

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

Lecture - 14
POS Tagging; Fundamental Principle;
Why Challenging; accuracy

Yesterday, we began discussing the challenges of part of speech tagging, the principle involved, in part of speech tag design and what kind of challenges, one could face when doing part of speech tagging and designing the part of speech tag. Just to remind everybody part of speech tagging is the first important disambiguation task, that takes place on language data. The text contents meaning, but the road to meaning is not smooth. There are many challenges and difficulties to be faced on the way and part of speech tagging is first such mile stone, which needs to be crossed.

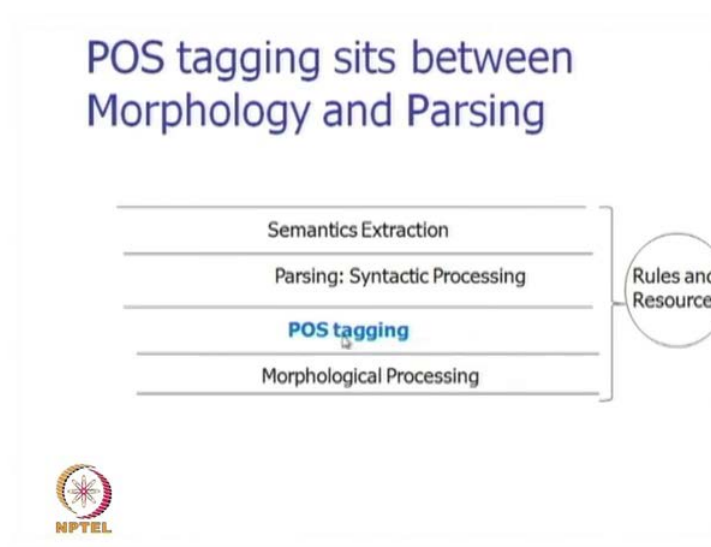
Part of speech tagging is a disambiguation task, because given any word, we would like to first find out, what the category of the word is languages, typically assigned multiple lexical categories for a word. For example, the word play in English can be a verb to play, it could also mean a play, which is game or could mean a dramatic performance. So, it could be noun or verb, the same thing applies to most languages, where there is category ambiguity for words.

So, when we do part of speech tagging, we are essentially solving a disambiguation problem and what is the clue to the disambiguation, the clue comes from the sentence itself, information on the word, the context in which the word appears. But the key thing to remember at this point, is that since part of speech tagging is one of the levels of the natural language processing, this is one of the first level, first we do morphological analysis segmentation etcetera on the word, try to extract as much information as possible from the word.

And then we find out the category of the word, if there is a question has to why the category should be disambiguated, the reason is that the lexical category can be multiple for a word, and the words category in a particular sentence comes from the context. Now in the context when we do the processing of the sentence, we have to first process word. So, the word process, the processing of the word is morphological processing, separating

the suffix, obtaining the word features for example, a word like place. Place can be analyzed in two different ways, it could mean play and plural s a number of place and it could also mean, he plays third person singular number present tense. So, these analysis needs to be done, when we want to obtain the part of speech tag of the word from the context, we have to look at the words around this particular word, the target word. This gives rise to some fundamental points about part of speech tagging, let proceed looking at the slides.

(Refer Slide Time: 03:40)



So, part of speech tagging sits between morphology and parsing, this is a diagram, which shows different levels of processing, for natural language data, natural language text. So, first comes morphological processing, as has been explained a word like [FL] in Hindi will be separated into [FL] has the root and [FL] has the suffix indicating, there by first person singular number future tense.

So, morphological processing obtains this kind of information, then comes of part of speech tagging, part of speech tagging will produce the categories on the word, then comes parsing, which obtains the structure of the sentence and divides the sentence into phrases. After that there is this level of semantics extraction, which obtains the meaning of the sentence in terms of the roles, the semantic roles played by the nouns. So, if we say Raam eats rice with a spoon, then Raam is the agent of eating activity, rice is the object of eating activity and spoon is the instrument.

So, this kind of information is obtained at different layers, part of speech tagging concentrates on simply identifying the grammatical category of the word. So, this is to be born in my, this has a very well define task to do, namely identifying the grammatical category. And the only the information, it can use is the lower level information of morphological processing, it cannot make use of any parsing information, nor can it make use of any semantic extraction, that is because the processing of semantics and parsing comes later.

I show here in a circle rules and resources meaning there by, that all these layers need rules of language constrains, language phenomenon, they also need resources, like the dictionary or a meaning database or properties of the words and so on. So, this forms again a very important component of the processing, which we will have occasion 2 discussed. Now, I will just repeat what, I said here at the part of speech tag level, we have to make use of only the contextual information around the word, the morphological processing information, we cannot assume syntactic structure to be available, nor can we assume semantic information to be available.

(Refer Slide Time: 06:23)

Morph → POS → Parse

- Because of this sequence, at the *level* of POS tagging the only information available is the word, its constituents, its properties and its neighbouring words and their properties

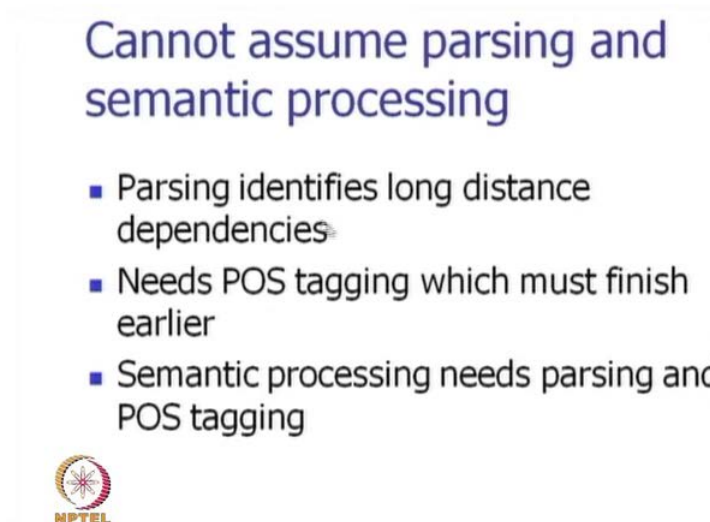
The diagram shows a sequence of words in boxes: w_0 , w_1 , w_2 , ..., w_i , w_{i+1} , w_{n-1} , w_n . An arrow points from the text 'Word of interest' to the box containing w_i . The NPTEL logo is visible in the bottom left corner of the slide.

