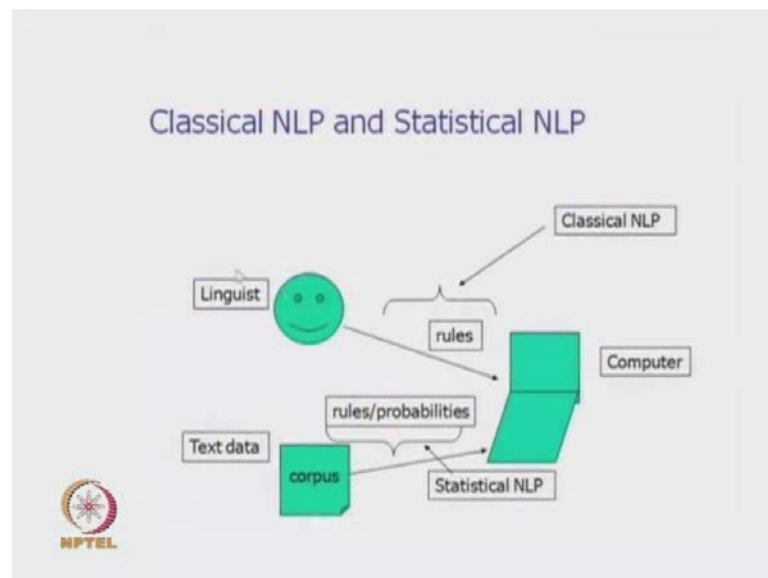


Natural Language Processing
Prof. Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay

Lecture - 5
Sequence Labeling and Noisy Channel

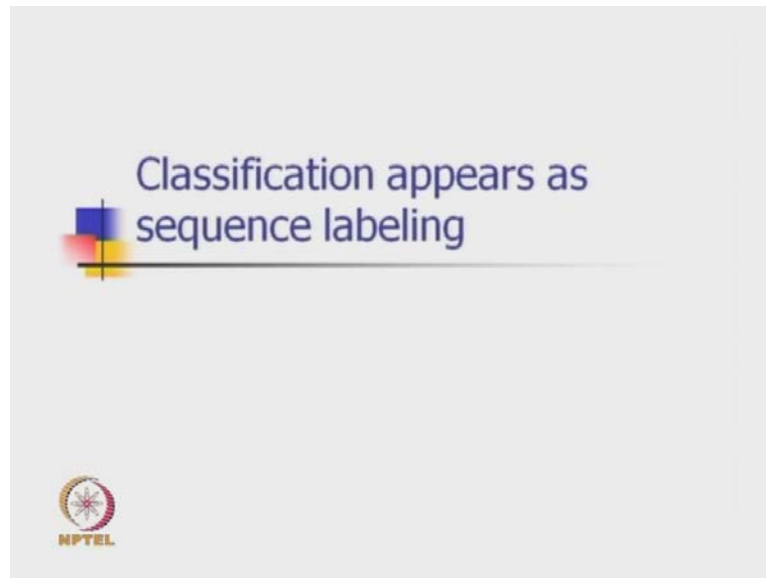
In the last lecture, we introduced statistical natural language processing, we will move ahead, with the discussion. And we will talk about a very, very fundamental point of statistical natural language processing. This lecture 5 is on noisy channel, and sequence labeling a very key concept, very key concepts of statistical natural language processing.

(Refer Slide Time: 00:45)



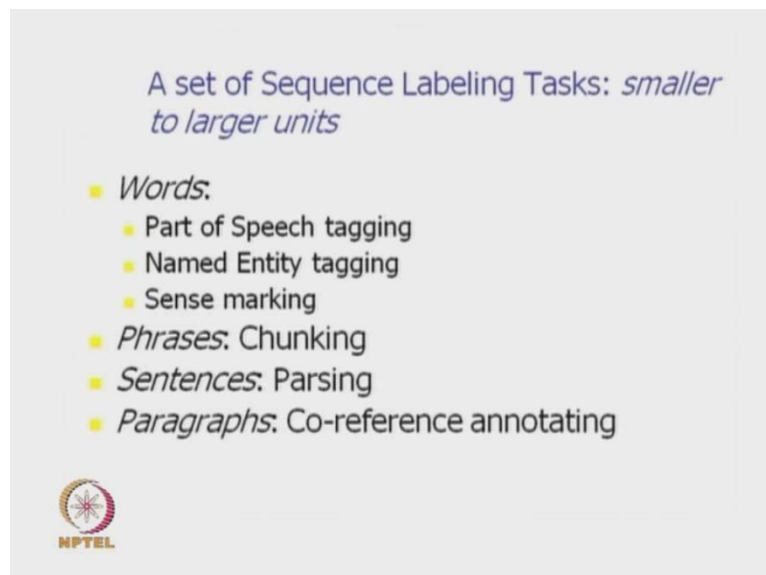
So, we have already remarked that, classical natural language processing, and statistical natural language processing differ in the way, the computer gets its knowledge of natural language processing, so to say the knowledge with which it begins to process, the natural language data. In classical natural language processing, the rules or the knowledge come from the linguistic, who is the human being, a language expert, whereas in statistical natural language processing the rules and probabilities are learned from the data or the corpus. So, the textual data provides the machine with rules and probability values, and this happens by the application of some machine learning technique.

(Refer Slide Time: 01:36)



So, we now look up on this whole business of classification, in natural language processing as a sequence labeling task. We would like to say that these whole of NLP tasks, various tasks at different levels of stages that we talked about, they are actually sequence labeling task.

