QUIZ 1: print notes now!
(if you want to :)

CS 333: Natural Language Processing

Fall 2025

Prof. Carolyn Anderson
Wellesley College

Reminders

- HW 1 is due on Thursday
- I have help hours from 4-5 on Thursday
- Caroline has help hours today from 7:30-8:30
- Ashley has help hours tomorrow from 4-5
- All help hours are in W423

Information Theory

What is language about?

- A critical function of language is to communicate information.

Information theory is the study of how information is stored and exchanged (communicated).

Today we will explore some hypotheses about language as an *efficient information communication system*.

Communicative efficiency

Communicative efficiency hypothesis:
More predictable meanings are expressed with shorter / faster forms because this leads to efficient communication.

Function words	
was	esophagus
is	asparagus
in	umbrella
to	sphinx

Communicative robustness

Communicative robustness hypothesis:
More predictable meanings are expressed with shorter / faster forms because it is important for infrequent meanings to be expressed in a way that is robust to error.

Probability review

- ***Probability: p(X)***

How likely an event is to occur.

- ***Probability distribution:***

A description of a phenomenon in terms of the probabilities of all possible outcomes. Sums to 1.

- ***Conditional probability: p(X|Y)***

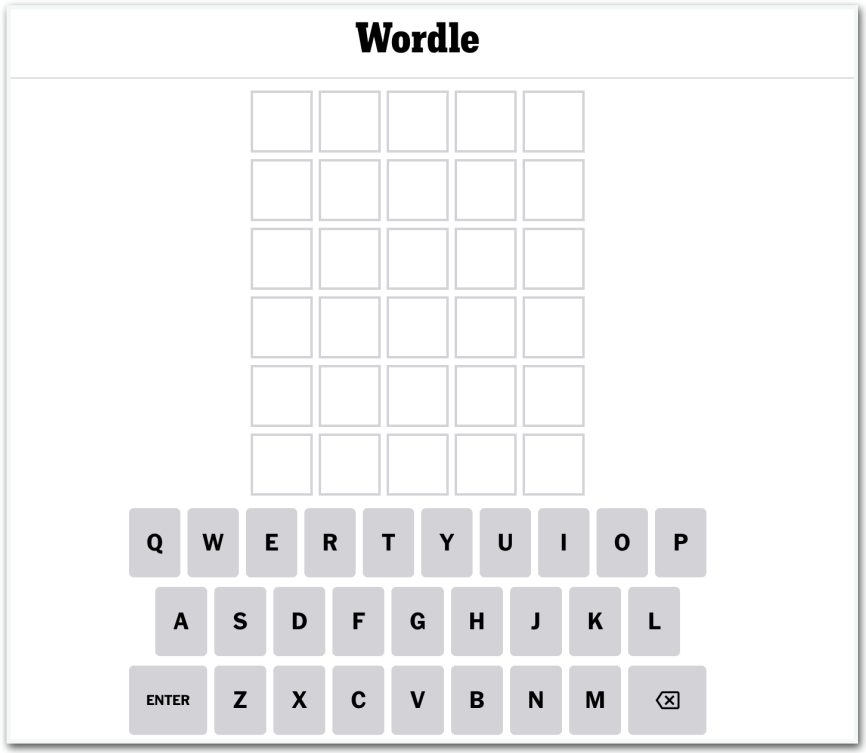
The chance of event X occurring given that event Y occurs. If events are truly independent, $p(X|Y) = p(X)$.

- ***Joint probability: p(x,y)***

The chance of event X and event Y both occurring. If events are truly independent, $p(X,Y) = p(X)p(Y)$.

Statistics in language

- You have implicit knowledge about the ***probability*** of letters in English.
- You also have implicit knowledge about the ***conditional probability*** of letters in English.



Estimating probability by

Sample text:

"on wednesdays, we wear pink."

