

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

**Lecture – 04**  
**Empirical Laws**

So, welcome back for the forth lecture of the first week. So, in the last lecture we discussed why is NLP hard and we took examples of various cases of ambiguities in the language and some non standard usage also. So, (Refer Time: 00:36) we will actually go into a text corpus and try to study some very simple empirical laws, that are very very universal about language.

(Refer Slide Time: 00:52)

*Function Words vs. Content Words*

Function words have little lexical meaning but serve as important elements to the structure of sentences.

*Example*

- The *winfy prunkilmonger* from the *glidgement mominkled* and *brangified* all his *levensers vederously*.
- *Glop* angry investigator *larm blonk* government harassed *gerfritz* infuriated *sutbor pumrog* listeners thoroughly.

*Function words are closed-class words*  
prepositions, pronouns, auxiliary verbs, conjunctions, grammatical articles, particles etc.

Pawan Goyal (IIT Kharagpur) Empirical Laws Week 1: Le 5

So, now before going to that, I will try to make a simple distinction clear to you. So, what are the difference between function words versus content words in language. So, whenever you see a language in its vocabulary, there are various kinds of words and they can be called either function words or content words what are the difference between those?

So, function words are mainly used to make the sentence grammatically. So, things like where is determine is, a and pronouns and all are function words; content words are various nouns and verbs that convey what are the important concepts in the sentence. So, content words are mainly used for determining the structure of the sentence. So, now,

this is an example to give you an idea about what are content words and what are function words.

So, you see here the same sentence has written in 2 different ways, in one of the cases we have replaced the content words by some garbage words. So, this mean (Refer Time: 02:03) the words that are not in the language.

In other example we have (Refer Time: 02:06) the function words by some arbitrary words that are not in the language. So, can you try to find out in which sentence we have replaced the function words with something else? So, we will see here in the second sentence, we have removed the word like the and replaced with words like glop and so on. In the first sentence we have replaced all the nouns like angry and investigator, with winfy prunkilmonger that is research to the garbage word that is not the language, now.

Now, try to see the 2 sentences end, in which sentence you can see the structure of the sentence clearly and in which of the sentence you can understand the topics of the sentence clear and you will find out that in the second sentence you know all the topics. So, this is talking about some investigator, government and infuriated and all that, the information you cannot get from the first sentence. From the first sentence you can understand the extraction of the sentence, the some x some noun from something and a verb and again a verb all his a noun and a adverb. And you do not know what are these nouns and adverbs, but you know what is the extracts of these sentence, there are certain fillers in the sense of various nouns and verbs.

So, in general function words in a language are closed class words, what I mean by that there are very specific grammatically categories that can be called as function words and they are generally closed, we do not have newer and newer functions what coming into a language. Some examples of function words are like prepositions, pronouns, auxiliary verbs, conjunctions, grammatical articles and various particles in the language.

(Refer Slide Time: 04:04)

Most Common Words in Tom Sawyer		
Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

The list is dominated by the little words of English, having important grammatical roles.

Pawan Goyal (IIT Kharagpur) Empirical Laws Week 1: L1 3

So, now let us take a corpus, and try to see what are the distribution of words, that we encounter in the corpus. So, here we have taken Tom Sawyer. So, that was a noble written by Mark Twain and what we are doing in the slide, we are just reporting which of the words occur with what frequency; so this is a sorted list a starting from the highest frequency to some top 15-20 words and what is the grammatical category. So, the top word you see here is 'the', so that occurs with the frequency of 3332 and this is very very common across different language corpus. So, if you pick up any arbitrary language corpus, you will feel that the distribution is roughly the same

So, now can you have a look at this list and find out, what is special about this list, what is the category of words that you see here? So, remember we talked about 2 categories: function words and content words, can you try to find out which of the category you see here? So, we see most of the words here are function words. So, the, and, a, to, of, and you can see also by the grammatical categories that is also written here. So, one thing like that we see in this distribution is that this list is dominated by the little words of English that are very important grammatical roles for the sentence.

Now, what is one exception that you see to this? So, we call them function words that we said earlier determiners, prepositions, complementizers like the word that here.

(Refer Slide Time: 05:54)

Most Common Words in Tom Sawyer		
Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

The one really exceptional word is *Tom*, whose frequency reflects the text chosen.

Pawan Goyal (IIT Kharagpur) Empirical Laws Week 1: Lecture 4 3 / 15

So, now what is one exception that we see in this list? So, that exception is probably the word Tom; so Tom is very specific to the topic of this text. So, this is about Tom Sawyer. So, you see the word Tom also has a very high frequency and this makes you to the list of top frequent words. So, again if you take an arbitrary text and try to find out some top frequent words, mostly they will be function words; also sometimes known as there is this top words, plus you can find some occurrences of words that convey the topic of that text. So, in this that is a word Tom.