So, this change of Morph POS and Parse produces a very fundamental an important constraint on what can be done, at the part of speech tag level. So, this is our concern, what can we do at the part of speech tag level. So, because of this sequence at the level of part of speech tagging, the only information available is the word, I emphasis again

the only information available is the word, it is constituents that means, it is suffixes and the morphemes, it is properties, lexical properties and it is neighboring words and their properties.


So, this picture shows, what we mean by context and the word properties, here is a word sequence $w_0 w_1 w_2 \dots w_i \dots w_n$, there are n words in this window may be a sentence, the word of interest is w_i . And the part of speech tag on w_i can be obtained only by information on w_i and around w_i , this point about around w_i is important, because we cannot assume an extremely long contextual window around w_i ok. So, it has to be necessarily limited, because part of speech tagging is only the preliminary level of processing.

(Refer Slide Time: 07:54)



Cannot assume parsing and semantic processing

- Parsing identifies long distance dependencies
- Needs POS tagging which must finish earlier
- Semantic processing needs parsing and POS tagging




So, if we proceed further, then we are stating that, we cannot assume parsing and semantic processing, parsing identifies long distance dependencies, it needs part of speech tagging, which must finish earlier, before parsing the part of speech tagging must be over. Semantic processing needs parsing and part of speech tagging, which means this again is a much later process. So, therefore, this sort of establishes the fact, that part of speech tagging, will not be able to make use of parsing and semantic processing.

(Refer Slide Time: 08:30)

Example

- *Vaha ladakaa so rahaa hai*
- *(that boy is sleeping)*
- *Vaha cricket khel rahaa hai*
- *(he plays cricket)*
- The fact that "vaha" is demonstrative in the first sentence and pronoun in the second sentence, needs deeper levels of information



We take examples, we discussed in the last class some examples of the following form [FL] so [FL] that boy is sleeping [FL] cricket [FL], he is playing cricket he plays cricket he is playing cricket. So, these two sentences were pointed out to be of very same similar structure [FL] cricket both nouns, so [FL] verbal roots [FL] auxiliary verb indicating present tense continuous tense. So, these 2 sentences are very similar, only changes are this noun [FL] 2 cricket and [FL].

Now, if we look at the translations that boy sleeping, so [FL] here is a demonstrative qualifying [FL] a particular boy and [FL] here is not a demonstrative, it is a pronoun it refers to a person, which came in the discourse before a particular noun, which came in the discourse before and this [FL] refers to this person therefore, it is a pronoun. So, even though the sentences have very similar structure, this [FL] is a demonstrative, this [FL] is a pronoun.


So, the fact that [FL] is demonstrative in the first sentence and pronoun in the second sentence needs deeper levels of information, if we think for 2 minutes why is it that this [FL] is a demonstrative and this [FL] is not a demonstrative, it refers to something outside, this sentence outside, this context. Then you must appreciate that, this requires knowledge of the word, this requires knowledge of the language and it is not a simple task. So, this [FL], which is demonstrative [FL] it refers to [FL], but this [FL], which is a

pronoun, it refers to something outside, the sentence and how we do this disambiguation is a complex task.

(Refer Slide Time: 10:38)

"vaha cricket" is not that simple!

- *Vaha cricket jisme bhrastaachaar ho, hame nahii chaahiye*
- *(that cricket which has corruption in it is not acceptable to us)*
- Here "vaha" is demonstrative
- Needs deeper level of processing




So, to further establish my point, I would like you to say that [FL] cricket when, we resolve, it saying that [FL] is a pronoun is not a simple problem, this can be understood by looking at another example. Let's take this sentence [FL] cricket [FL] that cricket, which has corruption in it is not acceptable to us. So, this is a sentence, which is along sentence, it is a complex sentence, because it has 2 verbs ho in a form of a main verb and [FL] is the other verb.

So, it is a complex sentence with 2 components in it [FL] and [FL] cricket [FL], so this [FL] again has become demonstrative, this [FL] is referring to, this cricket it is a demonstrative for, this noun, it is a qualifier for this noun and therefore, it should be detected as demonstrative. If you compare this with the earlier sentence [FL] cricket [FL] and [FL] cricket [FL], these 2 words have different roles to play one demonstrative the other pronoun. Even though the same noun cricket appears there, therefore, it is a complex processing, if I have been able to impress this on you, that this is a complex processing by which, we identify the grammatical categories and it needs deeper level of processing, it requires semantic information.

(Refer Slide Time: 12:28)

Syntactic processing also cannot be assumed

- *raam kaa yaha baar baar shyaam kaa ghar binaa bataaye*
JAANAA
mujhe bilkul pasand nahii haai
- (I do not at all like the fact that Ram goes to Shyam's house repeatedly
without informing (anybody))
- "Ram-GENITIVE this again and again Shyam-GENITIVE house
any not saying GOING I-dative at all like not VCOP"
- JAANAA can be VINF (verb infinitive) or VN (verb nominal, i.e., gerundial)

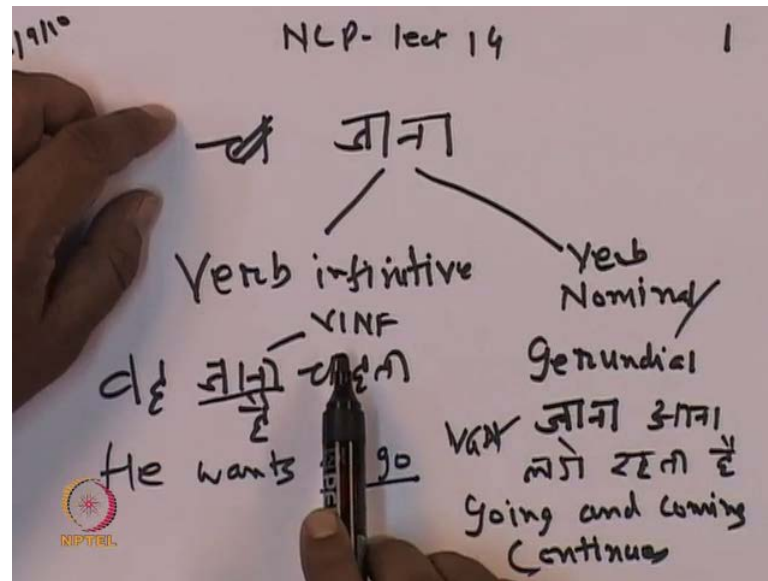


Proceeding further syntactic processing also cannot be assumed, at the level of part of speech tagging syntactic processing means identifying the phrases, the constitutions of a long sentence. And how we obtain those segments, it is a structural work uncover, the structure contained in the sentence syntactic processing also cannot be assumed, that the part of speech tag level, because parsing requires pos tagging. I give here an example sentence, which is interesting to analyze [FL], it is a long sentence, let me read the Hindi sentence again [FL]. The English translation is I do not at all like the fact that, ram goes to Shyam's house repeatedly without informing anybody again, I do not at all like the fact, that ram goes to Shyam's house, repeatedly without informing anybody.