Total count of letters:

p(w) = $3/22$

p(e) = $4/22$

p(e | w) = $\frac{\text{total count of e after w}}{\text{total count of w}} = 3/3$

p(w,e) = $p(e|w)\,p(w)$

$3/3 \cdot 3/22 = 3/22$

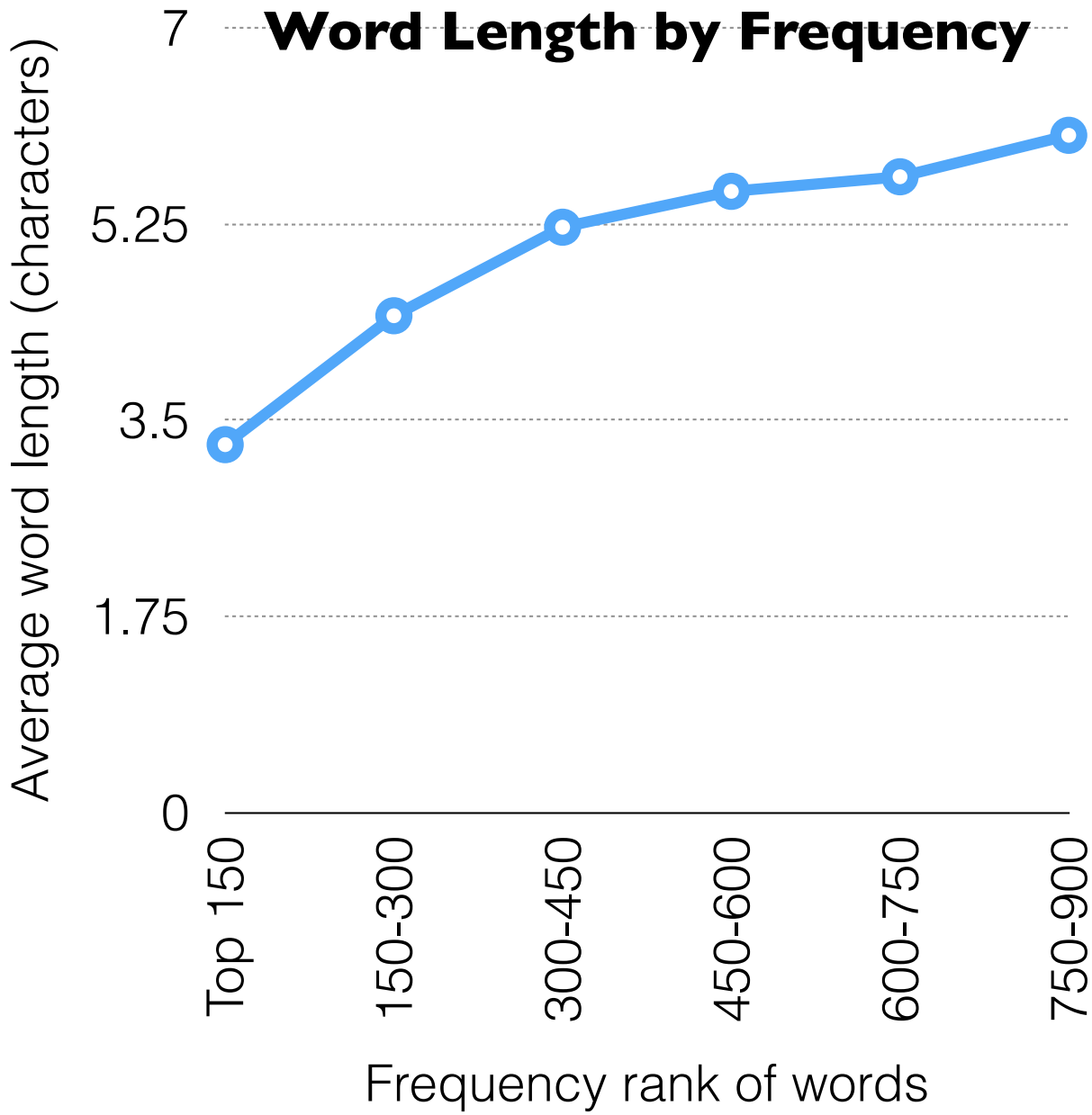
Zipfian hypotheses

Zipf's law: The frequency of a word is inversely proportional to its frequency ranking.

We'll look into this next class!