(Refer Slide Time: 06:34)

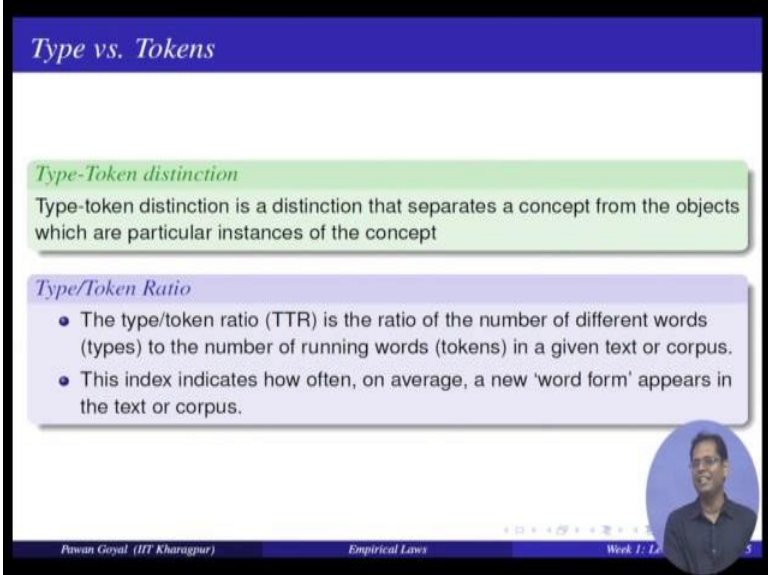
Most Common Words in Tom Sawyer		
Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

How many words are there in this text?

Pawan Goyal (IIT Kharagpur) Empirical Laws Week 1: Lecture 4 3 / 15

So, let us try to see something else, how many words are there in this text, what are the various technical terms by which these are indicated?

(Refer Slide Time: 06:48)



*Type vs. Tokens*

*Type-Token distinction*

Type-token distinction is a distinction that separates a concept from the objects which are particular instances of the concept

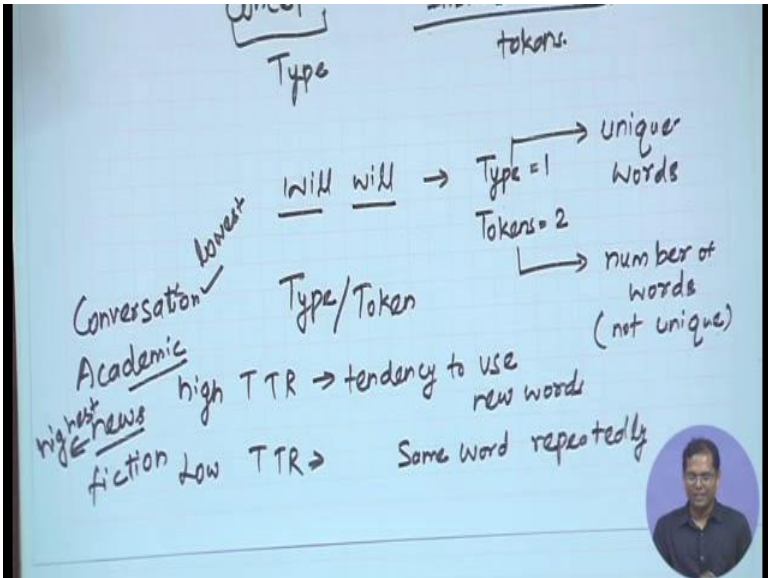
*Type/Token Ratio*

- The type/token ratio (TTR) is the ratio of the number of different words (types) to the number of running words (tokens) in a given text or corpus.
- This index indicates how often, on average, a new 'word form' appears in the text or corpus.

Pawan Goyal (IIT Kharagpur) Empirical Laws Week 1: L

So, for that we will talk about the type token distinction. So, type token astrology philosophical distinction between a concepts; concept is a type and a particular instances of the concept. So, concept is type and instance is concepts or tokens.

