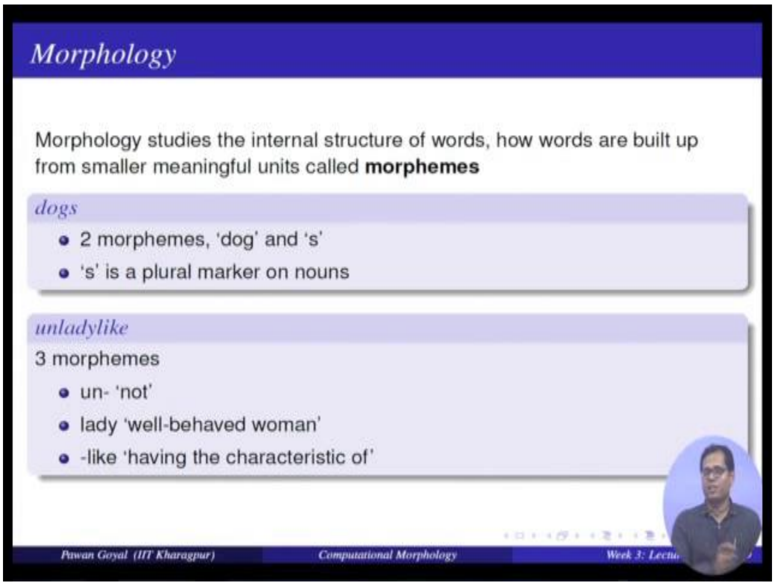
Natural Language Processing
Prof. Pawan Goyal
Department of Computer Science and Engineering
Indian Institute of Technology Karagpur

Lecture – 13
Computational Morphology

Hello everyone, so welcome back for the second module of third week. So, in the last week in the last module we were discussing about language modeling that is how we can effectively use the word ordering information for where is applications. We will go to the next topic, we will talk about the morphology information that what are the different modules that help us captures information and what are different linguistic terms that are involved. So, today we will start with the linguistic terms and then we will end with what are possible modules that can help us capture this in an automatic manner.

(Refer Slide Time: 01:00)



Morphology

Morphology studies the internal structure of words, how words are built up from smaller meaningful units called **morphemes**

dogs
- 2 morphemes, 'dog' and 's'
- 's' is a plural marker on nouns

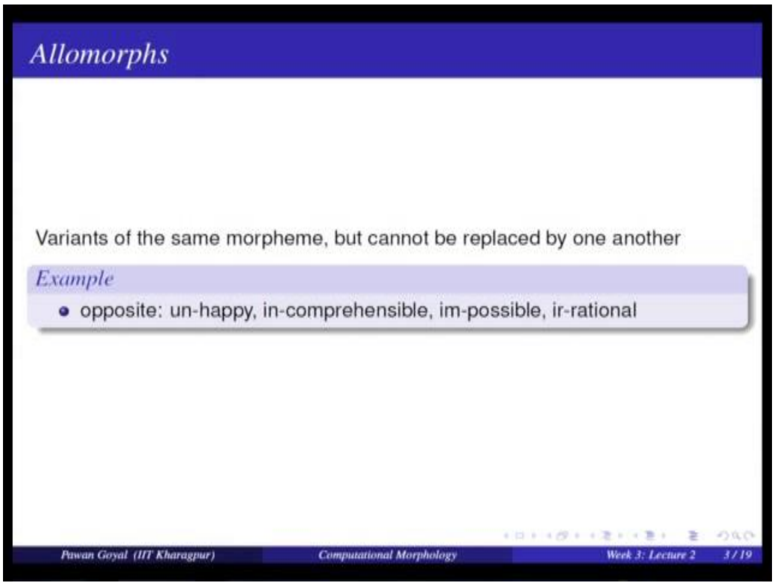
unladylike
3 morphemes
- un- 'not'
- lady 'well-behaved woman'
- -like 'having the characteristic of'

Pawan Goyal (IIT Kharagpur) Computational Morphology Week 3: Lectu

So, we are talking about computation morphology today, so what is morphology? In morphology we study what is internal structure of words. So, that is given a particular word in a language, what are the different meaningful units it is made up of and each small unit is called a morphemes. So, let us take some examples, so if I take the word dogs, so this is made up of two different units; one is dog the actual root word and another is affix that is applied to the word to make it a plural that is s. So, there are two morphemes in this single word dogs.

Similarly if I take a word like unladylike. So, we are having three different units here, so these are three different morphemes. So un; that corresponds to not, a negation or opposition, then lady well behaved women like that is having the characteristic of. So, there are three different morphemes that together constitute this single word.

(Refer Slide Time: 02:00)



Allomorphs

Variants of the same morpheme, but cannot be replaced by one another

Example

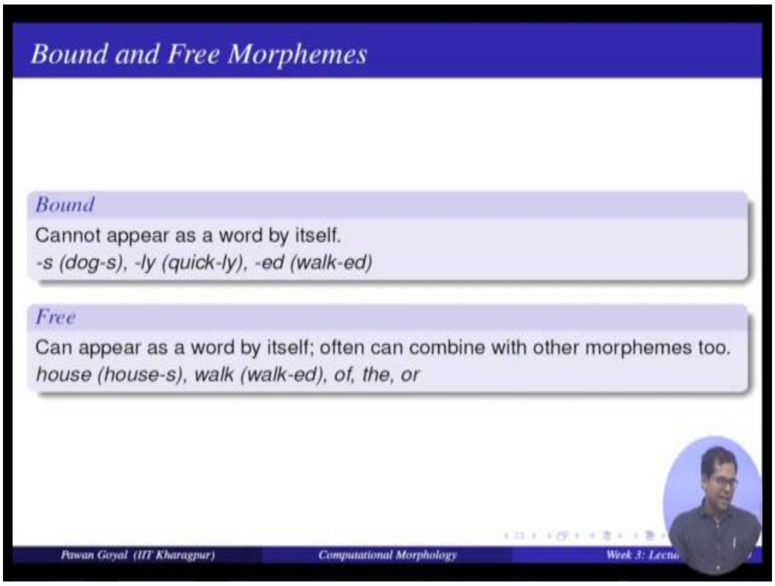
- opposite: un-happy, in-comprehensible, im-possible, ir-rational

Pawan Goyal (IIT Kharagpur) Computational Morphology Week 3: Lecture 2 3 / 19

So, now we will also see what are the various linguistic notions that you might want to be aware of when we are talking about morphology, so for example, what are allomorphs? So given a word like happy, if you have to convert it into its opposites; so we are making a word like unhappy, but if I take another word like rational, I will use i r; so I will make irrational. So, if they are different morphemes that can be used for the same purpose they are called allomorphs, so here is an example.