Zipf's hypothesis:
Shorter words are more frequent because languages maximize efficiency: they assign common meanings to words that take less effort to produce.

Zipfian hypotheses



Zipf's hypothesis: Shorter words are more frequent because languages maximize efficiency.

Statistics from the Brown corpus

Communicative efficiency

How can we code meanings efficiently?

Imagine I have a bag of marbles with three colors: blue, red, and green. There are twice as many red marbles as blue and twice as many blue as green.

I am going to close my eyes, pick a marble out of the bag, and I want you to yell out what color it is.

Here's the trick: the only words you can yell are **SNUFFLEUPAGUS**, **SHAMBLE**, and **SQUEAK**.

Communicative efficiency

How can we code meanings efficiently?

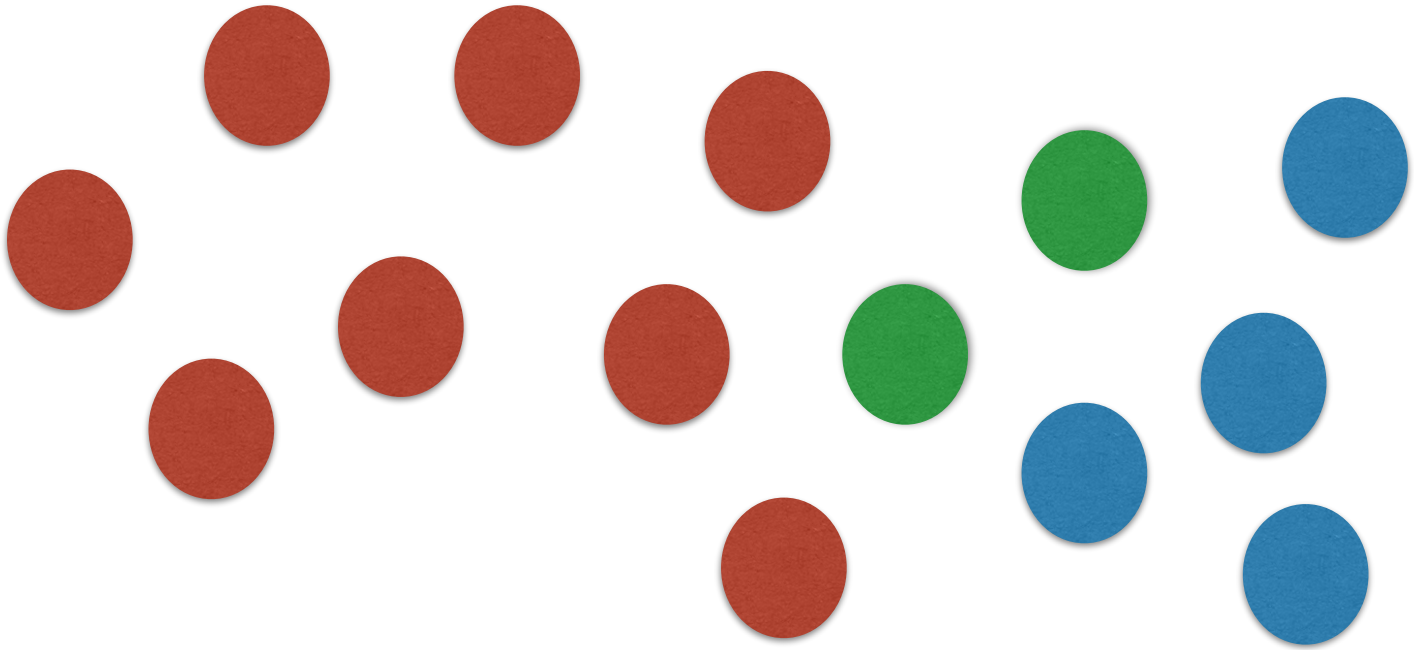
Imagine I have a bag of marbles with three colors: blue, red, and green. There are twice as many red marbles as blue and twice as many blue as green.

I am going to close my eyes, pick a marble out of the bag, and I want you to yell out what color it is.

