they are still noun phrases. So, what happens in the sentence? Joe appears in a place where you could probably put a larger noun phrase.

Now this gives a very nice idea about the structure of the sentence, in this sentence, I am having a noun phrase and a verb phrase, verb phrase contains a verb and noun phrase and noun phrase; you can either put a single word like Joe or you can put multiple words like the man from Amherst similarly in that noun phrase you can put potatoes or it becomes a noun phrase like beautiful russet potatoes. So, this gives me lots of idea about how words are grouped and arranged together.

(Refer Slide Time: 09:20)


### Evidence that constituency exists

*They appear in similar environments*

Kermit the frog comes on stage  
They come to Massachusetts every summer  
December twenty-sixth comes after Christmas  
The reason he is running for president comes out only now.  
 But not each individual word in the constituent  
 \*The comes out... \*is comes out... \*for comes out...

*Can be placed in a number of different locations*

Constituent = Prepositional phrase: On December twenty-sixth  
On December twenty-sixth I'd like to fly to Florida.  
 I'd like to fly on December twenty-sixth to Florida.  
 I'd like to fly to Florida on December twenty-sixth.  
 But not split apart  
 \*On December I'd like to fly twenty-sixth to Florida.  
 \*On I'd like to fly December twenty-sixth to Florida.



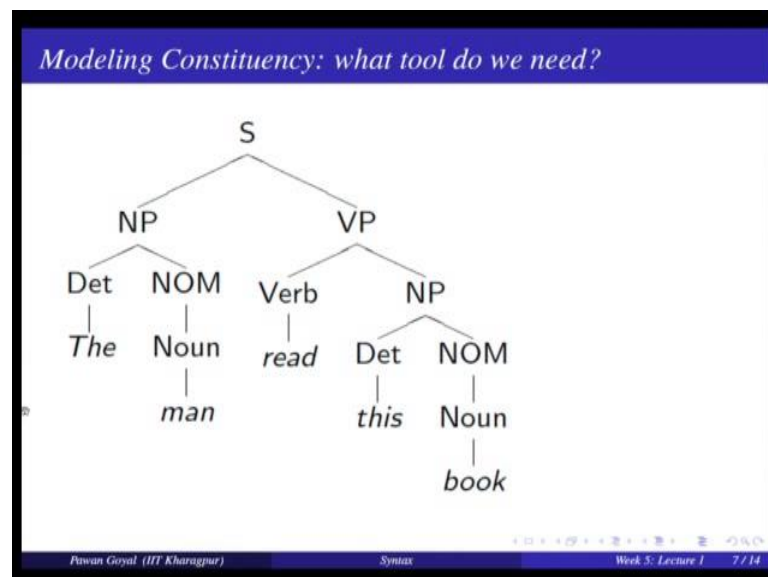
Pawan Goyal (IIT Kharagpur)
Syntax
Week 5: Le...

Now, there is some evidence that questions actually exist in language, yes, there are 2 different evidences, one is that this phrase appears in very very similar environment. So, far I have talked about the 4 phrases that noun phrases that discussed in one of the various phrases like let us see these examples Kermit, the frog comes on stage, note it, they come to Massachusetts every summer, December 26th comes after Christmas, the reason he is running for president comes out only now. So, all these 4 noun phrases are coming in a very similar context of the word say comes yes, but I cannot take any individual word from here and put that in the context. So, I cannot take the word the, and say the comes

out or I cannot say is comes out and I cannot say for comes out in this sentence, I can only say the reason he is running for president comes out, only now not an individual word. So, this whole thing behaves as a single unit and they occur in similar context all the noun phrase occurring in similar context, this is got evidence that is got the constituency, each has to be there.

What is the other evidence? So, the evidence is that this whole phrase is together can locate many locations in the same center. So, for example, if I take on December 26 that is the prepositional phrase and I can put it in many different phrases in the sentence on December 26, I would like to fly to Florida, I would like to fly on December 26 to Florida or I would like to fly to Florida on December 26, all 3 are valid sentences where this complete phrase on December 26 as we put in multiple location, but you cannot break this into 2 parts and put it in phrases. So, you cannot say on December I would like to fly 26th to Florida or on I would like to fly December 26th to Florida you cannot say that. So, this will as a single group it cannot be split of part. So, these are some evidence that constituency actually exists in the language.