(Refer Slide Time: 01:54)



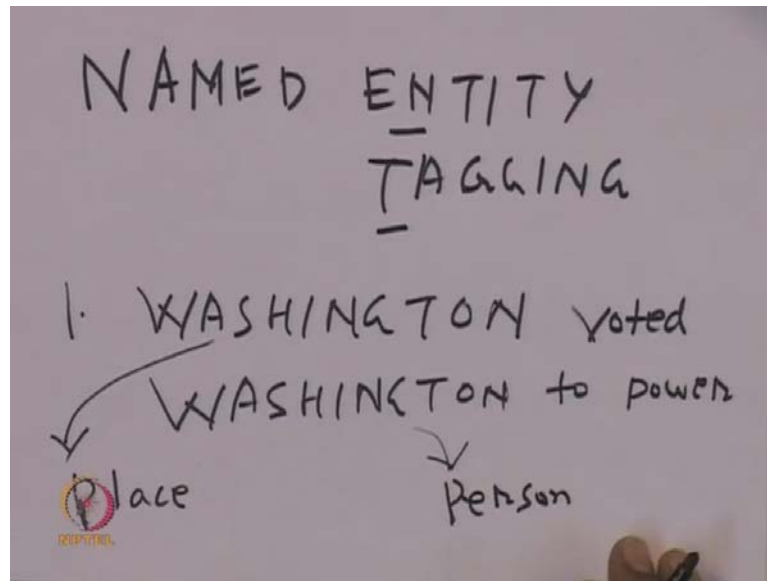
There are actually of sequence labeling task, which go from smaller to larger units, the smallest unit on which label has to be placed in natural language text is the set of words, and the task is called part of speech tagging. The first task on words is part of speech

tagging, the second task is named entity tagging, which is detecting the proper nouns, and understanding their category. So, a proper noun can be the name of a person or it could be the name of a organization, and finally, and probably the most difficult labeling task is, that of sense marking, the words are given sense labels. So, 3 kinds of labels on the words, part of speech named entity, what kind of names these are and senses. We move from words to phrases, which are bigger units of text, and these, kind of labeling task is called chunking.

We will explain what a chunk is, proceeding from phrases, we graduate to the level of sentences. And there we produce parsing, parsing actually produces a tree for the sentence, how is it a labeling task, we will see very soon. And when we deal with paragraphs connected sequence of sentences, we do co-reference resolution. That means we find out entities, which refer to the same external entity, whose reference are same, and that is again indicated by labels, this is called co-reference annotation.

So, you can see, how you go from small units of words to phrases to sentences. And finally, paragraphs we increase the size of the textual units, let us spend a bit of time on some of these entities, let me draw, your attention to named entity tagging. Let me write few things for named entity tagging. Named entity tagging is an important task, on which we will have some discussion.

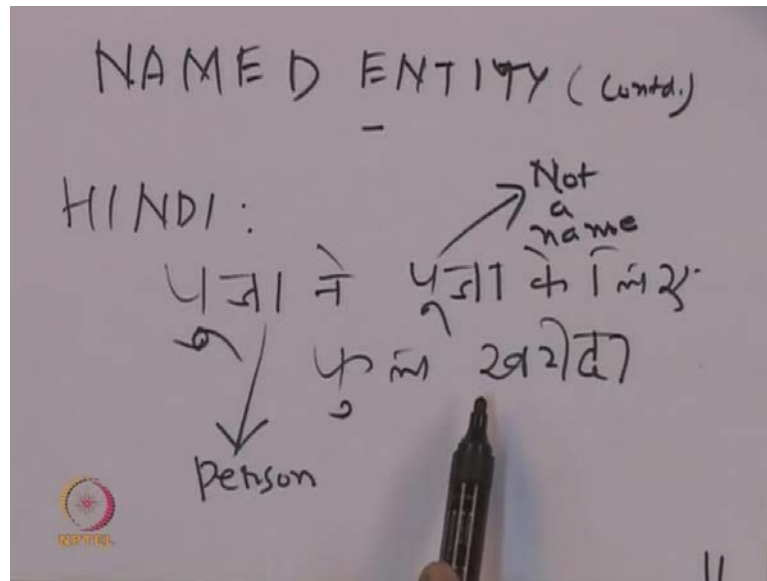
(Refer Slide Time: 04:16)



This is named entity tagging, let me give you 2 examples, first example is this Washington voted Washington to power; Washington voted Washington to power, what is the meaning of this sentence? The first Washington is the city of Washington, the capital of United States of America, Washington voted Washington to power. The second Washington is George Washington, who became the president of America long time back. So, this is saying that the people of the city of Washington, they cast votes in favor of George Washington the president of America, and brought him to power.

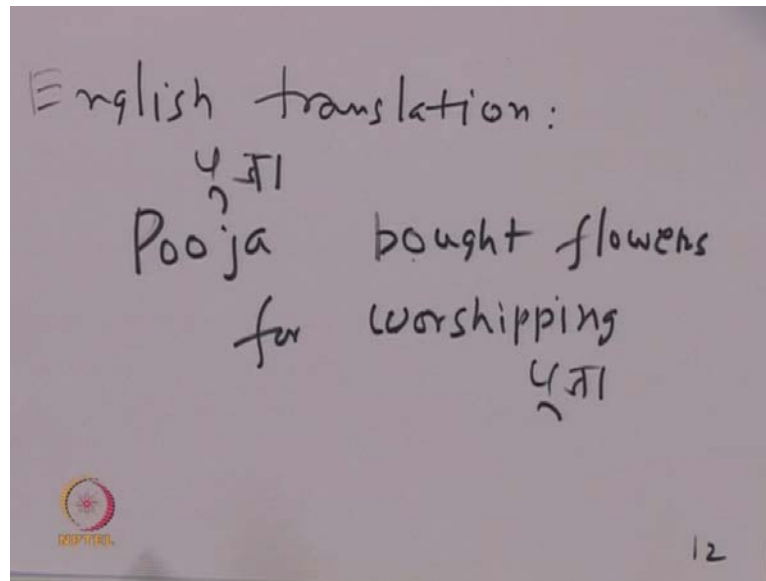
So, it is true that both these Washington's are proper nouns, in English we have the advantage, that proper nouns start with capital letter. This Washington will start with a capital letter, this Washington also will however, there are 2 different kinds of proper nouns, 1 the first Washington is a Place. So, the Named Entity Tagging will say that, it is a place Named Entity, and the Named Entity recognition system, for the second Washington will say it is a Person. So, there are 2 instances of Washington, actually are 2 different entities as far as, name of something is concerned. The first Washington is the name of a place, second Washington is the name of a person, let me give you a more dramatic example, this time from Indian language.

(Refer Slide Time: 06:41)



From Hindi, we take this example [f1], I will not translate this, because in the discussion, the transition will become apparent, [f1], the first Pooja is person. Now in Indian languages, we do not have the advantage of starting a proper noun, with a capital letter. We do not have capital small distinction, all the alphabets are of same kind, and the second Pooja is not a proper noun, it is not a name, it is not a proper noun, so[f1] the Pooja is the name of the person, second Pooja is worship.

