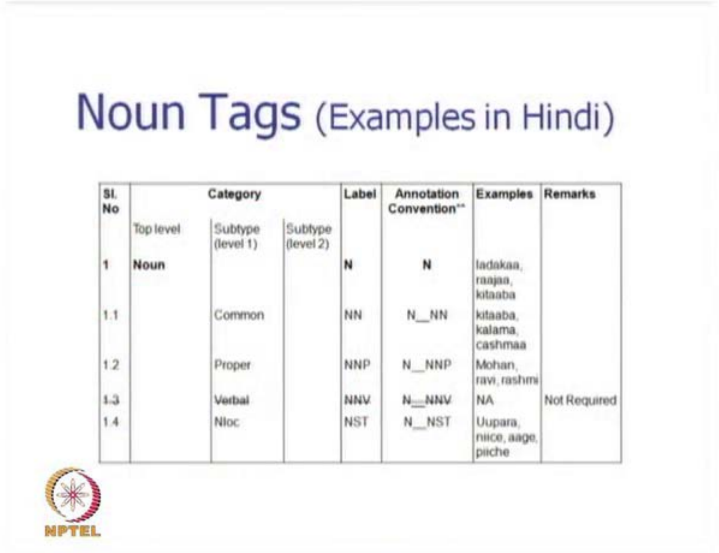
Natural Language Processing

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

Lecture - 15
POS Tagging; Accuracy Measurement; Word Categories

In the last lecture, we discussed why part of speech tagging is a challenging task, part of speech tagging happens to be the first non trivial task of statistical natural language processing, making use of machine learning techniques. Now, we will discuss today how to calculate accuracy of part of speech tagging, what is the way to report correctness of part of speech tags obtained, how to do error analyses. And then we will move on to discussing the categories that come from part of speech tagging, the linguistic foundation of part of speeches which are word categories.

(Refer Slide Time: 01:04)

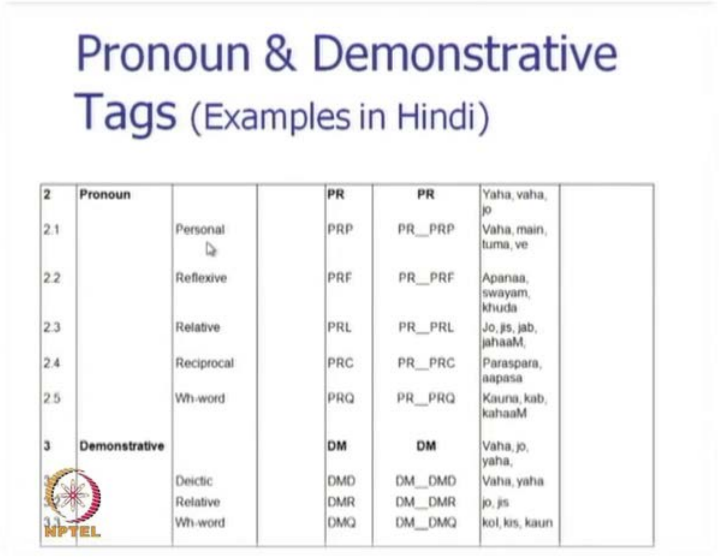


Noun Tags (Examples in Hindi)

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	kitaaba, kalama, cashmaa	
1.2		Proper		NNP	N__NNP	Mohan, ravi, rashmi	
~~1.3~~		~~Verbal~~		~~NNV~~	~~N__NNV~~	NA	Not Required
1.4		Nloc		NST	N__NST	Uupara, niice, aage, piiche	

So, just a recap of the tags that we had seen for Indian language processing, this is happening in the context of a large machine transition effort throughout the country involving many institutions. We saw that the words have higher level categories with the form of noun, adjective, verb, adverb, and so on. And within each large category there are sub categories, so under noun for example, we have common noun, proper noun, Nloc which are special categories, then verbal noun which is not there in Hindi, but is found to be important in Dravidian languages.

(Refer Slide Time: 01:55)

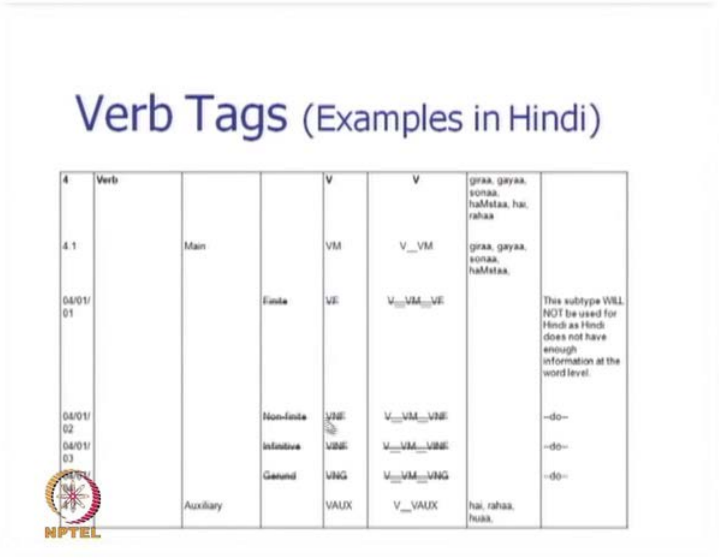


Pronoun & Demonstrative Tags (Examples in Hindi)

2	**Pronoun**			**PR**	**PR**	Yaha, vaha, jo	
2.1		Personal		PRP	PR__PRP	Vaha, main, tuma, ve	
2.2		Reflexive		PRF	PR__PRF	Apanaa, swayam, khuda	
2.3		Relative		PRL	PR__PRL	Jo, jis, jab, jahaaM,	
2.4		Reciprocal		PRC	PR__PRC	Paraspara, aapasa	
2.5		Wh-word		PRQ	PR__PRQ	Kauna, kab, kahaaM	
3	**Demonstrative**			**DM**	**DM**	Vaha, jo, yaha,	
[illegible]		Deictic		DMD	DM__DMD	Vaha, yaha	
3.2		Relative		DMR	DM__DMR	jo, jis	
3.3		Wh-word		DMQ	DM__DMQ	kol, kis, kaun	

Then in pronoun again we have many sub categories like, personal, reflexive, relative, reciprocal and W h word. Under demonstratives we have deictic relative and W h word, which again are quiet common with the pronoun words. So, we remarked that it is often a challenge to correctly distinguish pronoun, tags from demonstrative tags.

(Refer Slide Time: 02:25)

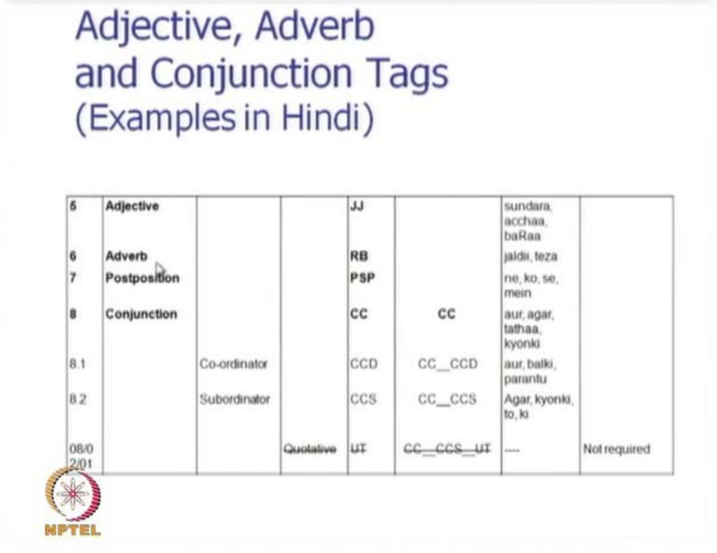


Verb Tags (Examples in Hindi)

4	Verb			V	V	giraa, gayaa, sonaa, haMstaa, hai, rahaa	
4.1		Main		VM	V__VM	giraa, gayaa, sonaa, haMstaa,	
04/01/01			~~Finite~~	~~VF~~	~~V__VM__VF~~		This subtype WILL NOT be used for Hindi as Hindi does not have enough information at the word level.
04/01/02			~~Non-finite~~	~~VNF~~	~~V__VM__VNF~~		--do--
04/01/03			~~Infinitive~~	~~VINF~~	~~V__VM__VINF~~		--do--
[illegible]			~~Gerund~~	~~VNG~~	~~V__VM__VNG~~		--do--
		Auxiliary		VAUX	V__VAUX	hai, rahaa, huaa,	

Under verb we have two large categories again, main and auxiliary under main there could be finite verbs, non finite verbs, infinity, gerams. For Hindi it is not possible to distinguish them an ambiguously from, the word itself are from a small window around the word. And therefore, this sub categories are dropped, we work with main verb and auxiliary verb.