(Refer Slide Time: 11:44)



Now, what is the formal tool by which we can model this constituency? That is how words are arranged together which words come together which words do not come together and what groups make a sentence and what groups make a verb phrase what group make a noun phrase; how can I model all that what is the formulation and if you have take any

course on formal language in automated theory all theory computation you might already know that the formulate that we can use that in context free grammar.

(Refer Slide Time: 12:21)

**Modeling Constituency**

*Context-free grammar*  
The most common way of modeling constituency

*Consists of production Rules*  
These rules express the ways in which the symbols of the language can be grouped and ordered together

*Example*  
Noun phrase can be composed of either a ProperNoun or a determiner (Det) followed by a Nominal; a Nominal can be more than one nouns  
 $NP \rightarrow Det\ Nominal$   
 $NP \rightarrow ProperNoun$   
 $Nominal \rightarrow Noun \mid Noun\ Nominal$

Pawan Goyal (IIT Kharagpur) Syntax Week 5: Lecture 1 8 / 14

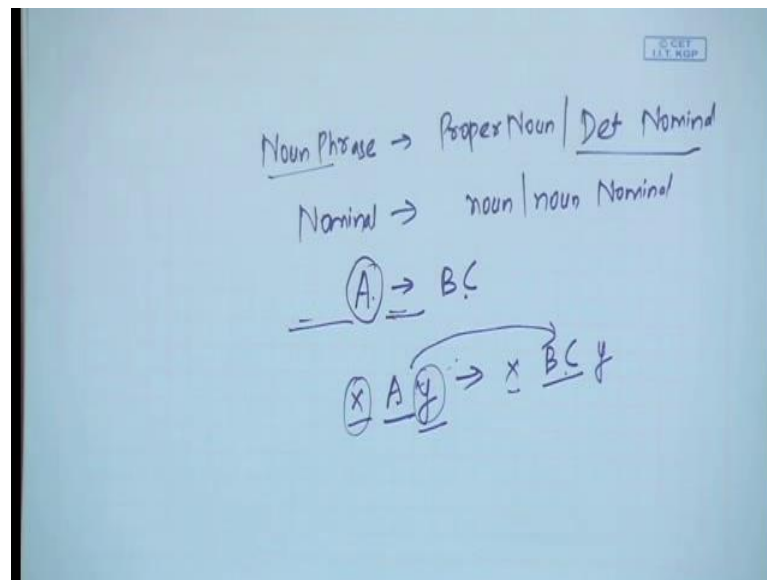
This is the most common way of modeling constituency. So, in one of the earlier lectures we talked about the (Refer Time: 12:28) languages by using deterministic finite automata of finite automata. So, this is context for context free languages by using by the context free grammar.

Now, so I will not going in very very basics, I will just talk briefly how do we use context free grammar, I will also defined the notions for the context free grammar. So, in the case of context free grammar what you will have you will have some sort of production rules that is what we mostly interested in. So, what they do the production rules will try to express what are the way in which various symbols of the language of group together and that is our main interest in using context free grammar which symbols are being group together that I can use that I can find out using or I can express using the production rules in context free grammars.

Let us see one simple example. So, I want to model this (Refer Time: 13:27) language that is noun phrase can be composed to either a proper noun or a determiner followed by a nominal where a nominal can be more than one nouns that is something I want to express about language. So, how do I use the context free grammar? So, I will say noun phrases proper noun or determiner nominal the nominal is noun or many nouns.



(Refer Slide Time: 14:03)



How I write in context free grammar, I will say noun phrase is proper noun or determiner nominal and what is a nominal? It is one noun or more than one noun. So, noun or noun by a nominal this is the regression here. So, I can allow as many numbers of nouns as I want by using this and that is my context free grammar for denoting the symbol that is the idea I can express all these facts about language how words are groups together which groups come together by using this production rules.