So, this is an interesting sentence from the view of part of speech tagging of a particular word, which is capitalized here [FL]. So, if we look at the features of the words in particular of the proper nouns here, ram is in a genitive form here [FL] also is in genitive form [FL]. So, the glass of the sentence would be ram-genitive, this again and again Shyam-genitive, house any not saying going I dative at all like not V COP.

So, this is a grammatical analysis of the words in the sentence and we have capitalized the word, which needs to be disambiguated with respect to part of speech, now we look at this [FL] and say that [FL] can be verb in it infinitive ok. So, or it could be a gerundial verb nominal, so let me explain this 2 terms to you, by writing on the paper.

(Refer Slide Time: 14:51)




The word [FL], it can be a verb infinitive or it could be a verb nominal also called gerundial. So, let me give an example to clearly distinguish, this 2 cases verb infinitive [FL], he wants to go. So, to go is the infinitive and this [FL] is the infinitive, this is [FL] how about gerundial, how can [FL] be gerundial [FL] going and coming continues. So, here going is [FL], this is the act of goings, it is a gerundial or verb nominal. So, here the tag would be let say V G N, so there are these 2 tags for [FL] one is the V infinitive tag or V gerundial tag. So, [FL] form is the same, we have to produce the correct tag V G N or V I N F depending on its role in the sentence.

(Refer Slide Time: 16:29)

Syntactic processing also cannot be assumed

- *raam kaa yaha baar baar shyaam kaa ghar binaa bataaye*
JAANAA
mujhe bilkul pasand nahii haai
- (I do not at all like the fact that Ram goes to Shyam's house repeatedly
without informing (anybody))
- "Ram-GENITIVE this again and again Shyam-GENITIVE house
any not saying GOING I-dative at all like not VCOP"
- JAANAA can be VINF (verb infinitive) or VN (verb nominal, i.e., gerundial)



So, let us look at the slide again and understand, why it is difficult [FL], we have to produce [FL] or V G N verb be infinitive or nominal verb or gerundial verb on this particular word, how do we do this. We look at the next slide and see where is the clue for this disambiguation, the correct clue for disambiguation for [FL] and this word group is for apart ok. So, you see this [FL] could be verb infinitive or verb gerundial, this is actually a gerundial, because [FL] ram's going this is the translation ram's going to Shyam's house he is not like by me.

So, this is actually going the gerundial form and it is gerundial, because of [FL]. So, this clue has to come from [FL], which is for away from [FL] there are many words 1 2 3 4 5 6 7 8, 8 words between [FL] and [FL] and this fact, that [FL]. This particular text segment has to be sort of masked and the dependency between [FL] and [FL] the relationship between [FL] and [FL] has to be un covered, to detect that, this is a gerundial.

So, this kind of long distance dependency can be detected only if, this particular structure here [FL], if this particular structure is masked, it is take care at different level and [FL] and [FL] or brought together then only we can identify that, this is a gerundial verb, nominal verb. And this means, it requires doing parsing on the sentence, you have detect the structure and then finally, comes mean sentence, which is [FL], this shows that, one has to take a sentence and peep into the structure of the sentence identify the segments and then only this kind of disambiguation is possible.

So, this is what is written, this needs parsing, which in turn needs correct tags, thus there is a circularity, which can be broken only by retaining 1 of V INF and V N. So, the point being made here is that, in general in the word's case scenario, it is impossible to identify V INF and V N correctly at the level of part of speech tag and therefore, there is no point in retaining these 2 tags, because at the part of speech tag level, you correct do anything to separate them.

(Refer Slide Time: 19:33)

Fundamental principle of POS tagset design

- IN THE TAGSET DO NOT HAVE TAGS THAT ARE POTENTIAL COMPETITORS AND TIE BETWEEN WHICH CAN BE BROKEN ONLY BY NLP PROCESSES COMING AFTER THE PARTICULAR TAGGING TASK.



So, this takes us to the fundamental principle of part of speech tag set design, which is this, in the tag set do not have tags, that are potential competitors and tie between, which can be broken only by N L P processes coming after the particular tagging task. So, this applies to any tag set design whenever, for annotation where, designing a tag set, we have to bear this principles in mind. If there are potential competitors and the tie between them can be broken only by subsequent N L P processes, which come after this particular annotation task, then there is no point keeping both the competitors together only one of them can be resolved. So, this is the fundamental principle of part of speech tag design.

So, I hope this particular point is brought home very clearly to you that part of speech tag design is essentially a leveling problem, identifying the correct level is a task, this needs disambiguation and the task cannot make use of any subsequent task, which comes after part of speech tag. So, having looked at the principle of pos tagging and in general the task of tag set design.

(Refer Slide Time: 20:58)

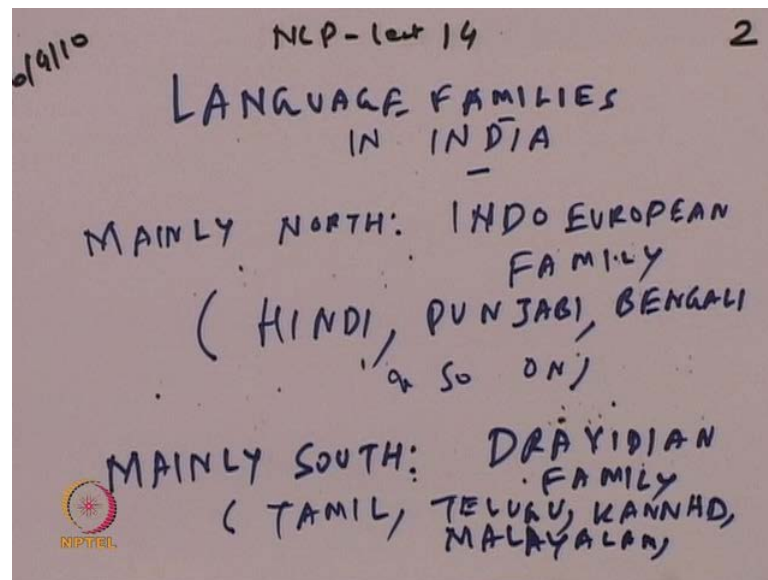
Indian Language POS tagging

- Happening under a large scale nation wide project called "Indian Language to Indian Language Machine Translation"
- Multiple institutes across the country: consortium mode
- ILILMT POS tags are accepted as standard for tagging of IL corpora