(Refer Slide Time: 02:55)

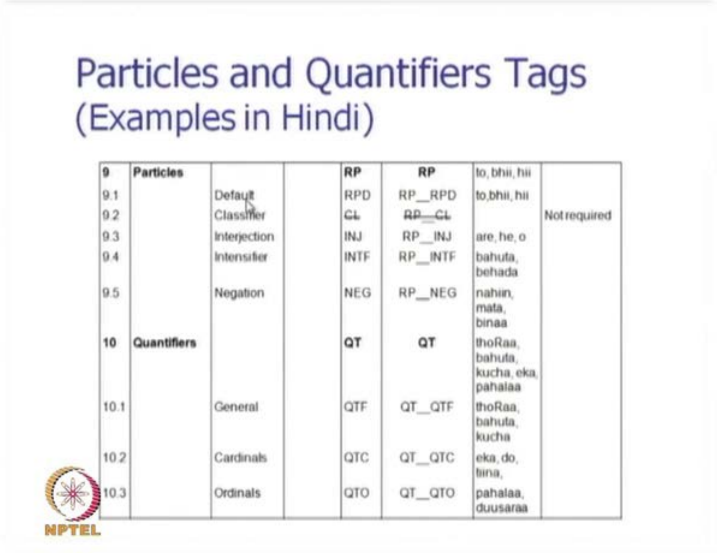


Adjective, Adverb and Conjunction Tags (Examples in Hindi)

5	**Adjective**			**JJ**		sundara, acchaa, baRaa	
6	**Adverb**			**RB**		jaldii, 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 2/01			~~Quotative~~	~~UT~~	~~CC__CCS__UT~~	----	Not required

Then, we move on to adjective and for Indian languages, we mention adjectives and adverbs cannot be many times distinguished in English, we have the le suffix which helps to identify the adjectives, identify the adverbs. Post positions again are like prepositions of English and they come after the noun, conjunctions are coordinators and subordinators, they help join noun phrases, nouns and even complete sentences.

(Refer Slide Time: 03:31)

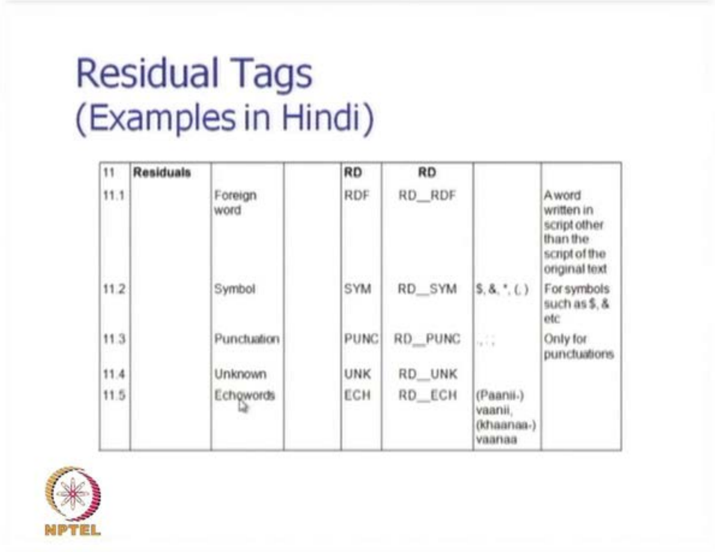


Particles and Quantifiers Tags (Examples in Hindi)

9	**Particles**			**RP**	**RP**	to, bhii, hii	
9.1		Default		RPD	RP__RPD	to,bhii, hii	
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	nahiin, mata, binaa	
10	**Quantifiers**			**QT**	**QT**	thoRaa, bahuta, kucha, eka, pahalaa	
10.1		General		QTF	QT__QTF	thoRaa, bahuta, kucha	
10.2		Cardinals		QTC	QT__QTC	eka, do, tiina,	
10.3		Ordinals		QTO	QT__QTO	pahalaa, duusaraa	

Then we come to particles, where we have default, particles, classifiers, interjection, intensifier and negation. Classifier does not hold for Hindi, Hindi does not have this particular category, but in Bengali they are classifiers. Then a quantifiers are used which are general quantifiers cardinals and ordinals.

(Refer Slide Time: 03:56)



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	

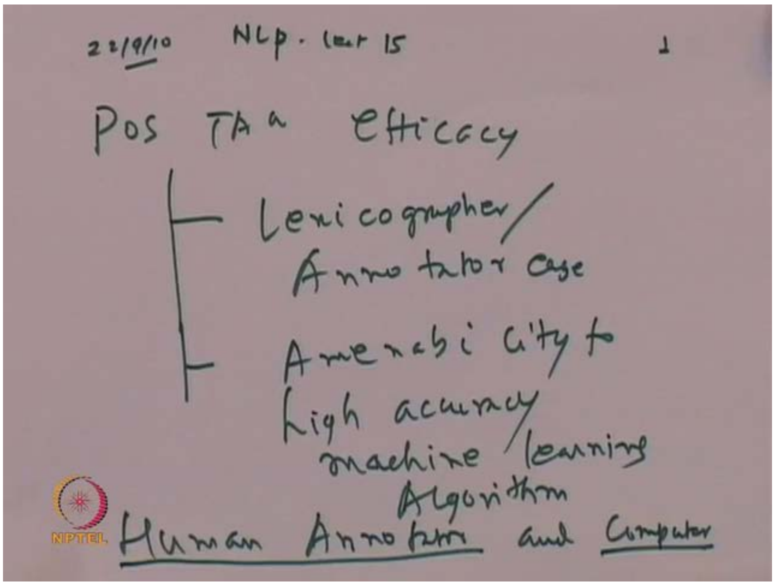
Finally, we have residual tags which are for foreign words, symbols, punctuation, un known words and equal words. So, all these form something like a miscellaneous bag whatever could not be put in large categories are placed here, in this miscellaneous bag. So, this was an account of part of speech tags, which has being used for Indian languages and this is looked up and as a pan Indian standard. Now, if you remarks are in order for this part of speech tags and their symbols.

So, as you can see this part of speech tags have sub categories and this sub categories are denoted by the main category underscore the subcategory. Now, the question of is this grouping of tags, good enough is it correct complete, so that will eventually emerge when we begin to actually part of speech a lot of data, the ministry funded project from government of India, stipulates that language processing groups create about 25000 strong corpora, tag to it part of speech.

And when lexicographers produce tags on the language document and then machine learning algorithms are used to learn this tags, then we would come to know the efficacy of this tag set. So, this means that we will be able to identify if the lexicographers are comfortable with this tags, at this tags described well enough, clearly enough, precisely enough. So, that the lexicographer has no problem as an a tag to a particular word and after that, when we submit this tag copula to a machine learning algorithm does the algorithm easily learn this tags with high accuracy, this is the ultimate test.

Let me, write down this two tests for the efficacy of a part of speech tag.

(Refer Slide Time: 06:21)



22/9/10 NLP. lect 15 1

POS TAG Efficacy

- Lexicographer/ Annotator ease
- Amenability to high accuracy machine learning Algorithm

Human Annotator and Computer

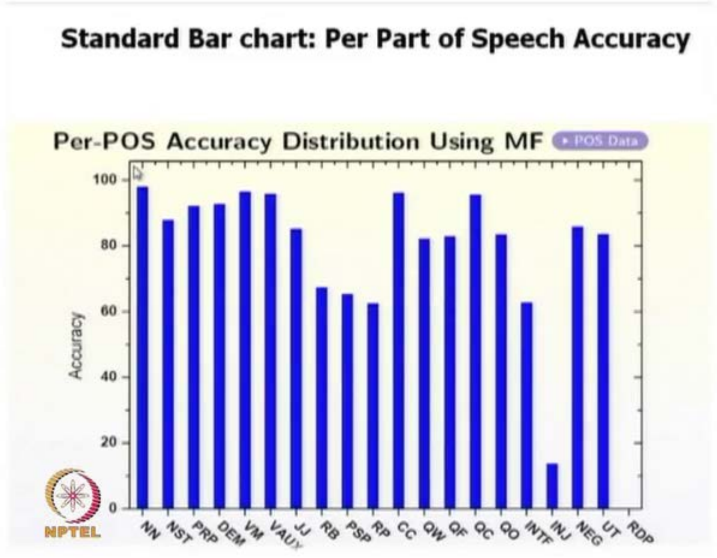
So, pos tag efficacy is a function of one lexicographer slash annotator is and amenability to high accuracy machine learning algorithm. So, we actually have for two entities here, human annotator and computer, so this will tell us how would our pos tag set is and when the pos tag set is, designed properly it serves both the sense let us annotator find it easy to annotate data with this tag sets. And high accuracy machine learning algorithms also can be designed.

(Refer Slide Time: 07:51)

Accuracy measurement in POS tagging

Let us move forward and we discuss now the accuracy measurement in part of speech tagging.

(Refer Slide Time: 07:58)



Standard Bar chart: Per Part of Speech Accuracy

First we show a bar chart, which is typically shown for part of speech tags systems. So, you can seen here, there are two access the accuracy is plotted on the y axis and part of speech tags are mention on the x axis. So, NN we have seen is the common noun tag, so for this particular system which is reporting accuracy, the common noun accuracy is close to 100 percent. And common nouns form a the largest chunk of the language data and therefore, high accuracy here is good news.