So allomorphs; we are saying variants of the same morphemes, but in general we cannot replace by one another. So, let us see here, so if I have the word happy; I can use u n, to make it unhappy, comprehensible; I can use i n, incomprehensible, if possible I can use i m; impossible and rational I will say irrational, but you can see that we cannot replace I r by u n here. So, I cannot say unrational; that is not a valid English word, so there are various reasoning for why a particular morpheme is used one of these is could be because of the phonemes that are there in the particular word. So, irrational you have the irrational with happy, we have unhappy, so for different words you have different sort of morphemes that you apply to make opposites. So, that is for one particular function, but for different functions like tenses and person and all that, you again have various categories of morphemes that are applied.

(Refer Slide Time: 03:36)



Bound and Free Morphemes

Bound

Cannot appear as a word by itself.
-s (dog-s), -ly (quick-ly), -ed (walk-ed)

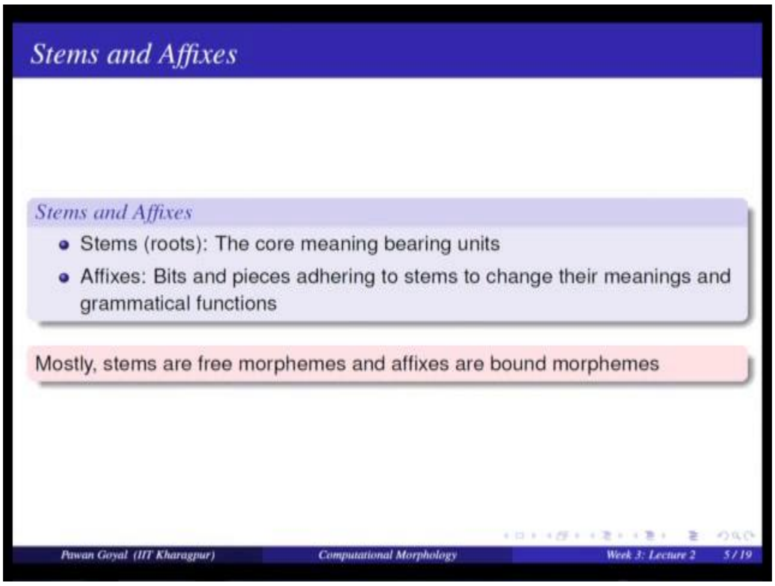
Free

Can appear as a word by itself; often can combine with other morphemes too.
house (house-s), walk (walk-ed), of, the, or

So, again when we talk about morphemes there is a distinction called bound morphemes verses free morphemes. So what are those? So bound morphemes are those morphemes that cannot appear as a word by itself. So, for example take the previous example unhappy; so I cannot take the morpheme un here and use it as a unique single word itself. So, this is bound by another word that is applied with this say unhappy, so it is bound with happy or some other words with this you can be use it.

On the other hand, if I take the word happy it is free morphemes, it can use on it is soon that is how we divide the morphemes into these two categories bound verses free. So, here are some examples, so bound morphemes are like s; so it can be used for making plural. So, with dog we have dogs; l y, so quick to quickly and e d; walk to walked. So, these are all bound morphemes and then there are free morphemes when they can appear as a word by themselves and they can also combine with other morphemes if needed. So, for example, if I take the morphemes like house, it can appear as a word by itself, but can also be combine with other morphemes like s to make plural houses. Similarly walk; it can combined with e d to make a past tense of walked and so on. Some words like of, in etcetera they are also free morphemes, they can appear on their own.

(Refer Slide Time: 05:05)



Stems and Affixes

Stems and Affixes

- Stems (roots): The core meaning bearing units
- Affixes: Bits and pieces adhering to stems to change their meanings and grammatical functions

Mostly, stems are free morphemes and affixes are bound morphemes

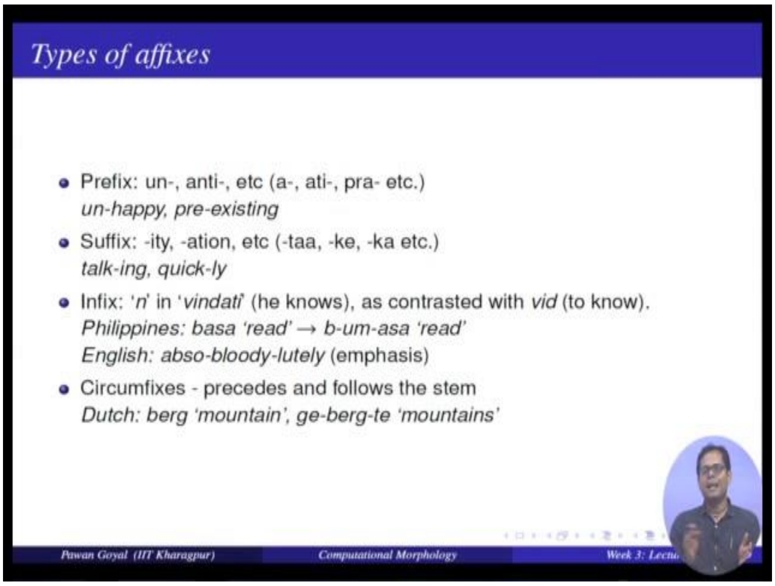
Pawan Goyal (IIT Kharagpur) Computational Morphology Week 3: Lecture 2 5 / 19

So, then there is another distinction and this is very very important when we talk about morphology. So in general when I talk about a word, it can have two main parts; one is the stem the main the root word and another is the affix and it can have more than one of affixes as well, so what are stem and affixes. So, when we talk about a word, so a stem is the core meaning bearing part. So, I take a word like boys, so the stem here will be boy. So, that is the meaning can bearing part and then there are various bits and pieces that can be applied to it that can ultra certain grammatical functions.

So, I can apply and as at the end of to make a plural of this, so these are called affixes. So, in general you can make a correspondence between stems and affixes and bound and free morphemes. So, you can say that the stems are kind of free morphemes and the affixes are kind of bound morphemes, so this kind of correspondence you can make.

So, now what are the different types of affixes that you can apply with the given stem? So we just discussed one simple example of applying s. So, this s apply after the word boy are make a plural with boys; so this is called suffix that is applied at the end of the word, but in general what are the different types of affixes that can be applied.

(Refer Slide Time: 06:40)



Types of affixes

- Prefix: un-, anti-, etc (a-, ati-, pra- etc.)
  un-happy, pre-existing
- Suffix: -ity, -ation, etc (-taa, -ke, -ka etc.)
  talk-ing, quick-ly
- Infix: '*n*' in '*vindati*' (he knows), as contrasted with *vid* (to know).
  Philippines: basa 'read' → b-um-asa 'read'
  English: abso-bloody-lutely (emphasis)
- Circumfixes - precedes and follows the stem
  Dutch: berg 'mountain', ge-berg-te 'mountains'

Pawan Goyal (IIT Kharagpur) Computational Morphology Week 3: Lectu

So, we start with prefixes, so prefixes are applied before the word. So, some example in English are un, so we make unhappy, anti; antinational, pre; preexisting and so on. We have these prefixes in other languages also for example, in Hindi and Sanskrit; we have the prefixes like [FL] that we can apply with various words or verbs, so these are of prefixes.