Once we know that what is the formulation of context free grammar very briefly said so, in context free grammar when we study when we talk about a quadruple so, therefore, 4 important variable set of variables.

(Refer Slide Time: 15:09)

*CFG for Languages*

*CFG:  $G = (T, N, S, R)$*

- $T$ : set of terminals
- $N$ : set of non-terminals
  - For NLP, we distinguish out a set  $P \subset N$  of pre-terminals, which always rewrite as terminals
- $S$ : start symbol
- $R$ : Rules/productions of the form  $X \rightarrow \gamma$ ,  $X \in N$  and  $\gamma \in (T \cup N)^*$

*Terminals and pre-terminals*

Terminals mainly correspond to words in the language while pre-terminals mainly correspond to POS categories

Pawan Goyal (IIT Kharagpur)      Syntax      Week 5: Lecture 1      9 / 14

Firstly, I have set of terminals we have the leaf nodes in my tree whenever we see. So, they will always come at the end and whenever I get a terminal I cannot derive anything from there. So, I have set of terminals. So, we will see what in the case of language what do they mean then we have set of non terminals that help me to do the derivations, these are the variable from which you can further derive in more (Refer Time: 15:36)

Now, so what is different in the case of NLP, what is some distinguish, we will make in the set of non terminals, we will also distinguish the set  $P$  that are pre terminals. So, pre terminals are those non terminals that will always derive terminals. So, they will always give me leaf nodes and with the example that will be clear what are what are they in the case of language.

Then I have start symbol from which I am starting my derivation. So, if I have to model the sentence I should be starting with  $S$  that is the sentence and then I have the rules and they always of form  $X$  going to  $\gamma$  and  $X$  has to be a non terminal, a single non terminal and  $\gamma$  can be any sequence of terminals and non terminals and that is the constraint that we see in case of context free grammar. So, this is the quadruple and we also seeing a pre terminal that is the set of non terminals in the language, what are terminals and pre terminals? Terminals in the language will mainly with the final words that I will see in the lexicon and pre terminals will be part of speech categories from the pre because from the pre terminals you can derive only terminals.

(Refer Slide Time: 17:02)

The slide is titled "CFG for Languages" in a blue header. Below the header, there is a green box containing the following text:

*Example*  
 $NP \rightarrow \text{Det Nominal}$   
 $NP \rightarrow \text{ProperNoun}$   
 $\text{Nominal} \rightarrow \text{Noun} \mid \text{Noun Nominal}$   
Now, these can be combined with other rules, that express facts about a  
lexicon.  
 $\text{Det} \rightarrow \text{a}$   
 $\text{Det} \rightarrow \text{the}$   
 $\text{Noun} \rightarrow \text{flight}$

Below the green box, there is a pink box containing the question:

Can you identify the terminal, non-terminals and preterminals?

At the bottom of the slide, there is a blue footer with the text "Pawan Goyal (IIT Kharagpur)" on the left, "Syntax" in the center, and "Week 5: Lec 1" on the right. A small circular inset image of a man is located in the bottom right corner of the slide.

Let us see one example and then we can point out what are terminals pre terminals and non terminals. So, this is what we are modeling earlier what is the noun phrase determiner for the nominal or a proper noun and where a nominal is a noun or a set of nouns which I model using a noun followed by a nominal now are you see some terminals here. So, there are no words. So, there is no terminal. So, I cannot use that to derive a phrase or a sentence it can only give me a set of grammatical categories. So, I have to include some facts on the lexicon to make it a complete grammar.

For example, I can include some determiners some nouns and some proper nouns. So, here I am including a and the has to determiners and flight as a noun now here can you identify what are the terminal non terminals and pre terminals. So, terminals are the words in my lexicon. So, a the and flight are my terminals pre terminals are the grammatical categories or parse category that will always give me terminals can you see that determiner noun are only giving me terminals. So, these are my pre terminals and apart from that all the variables like N P nominal they are my non terminals. So, proper noun no example is given, but proper noun is also a pre terminal it is part of speech category it will give you some words to the lexicon. So, these are my terminals non terminals and pre terminals.