NST we have seen is a special kind of tag required for Indian languages where, there are words which can act as pos positions and also can act as nouns things like [FL] and so on there the accuracy is less than 90 percent. On PRP which is the pronoun the accuracy go crosses 90 percent, demonstrative accuracy is also around that close to that, main verb accuracy is high, main verb accuracy is close to common noun accuracy. And this should be Indian languages typically are strong on morphology and verb forms are quiet an ambiguously, shown by a markings on the verbs themselves.

So, it is possible for a morphology analyzer to send the information of verb and therefore, a part of speech tagger can make use of this information to accurately predict that a word is a verb or not. Verb auxiliary also has similar kind of accuracy, close to the common noun accuracy may be about to 96 percent are shown, 95 percent. Adjective accuracy is not, so high that is slightly lower than 90 percent and the reason of course, is it is neighbor which is RB or adverb, adverb accuracy you can see is quiet low less than 70 percent.

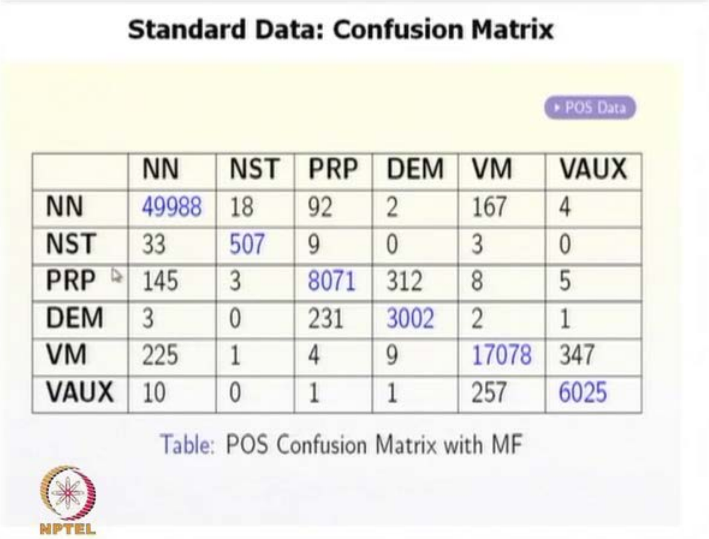
And the reason for this two categories low accuracy is because they can be quiet easily confused in Indian languages. So, ram [FL] here [FL] is adjective, ram [FL] here [FL] is adverb and it is possible that this adverb is far away from the verb and therefore, from the immediate vicinity of the word, it is not easy to disambiguate the category. PSP which are pos positions again have low accuracy, which is close to 70 percent and the reason should be that it is clashing with the NST's the words which can act as locative nouns, temporal locative or special locative nouns.

So, PSP and NST can easily confused and this reduces the accuracy of both RP's are particles, particles also have low accuracy and there are many particles like, negative and so on. And this accuracy is low which needs to be analyze, particles need not have low accuracy in general, the conjuncts which are like and or, but etcetera, corresponding to Hindi which is are [FL] than [FL]. So, these words are an ambiguous and they should be deducted with complete accuracy, one needs to investigate why there are not exactly 100 percent.

Quantified accuracy is again below 90 percent, which is also a matter to be investigated because quantifiers also are quiet unique and distinct for example, all each every [FL] this are quantifiers and they seem to have a low accuracy. This QW was, w h quantifier question quantifiers, than QF is again quantifier has similar accuracy has QW, QC is another kind of quantifier, QO is again another quantifier they have various kinds of accuracy. INJ is the interjection which has very low accuracy, very strangely and negative again has accuracy which is moderate close to 90 percent and so on UT is not used in the current tag set.

So, this is for some language and some of this tags are understandable they are in the tag set, which was discussed for Hindi and some are specific to that language and the accuracies are in the form of bars. So, from this diagram one can make out which category has, what kind of accuracy and the low accuracy tags are to be investigated further that is were one has to invest resources, ideas to improve the accuracy. So, that the accuracy of the overall system goes up. So, this is an extremely useful graph and extremely useful way of understanding, how the part of speech tagger is working and it is absolutely indispensible, for design of good quality part of speech taggers, one needs to have this kind of bar category accuracy report.

(Refer Slide Time: 14:12)



Standard Data: Confusion Matrix

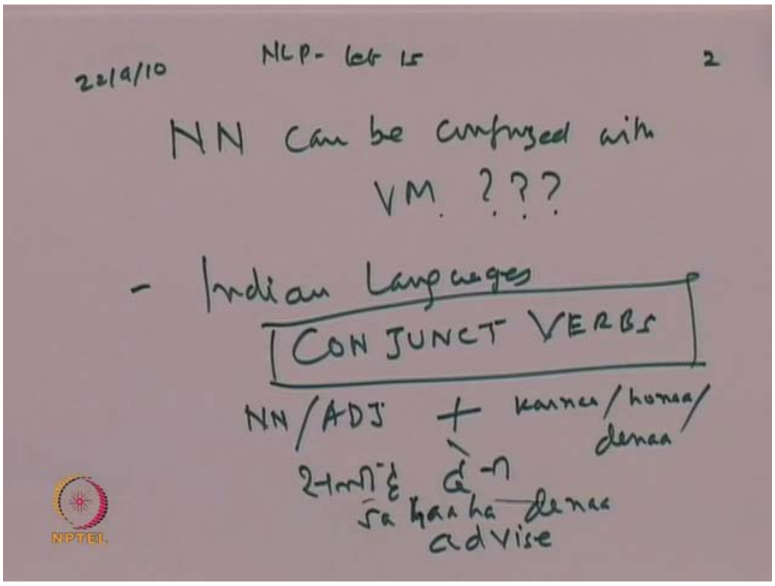
	NN	NST	PRP	DEM	VM	VAUX
NN	49988	18	92	2	167	4
NST	33	507	9	0	3	0
PRP	145	3	8071	312	8	5
DEM	3	0	231	3002	2	1
VM	225	1	4	9	17078	347
VAUX	10	0	1	1	257	6025

Table: POS Confusion Matrix with MF

Proceeding further, how did we come up with that kind of bar chart. So, here a particular structure called the confusion matrix is of great help, now what is done is that the part of speech tags are placed row wise, part of speech tags are also placed column wise the same part of speech tags. So, if I take a particular part of speech let us part of speech tag let say NN which is the common noun, we see the row of NN in the row of NN I find that for the column under NN, we have about 50000 entries.

NN has been confused with NST 18 times, with pronoun 92 times, then with demonstrative 2 times, main verb 167 times and auxiliary verb 4 times. One might wonder how can a noun and pronoun be confused, how can a noun and main verb be confused amongst to all, we have to remember that in Indian languages there is a phenomenon called conjunct verb formation let me write it down.

(Refer Slide Time: 15:42)



22/9/10 NLP- lect 15 2

NN can be confused with VM ???

- Indian Languages

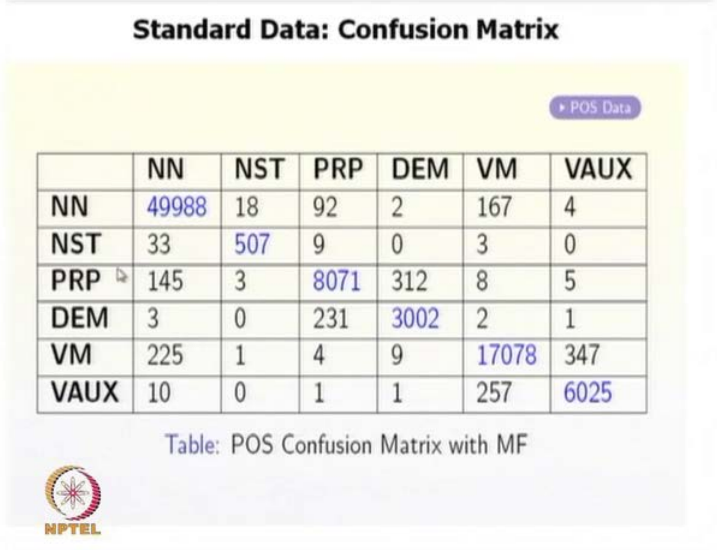
CONJUNCT VERBS

NN/ADJ + karnaa/honaa/denaa

सलाह देना
saalaha denaa
advise

So, we are wondering why NN can be confused with VM we wonder it. Now, how can a common noun we confused with a main verb, the reason is Indian languages have this phenomenon called conjunct verb. So, this is nothing but a noun or adjective plus a verb in the form of [FL] and so on. So, for example, [fl] for example, [FL] which means to advice, so in Hindi for the action of advising, we have a double word group which has [fl] and [fl] two parts in it. So, [FL] is the nominal part in this group this whole thing is a verb, though the first part is a noun. And the part of speech tagger should possibly tag the whole thing as a verb and this kind of problem can give rise to main verb noun confusion this is the explanation.