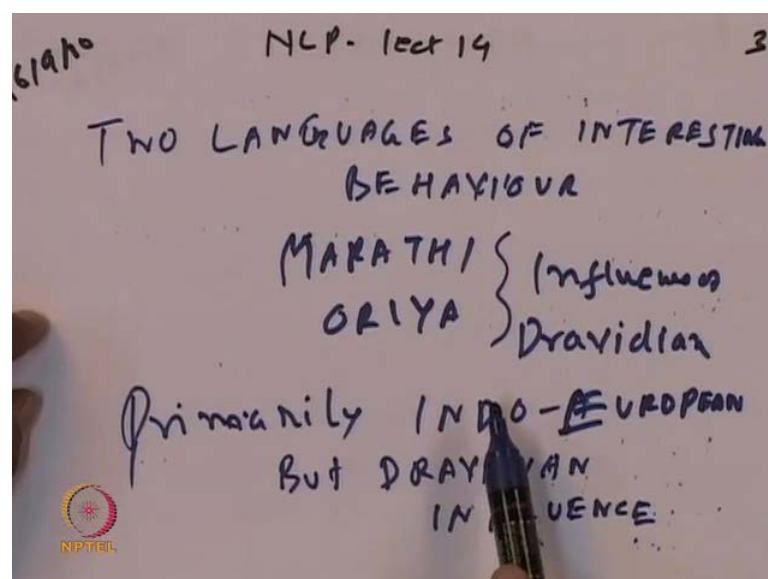
We now proceed, with discussion of Indian language part of speech tagging, we had started this, in the last class, we mention that Indian language part of speech tagging is happening, under a large scale India wide project called Indian language to Indian language machine translation. Multiple institutes across, the country are participating in this, so it is a consortium of institutes. The part of speech tags are the tags required for Indian language to Indian language machine translation and therefore, the attempt is to design a pan Indian tag set. The tag set, which will be applicable across languages, from languages in the European family, to the Dravidian family, to the Syno Tibitian family in the east.

(Refer Slide Time: 22:04)



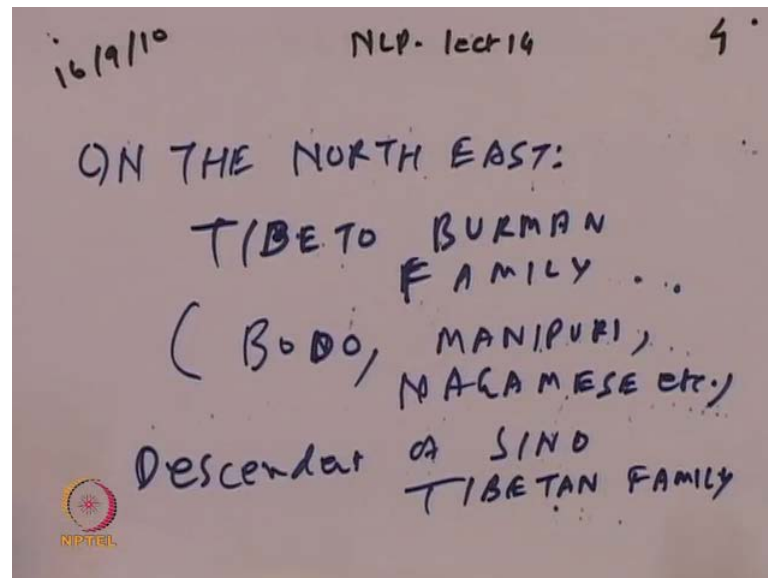
So, let me just make a remark on this particular point, the language groups in India and I would like to write it down, language families in India, this is important for any kind of natural language processing in the country. So, mainly in the north, mainly north, we have Indo European family, so important members in this are Hindi Punjabi Bengali and so on. Mainly south, we have Dravidian family the major members in this are Tamil Telugu Kannada and Malayalam alright.

(Refer Slide Time: 22:56)



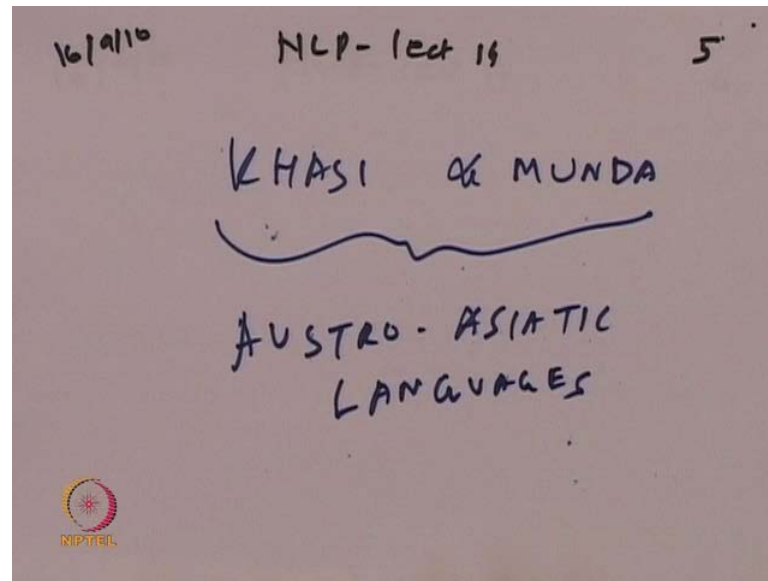
There are 2 languages of interesting behavior, these are Marathi and Oriya and both, these have influence of Dravidian, primarily they are Indo Aryan or Indo European primarily Indo European, but Dravidian influence. So, Marathi and Oriya being at the border of south India north India, have influence of Dravidian in them, though they are particularly Indo-European languages.

(Refer Slide Time: 23:50)



On the north east, we have the Tibeto Burman family, in this Tibeto Burman family the prominent members are Bodo Manipuri Nagamese etcetera. So, these are Tibeto Burman family, descendent of the larger Sino Tibetan family alright. So, this covers the major languages of the country, Indo European in the north Dravidian in the south and Sino Tibetan or Tibeto Burman in the north east, I must mention 2 interesting cases, 2 languages, which are Khasi and Munda, they are what are called Austro Asiatic languages.