(Refer Slide Time: 18:40)

*CFG as a generator*

$NP \rightarrow \text{Det Nominal}$   
 $NP \rightarrow \text{ProperNoun}$   
 $\text{Nominal} \rightarrow \text{Noun} \mid \text{Noun Nominal}$   
 $\text{Det} \rightarrow \text{a}$   
 $\text{Det} \rightarrow \text{the}$   
 $\text{Noun} \rightarrow \text{flight}$   
Generating 'a flight':  
 $NP \rightarrow \text{Det Nominal}^{\circ}$   
 $\rightarrow \text{Det Noun} \rightarrow \text{a Noun} \rightarrow \text{a flight}$

- Thus a CFG can be used to randomly generate a series of strings
- This sequence of rule expansions is called a derivation of the string of words, usually represented as a tree

Pawan Goyal (IIT Kharagpur) Syntax Week 5: Lecture 5

Now once you have the CFG, you can use that to generate various phrases or sentences in language. So, for example, I here this is the CFG for non phrases. So, can I can I generate a flight the phrase a flight is in this context free grammar. So, I will have to start with n p and I have to generate a flight. So, what is the first derivation I will to do from NP I will take this rule determiner followed by a nominal yes now determiner will give me a now from nominal I cannot go to flight directly. So, for nominal I will have to first get a noun and from noun I get the word flight. So, NP gives me determine nominal gives me noun determiner gives me a noun and a flight that is how it generate a sequence of words using this grammar and now you can do it for any sentence you can define a grammar for a sentence and generate a sentences from that.

What we are seeing here? So, there is a context free grammar is generating a series of strings and this sequence of rule expansions. So, the sequence that that you are using starting from NP going to determiner nominal then determiner noun then a noun then a flight this is called the derivation of the string using this grammar and we use it is tree structure to represent this derivation remember one of the very first tree that we had shown as a motivation. So, we will try to come up with such trees using these derivations and when we will be call the past tense.

(Refer Slide Time: 21:20)

The slide is titled "CFGs and Grammaticality" in a blue header. The main content area has a light green background. A green box contains the text: "A CFG defines a formal language = set of all sentences (string of words) that can be derived by the grammar". Below this, a pink box contains two bullet points: "• Sentences in this set are said to be **grammatical**" and "• Sentences outside this set are said to be **ungrammatical**". At the bottom, there is a blue footer with the text "Pawan Goyal (IIT Kharagpur)", "Syntax", and "Week 5: Lect". A small circular video inset of a man is visible in the bottom right corner.

*CFGs and Grammaticality*

A CFG defines a formal language = set of all sentences (string of words) that can be derived by the grammar

- Sentences in this set are said to be **grammatical**
- Sentences outside this set are said to be **ungrammatical**

Pawan Goyal (IIT Kharagpur) Syntax Week 5: Lect

Now, what is the notion of grammaticality using context free grammars the idea is that you defined one grammar for your language you have to assume that this is the only grammar. So, any tool that is not expressed in the in the grammar is not allowed in the language. So, now, when you are given a new sentence if you can see that there is a way to generate this sentence using my grammar the sentence is grammatical as per my grammar if the sentence cannot be generated in the grammar this is not grammatical this is a simple notion or using this grammar whatever sentence I can generate is grammatical and whatever I cannot is not grammatical. So, it depends on the grammar that you have designed the context free grammar that you have designed so yes. So, whatever can be derived is grammatical and others are ungrammatical.

(Refer Slide Time: 21:27)

The slide is titled "CFGs and Recursion" in a blue header. Below the header, there is a light purple box labeled "Recursive Definition" containing two bullet points: "• PP → Prep NP" and "• NP → Noun PP". Below this is a light green box labeled "Example Sentence" containing the sentence: "[<sub>S</sub> The mailman ate his [<sub>NP</sub> lunch [<sub>PP</sub> with his friend [<sub>PP</sub> from the cleaning staff [<sub>PP</sub> of the building [<sub>PP</sub> at the intersection [<sub>PP</sub> on the north end [<sub>PP</sub> of town]]]]]]]]].". In the bottom right corner, there is a small circular portrait of a man. The bottom of the slide features a blue footer with the text "Pawan Goyal (IIT Kharagpur)", "Syntax", and "Week 5: Lec 1".