Here's the trick: the only words you can yell are **SNUFFLEUPAGUS**, **SHAMBLE**, and **SQUEAK**. And we want to do this as **fast as possible**.

Communicative efficiency

READY?

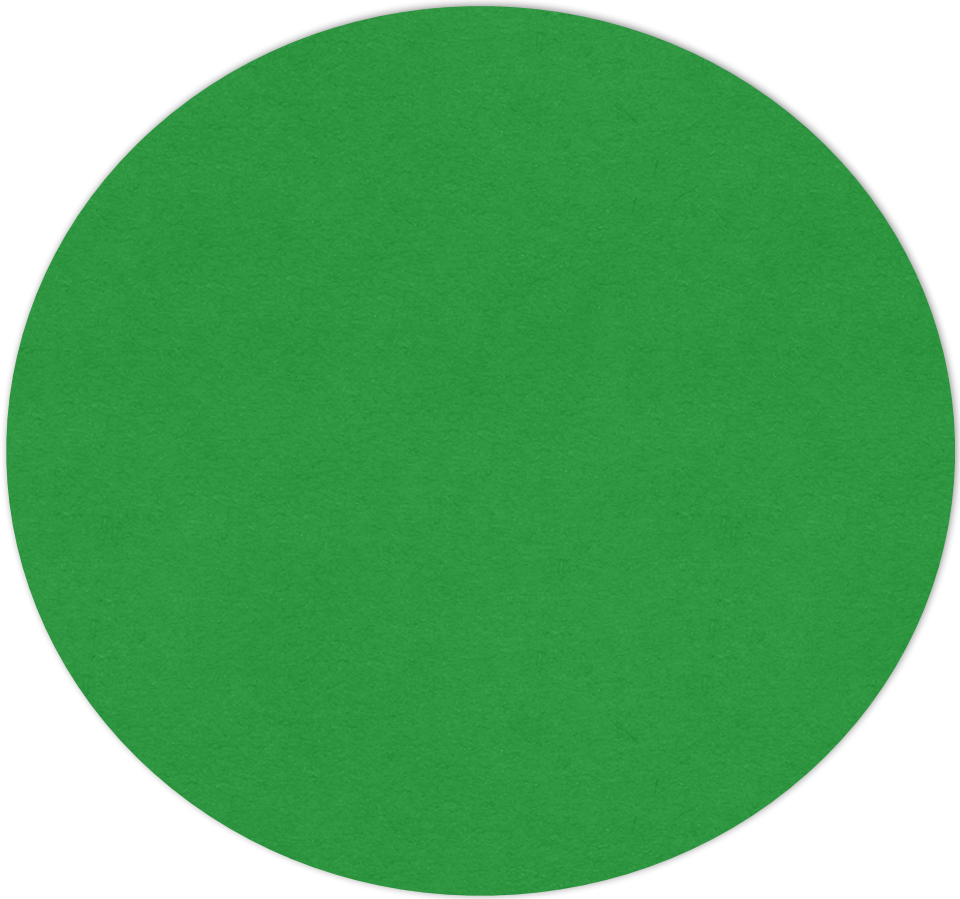


SNUFFLEUPAGUS

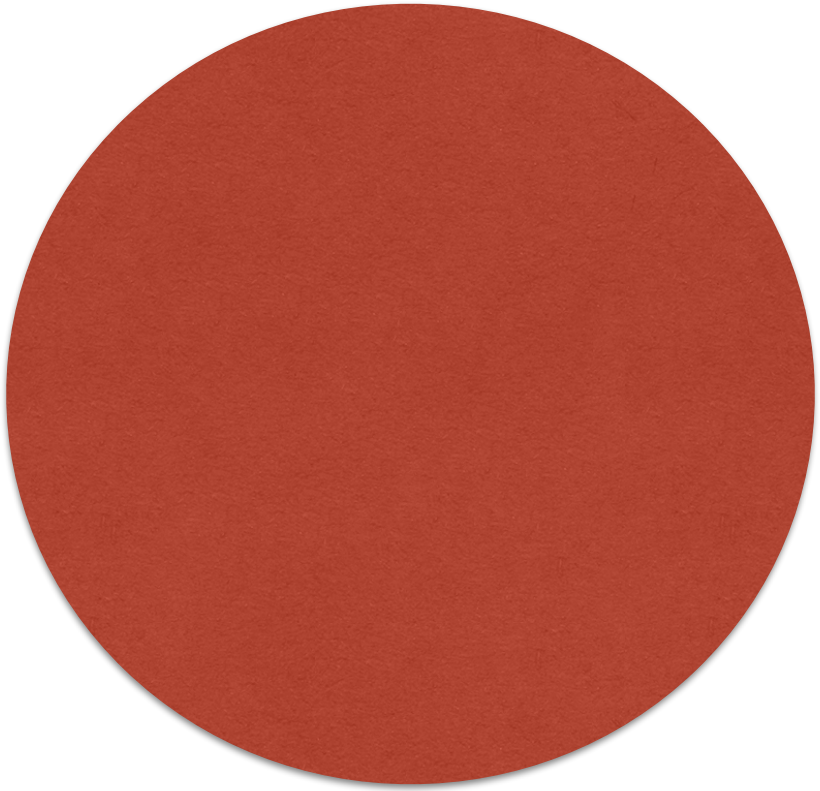