(Refer Slide Time: 24:45)



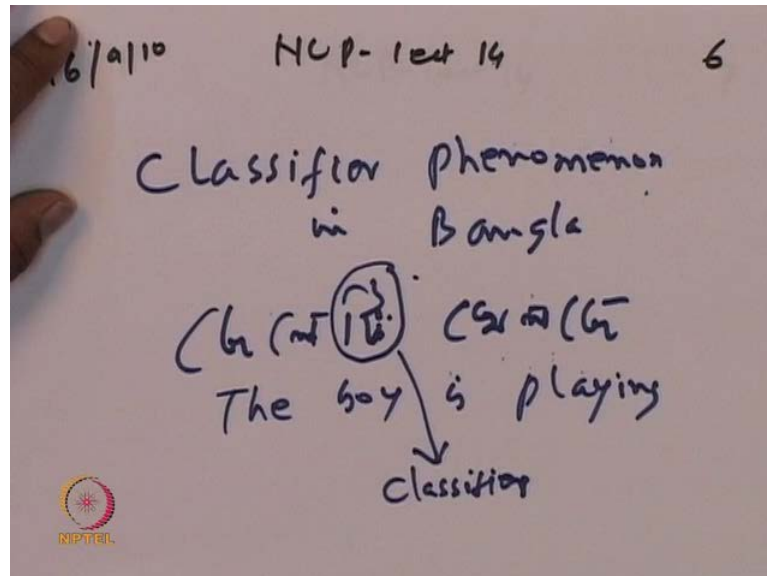
So, these are the only 2 languages, which belong to this family and they are the representatives of the Austro Asiatic family. So, to summarize, there are 4 language families in the country 2 very large groups are indo Arian containing mainly the languages of north India then we have Dravidian the important members are the 4 languages from the south. Third important group is the group of languages in the north east of the country namely Boro Manipuri Nagamese etcetera.

There are 2 languages, which are interesting cases Munda and Khasi, they belong to Austro Asiatic family, we have also said that Marathi and Oriya are the 2 languages, which have influence of Dravidian languages on them, even though, they are primarily indo-European languages alright. So, what is the relevance of this discussion is that, we have a very diverse set of languages with their complex language phenomena, which need to be resolved whenever we, do anything for the machine processing of this languages.

So, we cannot for example, have a part of speech tag set only for Hindi considering only Hindi language phenomena and expect there by that, it will work for other languages also. So, other languages have phenomenon, which are not present in Hindi and what should we do for part of speech tagging when such phenomena appear in the language. So, we will take some example of this, but one example that comes to my mind

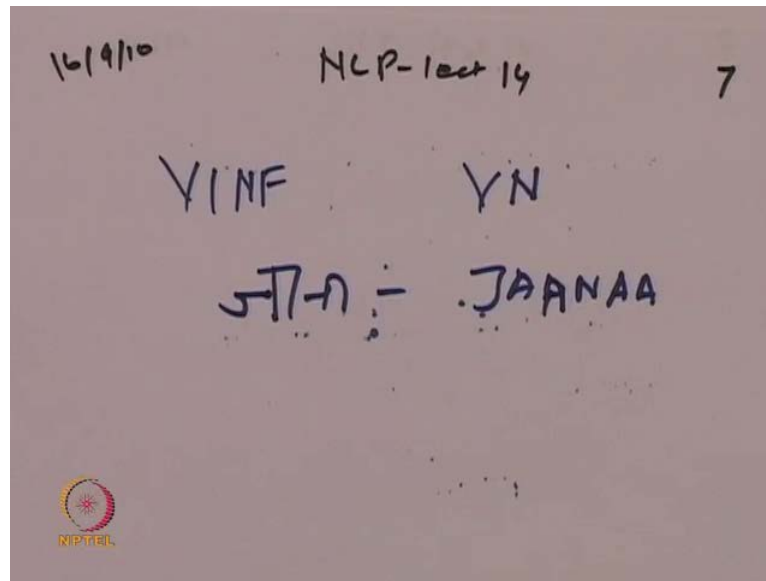
immediately is the phenomenon of classifiers in Bengali ok. So, in Bengali, we say Cheleti, I will write it down.

(Refer Slide Time: 26:59)



Classifier phenomenon in bangle, so Cheleti, this is Bengali script [FL], the boy is playing, this t is an interesting language element, which has a definite role to play, t is called a classifier, it denotes a particular boy. So, it can be looked up and has a discourse identifier, it is a discourse particle and this identifier is a particular boy, it is a classifier, so this phenomenon does not exist in Hindi. And therefore, it if do not do anything about this classifier then we are not doing justice to the processing of Bengali language, so this particular phenomenon is to be handled. Therefore, when we consider the fan Indian scenario all Indian languages and discuss their parts of speech tagging then all these phenomenon had to be have to be take into account. I will I would like to give another example, we have remarked.

(Refer Slide Time: 28:17)



That V INF verb infinite and nominal verb nominal or difficult distinguish for Hindi, so the example, which was discussed was [FL] and we found that, it is impossible to distinguish between them V N and V INF, in the most general case. However, if a language, shows a marking on the word itself, for this infinitive verb and gerundial verb or nominal verb, then we can written, these 2 tags V INF and V N, because they can be resolve from the property of the word.

So, Dravidian languages and even Marathi language can arrange for this, so verb in infinitive form and verb in gerundial or nominal form have different forms, they have completely different forms ok. And therefore, we can written the 2 tags for these languages, even if Hindi does not written the other languages like Dravidian and Marathi can written this tags and this considerations have to be kept in mind. So, let me summarize this is discussion by a important point, the point is that, if you are designing part of speech tag for one single language, then it is one task when, we are designing the part of speech tag set for a number of languages with different phenomena coming from different families. Then we have to be more designing and more encompassing with respect to language phenomena, so these considerations went into the part of speech design, which came from the I L I L M T.