(Refer Slide Time: 07:05)



Concept  
Type

tokens.

will will → Type = 1  
Tokens = 2

Type/Token

unique Words

number of words (not unique)

Conversation — lowest

Academic — high

highest ← news

fiction — low

high TTR → tendency to use new words

Low TTR → Some word repeatedly

So, what do I mean by type and token, when it comes to a language? So, language if the same word occurs multiple times, so I call them the same type, same concept, but

different tokens. So, if I write the same word say, will will the same word twice, I say this is a single type this is only one type, but there are 2 tokens. So, each individual occurrence is a different token, but they have the same type. So, in other sense by type I mean some sort of unique words in my vocabulary and token means the number of words they are not unique. So, if the word occurs 5 times I will count it as 5 different tokens, but the same type.

Now, so once we have understood what is a distinction between type and token, there is a particular concept that is called type token ratio that is used to should describe some text. So, what is type token ratio? Very simply, the ratio of the number of types that is number of different unique words to the number of running words so all the tokens. So, you find out the number of types in the text, divided by the number of tokens in the text and you get the TTR, Type Token Ratio.

So, what this ratio indicates? That tells on an average how often a new word appears in the text. So, if type divide by token that is my type token ratio if it is high; that means, in the text lot of new words are keep on coming, because the number of types is very close to the number of token, but if this is small; that means, the small words are getting repeated again and again in the corpus.

(Refer Slide Time: 09:37)

Comparison Across Texts	
Mark Twain's Tom Sawyer	<ul style="list-style-type: none"><li>71,370 word tokens</li><li>8,018 word types</li><li>TTR = 0.112</li></ul>
Complete Shakespeare work	<ul style="list-style-type: none"><li>884,647 word tokens</li><li>29,066 word types</li><li>TTR = 0.032</li></ul>

So, suppose we take 2 different text corpus: one is Mark Twain's Tom Sawyer, another we take complete Shakespeare work and we try to find out what is the TTR for each of

this. So, our Mark Twain's we find that there are 71,370 word tokens and 8,018 word types. So, how do we compute TTR? You just find out, you just divide types by tokens. So, 8018 divide by 71370 that will give me the TTR that is close to 0.112.

Now, if you take the Shakespeare work, you find similarly there are 88400 plus tokens and 29000 plus types and type TTR comes to be much more than that of Tom Sawyer, 0.03.

(Refer Slide Time: 10:30)

*Empirical Observations on Various Texts*

*Comparing Conversation, academic prose, news, fiction*

- TTR scores the lowest value (tendency to use the same words) in conversation.
- TTR scores the highest value (tendency to use different words) in news.
- Academic prose writing has the second lowest TTR.

*Not a valid measure of 'text complexity' by itself*

- The value varies with the size of the text.
- For a valid measure, a running average is computed on consecutive 1000-word chunks of the text.

Pawan Goyal (IIT Kharagpur) Empirical Laws Week 1: Le 5

Suppose I take 4 different types of media. So, 4 different types of media like conversations if 2 people are conversing; academic prose will be scientific articles and books; news and fiction; I take 4 different sorts of media in which language is communicated. Now, can you try to have a guess which one will have the highest TTR, and which will have the lowest TTR?

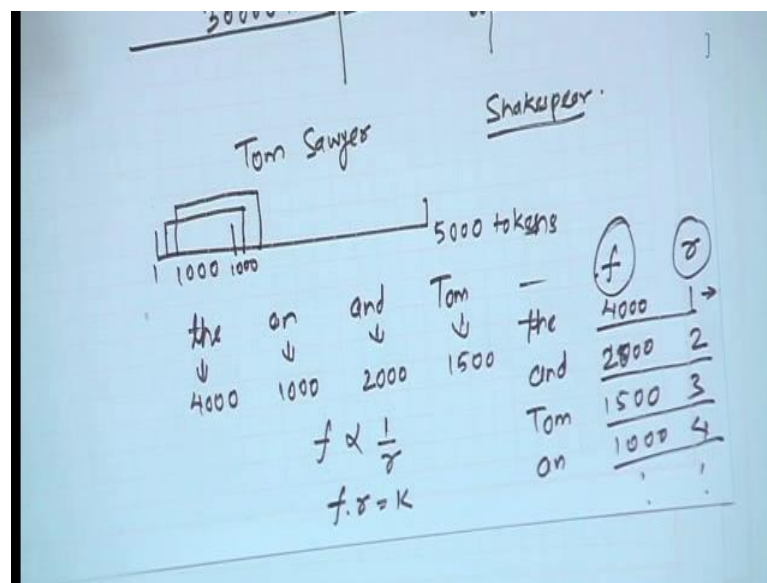
So, let us go to the definition of TTR. So, high TTR, it will be tendency to use the new words, what about low TTR? If you have low TTR, that is tendency to use same word repeatedly. So, now given that I have 4 different media: one are conversation, then I have academic prose, then I have news and fiction. Now which one do you think will have the tendency to repeat the same word again and again? So, we think over it, in conversation we try to repeat the same word again and again. So, whatever you are conversing, I will try to take it from there and repeat it.