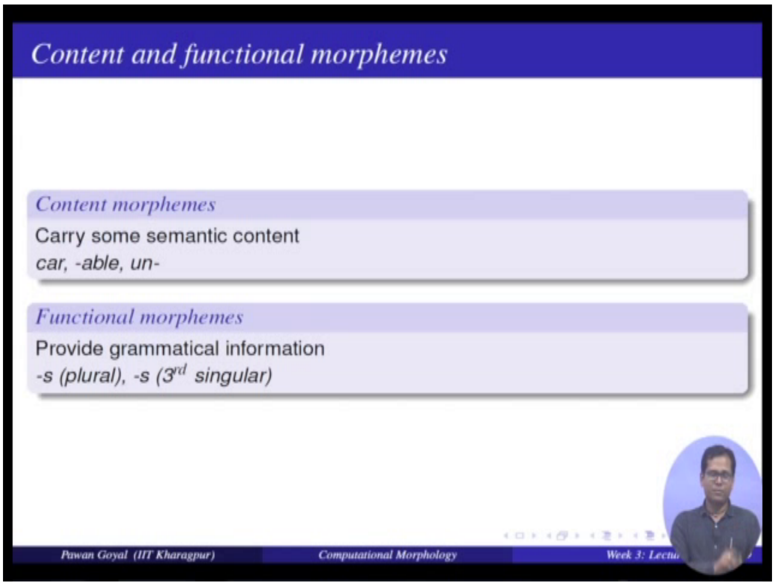
Then we have suffixes that are applied after the word, so like with the word talk; I can make a new word talking by applying the suffix by i n g. Similarly the word quick; I can make a new word quickly by applying the suffix l y; at the end and in the case of Hindi for example, you can have this exercise like [FL] and [FL] etcetera that can apply after the words.

Then, so the prefix and suffix are the major suffixes, affixes that we have in language, but certain language have some other kind of exercise also. So, one is called infix that is it is applied in between that stem not before or after that, so example is like in Sanskrit we have the word with to know and we make the present tense by putting vindati. So, the character n; the phonin n is applied in between; between v and the, so it becomes vindati. So, there are examples in other languages also like in the case of philippines here basa; b a s a is read and to make it; converted to the past tense read, the infix u m is applied in between, so we have b u m a s a, so u m is applied in between is called infix.

Now, so when you think of an infix in the case of English; English it is not common; in common language it is not infix is not used, but in general can we make a word in English that that uses infix. So, here an example is if I take the word absolutely and put a word bloody inside, so you can have a word like absolutely; bloodilutely. So, this is to put some emphasis that is not used in common English, but in certain movie dialogues and all you can find such kind of words.

Finally, so what you think can be the fourth category, we have prefix, suffix, infix; what is the meaning is that a suffix, an affix comes before as well as after, so this is called circumfix. So, this proceeds as well as follows the same word, so an example in Dutch; so you have a word for mountain berg and to convert it to plural we have an affix that is applied before g e as well as after. So, the whole affix is g e t e half of that is applied before and half of that is applied after, so this is called circumfix. So among this four kind of affixes, the first two are very very common prefix and suffix and the last two are expressive to certain languages only.

(Refer Slide Time: 09:54)



Content and functional morphemes

Content morphemes

Carry some semantic content

car, -able, un-

Functional morphemes

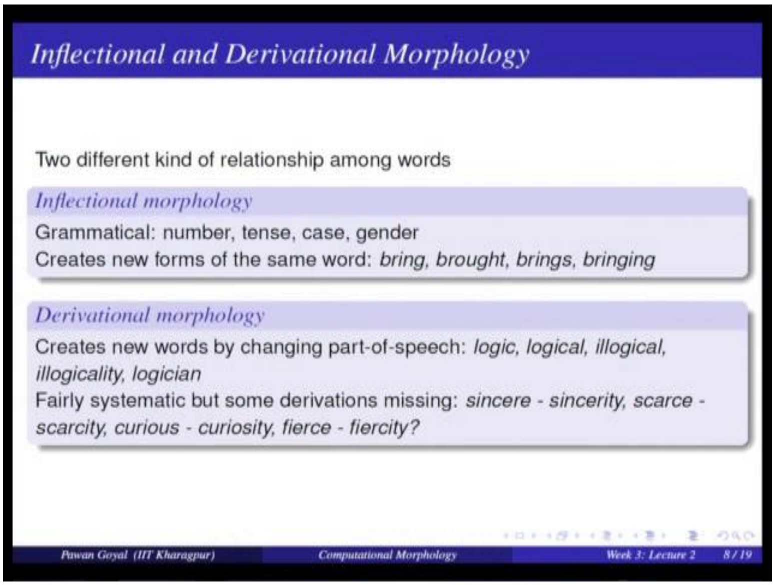
Provide grammatical information

-s (plural), -s (3^{rd} singular)

Now, morphemes can also be divided into some other categories like content verses functional morphemes. So, what are content morphemes that are those that will that will be a semantic meaning? So, for example, if even if I take any free morphemes like car a word, it is always a content morpheme; it contents some semantic meaning. Similarly I can take a morpheme bound morpheme like able also that is a semantic meaning that given a word, it keeps a sense of being able to do something.

So, on the other hand there are certain morphemes that are functional that do not take the semantic content, they are only used for certain grammatical functions like as for plural, as for third singular they are both functional morphemes; you apply it after a word and it has a particular grammatical meaning, but it may not be a semantic; it may not contain a semantic meaning as such.

(Refer Slide Time: 10:52)



Inflectional and Derivational Morphology

Two different kind of relationship among words

Inflectional morphology

Grammatical: number, tense, case, gender
Creates new forms of the same word: *bring, brought, brings, bringing*

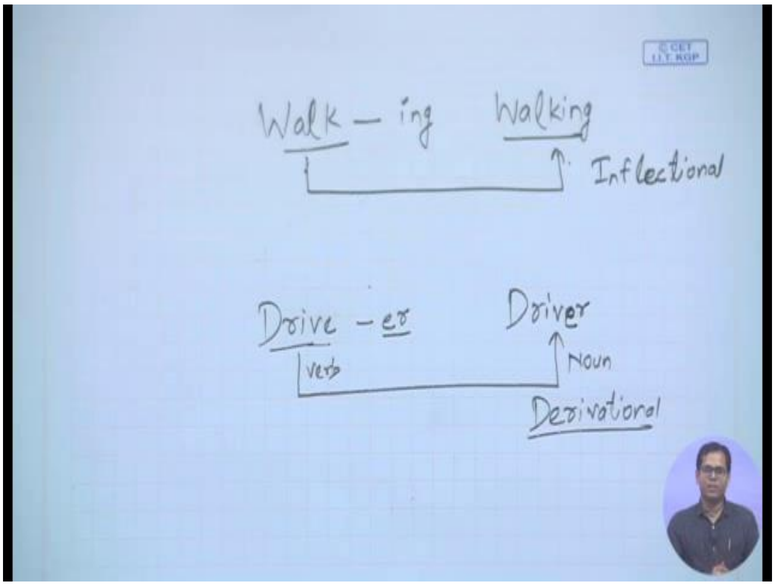
Derivational morphology

Creates new words by changing part-of-speech: *logic, logical, illogical, illogicality, logician*
Fairly systematic but some derivations missing: *sincere - sincerity, scarce - scarcity, curious - curiosity, fierce - fiercity?*

Pawan Goyal (IIT Kharagpur) Computational Morphology Week 3: Lecture 2 8 / 19

Now so based on whatever kind of affixes that you apply to a word, you will generate a new word.