(Refer Slide Time: 17:11)



Standard Data: Confusion Matrix

▸ POS Data

	NN	NST	PRP	DEM	VM	VAUX
NN	49988	18	92	2	167	4
NST	33	507	9	0	3	0
PRP	145	3	8071	312	8	5
DEM	3	0	231	3002	2	1
VM	225	1	4	9	17078	347
VAUX	10	0	1	1	257	6025

Table: POS Confusion Matrix with MF

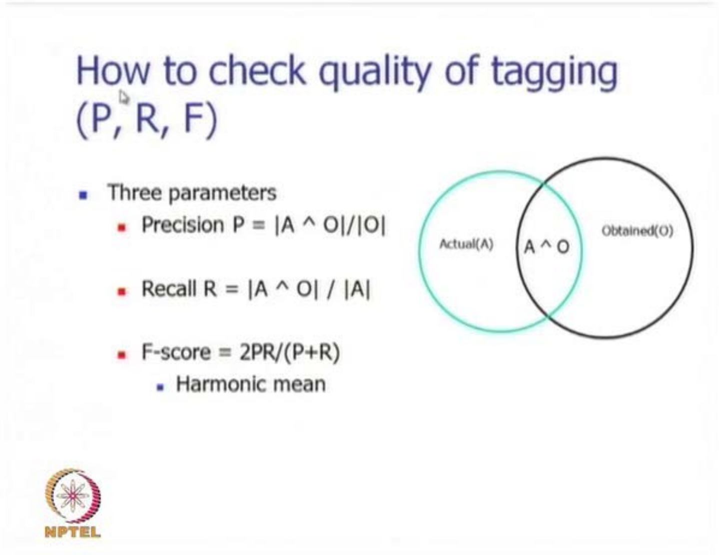
So, coming to the matrix once again the confusion matrix is a very important structure, which shows which part of speech tag has been confused, how many times with another category. So, noun we have analyzed then we find that NST has been confused it noun 33 times, this need not be symmetric NN has been confused with NST 18 times, but NST is confused with noun 33 times. So, this means that for NN we have placed NST wrongly 18 times.

And for NST we are placed NN 33 times wrongly. So, this need not be symmetric of course, as it is clear similarly if we take that them row them is a demonstrative, what does our expectation from the last few lectures, we understood that DEM and pronoun cannot be easily separated from each other. And that is bound out by the data here, we find that the DEM has been confused 231 times for pronoun and 231 times forms, a pretty significant percentage of the total number of demonstratives.

So, we have placed PRP for DEM 231 times and out of about 3000 such tags 200 times this error has occurred to which is about 6 percent of error more than 6 percent of error I would say. Similarly, the main verb has been confused with noun 225 times, so this is a matrix this is also known as the confusion matrix, it is an extremely useful data structure which tells us, which category has been confused, which other category how many times.

So, going back to the previous slide while this reports accuracy on each part of speech tag, this is the function of the bar chart, the confusion matrix gives more detail it shows how many times a tag has been confused with other tags. Now, it is not very difficult to see that the noun accuracy which was shown on the y axis of the bar chart, can be computed by dividing this number by the some of the numbers in this row. And we can see that this is close to 50000 other numbers are pretty small and that is why the noun accuracy is close to 100 percent. Now, if we take the DEM row then them has been arrow nastily marked in here, 3 times then 231 times as PRP, VM 2 times, VAUX 1 time and this kind of errors give rise to the them accuracy as 3002 divided by some of all this numbers. Now, one can understand how this numbers have been obtained and plotted on the bar chart, so confusion matrix very useful.

(Refer Slide Time: 20:30)



How to check quality of tagging (P, R, F)

- Three parameters
  - Precision P = |A ^ O|/|O|
  - Recall R = |A ^ O| / |A|
  - F-score = 2PR/(P+R)
    - Harmonic mean

We come to other parameters for accuracy checking, how to check the quality of tagging this is done through three parameters called, precision, recall and F score. Precision P recall or F score, which is a function of the precision and recall. Let us understand how this is computed, suppose we have this actual set of tags which should have happen. So, how do we visualize this, we visualize this these way that we have the goals standard data.

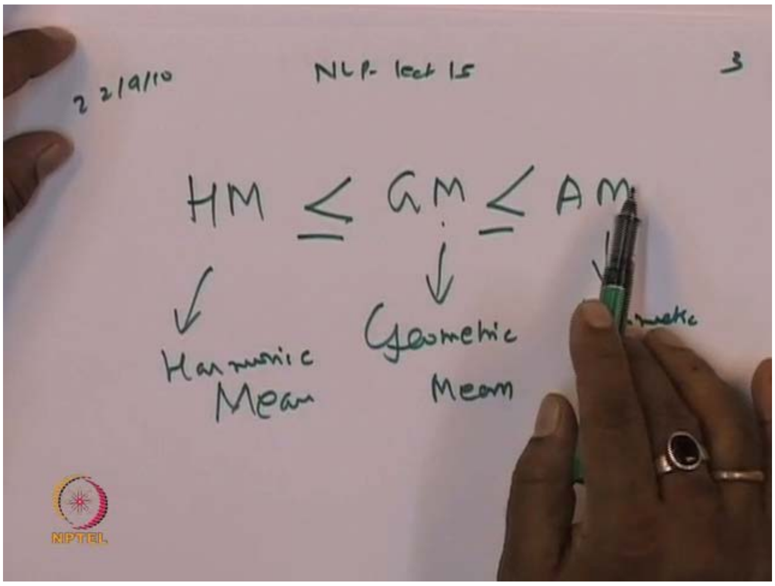
That means, we have language data the text which has been tagged by their actual tags and this data is available to us. So, we have a set of words with their corresponding tags, so these tags will form the actual set we call it A, now on this textual data we have got some tags by the part of speech tagger, they may or may not be same as the actual tag which is there in the goal standard data. However, we have this set of obtained tags which is O.

So, set of actual tags is A, set of obtain tags is O and then precision can be measured as A intersection O divided by the number of O. So, the number of times A and O agree divided by the number of O's, what is the meaning of this, the meaning of this is that precision measures, the accuracy as a percentage of what has been obtained. So, out of the things that have been obtained, cardinality of O how do the things which have been obtained, what percentage is correct that is A intersection O.

Recall, on the other hand is of the actually correct tags how many have been obtained which are correct. So, we already know that A intersection O is the agreement set between A and O this is the number of times the part of speech tagger has been correct, this is the proportion of how many things are correct, how many correct tags have been obtained from the actual correct set of tags this is recall.

So, precision measures how many are correct from whatever has been obtain, recall measures of the correct ones how many have been obtained and these two things can be combine to compute what is called the F score, F score is two into precision into recall divided by precision plus recall. So, this is nothing but the harmonic mean now the question that may arise is why F score is defined as a harmonic mean. So, this requires a bit of writing.

(Refer Slide Time: 23:57)



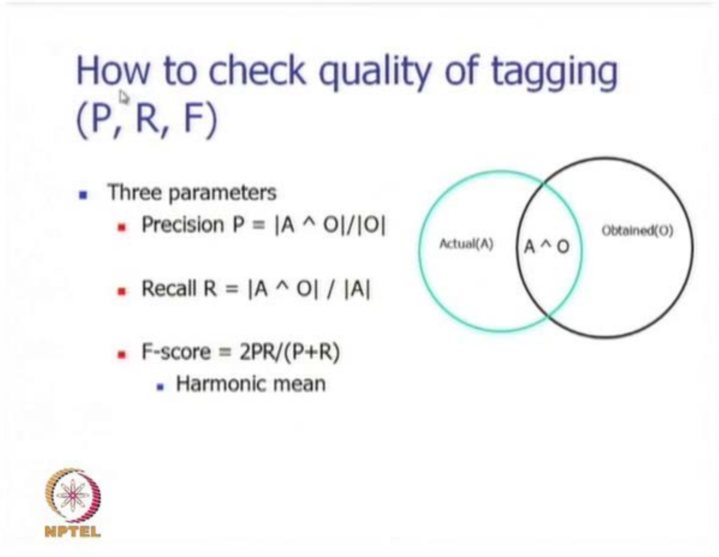
2 2/9/10 NLP- lect 15 3

$$HM \leq GM \leq AM$$

Harmonic Mean
Geometric Mean
...metic

So, we know that harmonic mean is less than or equal to geometric mean is less than or equal to arithmetic mean. So, this is harmonic mean, this is geometric mean and this is arithmetic mean right and this relationship holds, harmonic mean is the smallest of the three quantities, arithmetic mean, geometric mean, harmonic mean. So, it is clear that if we maximize the harmonic mean automatically the geometric mean and arithmetic mean also will go up, ensuring increase in harmonic mean, ensures increase in geometric mean and arithmetic mean also automatically.