So, conversations have the tendency to use the same word repeatedly so that means, they will have the lowest TTR, on the other hand it has been found in news you have tendency to use newer words, so news this is the highest. So, can you guess which of the remaining one will have the second lowest TTR, where the same words have been used repeatedly and that is actually the academic prose; because academic prose is very very formal, so they are very very particular kind of words and verbs that we use. So, that is why TTR is again found to be low in academic prose.

So, TTR scores the lowest value in the case of conversations and highest value in news and the academic prose as the second lowest TTR and this also make sense. Now one thing you must be conscious about that TTR in itself is not a very valid measure of finding text complexity, you might say that Ok, if TTR is high text might be complex, but this is not a valid measure in itself. So, why is that so? So think of this scenario, if you are using more and more text what would happen to your TTR ratio?

(Refer Slide Time: 13:45)



Suppose your text contains, you are taking a running text, so till here you have 30000 tokens, now you go forward in your text and you go to 60000 tokens, now what would you guess, can you see the TTR at this point will always be lower than the TTR at this point? Yes because whatever words were there in the corpus, most the words have already been introduced in the first half; second half most of the words will be again

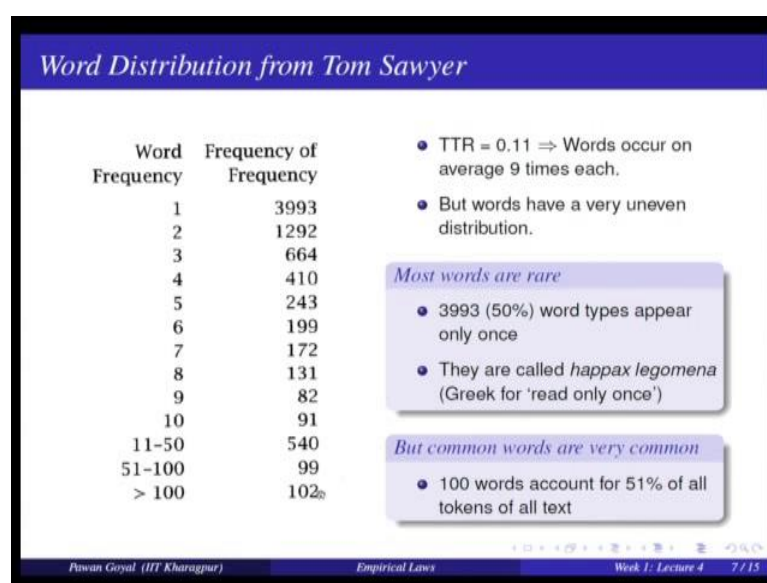


getting repeated. So, the number of tokens will increase linearly with the size of text, but numbers of words in a vocabulary do not.

So, that is why if I take say a Tom Sawyer versus Shakespeare's works, you will see the TTR is high in Tom Sawyer than Shakespeare work, because they since having a much larger number of tokens. So, in itself it cannot be a valid measure. So, when you take into account the size of the text. So, another way in which this is computed is by taking a running average. So, you take a running average on some consecutive 1000 word chunks, what I mean by running average? Suppose I have a text that as 5000 tokens.

So, now you first compute TTR for the initial 1000 tokens 1 to 1000, get a TTR for that; now you take start from 2 to 1001, I did not have TTR, 3 to 1002 compute TTR and then take an average. So, you compute TTR on moving window of 1000, length 1000 and then you take an average. So, that is considered to be a better measure for TTR.

(Refer Slide Time: 15:51)



So, now, you see that we will talk about some empirical laws, for that let us try to have a look at some of the frequency distribution that we see in Tom Sawyer. So, remember we said that the TTR on Tom Sawyer is roughly 0.11. So, what does it indicate? It says that on an average here what is repeated 9 times, yes try to token ratio is 1 by 9 roughly.

Now, what is that mean, does that mean that each word occurs 9 times in the text or is it something different? So, what we saying here that words do not have a very even

distribution; it is not happen that each word occurs 9 times in the text. So, what is the usual kind of distribution that we see in the corpus?

So, in the left hand side of the slide, you see I have 2 columns: first column gives me the word frequency, second column gives me frequency of frequency; what do I mean by frequency of frequency? How many words in this corpus have that frequency, so take the first row; that means, there are 3993 words in the corpus, that occur only once they have frequency 1 and there are 1292 words in the corpus that occur twice and so on.