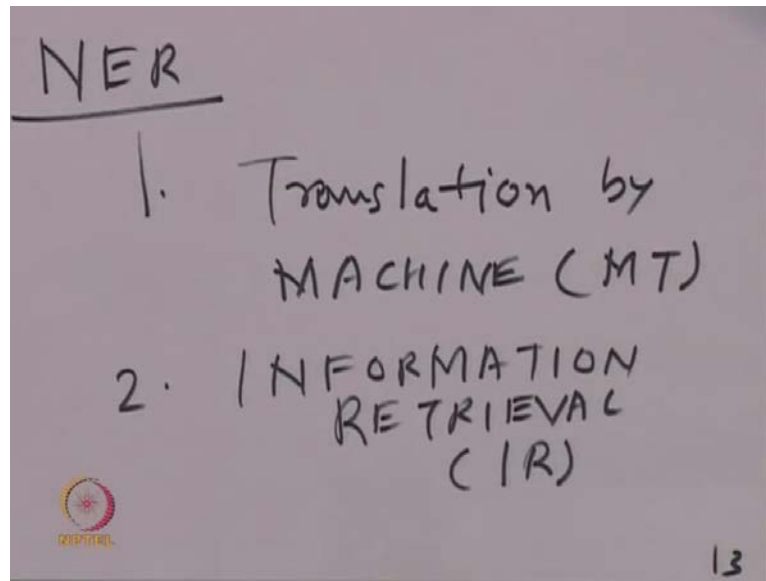
(Refer Slide Time: 08:08)



So, if we want to translate this sentence, we have to write the English translation as Pooja bought [fl] flowers for worshipping, so this was the name Pooja, and this is the Pooja in the sense of worshipping. So, you see, now if you had the task of translating from Hindi to English. And this sentence was giving to you [fl] and if you had to translate this sentence, if you do not detect that first Pooja is a proper noun, it is a named entity. Then the translation will be improper, you will say worship bought flowers for worshipping, which is absolutely strange or you could say Pooja bought flowers for Pooja, which is not so bad. You have mixed 2 languages here, Hindi and English, this is known as code mixing. When the words of 2 different languages are used together, to form a sentence, this is known as phenomenon of code mixing, code mixing, c o d e m i x i n g, code mixing.

So, Pooja being kept as Hindi string makes an acceptable translation, Pooja bought flowers for Pooja but Pooja being translated in both places, makes it completely strange, worship bought flowers for worshipping or worshipping bought flowers for worshipping is strange. So, this shows the importance of detecting proper nouns or named entities as they are called in the NLP for the purpose of translation.

(Refer Slide Time: 10:50)



So, we understand now, that Named Entities recognition as a task, NER as a say, NER is a famous task, it is important for translation, translation by machine, machine translation MT it is important for information retrieval which is known as IR a very, very important field of research and development these days, all of us use search engine.

We use information retrieval and name entity recognition is important for, both these tasks apart from many other tasks, like question answering, summarization, information extraction everywhere, Named entity recognition is very important, and a machine translation information retrieval in particular are important consumers of named entity recognition. Let us proceed with the slides, we have seen, that words have to be tagged with part of speech, named entity which we understood just now, Sense marking also called word disambiguation. We will have lot of things to say about sense marking, phrases, sentences, paragraphs.


(Refer Slide Time: 11:59)

Example of word labeling: POS Tagging

```
<s>
Come September, and the UJF campus is abuzz
with new and returning students.
</s>
```

↓

```
<s>
Come_VB September_NNP , , and_CC the_DT
UJF_NNP campus_NN is_VBZ abuzz_JJ with_IN
new_JJ and_CC returning_VBG students_NNS ._.
</s>
```

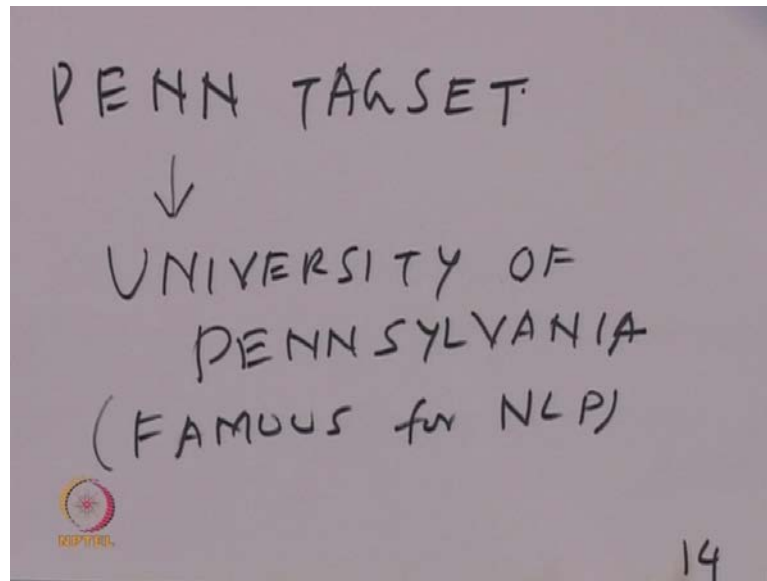


Here is an example of different stages of sequence labeling, first example of word labeling is part of speech tagging POS tagging. Here is this sentence come September, and the UJF campus is abuzz with new and returning students. That means when the month of September comes this university, University Josef Fourier campus becomes lively, becomes vibrant with new and returning students. Now, this piece of text as it is given is a piece of raw text, it is a raw piece of text, when we do the first level of processing on this text, we produce word labels. So, Come is a verb it is indicated by VB, V for verb, B for base, the word come is in its base form, that is why its VB.

It is not in the form coming or came, which are the present participle or the past tense from Come, Come is in its base root form, that is why its VB September is NNP. It is a proper noun, this is the symbol for proper noun, will come to where these symbols are coming from, comma is given the level comma itself, punctuation marks are given the same level. And is CC, CC means a conjunction, it is a conjunction, The is a determinant DT, UJF is again NNP that is a proper noun, campus is a noun the label is noun NN means a common noun, is VBZ which is an indication for auxiliary verb. Abuzz is auxiliary verb, abuzz is adjective, adjective is indicated by the symbol JJ, with is a preposition is indicated by IN, new is indicated by the adjective symbol JJ, and is again a conjunction CC, returning is verb in gerund form, so VB verb. And it is a gerund form G students is noun, and S for plural, so students is a plural noun, full stop is full stop again.