(Refer Slide Time: 24:55)

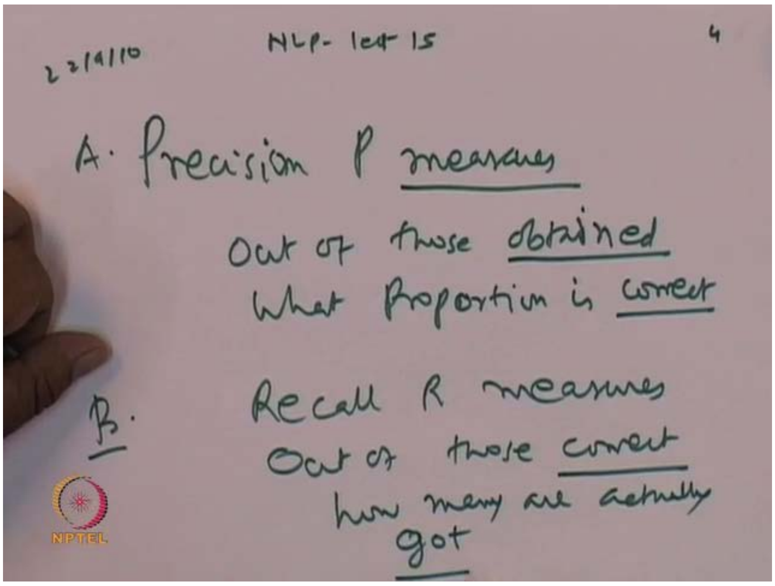


How to check quality of tagging (P, R, F)

- Three parameters
  - Precision P = |A ^ O|/|O|
  - Recall R = |A ^ O| / |A|
  - F-score = 2PR/(P+R)
    - Harmonic mean

So, if we look at the slide again when we improve the F score, we improve the arithmetic mean of both precision and recall the average of precision and recall, we also improve the geometric mean of precision and recall. So, this is a good idea, now let me just write down the intuitive meaning of precision and recall.

(Refer Slide Time: 25:21)



22/9/10 NLP- lect 15 4

A. Precision P measures

Out of those obtained

What proportion is correct

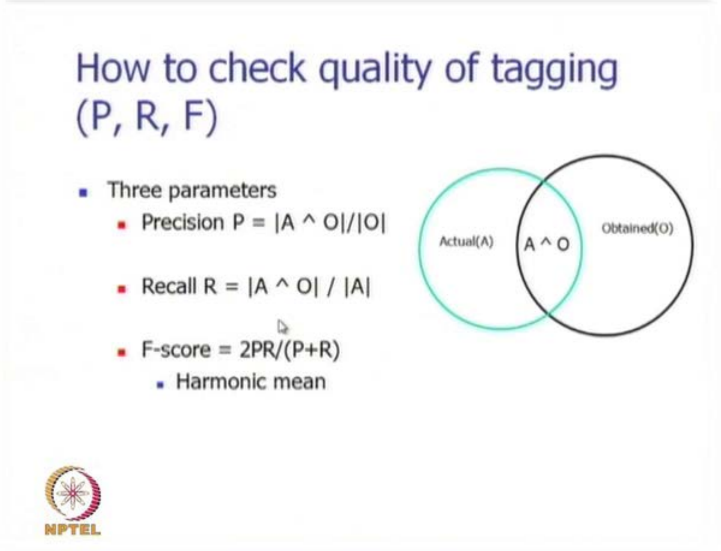
B. Recall R measures

Out of those correct

how many are actually got

Precision P this measures out of those obtained what proportion is correct, out of those obtained of what proportion is correct this is precision. And recall or measures out of those correct how many are actually we got.

(Refer Slide Time: 26:04)

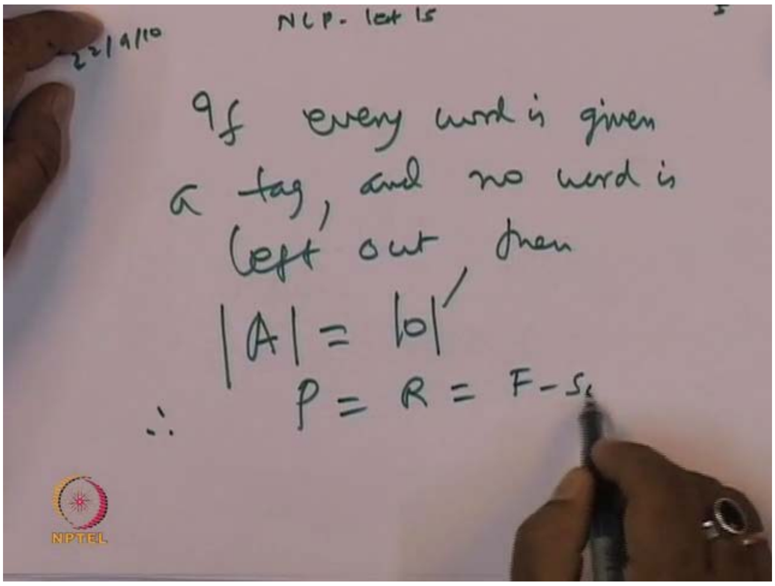


How to check quality of tagging (P, R, F)

- Three parameters
  - Precision P = |A ^ O|/|O|
  - Recall R = |A ^ O| / |A|
  - F-score = 2PR/(P+R)
    - Harmonic mean

So, this the measure look at slide again this intersection area is the agreement between A and O. So, this intersection divided by this whole circle gives out of those obtained what proportion is correct and this intersectional area divided by this green circle, measures what proportion of correct things have been obtained. So, I spent some time on precision and recall and F score because this is a very frequently used measure everywhere. Now, we note an interesting thing and we find that the precision recall and F score they are different under certain condition and they are saying in other conditions.

(Refer Slide Time: 27:00)



22/9/10 NLP- lect 15

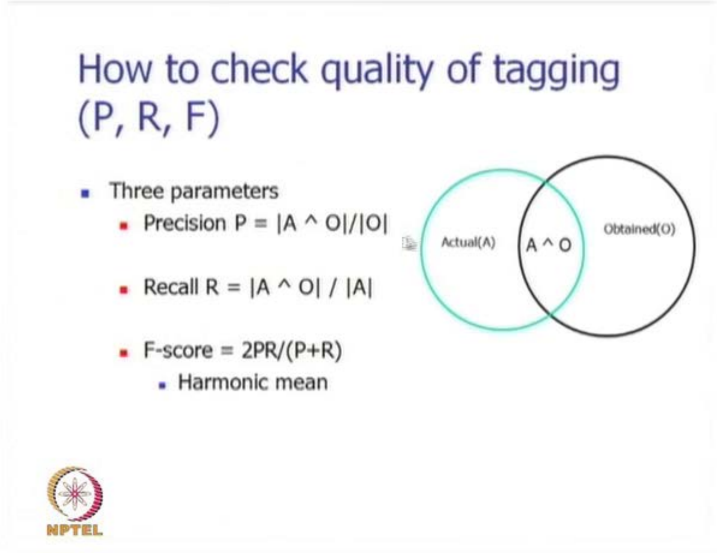
If every word is given a tag, and no word is left out, then

$|A| = |O|$

$\therefore$ $P = R = F\text{-}Sc$

So, if every word is given a tag and no word is left out, if every word is given a tag and no word is left out. Then we know that, the size of A is equal to size of O therefore, precision equal to recall is equal to F score.

(Refer Slide Time: 27:35)



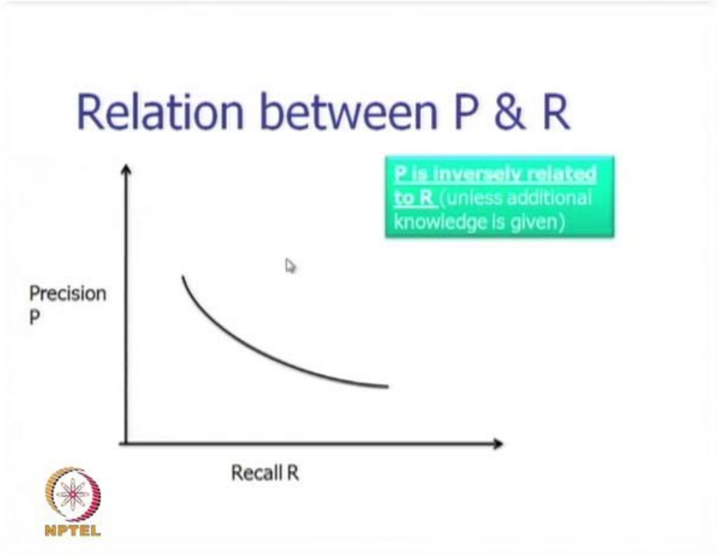
How to check quality of tagging (P, R, F)

- Three parameters
  - Precision P = |A ^ O|/|O|
  - Recall R = |A ^ O| / |A|
  - F-score = 2PR/(P+R)
    - Harmonic mean

Actual(A)
A ^ O
Obtained(O)

Why, so if you look at the slide again the reason is that the numerator A intersection O is same for both precision and recall, only the denominator is different. But, if we a place a tag for very word then this cardinality of O and cardinality of A become same, only when the part of speech tag are misses out some words because let say it is un known it is not there in the dictionary then we may have different values of precision and recall. And there precision recall and F score may be different from each other, but in general when we obtain a tag for every word, these quantities are same.