Similarly, there are 102 words in the corpus that occur more than 100 times. So, now, once you see this statistic then immediately see that the distribution is not very very uniform. So, what are the 2 different observations that we have here? So, first observation that we make are most words are real. So, which are the real words? The words that occur with frequency 1 or 2, so we see here roughly 50 percent of the word types; remember in Tom Sawyer roughly 8000 word types where there. So, roughly 50 percent of them occur only once. So, they are the frequency of 1 and they are also called happax legomena. So, that is a Greek term for read only once, they occur only once in my corpus.

What is the other observation that we have? 100 words account for 51 percent of all the tokens of all text. So, you had only 102 words here that were having frequency greater than 100, but the together account for 50 percent of the tokens of the corpus. So, this is again a strange.

(Refer Slide Time: 18:27)

**Zipf's Law**

- Count the frequency of each word type in a large corpus
- List the word types in decreasing order of their frequency

**Zipf's Law**

A relationship between the frequency of a word ( $f$ ) and its position in the list (its rank  $r$ ).

$$f \propto \frac{1}{r}$$

or, there is a constant  $k$  such that

$$f \cdot r = k$$

i.e. the 50th most common word should occur with 3 times the frequency of the 150th most common word.

Pawan Goyal (IIT Kharagpur) Empirical Laws Week 1: Lecture 4 8 / 15

So, now this particular observation gives the first empirical law that we will study that is called Zipf's law. So, very very popular law in the case of language corpus, how the vocabulary's distribution behave; so you see Zipf's law what do we need to do? First take a large corpus, now take find out what are the individual word types and count the frequency of each word type.

Now what you need to do? You list the word types in their decreasing order of frequency. So, for example, suppose in the previous case we found there are various types like the, an, and, Tom and so on in the corpus, I will find out the frequency. So, this occurs with 4000 frequency, this occurs with 1000, this occurs with say 2000 and this occurs with 1500 and so on there are various first occurring the different frequencies.

Now, the first thing I will do, I will try to sort them in the decreasing order. So, first I will have - the, then I will have - and, then I have - Tom, then I will have - an, and so on all that the words in the vocabulary. So, they will have their frequency this is 4000, 1000 sorry 2000, 1500 and 1000. Now I will give them the ranks, because I have already sorted them in decreasing order I put the rank. So, there is this rank 1, rank 2, rank 3, rank 4 and so on.

So, now what is Zipf's law say? So, Zipf's law gives me a relationship between the frequency of a word, the  $f$  in this list and the rank of this word, in this sorted list and what is the relation that Zipf law gives? A relation is very very easy. So, relation is the

frequency, it is inversely proportional to the rank of the word, inversely ratio between the frequency and the rank. So, that is I can say that if I multiple  $f$  dot  $r$ , I will get some sort of the constant  $f$  dot  $r$  is a constant; this is for our arbitrary example that we took in this case we can even see that this holds. So, 4000 into 1 is. So, this is 4000 this is again 4000, this is 4500 and this 4000 they are roughly the same. So, what it means suppose I make a list like that and I have 2 words.

(Refer Slide Time: 21:32)

Handwritten notes on a grid background:

$$50 \quad f_1$$

$$150 \quad f_2$$

$$f_1 = 3f_2$$

$$P_r = \frac{f_r}{N}$$

$N \rightarrow$  Total number of tokens

$$f \propto \frac{1}{r}$$

$$\frac{f}{N} = \frac{A}{r}$$

$$k = \frac{A}{N}$$

So, I have a word with as a rank of 50 and another word that has a rank of 150 and I find it frequency  $f_1$  it is frequency  $f_2$ . So, I can see that  $f_1$  is 3 times  $f_2$ . So,  $f_1$  is 3 times  $f_2$ . So, the 50th most count word should occur with 3 times the frequency of the 150th most convert, this is what Zipf's law tells me. Now, we can write it in also a different manner. So, we talked about frequency, we can also talk about probability. So, let  $P_r$  denote the probability of a word with rank  $r$ .

So, what is  $P_r$ ?  $P_r$  is nothing but frequency of the word of rank  $r$ , divide by the number of tokens that I seen the corpus, that is how I will compute the probability of that. So,  $N$  is the total number of word account says tokens. So, what does Zipf law tell? So, it says frequency is inversely proportional to  $r$ , I can also write  $f$  divide by  $N$  with some  $A$  by  $r$ . So, what I have done? I taken the constant  $k$ , I have written that it as  $A$  times  $N$ .  $N$  is a number of tokens in my vocabulary  $N$   $A$  is from other constant and it is fixed given a corpus and using in the constant. So now, why we are doing that? Because it has been

seen in that in the corpus,  $A$ 's when we take a valid close to 0.1, this is generally more fix than the value of  $k$ .