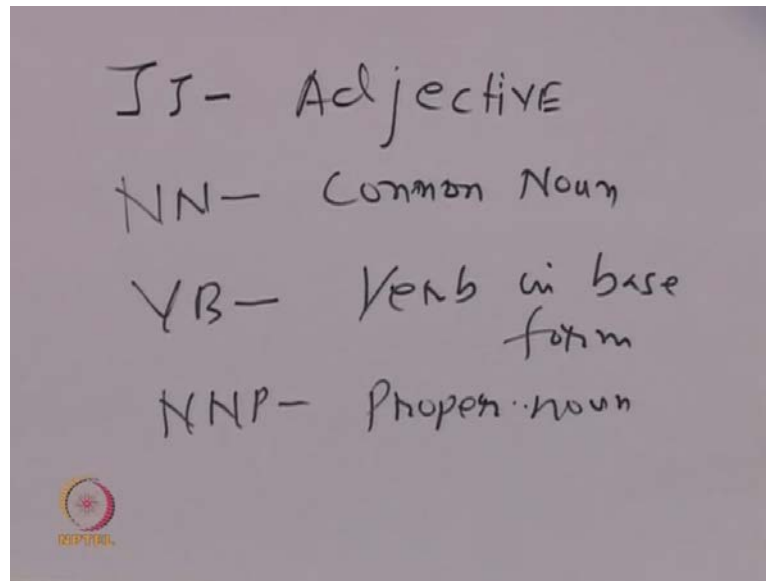
So, these levels which are indicated by an underscore, between the verb and the label are produced by what is called POS Tagger very important component of natural language processing system, POS TAGGING POS. So, you can see how these words are given the labels, and we have detected here, the proper nouns, the conjuncts and adjectives and so on. Just a point about, where these labels are coming from.

(Refer Slide Time: 15:52)



Let me write down, these labels, the name of these labels and the source their source, so these levels come from what is called the Penn Tagset, Penn is comes from University of pennsylvania, which is a famous place for natural language processing work, University of pennsylvania has a very strong natural language processing group. And these particular symbols, these set of symbols VB JJ and NN proper noun, which you have shown, they come from Penn tag set, so we have seen that JJ is...

(Refer Slide Time: 16:50)



Actually Adjective NN is a Common noun, VB is a verb in Base form, NNP a Proper noun. So, you can do some Google search on Penn Tagset, go to Google and type these query pen tag set, it should take you directly to the repository of Penn Tagset, and these symbols are explained there. So, this kind of tags are crucial for natural language processing, one of the important concerns here is, how do we come up with these tag set.

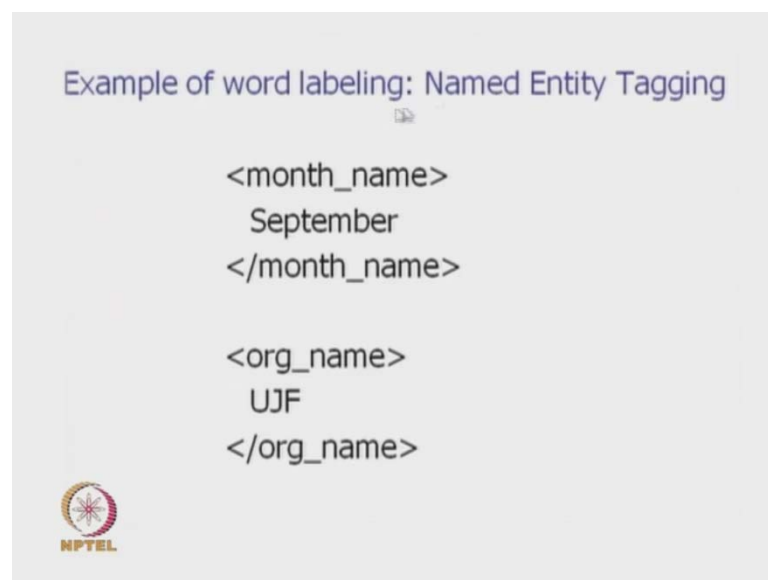
These tags which are used for marking the raw text to produce the levels on words, how do we come up with this tag sets now. We should spend some time on understanding, how a tag set is designed, how does one form this whole repository of tags. It is a very complicated and intricate question. Long time is required to arrive at a tag set, which is feasible annotation wise, which is easy for the annotators to handle annotators, human annotators use those levels to produce tags on the words.

So, annotators should find it easy to use this tag set, at the same time, the tag set should be useful for natural language processing. If the tag set is very intricate, it has labels, which are very fine grained, that means with the noun, you have various sub classes of very fine categories of nouns; within adjective you have very fine adjectives, very fine categories of adjectives. Then we will see later, that it becomes very difficult for a machine to uniquely identify, those tags from the words from the context, it becomes difficult to disambiguate, the tag label. This broad level classification of noun, verb

adjective etcetera are not so difficult, but within each category, producing very refine labels, becomes a challenging problem.

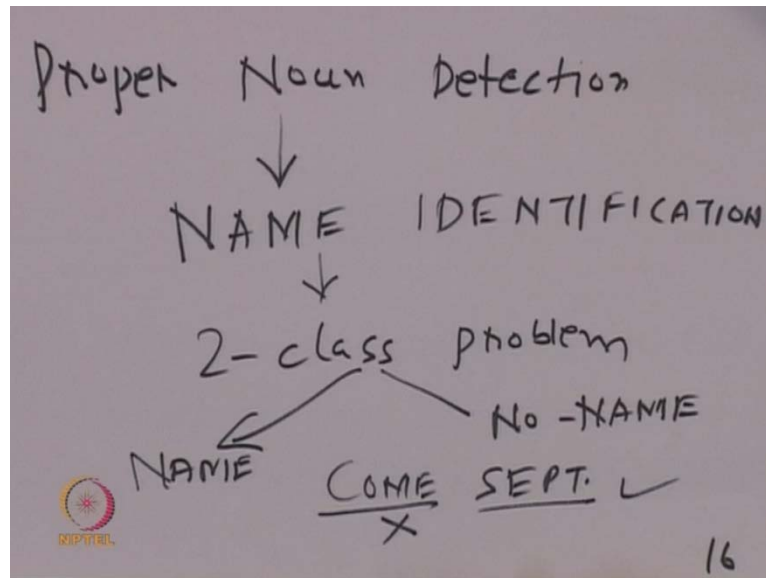
So, when we do part of speech tagging, we would like to make some remarks on them. So, we have to produce tags, which human beings find easy to use, we have produce tags, which a machine find easy to label with. Now, these are difficult and intricate question, we will spend some more time, understanding their new answers, understanding their complexities but we will proceed with the main theme of the lecture. Now, so looking at the slide, we see that, we have these raw sentence, and this part of speech levels have been produced.