SHAMBLE

SQUEAK

Communicative efficiency

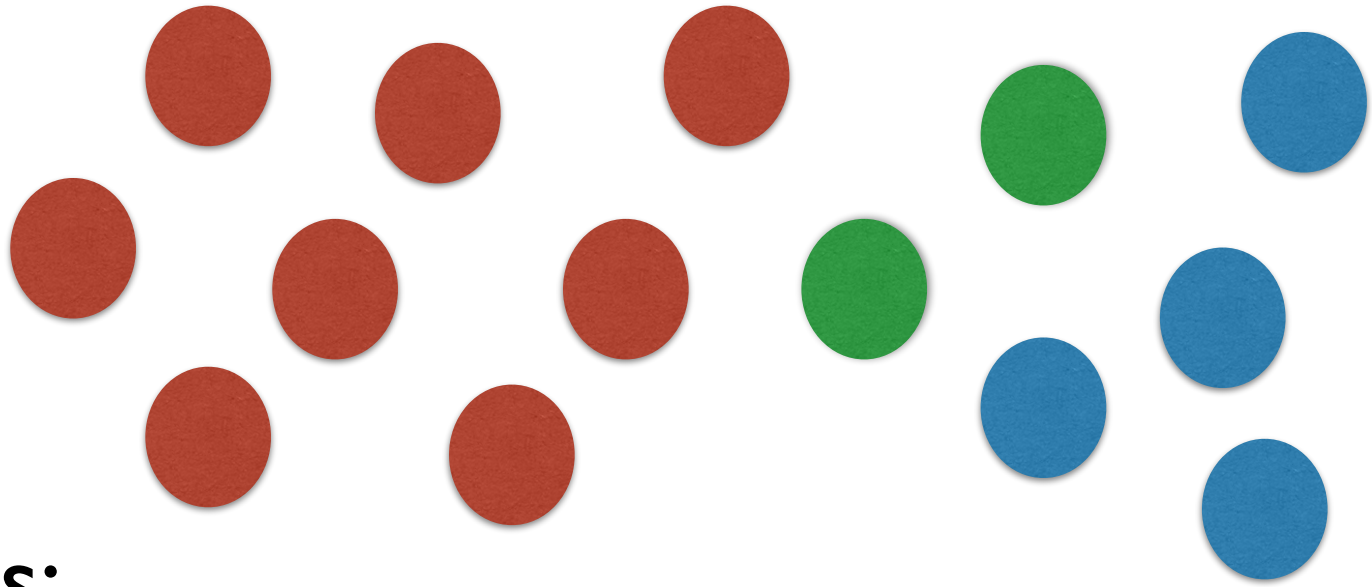


Communicative efficiency



Communicative efficiency

Distribution of marbles:



Available words:

SNUFFLEUPAGUS **SHAMBLE** **SQUEAK**

What is the most **efficient** assignment?