(Refer Slide Time: 28:19)

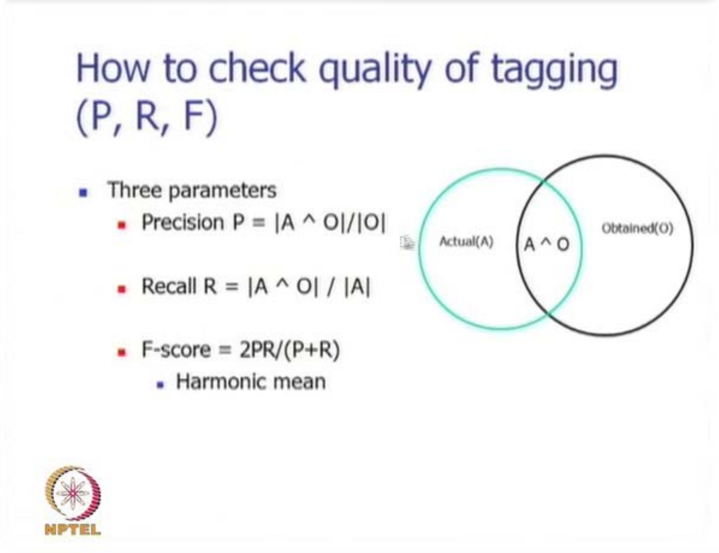


Relation between P & R

P is inversely related to R (unless additional knowledge is given)

We proceed further on the slides and interesting thing is the relation between precision and recall, typically precision is inversely related to recall why is it, so. Let us see, the graph here recall is being plotted on the x axis, precision on the y axis, we see that as recall improves precision tense to come down.

(Refer Slide Time: 28:48)



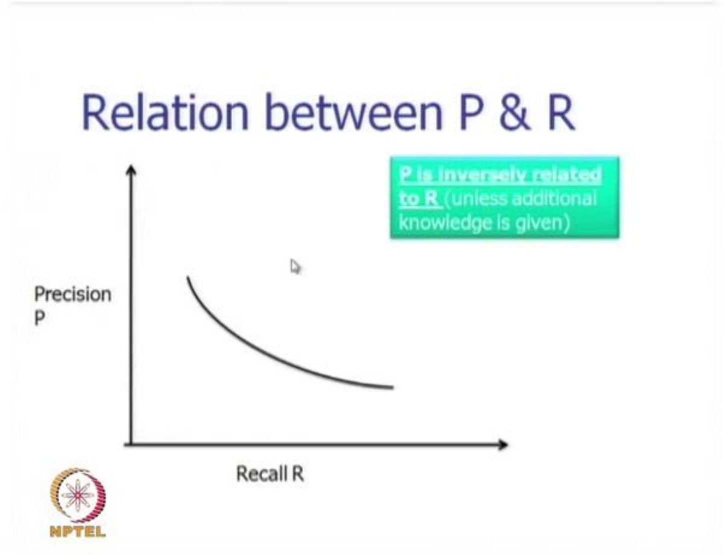
How to check quality of tagging (P, R, F)

- Three parameters
  - Precision P = |A ^ O|/|O|
  - Recall R = |A ^ O| / |A|
  - F-score = 2PR/(P+R)
    - Harmonic mean

This can be understood by going back to the previous slide, where the precision and recall have been defined in terms of the intersection area. So, what happens is that to increase the precision we may do the following, one of the two one is that we can increase the numerator. That means, increase the overlap between the actual tag and the obtain tag or we may decrease the size of the denominator, the size of O. Now, when we decrease the size of O, the precision can go up, but actually we are obtaining less and less tags.

We are possibly missing out on words not tagging them and that has an adverse effect on recall, if I want to improve recall then we again may increase the numerator and decrease the denominator. Now, decrease the denominator is not really possible because the actual set of tag towards are already given, we do not have much control on this, but when we increase this value, when we increase this try to increase this overlap, then we find that we are sacrificing on the precision. So, we can have more detail discussion on recall and precision later.

(Refer Slide Time: 30:27)

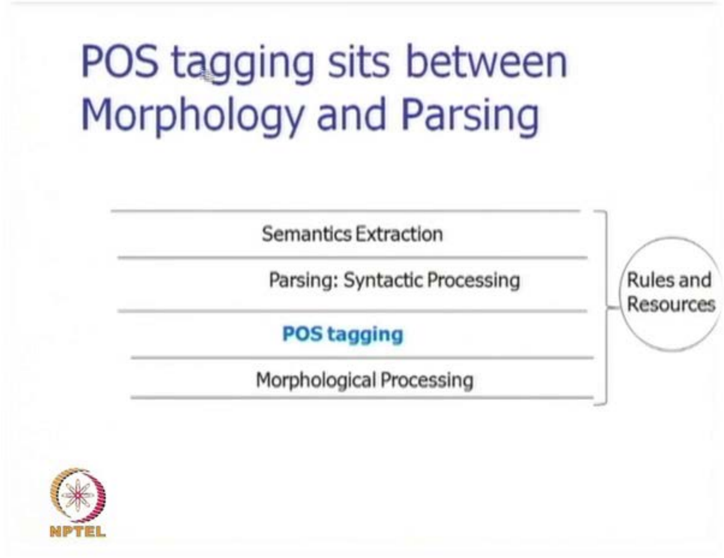


Relation between P & R

P is inversely related to R (unless additional knowledge is given)

So, in general the behavior is this that recall increase, typically entails precision fall. Now, this situation is quite typical and one can come out of this situation, only by injecting knowledge when this point is a deeper issue which we can understand later.

(Refer Slide Time: 30:46)



POS tagging sits between Morphology and Parsing

Semantics Extraction

Parsing: Syntactic Processing

POS tagging

Morphological Processing

Rules and Resources

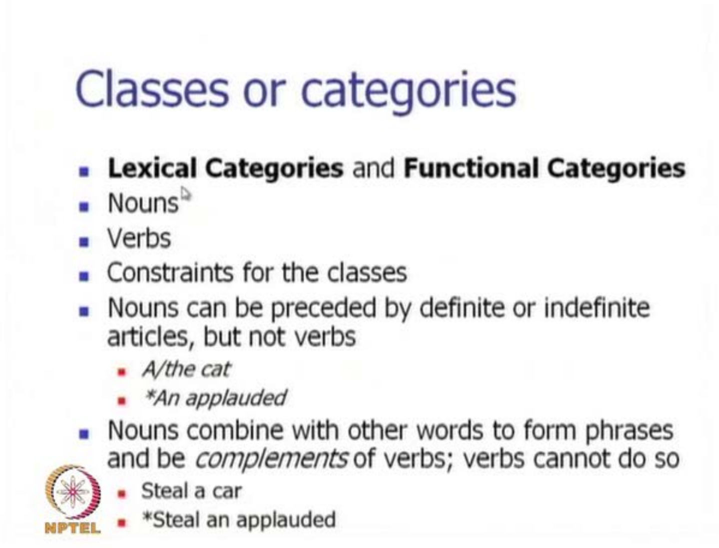
Now, we move on to the tags onwards and we remind our self that part of speech tagging is a layer, which sits between syntactic processing and morphological processing is a very useful diagram a classical diagram of natural language processing, where we say that the processing of language happens in layers. Now, each of this layer makes is a rules and resource.

(Refer Slide Time: 31:17)

Words and their categories

We move on to understanding words and their categories because we have been discussing part of speech tag, so long, but what is the linguistic foundation for this part of speech tags do they really exist does the language entail the existence of this tags. So, we discuss words and their categories.

(Refer Slide Time: 31:39)



Classes or categories

- **Lexical Categories** and **Functional Categories**
- Nouns
- Verbs
- Constraints for the classes
- Nouns can be preceded by definite or indefinite articles, but not verbs
  - *A/the cat*
  - **An applauded*
- Nouns combine with other words to form phrases and be *complements* of verbs; verbs cannot do so
  - Steal a car
  - *Steal an applauded

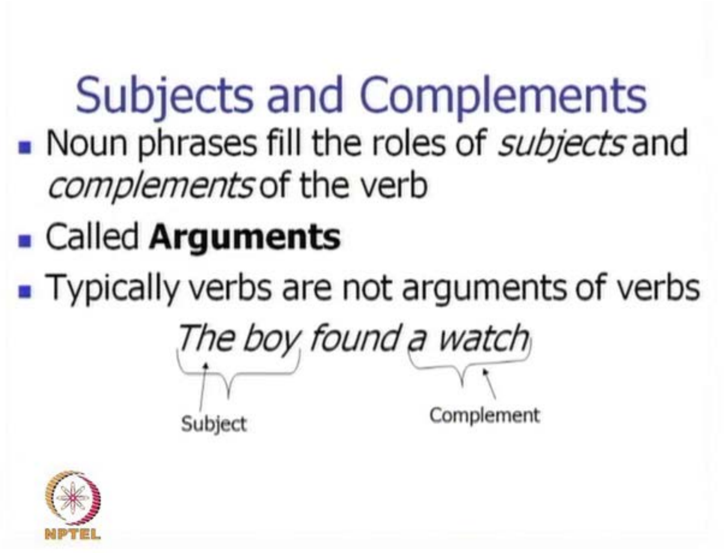
And the first discussion point is the categories themselves. We have this classical conflict in part of speech tagging, between lexical categories and functional categories. Lexical category means, the most frequent are most accepted category of a word which is a record in the dictionaries, so for example, if we have the word quick for example, quick we know is an adjective. And typically people's perception about quick is an adjective, it qualifies a noun many times for example, a quick dog a quick finish and so on.