(Refer Slide Time: 23:40)

Empirical Evaluation from Tom Sawyer							
Word	Freq. ( $f$ )	Rank ( $r$ )	$f \cdot r$	Word	Freq. ( $f$ )	Rank ( $r$ )	$f \cdot r$
the	3332	1	3332	turned	51	200	10200
and	2972	2	5944	you'll	30	300	9000
a	1775	3	5235	name	21	400	8400
he	877	10	8770	comes	16	500	8000
but	410	20	8400	group	13	600	7800
be	294	30	8820	lead	11	700	7700
there	222	40	8880	friends	10	800	8000
one	172	50	8600	begin	9	900	8100
about	158	60	9480	family	8	1000	8000
more	138	70	9660	brushed	4	2000	8000
never	124	80	9920	sins	2	3000	6000
Oh	116	90	10440	Could	2	4000	8000
two	104	100	10400	Applausive	1	8000	8000

So, now going back to the example of Tom Sawyer, let us take some words their frequency and their rank and see whether the relation that  $f \cdot r$  is roughly constant holds to it. What you are seeing here, we have arrays words like the, and, he, there and so on, we have written their frequencies and their ranks and what is  $f \cdot r$ . So, what we are seeing here  $f \cdot r$  if you start from the forth word say he, this remains roughly in the same range, it starts, it remains in 8000 to some 10400 and 200, the range does not vary a lot.

So, that is very nice empirical evidence that Zipf's law holds and let us see we will see in various corpus, that  $f \cdot r$  remains constant from most of the words that you will see.

(Refer Slide Time: 24:32)

**Zipf's Other Laws**

*Correlation: Number of meanings and word frequency*

The number of meanings  $m$  of a word obeys the law:

$$m \propto \sqrt{f}$$

Given the First law

$$m \propto \frac{1}{\sqrt{r}}$$

*Empirical Support*

- Rank  $\approx 10000$ , average 2.1 meanings
- Rank  $\approx 5000$ , average 3 meanings
- Rank  $\approx 2000$ , average 4.6 meanings

Pawan Goyal (IIT Kharagpur) Empirical Laws Week 1: Lex 3

So now, Zipf's law is one of the very very important law, that we have just seen, but Zipf has given some other laws also there are not so popular, but still we will try to have a look quick look at this. So, one of the laws is relating the number of meanings a word has with it s frequency. So, what this law says, the number of meanings  $m$  that a word as obeys the simple law, that  $m$  is proportion to the square of roots it is frequency.

So, that is as we have words with higher and higher frequencies, the number of the different senses or the meanings will which they can be used also increases. So now if I use it with the first law, what do we get?

(Refer Slide Time: 25:25)

$f \propto \frac{1}{r}$  - Law 1

$m \propto \sqrt{f}$  → Law 2

$\Rightarrow m \propto \frac{1}{\sqrt{r}}$  → bad post

Most words occur once [50% of the vocabulary types occur only once.]

Common words are very common [2/100 words account for 50% of the tokens.] → Good post

So, the first law says that frequency is inversely proportional to the rank, this is law 1 and the second law says the meanings are directly proportional to the square root of frequency, this is law 2. If we combine these 2 laws together, we can get meaning  $r$  inversely proportional to the rank of (Refer Time: 25:50); that means, if I put the words in the decreasing order of your rank. So, you are starting from rank 1 to rank 2 and so on. So, they mean number of meanings will also keep on decreasing.

So, these clause hold an average, what is the empirical support? So as I was saying the hold on an average, if we saw the words that are having rank roughly close to 10000. So, they have roughly 2; 2.1 meanings, if they have words that are having rank roughly 5000, they are found to have roughly 3 meanings and the words that are having rank plus 2000, they were having roughly 4.6 meanings and you can do the math and you will find out this roughly follows whatever we have seen in this law.



(Refer Slide Time: 26:42)

*Impact of Zipf's Law*

*The Good part*  
Stopwords account for a large fraction of text, thus eliminating them greatly reduces the number of tokens in a text.

*The Bad part*  
Most words are extremely rare and thus, gathering sufficient data for meaningful statistical analysis is difficult for most words.