Communicative efficiency

Distribution of marbles:

p(red) = 8/14 = 57.1%
p(blue) = 4/14 = 28.6%
p(green) = 2/14 = 14.3%

Available words:

SNUFFLEUPAGUS **SHAMBLE** **SQUEAK**

green blue red

What is the most **efficient** assignment?

Communicative efficiency

Assuming longer words are harder to say, the best arrangement is:

p(red) = 8/14 = 57.1%
p(blue) = 4/14 = 28.6%
p(green) = 2/14 = 14.3%

Communicative efficiency

Communicative efficiency hypothesis:
More predictable meanings are expressed with shorter / faster forms because this leads to efficient communication.

Communicative robustness

Communicative robustness hypothesis:
More predictable meanings are expressed with shorter / faster forms because it is important for infrequent meanings to be expressed in a way that is robust to error.

Communicative robustness

Hypothesis: the more unlikely a word is, the worse it is to make a speech error.

Imagine we're playing the same weird marble game.

But this time, Seojean is standing there with an airhorn, making earsplitting noises at random intervals.

Communicative robustness

I'd argue that the strategy of assigning longer words to rarer colors is still a good one, but for a different reason.

Why?

Communicative robustness

I'd argue that the strategy of assigning longer words to rarer colors is still a good one, but for a different reason.

Why?

If you hear nothing but airhorn on a particular turn, what color should you guess?

Communicative robustness

If you shout SNUFFLEUPAGUS, and the airhorn blocks out one syllable, I'll probably still hear enough to know what you said.

But if you shout SQUEAK, I might not.

If your message is **rare**, and the **channel is noisy**, then it makes sense to build some **redundancy** into your message.

Noisy channel model

The Noisy Channel model:

prior beliefs

$$p(\text{meaning} \mid \text{signal}) = \frac{p(\text{signal} \mid \text{meaning})p(\text{meaning})}{p(\text{signal})}$$

my model of the speaker

listener

Bayes' rule!

Noisy channel model

$$p(\text{meaning} \mid \text{signal}) = \frac{p(\text{signal} \mid \text{meaning})p(\text{meaning})}{p(\text{signal})}$$

The girl put out a bowl of milk for her _____

$$p(\text{🐱} \mid \text{hat}) = \frac{p(\text{hat} \mid \text{🐱})\, p(\text{🐱})}{p(\text{hat})}$$

$$p(\text{🎩} \mid \text{hat}) = \frac{p(\text{hat} \mid \text{🎩})\, p(\text{🎩})}{p(\text{hat})}$$

Noisy channel model

$$p(\text{meaning} \mid \text{signal}) = \frac{p(\text{signal} \mid \text{meaning})p(\text{meaning})}{p(\text{signal})}$$

The girl put out a bowl of milk for her _____

Probabilities of meanings:

p(😺) = 0.99

p(🎩) = 0.01

Implicitly, these are **conditioned** on the context, but we'll ignore this for now.

Noisy channel model

$$p(\text{meaning} \mid \text{signal}) = \frac{p(\text{signal} \mid \text{meaning})p(\text{meaning})}{p(\text{signal})}$$

p(signal|meaning): probability of pronunciation given meaning (speech error rate)

Let's assume a 5% error rate

p("mat" | 😺) = 0.05

p("cat" | 😺) = 0.95

p("cat" | 🐈) = 0.05

p("mat" | 🐈) = 0.95

Noisy channel model

$$p(\text{meaning} \mid \text{signal}) = \frac{p(\text{signal} \mid \text{meaning})p(\text{meaning})}{p(\text{signal})}$$

p(signal|meaning): probability of pronunciation given meaning

Let's assume a 5% error rate:

p("hat" | 😺) = 0.05 p("hat" | 🎩) = 0.95

p("cat"| 😺) = 0.95 p("cat" | 🎩) = 0.05

Noisy channel model

$$p(\text{meaning} \mid \text{signal}) = \frac{p(\text{signal} \mid \text{meaning})p(\text{meaning})}{p(\text{signal})}$$