(Refer Slide Time: 11:07)



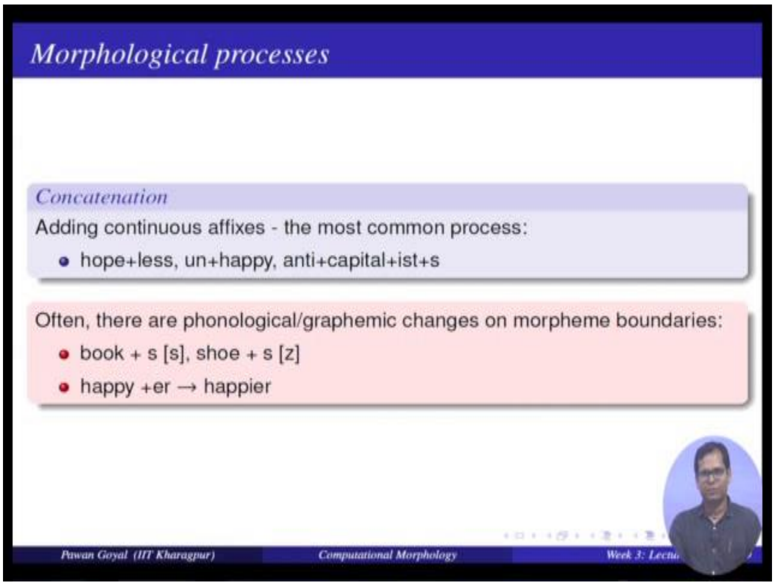
So, for example, if I have a word walk and I apply a suffix i n g; I get a new word like walking. So, there is a process that converts a word walk to walking, so by putting certain morphology here; now take another example I take a verb like drive and I add some affix and make a new word like driver, so now what is the difference that you see in the two processes. So, we can call it may be e r; drive plus e r that gives you driver, so from drive we can creating driver verses from walk you are creating walking. So, what is the difference that you see in the two processes, so if you think about it; in the first case we not changing the grammatical category of the word as such.

So, this is a verb walk, it remains the verb here also walking but we add some grammatical information as a continuous text. On the other hand here, you start with the verb drive and we end up making a noun, you covert the category of the verb itself by this process of morphology. So, there are two different kind of morphological processes, so this is called inflectional morphology and this is called derivational morphology. So, so in this slide we are seeing the difference between these. So, they give you the relation between two words that you have created; first case you have created from walk to walking, so what is the relation between these two words; second case you created drive to driver; what is the relation between these two words.

So in flectional morphology it makes changes in the terms of numbers, tens, case and gender. So, for example, if you start with the word verb bring you can alter the case, you can have brought, you can have third person singular brings and so on. So, they do not change the category of this word so, but you add some more grammatical information that is all. So, this is called inflectional morphology, you are adding certain inflectional information only in the grammatical sense that is not change in the category of the word. On the other hand, you have derivational morphology where you create some new words and you also change the change the category of the word. So, this is called part of speech that we will take up in the same week in the fourth module.

So, it changes the category of the word, so for example, from logic you can make logical and illogical, logicality, logician and all that and there you can see the do not have the same category all these words, they have they are different categories one is noun, adverb and so on adjective. In general derivational morphology is also fairly systematic like inflectional morphology, but sometimes certain derivations will be missing. So, for example, here is some nice examples, so if I take the word sincere I can make sincerity, scarce - scarcity, curious - curiosity, but from fierce you should not use like fiercity; although it is very very; if it looks very very regular, if you starts seen from the previous words. So, they are pretty regular, but some words do not have the corresponding derivational words.

(Refer Slide Time: 14:46)



Morphological processes

Concatenation

Adding continuous affixes - the most common process:

- hope+less, un+happy, anti+capital+ist+s

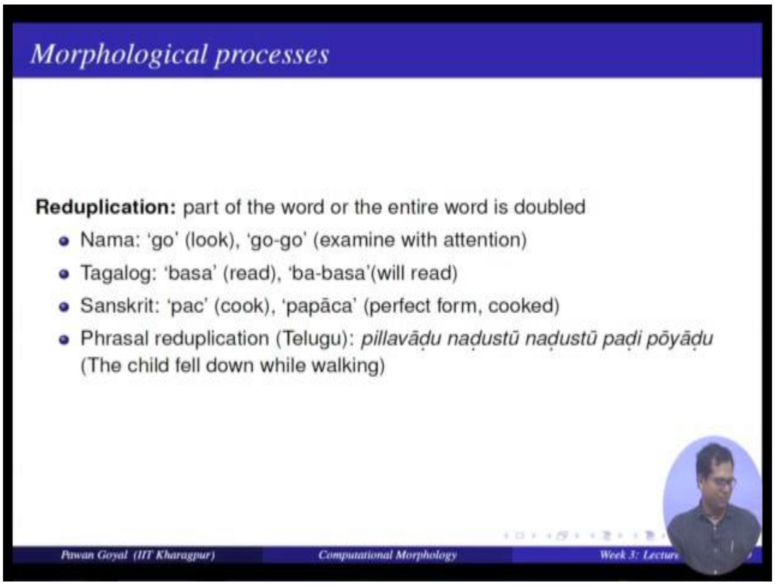
Often, there are phonological/graphemic changes on morpheme boundaries:

- book + s [s], shoe + s [z]
- happy +er → happier

Now so we have talked about whatever the various kind of morphemes and what are the process, what are the different types of affixes that you can apply, but in general what are the various morphological process that are involved to convert one word into another word, so we will talk about this. So, one simple process is called concatenation, so we will take various morphemes and concatenate together to make a new morpheme. So, for example, happy; un you make unhappy simple concatenation. Similarly here hope and less together make hopeless and un happy make unhappy and anti capital; i s t and s these four morphemes if you concatenate together and then make a single word or singular morpheme.