Pawan Goyal (IIT Kharagpur) Empirical Laws Week 1: Lecture 4 13 / 15

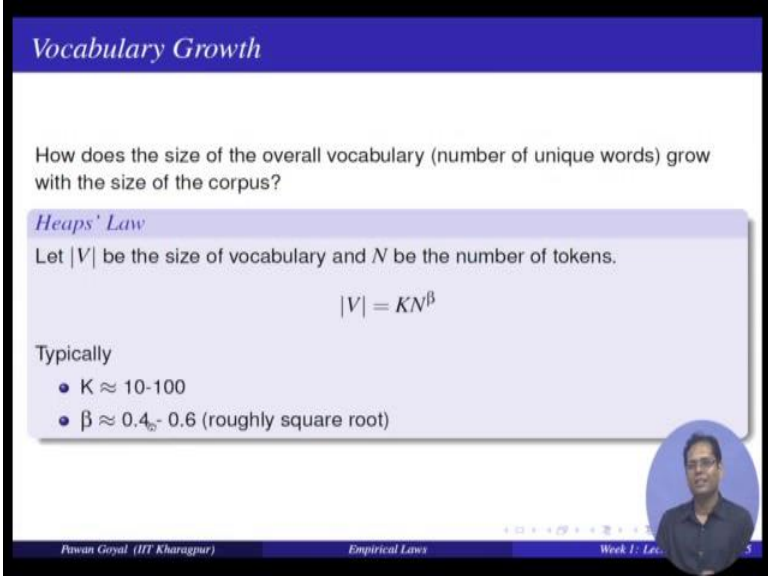
And these are the law by Zipf that tries to correlate the length of the word; that means, the number of characters the word has with the frequency of the word and it says the frequencies of the word is inversely proportional to its length and can you try to see that in the case of English? So, the very very long words are very very really used, so their frequency is very very small, but some of the very small words like some simple examples like function words, they are used very very often and that also make sense because language is used for having some efficient communication. So, you cannot have long words that are very very common then you will have to write a lot. So, the words that are very very common are generally short.

So, we talked about Zipf's law. So, for doing the processing of language what is the impact of Zipf's law? So, we said there is a good part in and there is a bad part. So, what is a good part? So, in the language we saw that, remember there were 2 important observations that we made from a Zipf's law what were those? So, one observation was that 50 percent of the vocabulary; that means, the types occur only once. So, they are very rare words and this is most words are rare words, this 50 percent of the words occur only once and what is their observation? You said that roughly 100 words account for 50 percent of the tokens. So, that is from common words are very very common, they occur a lot. So, words are very common they (Refer Time: 29:04) of 50 percent of your text.

So, now given these observations, what do you think is the good part, what is a bad part when it comes to an end? So this is a bad part, why is it the bad part? What it says is that, so 50 percent of my words in the vocabulary they can only once, it is not easy for me to gather some statistics to do some meaning analysis. So, there are only once that is all the knowledge I have about these words, so that is why this is called the bad part, I cannot do lot of modeling for these words. And the other part that roughly 100 words account for 50 percent of token, it is we call sometimes the good part why? Because for most of the analysis you can always remove these words and that will reduce your total number of tokens by half; so we will have to process roughly half of the text, because half of the text is simply these stop words and the function words.

So, that is what I mean here, the good part is that the stop words account for a large fraction of text, thus if I eliminate the stop words it reduces the size of my text roughly by half and the bad part is most of the words are extremely rare and it is very difficult to gather data to do some meaningful analysis with each verb.

(Refer Slide Time: 30:33)



**Vocabulary Growth**

How does the size of the overall vocabulary (number of unique words) grow with the size of the corpus?

*Heaps' Law*

Let  $|V|$  be the size of vocabulary and  $N$  be the number of tokens.