(Refer Slide Time: 20:19)



Now, comes the next stage if marking levels on the words, so September and UJF were detected as proper nouns. So, when they detected as proper nouns, we have taken the first step of named entity recognition, we have identified these as proper nouns. We have identified those words has proper nouns, when we have identified the proper nouns, we have done, what is called the name identification.

(Refer Slide Time: 20:57)




So, proper noun detection is also called name identification, name identification is actually a 2-class problem, namely name or no name, so come September, come September here, this is a name; this is not a name. So, just detecting a word is Name or not, this is a 2 class problem its binary problem; this is called name identification, but from that we graduate to the actual name recognition problem. And this is what is happening in the example as we see September and UJF have been detected as proper noun. But September is a month name UJF is a is an organization name, it is important to detect the names to that level of categorization, September is a month name, it as a time property associated with it.

Whereas, UJF is an organization name, and it has a place property, organization property associated with it, so why is this important the machine has to be told, these a points. So, that it behaves intelligently, so for example come September and the campus is UJF campus, is abuzz with students when this sentence, if the question that one asks is what is abuzz with students. September is not the answer, September does not have the place property students cannot go to September, students can go during September or students can go in September, but students cannot go to September as a place. So, it is important for the machine, to have this property.

(Refer Slide Time: 23:28)

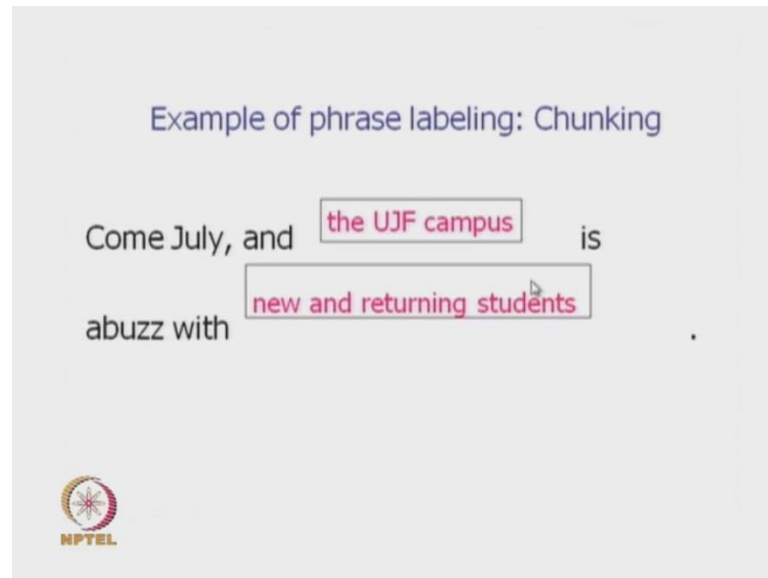
Example of word labeling: Sense Marking

<u>Word</u>	<u>Synset</u>	<u>WN-synset-no</u>
<i>come</i>	<i>{arrive, get, come}</i>	<i>01947900</i>
.	.	.
<i>abuzz</i>	<i>{abuzz, buzzing, droning}</i>	<i>01859419</i>



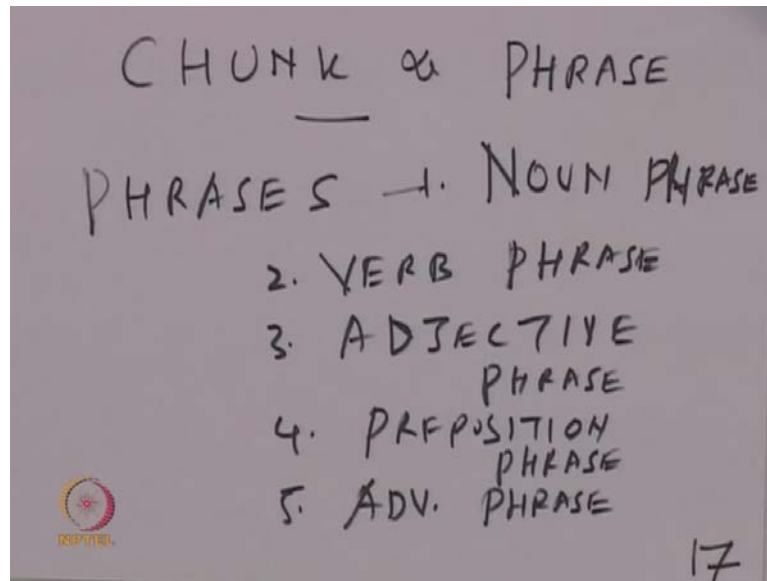
Next tagging, which is done on the word is called Sense Marking and extremely challenging problem natural language processing. We discussed this in our first and second lecture, when we dealt on sense ambiguity, the word sense ambiguity disambiguation. So, we see here, this word come and come has the sense of arriving. So, this is shown here, there is a very very important repository lexical knowledge called word net. And come in the sense of arriving, getting somewhere is shown here in the form of what is called a synset, we will discuss synset, when we cover lexical semantics of word net. And this is the identification of the sense number, word net is the sense repository abuzz similarly, has the sense of being alive, being vibrant and the synonyms words are a buzzing, and droning the word net synset number is the 01859419. So, these are the identification of the senses, now award have to be given, this level after the word has been disambiguated properly.