So, now then you are combining the different morphemes together; at the boundary or when they are combined there may be certain changes. The changes can be in the way the final word is pronounced or in the way the final word is written. So, there is two are called fundamic changes and graphemic changes; graphemic in the terms how the word is written and phenomic in the way word is pronounced. So, here some examples, so if I take a word book and if morpheme book and a morpheme s; if I combined together I get books. So, s see the pronunciation if I take shoe and s; I get shoes, so here it is not s, it is za; it is a different phonyms that terms. So, there are certain changes that happen at the boundary in terms of how the word is pronounced. Similarly if I take word like happy and e r; these change at the graphemiclable. So where in the grapheme y changes to i, so this is also called simply spelling change at the morpheme boundary, so this can happen.

(Refer Slide Time: 16:47)



Morphological processes

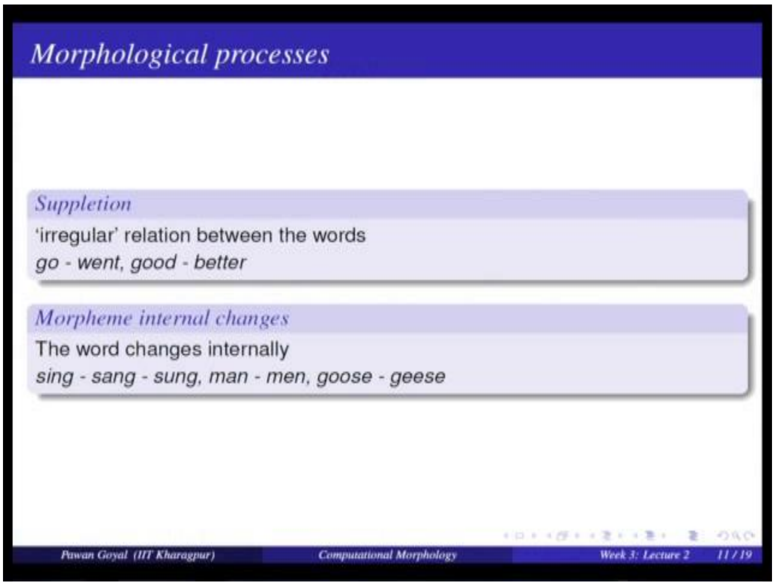
Reduplication: part of the word or the entire word is doubled

- Nama: ‘go’ (look), ‘go-go’ (examine with attention)
- Tagalog: ‘basa’ (read), ‘ba-basa’(will read)
- Sanskrit: ‘pac’ (cook), ‘papāca’ (perfect form, cooked)
- Phrasal reduplication (Telugu): *pillavāḍu naḍustū naḍustū paḍi pōyāḍu* (The child fell down while walking)

Pawan Goyal (IIT Kharagpur) Computational Morphology Week 3: Lecture

There are some morphological processes also like reduplication. So, then we adding in a suffix you might do reduplication somewhere in the stem. So, example is like in number language if I take a word go or look; if I want to say examine with attention, I will just repeat the go; similarly in tagalong; you have basa; for read and if I want to say will read I will read duplicate ba ba ba sa, so you reduplicating one part. In Sanskrit again this is very common phenomena or reduplication; I take the word like much that is to cook and if I want to convert it to the perfect form cooked, I will reduplicate the pa and I say papaca. So, you see the word pa is reduplicated they can be reduplication also in certain languages also example in Telugu, so certain phrase is repeated again. So, for example, I want to say the child fell down while walking, so here I say [FL]. So, here [FL] is being repeated twice for saying while walking.

(Refer Slide Time: 18:08)



Morphological processes

Suppletion

'irregular' relation between the words

go - went, good - better

Morpheme internal changes

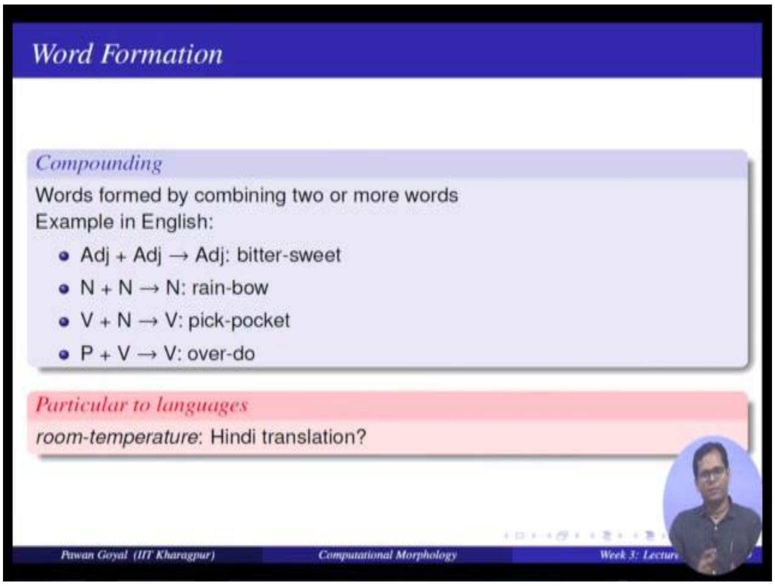
The word changes internally

sing - sang - sung, man - men, goose - geese

Pawan Goyal (IIT Kharagpur) Computational Morphology Week 3: Lecture 2 11 / 19

There are other morphological like suppletion, where a word is completely replaced by something that has no connection at the surface label. So, I have a word like go and I am converting into the past tense and it becomes went. So, you cannot find any connection go and went and this are surface label; this is simple suppletion. Similarly with good I can make better and there are many examples in English for this, sometimes there are some internal change in the morpheme. So, from sing you want convert the plural you will; so from sing you can get sang and sung for different tenses and from men to converted to plural and you get m e n and from goose; you get geese for converting to plural. So, you see their changes that are in internally in the word, you not adding any new suffix.

(Refer Slide Time: 19:11)



Word Formation

Compounding

Words formed by combining two or more words
Example in English:

- Adj + Adj → Adj: bitter-sweet
- N + N → N: rain-bow
- V + N → V: pick-pocket
- P + V → V: over-do

Particular to languages

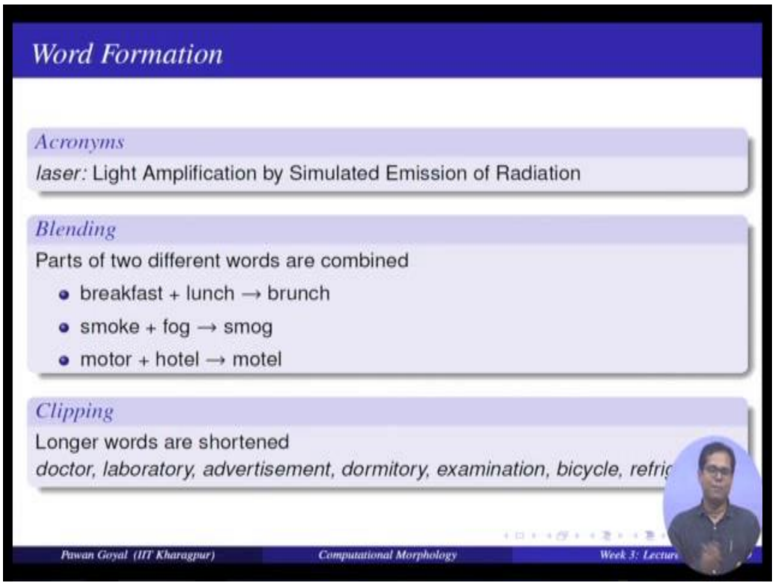
room-temperature: Hindi translation?