(Refer Slide Time: 30:07)

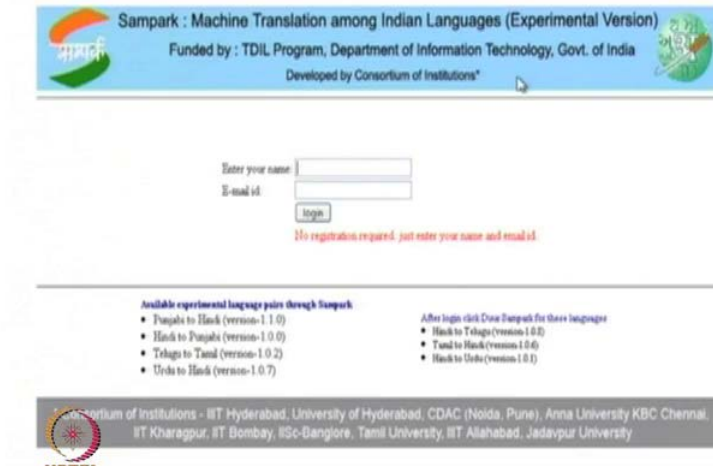
Indian Language POS tagging

- Happening under a large scale nation wide project called "Indian Language to Indian Language Machine Translation"
- Multiple institutes across the country: consortium mode
- ILILMT POS tags are accepted as standard for tagging of IL corpora



POS tag designing task, so this task was to create a part of speech tag set for the purpose of machine translation from one Indian language to another.

(Refer Slide Time: 30:16)



Sampark : Machine Translation among Indian Languages (Experimental Version)
Funded by : TDIL Program, Department of Information Technology, Govt. of India
Developed by Consortium of Institutions*

Enter your name:
E-mail id:

No registration required, just enter your name and email id.


Available experimental language pairs through Sampark

- Punjabi to Hindi (version-1.1.0)
- Hindi to Punjabi (version-1.0.0)
- Telugu to Tamil (version-1.0.2)
- Urdu to Hindi (version-1.0.7)

After login click One Sampark for these languages

- Hindi to Telugu (version 1.0.0)
- Tamil to Hindi (version 1.0.0)
- Hindi to Urdu (version 1.0.1)

Consortium of Institutions - IIT Hyderabad, University of Hyderabad, CDAC (Noida, Pune), Anna University KBC Chennai, IIT Kharagpur, IIT Bombay, IISc-Bangalore, Tamil University, IIT Allahabad, Jadavpur University




And this is the website of Sampark, the machine translation system amongst to Indian languages. Now, we discussed this yesterday.

(Refer Slide Time: 30:29)

Noun Tags (Examples in Hindi)

Sl. No	Category		Label	Annotation Convention**	Examples	Remarks
	Top level	Subtype (level 1) Subtype (level 2)				
1	Noun		N	N	ladakaa, raajaa, kitaaba	
1.1		Common	NN	N__NN	kotaaba, kalama, cashmaa	
1.2		Proper	NNP	N__NNP	Mohan, ravi, rashmi	
1.3		Verbal	NNV	N__NNV	NA	Not Required
1.4		Loc	NST	N__NST	Uupara, niice, aage, piiche	



And just now a point, which was made showed that, it make sense to have what is called a hierarchical tag, we said that V INF and V N will not to be distinguishable in Hindi, but they are distinguishable in Marathi and Dravidian languages. So, it make sense to have a hierarchy you have, we have a general verb tag, which applies across languages, but the sub categories like V INF and V N apply to certain languages and not all.

So, these leads us to a hierarchical categorization of tags, we discussed yesterday that, there is this general category of noun and under that, we have common noun proper noun and locative nouns. And this was discussed carefully with emphasis on the last category, which is the interesting case of words like [FL] niche [FL] etcetera, which can function both as noun and also has pos positions.

(Refer Slide Time: 31:29)

Pronoun & Demonstrative Tags (Examples in Hindi)

2		PR		PR	
2.1	Personal	PRP	PR_PRP	Yaha, vaha, jo	
2.2	Reflexive	PRF	PR_PRF	Vaha, main, tum, ve	
2.3	Relative	PRL	PR_PRL	Apanaa, swayam, khuda	
2.4	Reciprocal	PRC	PR_PRC	Jo, jis, jab, jahaam	
2.5	Wh-word	PRQ	PR_PRQ	Paraspara, aapasa	
3	Demonstrative	DM	DM	Kauna, kab, kahaam	
	Deictic	DMD	DM_DMD	Vaha, jo, yaha,	
	Relative	DMR	DM_DMR	Vaha, yaha	
	Wh-word	DMQ	DM_DMQ	jo, jis	
				kol, kis, kaun	

In case of pronouns and demonstratives, we again have some different categories like personal reflexive relative reciprocal, Wh word for pronoun for demonstrative deictic relative Wh word. What was pointed out is that demonstratives in most cases the words, which are demonstratives are also pronouns and we have already discussed in examples extensively to show the difficulty of this identification, because a word can have both pronoun and demonstrative tags. So, the hierarchy here is that pronoun is the high level tag and under it, you have these sub-levels, personal reflexive relative reciprocal Wh word.

(Refer Slide Time: 32:11)

Verb Tags (Examples in Hindi)

4		V		V	
4.1	Main	VM	V_VM	gaa, gayaa, sonaa, haMtaa, hai, rahaa	
04/01/01	Finite	VE	V_VM_VE	gaa, gayaa, sonaa, haMtaa	
04/01/02	Non-finite	VSE	V_VM_VSE		This subtype WILL NOT be used for Hindi as Hindi does not have enough information at the word level.
04/01/03	Infinitive	VSE	V_VM_VSE		--do--
	Gerund	VNG	V_VM_VNG		--do--
	Auxiliary	VAUX	V_VAUX	hai, rahaa, tha	


Verbs are we have been discussed and we spend some time discussing infinite category and gerundial category, we saw that for Hindi, there is no point keeping these 2 sub categories. But, of course, we have to keep the verb category, which is the top level and then the main verb and auxiliary verb, which is a very universal phenomena across all languages. Auxiliary verbs are helping verbs and this is indicated by underscore V AUX. So, it is verb and then auxiliary verb for example, [FL], these are auxiliary verbs.

Similarly the main verb is where indicated by V underscore V M [FL] etcetera. Somewhere languages a most Dravidian languages will be able to keep, these categories the sub categories and these symbols. So, they will have another level of refinement in the tags and this is a quite desirable, because this contains information, it is important to encode this information fine. So, I suppose it is becoming clear to you why it make sense to have a hierarchical categorization of POS tags.