(Refer Slide Time: 24:56)



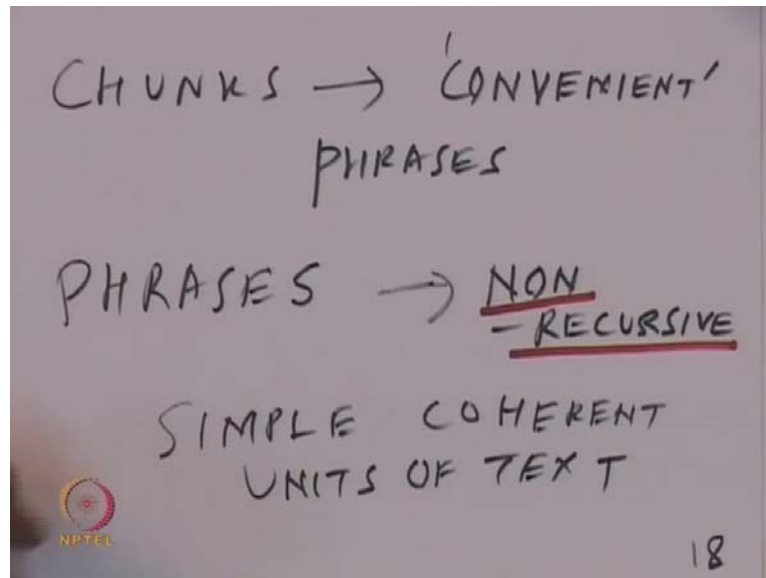
Next, we Come to bigger levels, bigger units of text, and here, we are looking at phrase labeling or chunking, example Come July and the UJF campus is abuzz with new and returning students. So, things in boxes and in pink alphabet, are in some sense, some kind of closely held together pieces of text. The UJF campus in some sense is take one coherent entity, even though there are 3 things. There are words expressing the concepts, it is a 1 unit concept, new and returning students is also in some sense, a an entity, where students are being qualified with adjectives, new and returning students. So, Come July, and some things is abuzz with something else. So, these are the noun phrases, and these nouns phrases are called chunks for the purpose of chunking. Let us now, spend a few minutes on understanding chunking, what is a difference between chunk and phrase, I will write it down.

(Refer Slide Time: 26:24)



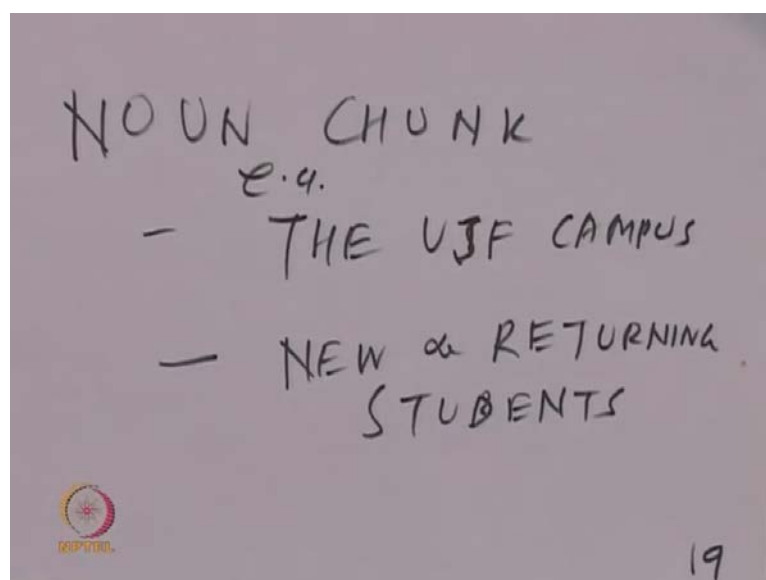
So, chunk and phrase, phrases are very famous in language. So, you have noun phrase, we have verb phrase that the most famous phrases; we have adjective phrase; we have preposition phrase; we have adverb phrase and so on. So, phrases are very important in language, combination of phrase produces a sentence. It is important to detect phrase boundaries, where does the phrase begins, and where does it ends. It is very important, what are chunks then chunks are somewhat convenient phrases.

(Refer Slide Time: 27:33)



So, chunks are convenient for whom convenient, for the machine, so convenient phrases. These are phrases which are non recursive, these phrases are non recursive, let me highlight that, non recursive phrases. That means, these are simple coherent units of text. So, what would be the most important chunk?

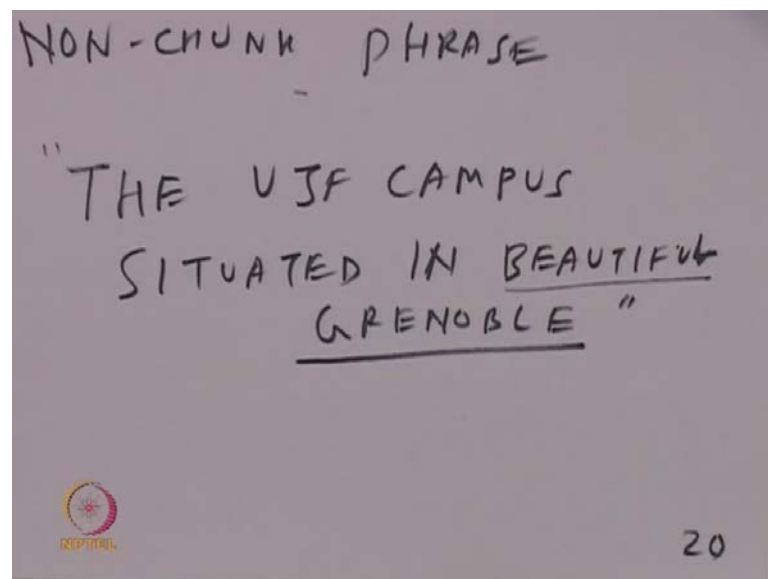
(Refer Slide Time: 28:23)



In noun chunk, so the example would be the UJF campus, new and returning, students, these are noun chunks. So, they are non recursive in the sense that in noun chunk, we

will not contain another noun chunk in it, in noun chunk, we will not contain, another noun chunk in it. I am repeating this statement these are non recursive. So, they would be very simply units of text, with a small window over them, for the purpose of boundary detection. And these are very convenient units, which a machine can pick up, for different purposes. So, the question that will natural arise is, what is an example of a phrase, noun phrase which is not a chunk. Because it has got recursion in it, the chunk will contain other chunk, so let me an give an example, of non- chunk phrase.