Pawan Goyal (IIT Kharagpur) Computational Morphology Week 3: Lecture

Then while we are talking about word formation, there is another process called compounding. So, that is I can take two different words and make a compound out of that and this compounding can be from various part of speech. For example, in English I can two take two adjectives, so bitter and sweet; I can make adject single compound bitter sweet, I can take two nouns rain and bow; I can make a noun rainbow, I can take a noun and verb pick and pocket and this becomes a verb pickpocket, similarly over do and what is interesting is that these compounds are very very particular to languages.

So, let us take an example like room temperature that is a compound in English, but is there a corresponding compound in Hindi. So, you never say [FL] as a single compound you will say [FL] so; that means so finding this compound is in interesting problem for going to application like machine translation. So, where you cannot just directly convert the words into the equivalent target languages, so you cannot say room with combined temperature with [FL], you need to find out with there is a compound there is relation between the two words and according to the translation.

(Refer Slide Time: 20:36)



Word Formation

Acronyms

laser: Light Amplification by Simulated Emission of Radiation

Blending

Parts of two different words are combined

- breakfast + lunch → brunch
- smoke + fog → smog
- motor + hotel → motel

Clipping

Longer words are shortened

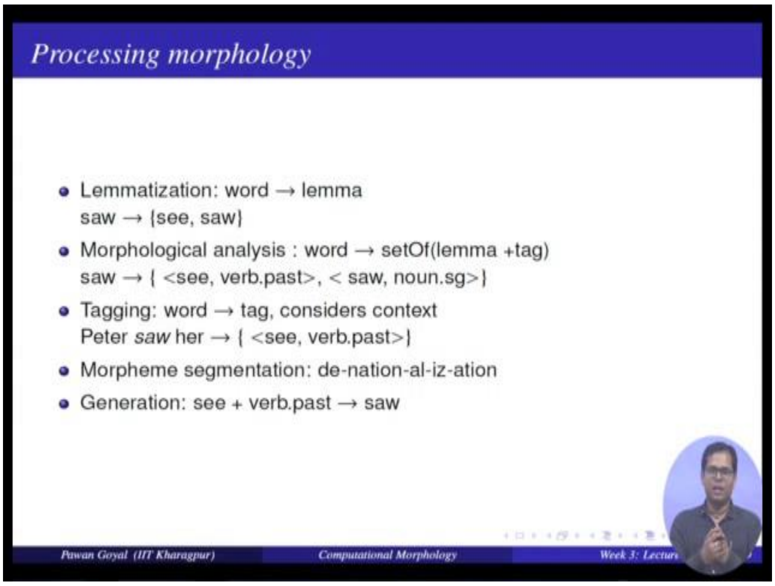
doctor, laboratory, advertisement, dormitory, examination, bicycle, refrig

Pawan Goyal (IIT Kharagpur) Computational Morphology Week 3: Lecture

Then there is another process that is called acronyms for example, the word laser is an acronyms then there is blending, so your take. So, this is again a very common process linguistic, so where you take two words and then you combine together, blend them together to make a new word, but the you are taking from both the words, but you not taking the full words. So, example is breakfast and lunch you can take together and make a new word like brunch.

Similarly smoke and fog, you make a new word a smog; motor and hotel make a word like a motel and there is a process of clipping. So, that (Refer Time: 21:13) used a lot for example, I have doctor becomes doc, laboratory becomes lab and advertising because becomes add, dormitory becomes dorm and examination becomes exam, bicycle becomes bike and refrigerator becomes fridge, so this is clipping; so long words are shortened

(Refer Slide Time: 21:34)



Processing morphology

- Lemmatization: word → lemma
  saw → {see, saw}
- Morphological analysis : word → setOf(lemma +tag)
  saw → { <see, verb.past>, < saw, noun.sg>}
- Tagging: word → tag, considers context
  Peter *saw* her → { <see, verb.past>}
- Morpheme segmentation: de-nation-al-iz-ation
- Generation: see + verb.past → saw

Pawan Goyal (IIT Kharagpur) Computational Morphology Week 3: Lectur

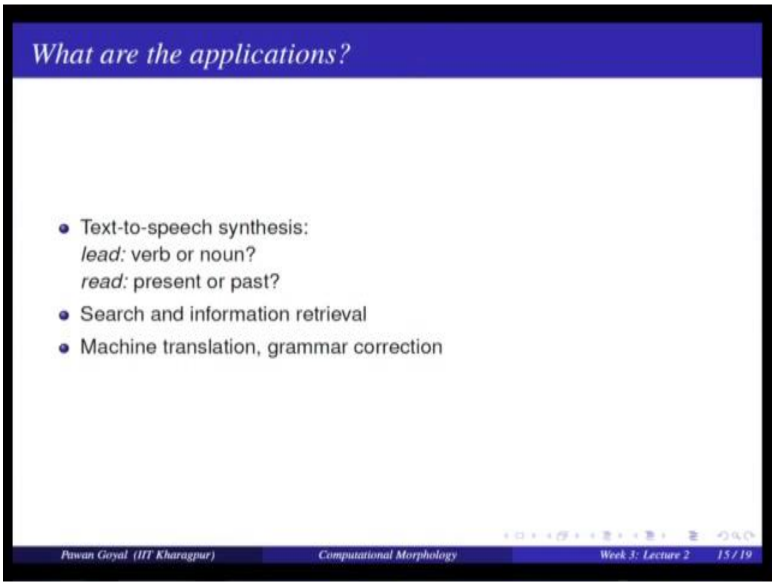
So, this is all the different processes that happens in happen in morphology now, so from the NLP prospective. So, we will talk about how do we process of morphology, so what do a; what does it and until what processing of morphology. So, what are different things that are done, so one simple thing is lemmatization that is given a word can identify what is the root word what is the lemma; an example is if you give me a word like saw; can I tell what is the root word, if it is a verb the root word what will be see, but if it is a noun the root word will be saw itself.

So, you can find what is the root word silly; it is lemma. There is a morphological analysis also. So, that is even a word can I find out what is the corresponding lemma along with morphological category of that word, so it is particular tag. So, example is I take a verb word like saw; can I tell the lemma is see and this a, past tense of that verb or saw is the lemma and it is a noun singular noun. So, this is morphological analysis of the given word.