But, quick and also have adverbial category sometimes not be frequently many times quite really it can act as adverb for example, in a colloquial setting you may say come here quick. So, come here quick, this quick is a qualifier for the action coming, so there it has adverbial function. So, quick has as lexical categories adjective this is what will be recorded in the dictionary, how ever in the sentence come here quick, quick has a functional role which is adverb. So, this is the functional categorization of quick.

So, this is an important issue we have to understand in the sentence what is the tag of the word, and in the dictionary what is the tag of the word. So, now, we have this part of speech tags part of speech categories which are nouns, verbs this are very large categories. Now, there are constraints for this classes how do we know that a word is a noun, what kind of test exists for this. So, here there are some simple observation, nouns can be preceded by definite or indefinite articles, but not verbs.

So, for example, verbs cannot be preceded by an article typically unless in poetry let us say. So, A or the cat is fine, but an applauded, applauded is a verb in past tense verb and before that would be ungrammatical, additionally nouns combine with other words to found phrases and they can be complements of verbs and verbs cannot do, so. So, what we are discussing is, what kind of constraints exist on the waivers of nouns and verbs, so that they can be distinguished. Now, steal is a verb, steal a car now a car is the object of steal, we also call it the compliment of the verb. Now, a verb cannot be the compliment of a verb without some changes in the morphology of the verb itself, so in general to let us keep things simple and say that nouns, typically form complements of other words. So, steal a car, but steal an applauded is wrong.

(Refer Slide Time: 35:14)



Subjects and Complements

- Noun phrases fill the roles of *subjects* and *complements* of the verb
- Called **Arguments**
- Typically verbs are not arguments of verbs

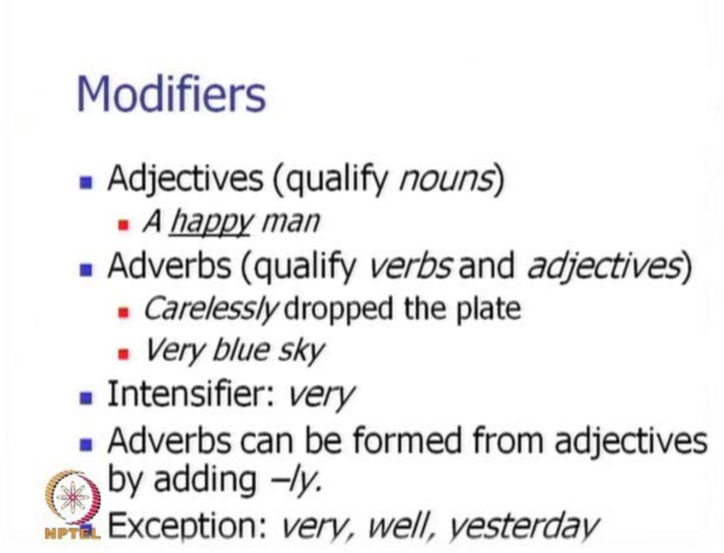
The boy found a watch

Subject

Complement

Now, at this point it is useful to distinguish between subjects and complements, in a sentence there are subjects and complements of a verb, a verb can be looked up on I was having subject role and complement role, which have to be fill with nouns. So, these are also known as arguments, typically verbs are not arguments of verbs it is nouns which form the arguments. Here, in this sentence the boy found a watch, a watch, watch is a noun preceded by a article. So, this is a noun phrase the boy, boy is the noun preceded by determiner the boy is a noun phrase. So, found has this subject as lot which needs to be filled, the boy fills it found has a compliment slot which needs to be filled a watch fills it.

(Refer Slide Time: 36:08)



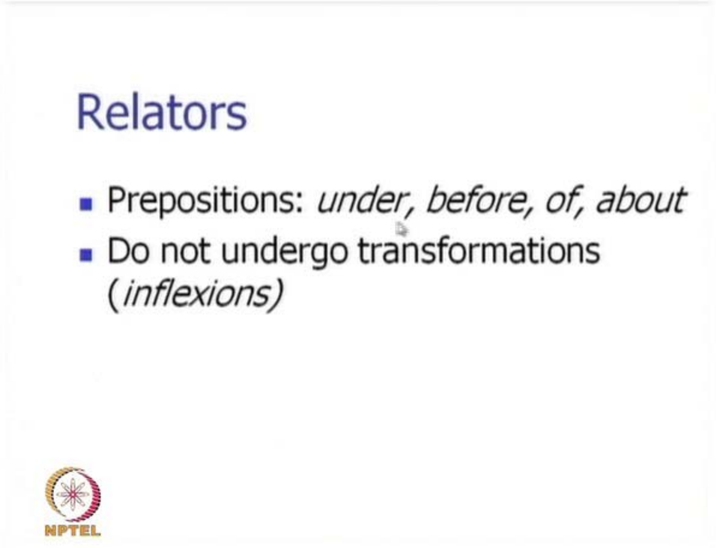
Modifiers

- Adjectives (qualify *nouns*)
  - *A happy man*
- Adverbs (qualify *verbs* and *adjectives*)
  - *Carelessly* dropped the plate
  - *Very blue sky*
- Intensifier: *very*
- Adverbs can be formed from adjectives by adding *–ly*.
- Exception: *very, well, yesterday*

Now, what did we are learn then nouns can combined with determiners, articles, etcetera to form a phrase. Nouns can be the arguments of verbs and that these distinguish nouns from the category called verbs, nouns and verbs both are large categories, verbs typically do not be arguments of other verbs, nouns are the arguments. And nouns are typically preceded by determiners, articles to form phrases. The next category is the set of modifiers, under that we have adjectives which qualify nouns a happy man is an adjective, happy is the adjective this is the example, adverbs they qualify verbs and adjectives.

So, he carelessly drop the plate carelessly is the adverb, very blue sky very is adverb because it is qualifying blue, very is also called an intensifier. In English adverbs can be formed from adjectives by adding ly and exception to this rule is a very well yesterday, yesterday is a temporal adverb, adverb of time yesterday he came, well is an adverb of manner he plays well and very is an intensifier adverb, which can go before an adjective.

(Refer Slide Time: 37:43)

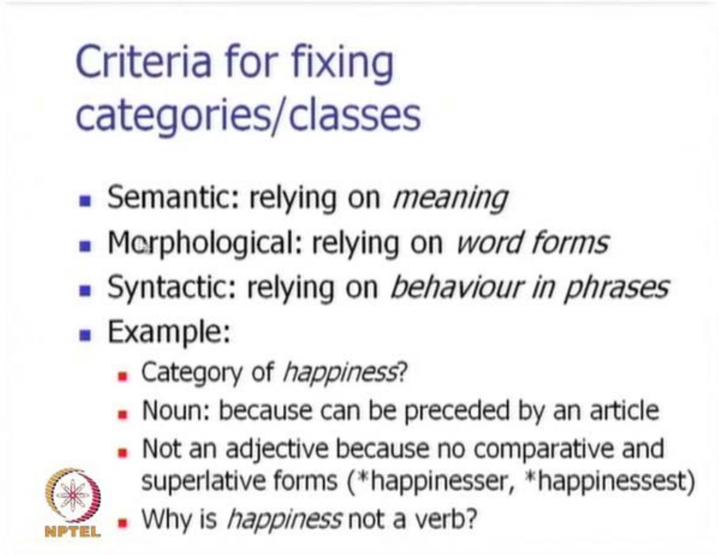


Relators

- Prepositions: *under, before, of, about*
- Do not undergo transformations (*inflexions)*

Then there are relaters which join entities in a sentence. So, we have prepositions like under before of about, so dust under the carpet, horse before the cart, capital of India a man above town, a story about Krishna. So, these are prepositions they do not undergo transformations in English, they do not take what is called inflexions.

(Refer Slide Time: 38:18)



Criteria for fixing categories/classes

- Semantic: relying on *meaning*
- Morphological: relying on *word forms*
- Syntactic: relying on *behaviour in phrases*
- Example:
  - Category of *happiness*?
  - Noun: because can be preceded by an article
  - Not an adjective because no comparative and superlative forms (*happinesser, *happinessest)
  - Why is *happiness* not a verb?