(Refer Slide Time: 29:49)



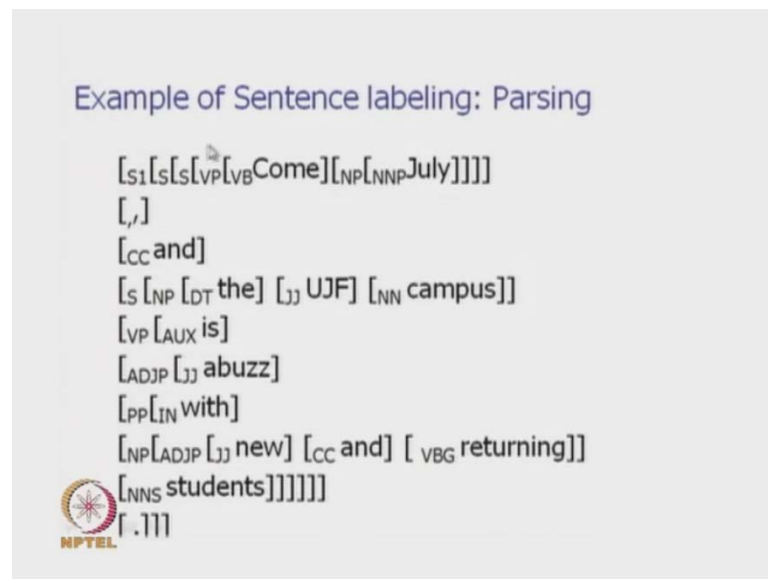
Non-chunk phrase of course, there are lots of different definitions of chunk, we are sticking with one definition, which is most commonly acceptable, and that actually insists on a simple description of a chunk. So, let me give you an non-chunk phrase, The UJF Campus is a chunk we have seen, situated in beautiful Grenoble. So, this is a non-chunk it is not a noun chunk it is a phrase, it is a noun phrase. The UJF campus situated in beautiful Grenoble is a noun, because you can make these a subject of a sentence.

For an example, UJF campus situated in beautiful Grenoble was visited by the Prime minister of India for example, so the UJF campus situated in beautiful. This whole thing has become the subject of a sentence, and it is actually noun phrase, why is it not a chunk? It is not a chunk, because it contains another non-chunk the UJF campus situated beautiful Grenoble is a non-chunk. This whole thing; this beautiful GRENOBLE is a

non-chunk, and we are not allowed to have a non-chunk that contains another non-chunk that is why, this is an non-chunk phrase, I hope this concept is clear.

So, let me emphasize this point once again, noun chunks will form a very important entity, in our discussion especially, when we come to information extraction, chunks are important for, information extraction. So, chunks are those very simple noun phrases, with a few words in them, and that is a noun chunk. And it is it is essentially a noun, where there is a noun being modified by other entities, in the whole window, so I hope it is clear to you. Similarly, there are other kinds of chunk, for other parts of speech and chunks can be detected, with lot of accuracy. We proceed further, looking at the slide this entity, the UJF campus is abuzz with new and returning students, illustrated the concept of chunking.

(Refer Slide Time: 32:46)

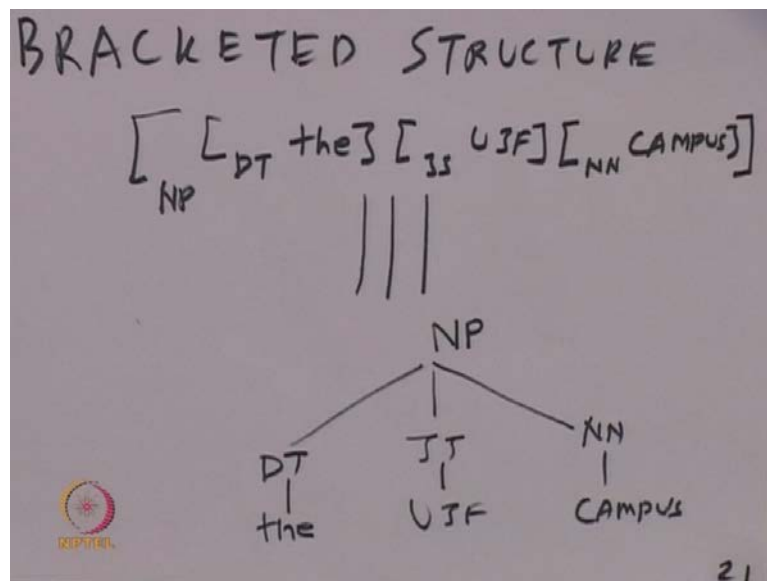


We come to some more complicated sequence level entities, this is the example of sentence labeling, after word labeling, and phase labeling, this is sentence labeling. And the NLP name or the linguistic name, for this kind of labeling task is called Parsing. Look at this complicated structure here, the sentence is the same, come July, and the UJF campus is abuzz with new and returning students. In this case, the sentence has been marked with a large number of brackets you can see a large number of brackets here. So, these are brackets, and how are these brackets placed, these brackets actually define trees and sub trees, and sub trees, within them how is this done. Let us look at the structure,

and let us take a small part of these sentence. Suppose we take the UJF campus is abuzz, with new and returning students. Suppose when we had considering this particular clause come July, and the UJF campus is abuzz, with new and returning students.

Now here, let us look at the structure minutely, the, is enclosed in 2 brackets, so this is a structure in itself, this is a determiner DT. Similarly, UJF is given the label JJ, so JJ is UJF NN is a noun, and campus is a noun. So, for this the UJF campus, the 3 entities have these level DT JJ and n f what we do is that, we draw arrows, from the labels to the words. So, DT to the there is an arrow JJ to UJF another arrow, NN to campus another arrow. Now, these have the brackets own brackets, enclosing brackets, look at the outer bracket here, this is the other bracket, the right bracket and corresponding to that, there is a bracket here this is given the label NP. So, this whole thing, starting with a left with NP, ending with a right bracket, means that there is an noun phrase here, and the noun phrase is composed of a determininal, and the adjective and a noun.

(Refer Slide Time: 36:11)



So, this can be expressed by means of a tree which I draw here, the bracketed structure this is known as the bracketed structure, which is in front of us is DT, the JJ UJF and NN campus, we have already seen it they have their own brackets. Then there is this outer bracket with label n p n and on the right enclosed by another right bracket. So, these bracketed structure is actually equivalent to, I draw this symbol here, equivalent to, this

entity which is a tree. So, the tree is being drawn, in a bottom-up manner, we have DT going to the, we have JJ going to UJF, we have NN going to campus.

So, we have the, these small sub trees DT the JJ UJF and NN CAMPUS, and the whole thing gathered together, into the noun phrase NP. So, do you see this correspondence, between the tree, and the bracketed structure. So, whole thing becomes a very nice equivalence, we produce a bracketed structure, as a linear sequence on the paper, and this linear sequence is actually equivalent to a 2-dimensional structure in the form of a tree. The UJF campus, the UJF campus had the leaf of the tree, and the leafs being marked with NP as a structure, so the NP tree has sub trees DT JJ NN, this is expressed by the bracketed structure here, I hope it is clear. So, the bracketed structure actually defines a tree, and the bracketed structure is a very famous kind of construct in natural language processing, for the purpose of parsing.