Then there is a process called tagging. So, where I find out what is the actual, so what is the actual category of this word? So, the difference here is from the morphology analysis that I also have to disambiguate. So, morphological analysis you saw I was giving two different possibilities; in tagging I have to further find out what is the actual correct grammatical category here. So, I have a sentence is Peter saw her; I know the word saw can be either a noun or verb and we also saw their lemmas, but can I say can I tell in this particular context, the word saw will only be a verb and not a noun. So, can I do the actual disambiguation also; this is that is how tagging is different from the morphological analysis and then we have a morphological segmentation where given a word I can segment to with different morpheme that involved. So, I have a word like denationalization, I said de nation; a l i z ation they are different morphemes that are in this word. So, in general among all these processes, so part of speech tagging is very very popular. So, will words sometime for that and before that we will quickly see what is morphological analysis lemmatization; how it is done?

And finally, there is also these process of generalization where I can take a word, a root word and a particular grammatical category and I have to generate a new word from this. So, see and I want to generate past tense; I want to find out saw in NLP; it is not very popular unless you are talking about national language generation, where you might want to use this.

(Refer Slide Time: 24:39)



What are the applications?

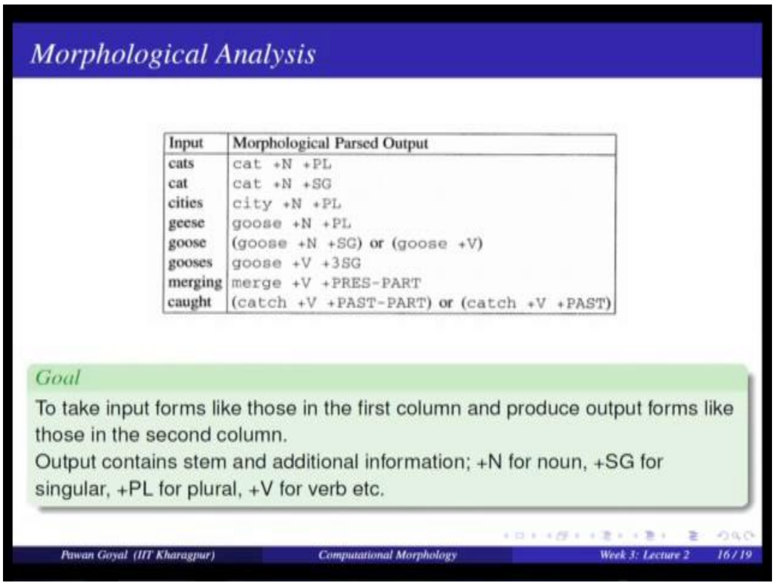
- Text-to-speech synthesis:
  lead: verb or noun?
  read: present or past?
- Search and information retrieval
- Machine translation, grammar correction

Pawan Goyal (IIT Kharagpur) Computational Morphology Week 3: Lecture 2 15 / 19

Before going into the process, what might be the application for doing this? Why should we be interested in doing morphological analysis? For example, text to speech synthesis, so what is text to speech synthesis you have seen something that is written and you have to find produce the corresponding is speech for that. So, now, if you have a word like lead written somewhere, now depending on whether it is a noun or a verb, you will have a different pronunciation for that lead verses led depending on noun or a verb. So, it is important of find out from the text whether it is a noun or verb. Same thing goes with read verses red; in general it is very important for other things likes search and information retrieval.

So, might want to use the a morphological category to reduce the certain space and also machine translation we saw some example and especially because if you know the morphological category, you can find out the for the target languages what is corresponding affix is or different words that is being used in that languages and grammatical error correction and all that. So, if you know what is the morphological category of this word in the whatever is written, if you can find out this is not the correct morphology that is use you can try to correct it.

(Refer Slide Time: 26:04)



Morphological Analysis

Input	Morphological Parsed Output
cats	cat +N +PL
cat	cat +N +SG
cities	city +N +PL
geese	goose +N +PL
goose	(goose +N +SG) or (goose +V)
gooses	goose +V +3SG
merging	merge +V +PRES-PART
caught	(catch +V +PAST-PART) or (catch +V +PAST)

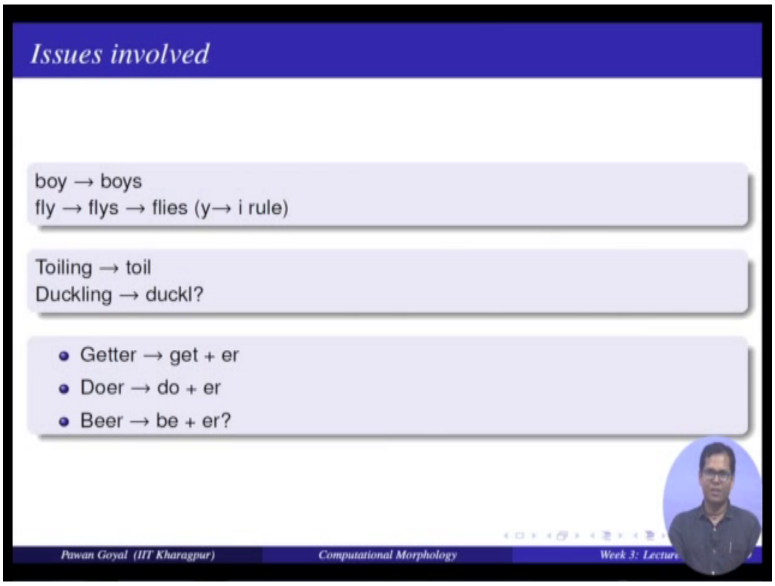
Goal

To take input forms like those in the first column and produce output forms like those in the second column.
Output contains stem and additional information; +N for noun, +SG for singular, +PL for plural, +V for verb etc.

Pawan Goyal (IIT Kharagpur) Computational Morphology Week 3: Lecture 2 16 / 19

So, now what is morphological analysis, so this is have seen earlier. So, as I input we have words like cats, cat, citizen so on and as an output I want to find out given a word like cats, what is the root verb here cat? What is category noun? And it is a plural cat plus n plus p l something like that. So, in this table, so if I am a given input in the left as per the left column, I want output like the right column. So, the output that I will generate will contain additional information like this is a noun, this is singular for s g for singular p l for plural and v for verb like that. So, this is all the information that I want to get from the given word.

(Refer Slide Time: 26:57)



Issues involved

boy → boys
fly → flys → flies (y→ i rule)

Toiling → toil
Duckling → duckl?