Now, CFGs are interesting because they can also model some very interesting phenomena in language syntax like recursion. So, in language you make lot of big sentences by doing recursion. So, for example, a preposition phrase can be written as a preposition followed by a noun phrase and noun phrase can be written as a noun phrase followed by a preposition phrase. So, you see there is a recursion here yes. So, I can encode the noun phrase noun and prepositional phrase is here preposition phrase can again encode a noun phrase which can encode a proportion phrase. So, this is the recursion is very very common in language.

Let us see 1 example. So, the set example is the mailman ate his and this is the complete noun phrase is starting from lunch till the end of the sentence, lunch and noun phrase is a noun followed by a preposition phrase with and this is not shown here, but this is again starts a noun phrase preposition phrase is a preposition followed by a noun phrase with his friend from the cleaning staff and all this is a noun phrase and what is this noun phrase noun his friend followed by a preposition phrase and so on the recursion is very nicely captured by using context free grammar.

Now, shall end this lecture by just saying, what is the context denotes in context free grammar? What is the meaning of context? So, language we talk about context as such. So, we say context is given a word find out what is the context what are the previous words and what is the topic and all that. So, these are all define the context, but this

context is nothing to do with what is context in the case of context free grammar this is very very formal notion.

(Refer Slide Time: 23:35)

*What does Context stand for in CFG?*

- The notion of *context* has nothing to do with the ordinary meaning of word context in language
- All it really means is that the non-terminal on the left-hand side of a rule is out there all by itself (free of context)

$A \rightarrow BC$

- I can rewrite  $A$  as  $B$  followed by  $C$  regardless of the context in which  $A$  is found
- Or when I see a  $B$  followed by  $a:C$ , I can infer an  $A$  regardless of the surrounding context

Pawan Goyal (IIT Kharagpur) Syntax Week 5: Lecture 5

This is nothing to do with the ordinary meaning of word context in language. So, all this means is that in context free grammar whenever I am doing a derivation. So, whenever I am writing  $A$  gives the  $BC$  or say noun phrase, give me determiner nominal whenever I am writing a rule like that it means is that the non-terminal left it all in its own by itself it does not need any context around it. So, I can always write  $a$  goes to  $bc$  irrespective of whatever is around my word  $a$ . So, even if I have  $x$   $a$  earlier I can use  $a$  to derive  $xbc$  and if I have  $y$   $a$  I can always write like this and this  $x$  and  $y$  are immaterial they do not matter it can be null it can be whatever.

Similarly, if I inferring whenever I see  $ABC$ , I can always infer  $A$  so, this; it might have come from  $A$  independent of the context of  $BC$  and this is what context free grammar. So, if you go would go the next label of context sensitive grammars there are you need the context this word  $a$  can derive  $bc$  in this particular left context in a particular right context we do not need this left and right context in the case of context free grammars this what the context the word context means in this case.

Whenever I have a rule,  $A$  goes  $BC$ , it means that I can write  $A$ , I can always write from  $A$   $B$  followed by  $C$ , regardless of the context in which  $A$  is found or whenever I find and  $B$  followed by  $AC$ , I can infer  $a$  regardless of the context in which  $B$  and  $C$  is found. So,

this is CFG for us, I am not going to lot of basics of context free grammars and I suggest that you can quickly look at any of the chapters in the basic books of formal languages and automated theory so, but whatever is required for our tasks of doing parsing, I have covered in this lecture and I will cover the necessary things in the next lecture.

In the next lecture, what we will do? We will start from CFGs and we will see how we can use that for actually doing the parsing for a given sentence in the in the language. So, we will take a various approaches for parsing so that will be the next topic for the next lecture.

Thank you.