(Refer Slide Time: 33:20)

**Adjective, Adverb
and Conjunction Tags
(Examples in Hindi)**

5	Adjective			JJ		sundara, acchaa, baRaa	
6	Adverb			RB		jaldi, teza	
7	Postposition			PSP		ne, ko, se, mein	
8	Conjunction			CC	CC	aur, agar, tathaa, kyonki	
8.1		Co-ordinator		CCD	CC__CCD	aur, balki, parantu	
8.2		Subordinator		CCS	CC__CCS	Agar, kyonki, to, ki	
08/0 201			Quotative	UT	CC__CCS__UT	---	Not required

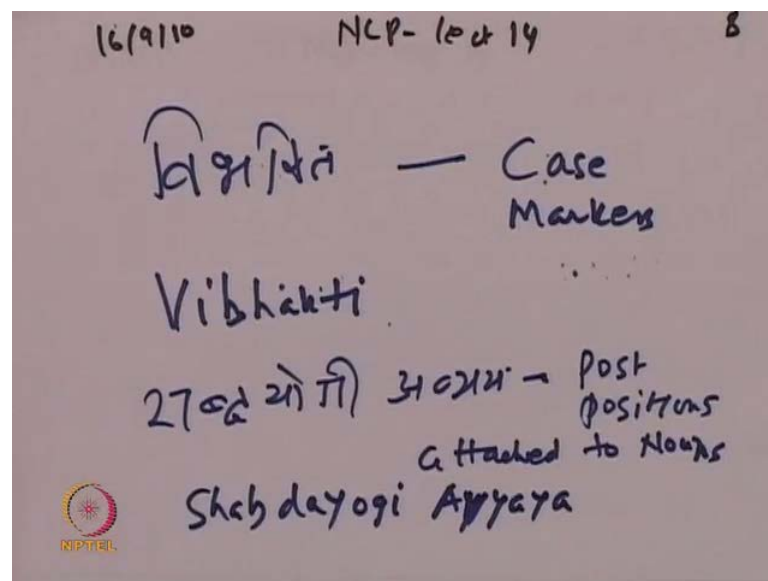
 NPTEL

Now, we go to adjective and adverb in the last lecture we remarked that, in Indian languages, especially adverbs are do not have any special marking ok. So, for example, in English, the word quick can be adjective, but if you want to have an adverb for from quick, you have to insertly quickly. So, English explicitly has this marking on the word in the form of ly, which shows that, the word is an adverb however, in Indian languages adverbs can act as adjectives and vice versa, the primary categories adjectives.

So, [FL] for example, is typically an adjective [FL] he is a good boy, but [FL] here [FL] is a manner of playing, he plays well therefore, this is an adverb. So, adjective adverb distinguish again becomes a difficult problem in Indian languages and there are can be debate about in general, if you cannot distinguish between adverb and adjective ok. Then why written both, the reason is historical, because adjectives and adverbs are very large, content word categories and one needs symbols on them, but when we do part of speech tagging actually on practical systems, we often find, that adjective and adverb are not resolve accurately, which beings down the accuracy of the part of speech tagger. Then we have this important category call postposition P S P [FL] etcetera, so ne is the argotic marker [FL] is the accusative markers se is instrumental marker [FL] is the locative marker, so these are from Hindi and these are postpositions.

Now, an important point of discussion here is that, in Hindi post positions are typically separate from the noun, that they are attach to, there is a blank space between them. But, in languages like Marathi and Dravidian languages, these words, these P S P is get attached to the noun itself. And therefore, they become what are called Vibhaktees and Shabdayogi avyaya, I will write these 2 terms, which are very important for Indian language processing Vibhakti, which are something like case markers Vibhakti and the other is Shabdayogi avyaya, these are postpositions, but attached to nouns test nouns.

(Refer Slide Time: 35:50)



So, I will write it in English Shabdayogi avyaya. So, just to given example in Hindi [FL], let us say [FL], this is ram to call, so this [FL] is an accusative marker and here it is separate from ram, there is blank space, but in Marathim this will [FL] etcetera and this Vibhakti is accusative Vibhakti is attached. So, this means that, the case marker in Marathi is attach to the word itself going to the case of Shabdayogi avyaya in Hindi.

(Refer Slide Time: 37:14)

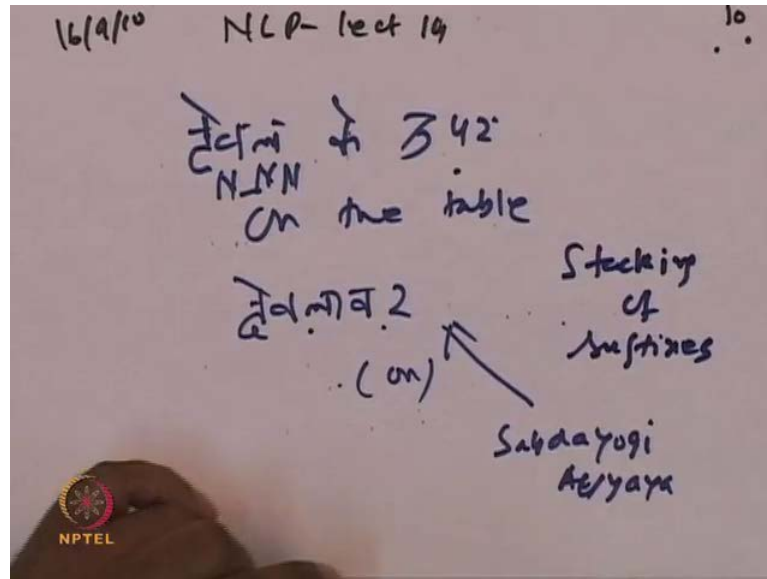
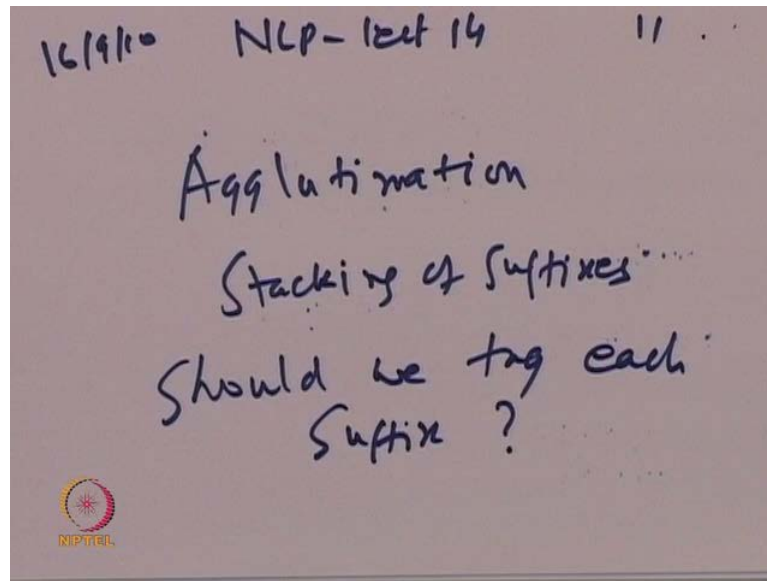