$$|V| = KN^\beta$$

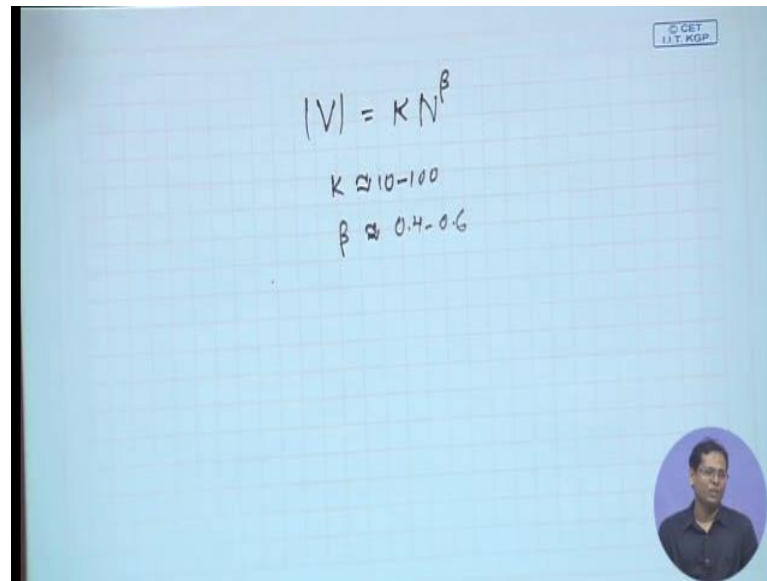
Typically

- $K \approx 10-100$
- $\beta \approx 0.4-0.6$  (roughly square root)

Pawan Goyal (IIT Khargpur) Empirical Laws Week 1: Lex

So, now we will talk about one more law here. So, that is the relation between the size of my vocabulary and the number of tokens in my text, so that is how does the size of my vocabulary go with the size of my corpus. So, we talked about type token ratio, so that is one sort of analysis that we can do, but that does not give me a particular law. So, what will be the rough distribution ought, rough relationship between them.

(Refer Slide Time: 31:27)

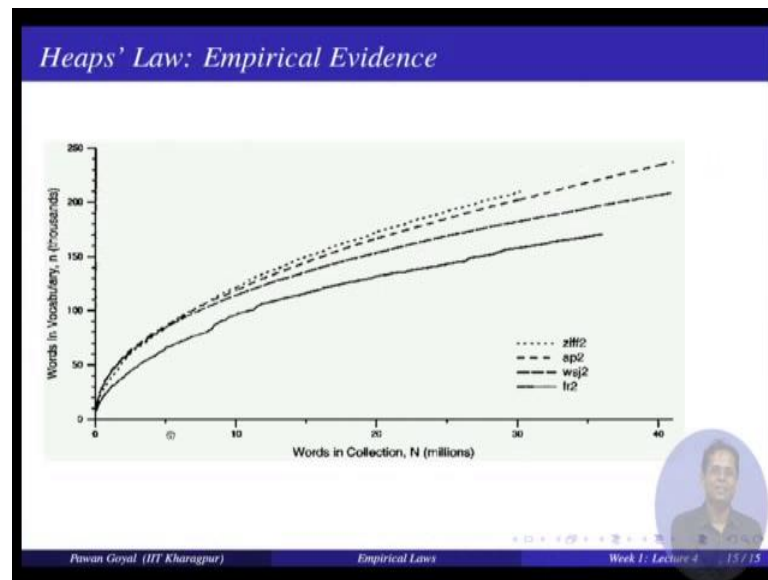


The slide displays the equation  $|V| = KN^\beta$  on a grid background. Below the equation, the ranges for the constants are given:  $K \approx 10-100$  and  $\beta \approx 0.4-0.6$ . A small logo in the top right corner reads "© CET IIT KGP". A circular inset in the bottom right corner shows a man speaking.

So, we have Heap's law they tries to give the relationship between the size of my vocabulary that we call as  $V$  and the number of tokens  $N$ ; what this law says that it gives a symbol relationship, it says the size of my vocabulary is some constant times  $K$ ,  $N$  is a number of tokens to the power beta and the ranges are also observing the corpus, so  $K$  has been found to be roughly in the range of 10 to 100, although this can vary this is our roughly this has been found to be in this range and beta is very close to the square root. So, it is found in the range of 0.4 to 0.6.

So, what it says, vocabulary generally grows as per the square root of the number of tokens. So, that also make sense remember we are talking about what happens if I keep on teaching my vocabulary, I will tend to use the sorry, keep on teaching my tokens, I will tend to use the same word again and again. So, number of unique tokens or unique types in each roughly to the square root of my corpus size.

(Refer Slide Time: 32:22)



So, what is some sort of empirical evidence for that? So, there are 4 different corpus that has been taken. So, like moisture (Refer Time: 32:30) wsj and there are some other different new corpus and what this plot shows on the x axis, you have the words in collection in millions. So, how many words in my collection? So, there are 10 million, 20 million, 30 million, 40 million. On the y axis it is showing the number of words in a vocabulary in thousand, so 50000, 100000, 150000 and so on. And the plot shows the relationship between the vocabulary and the number of tokens and you see at the plot shows roughly a square root behavior and this has been found to match with whatever he proposed. So, this is again an empirical law that is very very universal to the language.

So, that is for this lecture and in the next lecture we will start talking about some actual task related to the preprocessing of (Refer Time: 33:28) the corpus.

Thank you.