We will see its immense utility, when we do a rule based parsing, and also probabilistic parsing, now I hope the correspondence bracketed structure and tree is clear. We will continue with the slide here, and see how many more trees, how many sub trees and trees, we can see in this structure. So, we have already seen the UJF campus is a noun phrase tree, now look at is an auxiliary verb as indicated here aux is, abuzz is an adjective. So, JJ with is a preposition so in, so from the previous discussion, you would understand that, these are actually sub trees, aux with child as is, JJ with child as abuzz in with child as with.

Now, let us see what kind of higher level structures are coming in, because we see here a VP adjective, ADJP which is adjective phrase, PP which is a preposition phrase, and VP is word phrase. Let us see what kind of higher level structures come, we come to this entities here, new and returning, you see JJ has new CC as and VBG as returning. So, just like before, just like UJF campus, new and returning these entities have been bracketed individually. And the 3 entities new and returning have these 3 words forms a bigger structure, this is the adjective phrase, you see the left bracket here, the left bracket here starts with ADJP level. And the whole scope is the 3 words, the corresponding enclose, the corresponding closing bracket is the right bracket here. What is the meaning of these? The meaning of these is that new and returning is an adjective phrase, and that is indicated here, the whole thing is an adjective phrase.

Now, we come to the students is a plural noun as indicated by NNS, so NN have S has a child students. Now returning an new and returning students, together form a bigger phrase, a noun phrase, so adjective phrase here and a plural noun here. These two together form, what is called a noun phrase, would you call it an noun chunk no. Because it becomes it is a recursive phrase, new and returning students is a noun chunk. There is no recursion here, this whole thing is a noun chunk, and it is a noun phrase also. We will distinguish noun phrase and noun chunk for this particular example, as we move up the tree structure.

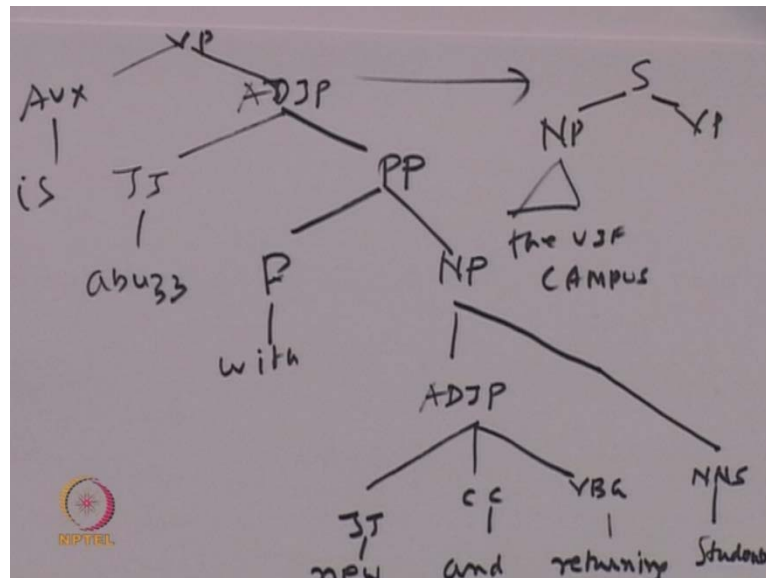
Now, we come to a the label PP, this we can see is a complex phrase, it is a bigger phrase, it has with as a preposition, and a noun phrase, which is composed of adjective phrase. And prolonged adjective phrase is composed of 2 adjectives JJ and VBG new and returning link, with a conjunct CC and so this preposition phrase PP is composed of a preposition with, and a noun phrase. Now these, a whole thing PP is combined with abuzz, which is an adjective actually. So, adjective and preposition phrase together, is forming an adjective phrase, an adjective and an preposition phrase together is forming an adjective phrase.

So, we find that abuzz with new and returning students is an adjective phrase. And this is in 2 also you will agree with this abuzz with new and returning students, abuzz is an adjective. After that we have a preposition with, a so abuzz is an adjective preposition with, and new. And returning is an adjective phrase, which qualifies a noun plural, noun student, so abuzz with new and returning students. This whole thing is an adjective, because the head of this whole unit is, and adjective namely abuzz. It is an adjective with a with something, that something is a preposition phrase.

So, we are saying that something is abuzz with whatever but the most important piece of information here is that something is abuzz, abuzz is an adjective. So, abuzz with new and returning students is an adjective phrase this is intuitive, and I hope you agreed to this kind of description. Let us go to the slide once again, and see that abuzz with new and returning students is and adjective phrase. Before that we have the verb auxiliary is abuzz with new and returning students, these whole thing forms a more complex phrase. This is a word phrase now, so is a abuzz with new and returning students is a verb phrase VP here, and then we had noun phrase, already is the UJF campus. So, the UJF campus is a noun phrase followed by is a verb auxiliary. And then we have the adjective phrase,

so noun phrase verb auxiliary, adjective phrase is forming the whole sentence here. Now, we can very quickly, draw the tree and finish the lecture. With that, I will draw the tree in front of you, with this understanding, so we will do it bottom up, so we have...

(Refer Slide Time: 45:35)



New which is a JJ, and which is CC, we have returning which is again VBG, these whole thing actually is an adjective phrase ADJP. And then we have students it is a with small students, but you can make out this is NNS, adjective phrase and NNS together have formed a noun phrase. And then you have a preposition here, which is with now P and NP together form a PP which we have seen already. And this PP along with a JJ, which is abuzz, these JJ and PP together form an ADJP adjective phrase. And we have an auxiliary which is aux and ADJP form auxiliary P form a VP. So, the space should have been managed a bit better, so VP auxiliary adjective phrase is abuzz with JJ, new and returning students. This whole thing has been given a structure, which is same as the bracketed structure. And now with this VP, and with an NP, we have the SNP is nothing but the UJF campus.

So, we have got the whole tree, this is equivalent to the structure on the slide, a part of the structure on the slide, the UJF campus is abuzz, with new returning students. This is a bracketed structure, and linear structure, this corresponds this to the 2-dimensional structure the tree. So, with this we finished the lecture, the summarizing comment here is that, we have understood part of speech labeling, named entity labeling, sentence

labeling as sequence labeling task. Bigger than that is that the chunk labeling task, bigger than that is the bracketed labeling task which is nothing but the tree. In the next lecture, we will continue with the sequence labeling algorithm.