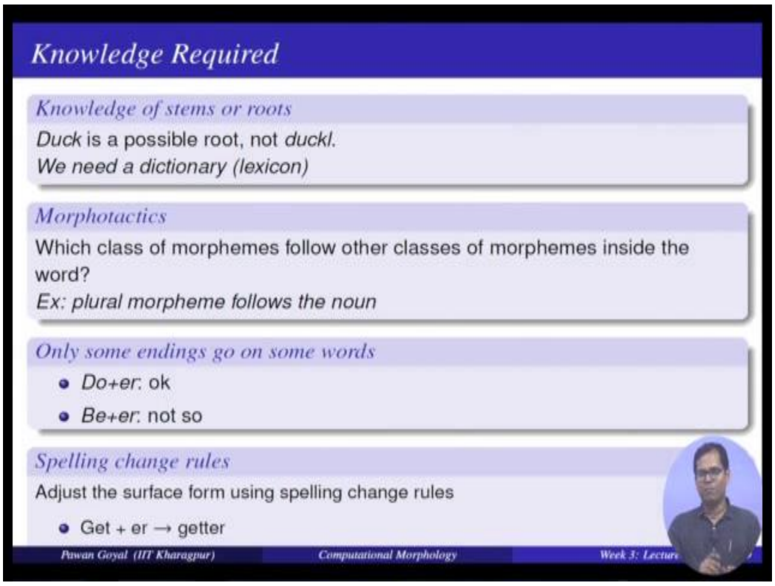
- Getter → get + er
- Doer → do + er
- Beer → be + er?

Pawan Goyal (IIT Kharagpur) Computational Morphology Week 3: Lecture

Now, what might be the issues involved while doing morphological analysis one particular problem is that it is not very very regular. For example, from the word boy I can get plural boys, but what happens if I taken input like fly; I get flies f i l e s. So, this you see that they are following two different sort of tools for doing changes at the bounding. Similarly if I take the word like toiling I can get toil, but what happens if I give an input like duckling, should I used same sort of full to get duckl that is not correct of English word.

So, how do I know the duckling is not a correct English word when an input like a duckling is provides to my system. Similarly from getter I get g e t plus e r from doer d o plus e r, but what happens if given input like beer, do I output like b e plus e r? So, these are some of the issues that involved in processing morphology, now if I have to solve these issues what is what are the different knowledge that I need to have? .

(Refer Slide Time: 28:03)



Knowledge Required

Knowledge of stems or roots

Duck is a possible root, not *duckl*.
We need a dictionary (lexicon)

Morphotactics

Which class of morphemes follow other classes of morphemes inside the word?
Ex: plural morpheme follows the noun

Only some endings go on some words

- *Do+er*: ok
- *Be+er*: not so

Spelling change rules

Adjust the surface form using spelling change rules

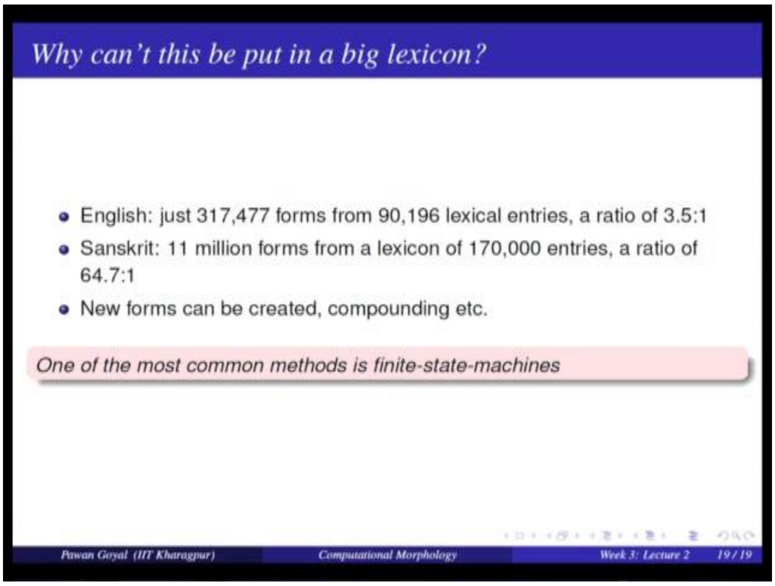
- Get + er → getter

Pawan Goyal (IIT Kharagpur) Computational Morphology Week 3: Lecture

So, I need to have some knowledge on what are the different words or fruits in English. So, for example, I need to know that duck is a possible root, but duckl is not a possible root in English. So, we need some sort of dictionary or lexicon of English what are different nouns and verbs etcetera in English, what else; I need to have some knowledge of morphotactics what is morphotactics? That is which kind of morphemes follow other kind of morphemes for example, if I have to convert a noun to plural; I know plural morphems always follow the noun. So, this is the morphotactics information which morpheme follows are the morphemes, then I also need to have this information that some endings go only on certain words not on everything.

So, on d o I can apply e r to get doer, but on b e on the verb b e; I cannot apply e r to get a word like beer. So, this is again these constraints also I need to have and then I need to some worry about spell spelling change rules. So, then whenever I have word like get I apply e r it converts to getter. So, there is just duplication on of duplication of the ending t.

(Refer Slide Time: 29:24)



Why can't this be put in a big lexicon?

- English: just 317,477 forms from 90,196 lexical entries, a ratio of 3.5:1
- Sanskrit: 11 million forms from a lexicon of 170,000 entries, a ratio of 64.7:1
- New forms can be created, compounding etc.

One of the most common methods is finite-state-machines

So, one question you might have is that why cannot I put everything together in a big lexicon. So, that is all the root words all their morphological variants why cannot they already put in a big lexicon and there I can search even new word, I can search in the lexicon find out what is the actual word and its category if they are finite by country do that. So, there are two different reasons why this may not be very good solution.

So, one is if you take any language like English it is very easy, so where for a given, but they are not many variations in terms of morphology. So, this is some strategies from a data set that in English there are roughly you take 90000 lexical entries and you will find out all the possible morphological forms that you can generate then you find a ratio of 3.521 that is from word you can generate roughly 3.5 more very to variants including that word in English, but this is not true for other languages.

For example, the same thing if you do for Sanskrit, you have a lexicon of 170000 entries, but if you try to derive the forms. So, the ratio that you come up with is something like 64.7 is to 1. So, the 11 million forms that are generate, so these are huge in number. So, 64 is a very big ratio in comparison to 3.5, but again this may for now it may not be a very big number again you can still argue that this might this will it be put in a big lexicon and you can search over that, but there is another problem that is you can keep on coining new words and apply the same morphological processes to generate its forms and you do not know these words a prior. So, you cannot store their morphological variants in the lexicon, so on the other hand if you can store the what kind of rules in valued for mailing from a word to its plural given a new plural form, you can try to find out its original root word.

So, that is why we will be studying what kind of methods are possibly used and the one of the most popular method in this field of computation morphology. Finally, state methods so this was very very popular earlier first the language like English also and later on for other languages like Indian languages another European language. So, even now if you talk about processing the morphology (Refer Time: 31:54) methods are one of the most popular choices.

So, in this lecture we talked about composition morphology, what are the linguistics terminologies and what is the process as such what is the NLP perspective there and in the next lecture we will talk about how do I use final state methods for this processing.

Thank you.