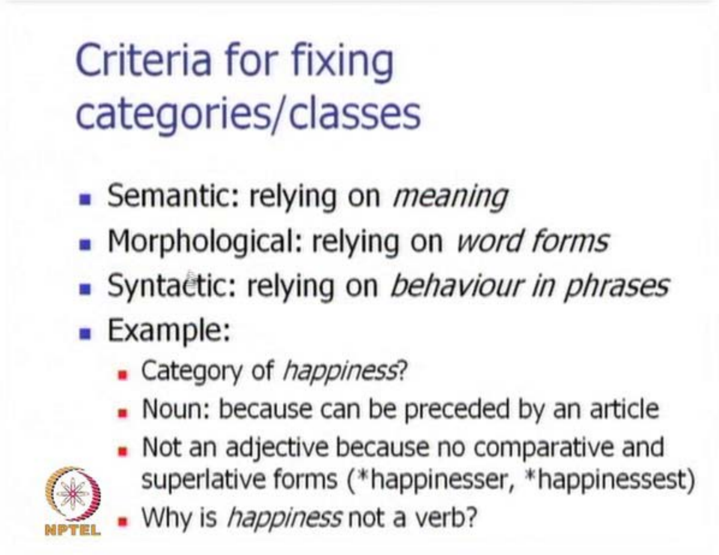
Now, we discuss the criteria for fixing the categories or classes, first important clue which comes for fixing the category is semantic, then morphological categories which realize on the word form and syntactic which realize on the behavior in phrases or structural behavior. So, this is an important point in all natural language processing, whenever we carry out a test for some phenomenon or some behavior, the test can be either semantic which requires deep meaning or syntactic which is a structural consideration or morphological which is based on the property of the word.

So, this is almost like a running theme across natural language processing, whenever we design algorithms for natural language processing, we and we test something any algorithm for natural language processing, which tests something. Is either semantic in nature or syntactic or morphological depending on whether it is relaying on meaning or on the structure or on the property of the word.

So, it is important to understand that semantic test is extremely reliable, this is extremely robust once you get at the meaning there is no doubt left, we have the decision very clearly fixed, but semantic test is also very hard to implement computationally, this is difficult to design semantic tests, syntactic tests are easier they relay on the structure the way the words relate to each other to form larger phrases.

So, this requires parsing and this is less complex then a semantic test, lower than the syntactic test is the morphological test, which realize on the word it is form and its properties and this is the simplest. So, algorithms typically attempt to remain at the level of morphological processing or word level processing because the moment it is tries to ascend to the level of syntax or semantics, then there are deep waters there are complex things to be handled.

(Refer Slide Time: 40:54)



Criteria for fixing categories/classes

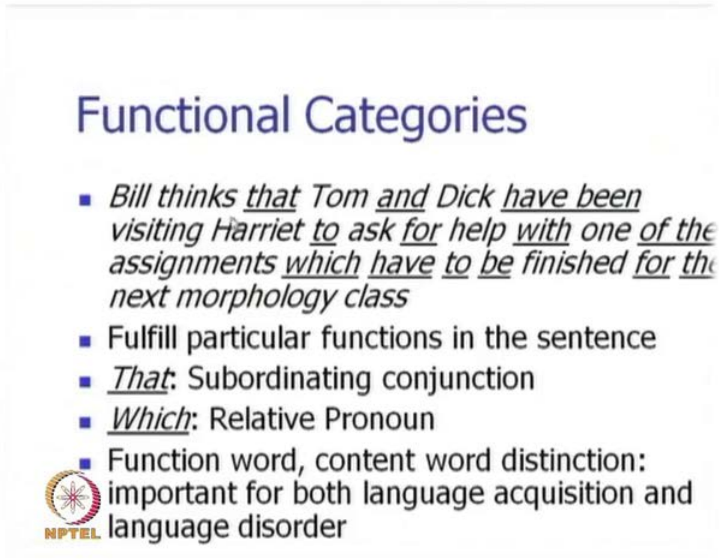
- Semantic: relying on *meaning*
- Morphological: relying on *word forms*
- Syntactic: relying on *behaviour in phrases*
- Example:
  - Category of *happiness*?
  - Noun: because can be preceded by an article
  - Not an adjective because no comparative and superlative forms (*happinesser, *happinessest)
  - Why is *happiness* not a verb?

So, we remark here semantic realize meaning, morphological realize on word forms, syntactic realize in where in many phrases for example. Let us take the word happiness and we ask, what is the categories of happiness, happiness we say is noun because it can be preceded by an article. So, this is a syntactic observations syntactic test, so a noun when it participate in a noun phrase, it can be typically preceded by an article, happiness is not an adjective because it does not take any comparative and superlative forms.

So, for example, we cannot say happinesser or happinessest, like for quick we can say quicker and quickest, but not, so for happiness. There is not comparative and superlative form for happiness, nor can we make such forms by introducing more and most prefer them, more happiness, most happiness that will make it a noun phrase, it is not degree in the sense of an adjective. So, this shows why this is not an adjective, this is from purely morphological consideration and the fact it is a noun is given by the fact that it is preceded by an article it can preceded by an article.

Why is happiness not a verb. So, this we leave as an exercise for the participant or the student, please think about why happiness cannot be a verb. So, see if you can make use of simple morphological test or we have to go to the level of syntax and semantics to decide, why happiness is not a verb. One simple thing I can say is that, verbs typically have a different form depending on the tense, past, present and future here, happiness will not change it is form depending on whether it happen in the past or will happened in the future or will happening in the present, which is a morphological consideration.

(Refer Slide Time: 43:06)



Functional Categories

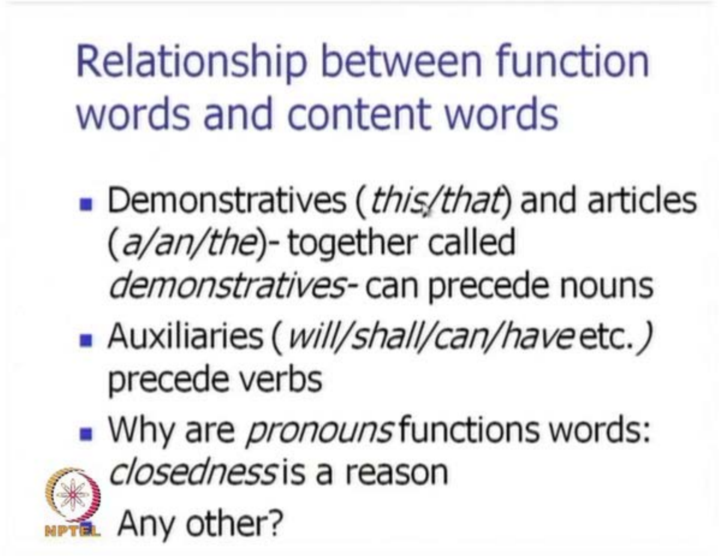
- *Bill thinks that Tom and Dick have been visiting Harriet to ask for help with one of the assignments which have to be finished for th next morphology class*
- Fulfill particular functions in the sentence
- *That*: Subordinating conjunction
- *Which*: Relative Pronoun
- Function word, content word distinction: important for both language acquisition and language disorder

Then there are this functional categories here is a long sentence, where which is full of functional category words. Bill thinks that tom and dick have been visiting Harriet to ask for help with one of the assignments which have to be finished for the next morphology class. This is a sentence from one of the celebrating linguistic text books and this is a long sentence, which has been obtained by gluing together, phrases and even sentence parts and this gluing has been achieve through, the functional categories.

The functional categories are shown here, with underline that and to for with of the which to have be for the these are underline words and their functional categories words they full fill particular functions in the sentence. So, that is a subordinating conjunction when you have that, this that can be either a relative pronoun the boy that lives in Delhi it is a relative pronoun or that introduces as a new sentence. I say it that I will go I will go is a new sentence, this is the subordinate class and that introduce a that class, which is a relative pronoun, the city which was devastated in the verb, here which is a relative pronoun. So, we have different categories of words, functions words and content words this is important for both language acquisition and language disorder, typically we find that when the brain is damaged. So, that the language faculty is also badly hampered, we typically find that the patient does not deal effectively with function words though content words are not affected, so easily.

Content words carry the load of information, content words carry the load of information and they have to be really robust and very error free, the function words play the role of putting together sentence parts and phrases. And even if they are wrong fairly accurate amount of meaning can still be transferred, from the speaker to the listener or from the writer to the reader.

(Refer Slide Time: 45:48)



Relationship between function words and content words

- Demonstratives (*this/that*) and articles (*a/an/the*)- together called *demonstratives*- can precede nouns
- Auxiliaries (*will/shall/can/have* etc.*)* precede verbs
- Why are *pronouns* functions words: *closedness* is a reason

Any other?

So, content words do play the most important role, the relationship between function words and content words is an important one, there are demonstratives these that an articles a, an, the. So, they precede nouns demonstratives precede the noun, these that denotes a particular noun, a, an, the they also denote noun the makes the noun a definite one, in a discourse scenario. Auxiliaries can precede verbs, we are talking about English language then we have pronouns these also looked up on a function words, function words and pronouns are function words because they are pretty close categories.

That means, we do not add new pronouns everyday, it is only after centuries or 200 years, 300 years that pronouns get added to a language repository. So, in old English for example, there was this pronoun call thou which got deleted later in modern English. So, closeness there’s fact that the set forms are close cultic class is possibly one of the reason for calling pronouns has functions words, they also function has reference for nouns, they do not carry the information themselves, but they refer to noun which carries information any way. So, we will discuss this, what categories in the next lecture.