p(signal) : ignore because we care about the relative probability

Noisy channel model

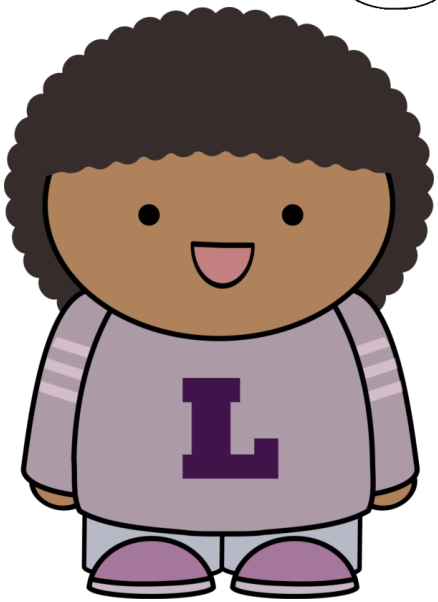
$$p(\text{meaning} \mid \text{signal}) = \frac{p(\text{signal} \mid \text{meaning})p(\text{meaning})}{p(\text{signal})}$$

p(signal) : we can ignore this, because we care about the relative probability of the meanings given the same signal.

Noisy channel model

The girl put out a bowl of milk for her hat.

hmm...



Noisy channel model

$$p(\text{meaning} \mid \text{signal}) = \frac{p(\text{signal} \mid \text{meaning})p(\text{meaning})}{p(\text{signal})}$$

p(🐱| "hat") = $\frac{p(\text{"hat"} \mid 🐱)\, p(🐱)}{p(\text{"hat"})}$

p(🎩| "hat") = $\frac{p(\text{"hat"} \mid 🎩)\, p(🎩)}{p(\text{"hat"})}$

Noisy channel model

$$p(\text{meaning} \mid \text{signal}) = \frac{p(\text{signal} \mid \text{meaning})p(\text{meaning})}{p(\text{signal})}$$

p(😺 | "hat") ∝ $0.05 \cdot 0.99 = 0.0495$

p(🎩 | "hat") ∝ $0.95 \cdot 0.01 = 0.0095$

Noisy channel model

$$p(\text{meaning} \mid \text{signal}) = \frac{p(\text{signal} \mid \text{meaning})p(\text{meaning})}{p(\text{signal})}$$

$$p(🐱 \mid \text{"hat"}) \propto (0.05)p(🐱)$$

$$p(🎩 \mid \text{"hat"}) \propto (0.95)p(🎩)$$

Noisy channel model

$$p(\text{meaning} \mid \text{signal}) = \frac{p(\text{signal} \mid \text{meaning})p(\text{meaning})}{p(\text{signal})}$$

p(🐱| "hat") ∝

p(🎩| "hat") ∝

In a noisy channel model, our prior belief can overcome the signal we receive.

Noisy channel model

The girl put out a bowl of milk for her _____

p(🐱| "hat") ∝ 0.0495

p(🎩| "hat") ∝ 0.0095

If our intended message is 🐱, making a speech error isn't so bad— our listener will land on the correct message anyway.

Noisy channel model

The girl put out a bowl of milk for her _____

p(🐱| "hat") ∝ 0.0495

p(🎩| "hat") ∝ 0.0095

If our intended message is rare, making a speech error is bad — our listener's prior belief in the unlikeliness of the message makes it hard to communicate that message, even if we produce it perfectly.

Noisy channel model

Another reason that assigning longer words to rarer meanings makes sense is for *communicative robustness*:
a longer word is more robust to error on a single phoneme, because there are more phonemes.

Summary

→ **Zipf's Law:** the frequency of a word is inversely proportional to its rank in the frequency table

→ **Zipf's Hypothesis:** shorter words are used for more frequent meanings because they are more efficient.

→ **Communicative efficiency:** languages evolve to express information efficiently

→ **Communicative robustness:** languages evolve to express information in a noise-tolerant way

Why Do We Care About This?

Understanding the information theory of natural language helps us design efficient techniques for text encoding and processing.

Next class: **TOKENIZATION**

How do we efficiently split text into useful chunks?