Table [FL] on the table in Marathi, it will be Tabulaavar, so you can see uper is separate from table, but here verb, which means on is attach to table, so this is the Shabdayogi avyaya. So, why this discussion, this discussion is to see, what is involved in part of speech tagging, this 2 languages in a Hindi, if you look at the sentence here in Hindi, we will have to produce tags for table, which is noun, common noun n underscore N N, K is a case marker one tag for this, [FL] is postposition another tag for this. So, they will need to have 2 tags, but you see in Marathi K [FL] or not, they are separately there are not, there with white spaces, they are attached with the noun itself Tabulaavar and this what should, we do for tagging here. Now, in Marathi you can have a stacking of suffixes, there can be suffix stacking a number tags suffixes can get attached.

(Refer Slide Time: 38:41)



So, there this phenomenon of agglutination stacking of suffixes, in which case it is important to identify the suffixes separately, but should we tag each suffix, this is an important question, which needs a careful analysis. And this needs careful analysis from the point of view of lower level processing, subsequent processing and so on. It is very clear that, when there is stacking of suffixes and a larger word is formed, we have to segment the word carefully and identify these suffixes.

(Refer Slide Time: 39:32)

Adjective, Adverb and Conjunction Tags (Examples in Hindi)

5	Adjective			JJ		sundara, acchaa, baRaa	
6	Adverb			RB		jaldi, teza	
7	Postposition			PSP		ne, ko, se, mein	
8	Conjunction			CC	CC	aur, agar, tathaa, kyonki	
8.1		Co-ordinator		CCD	CC__CCD	aur, balki, parantu	
8.2		Subordinator		CCS	CC__CCS	Agar, kyonki, to, ki	
08/0 2011			Quotative	UT	CC__CCS__UT	---	Not required

NIPTEL


Anyway so we proceed with the tags, so postposition or ne Ko etcetera, English on the other hand has preposition. So, with for example, we will come before noun eat with spoon, so with come before noun spoon in Hindi, it will be spoon say come after spoon, which is postposition. Then, we have category called conjunctions and conjunction has 2 sub categories coordinator and subordinator, in coordinator, we have conjuncts like and are, but etcetera, which link 2 main sentences ram sings and shyam dances. So, this and is a coordinator and then we have a subordinator, because there is a subordinated class and the main class, there is a main class and the class that, it is subordinating.

So, this is the C C S category things like agar [FL] that means, if while because so this kind of this kind of entities link a main class with a subordinate class. And when we do so then we have to identify that, these are sentence linkers class linkers and we have 2 different categories of the word C C D and C C S. Typically, it is not at all difficult to identify those categories C C underscore C C D underscore C C S is not any difficult, because the words are distinct.

(Refer Slide Time: 41:17)

**Particles and Quantifiers Tags
(Examples in Hindi)**

9	Particles		RP	RP	to, bhi, hi	
9.1	Default		RPD	RP_RPD	to, bhi, hi	
9.2	Classifier		CL	RP_CL		Not required
9.3	Interjection		INJ	RP_INJ	are, he, o	
9.4	Intensifier		INTF	RP_INTF	bahuta, behada	
9.5	Negation		NEG	RP_NEG	nahin, naha, binaa	
10	Quantifiers		QT	QT	thoRaa, bahuta, kucha, eka, pahataa	
10.1	General		QTF	QT_QTF	thoRaa, bahuta, kucha	
10.2	Cardinals		QTC	QT_QTC	eka, do, teen	
10.3	Ordinals		QTO	QT_QTO	pahataa, duusaraa	



Then, we come to a category called particles, particles are very interesting language elements, things like to [FL] is an emphasis marker indeed, b is also to is a discourse element, it has many different roles to play for example, you could say in a discourse, you could say to [FL] then you will come. So, this to is something like a something like a

discourse linker may not have any meaning in itself, but it functions in linking 2 different sentences.


Then we have a classifier, classifier is not required in Hindi, but in Bengali it is a classifier, which is a definite entity identifier, interjections are exclamation symbols are [FL] ok. [FL] in Hindi [FL] in English aha in English, these are interjections, they do not change their form, but they are there in the language to emphasizing point or say something interesting. Intensifiers are sometimes regarded as adverbs, because they qualify and adjectives, so [FL] and these language particles are elements, which play the role of intensifying an adjective.

So, something like very in very intensely and so on, which are the examples in English, then we have negations [FL] etcetera, which are there in Hindi and most languages have these elements. But, the interesting point here is that, the negation particles sometimes are reflected through a change in the verb form or through a change in the form of the word, which is negated. So, negation identification can become a complex problem, because it may or may not be present as a separate particle this at this point, we will discuss as we go ahead, then we have the quantifiers. So, [FL], etcetera, etcetera, which are quantifiers and there are different kinds of quantifiers, general quantifiers [FL] [FL] cardinals like [FL] ordinals like [FL]. So, ordinals are first second cardinals are 1, 2, 3, etcetera, etcetera, then a [FL] is little [FL] is much many, these are quantifiers again. And one does not see much difficulty in discussing them or identifying them, because they forms are very definite.

(Refer Slide Time: 44:24)

Residual Tags (Examples in Hindi)

11	Residuals		RD	RD		
11.1		Foreign word	RDF	RD_RDF		A word written in script other than the script of the original text
11.2		Symbol	SYM	RD_SYM	\$. & . ' ()	For symbols such as \$, & etc
11.3		Punctuation	PUNC	RD_PUNC	, ;	Only for punctuations
11.4		Unknown	UNK	RD_UNK		
11.5		Echowords	ECH	RD_ECH	(Paanii-) vaanii, (khaanaa-) vaanaa	



And finally, you have these residuals duals, which are foreign words symbol punctuations unknown words echo words, which need to be given a tag in the language. So, these categories and the subcategories form, the pan Indian tag set for processing of Indian languages, we will proceed further and discuss in the next class about, accuracy of POS tagging.