
Basics in Language and Probability

Philipp Koehn

31 August 2023



Quotes



1

It must be recognized that the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term.

Noam Chomsky, 1969

Whenever I fire a linguist our system performance improves.

Frederick Jelinek, 1988

Conflicts?



rationalist vs. empiricist

scientist vs. engineer

insight vs. data analysis

explaining language vs. building applications

language

A Naive View of Language



4

- Language needs to name
 - nouns: objects in the world (**dog**)
 - verbs: actions (**jump**)
 - adjectives and adverbs: properties of objects and actions (**brown**, **quickly**)
- Relationship between these have to specified
 - word order
 - morphology
 - function words

A Bag of Words



quick

fox

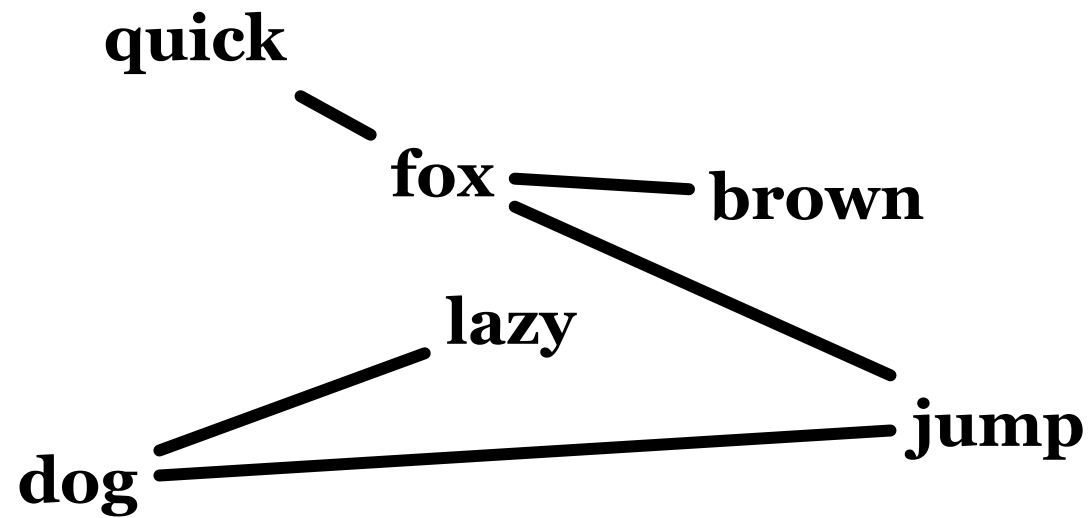
brown

lazy

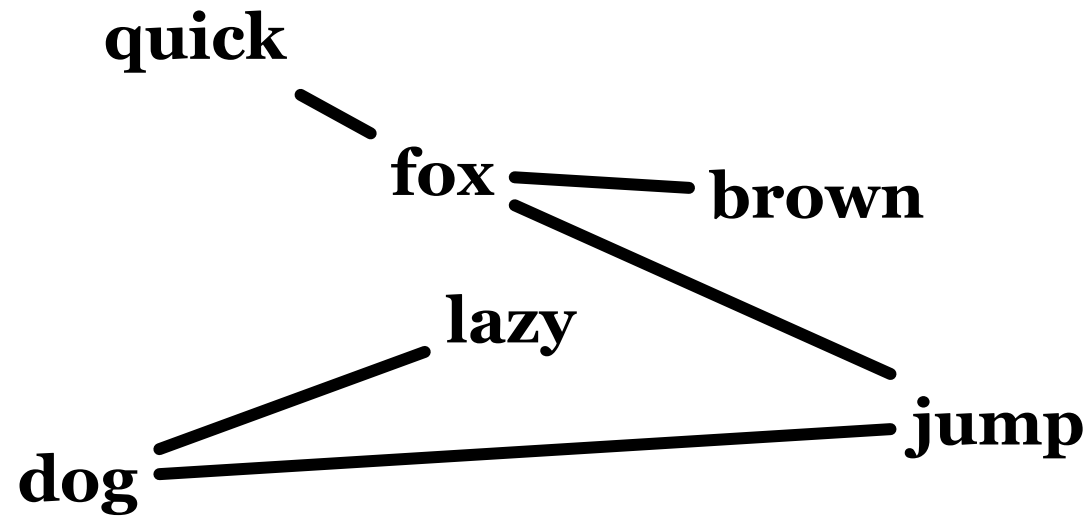
jump

dog

Relationships



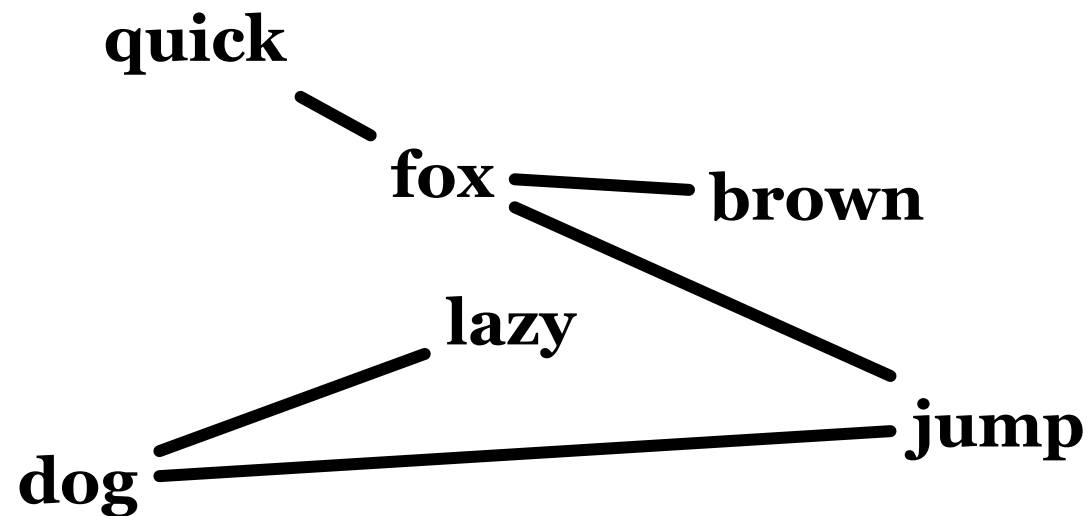
Marking of Relationships: Word Order



quick brown fox jump lazy dog

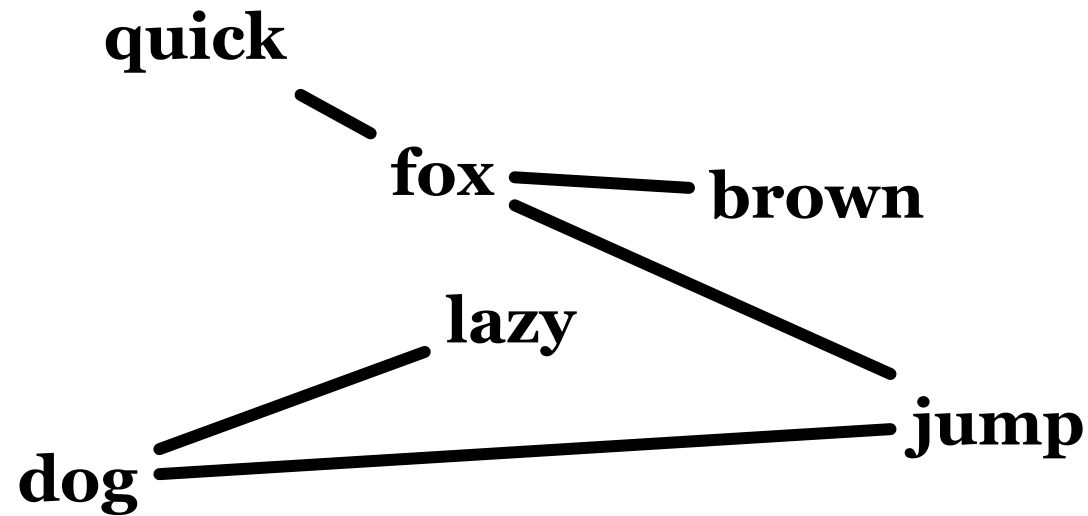
Marking of Relationships: Function Words

8



quick brown fox jump **over** lazy dog

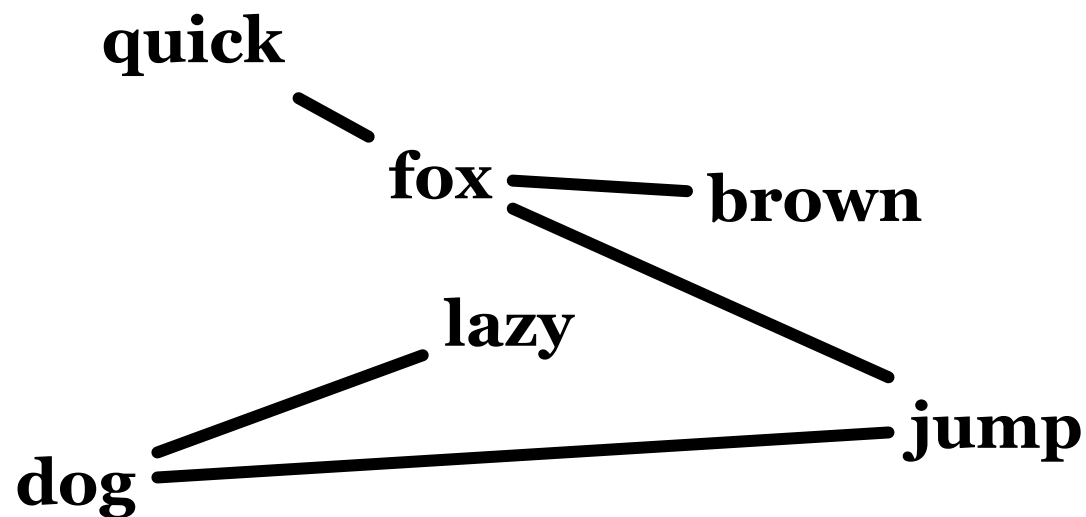
Marking of Relationships: Morphology



quick brown fox jump_s over lazy dog

Some Nuance

10



the quick brown fox jumps over the lazy dog

Marking of Relationships: Agreement

11



- From Catullus, First Book, first verse (Latin):

Cui dono lepidum novum libellum arida modo pumice expolitur ?
Whom I-present lovely new little-book dry manner pumice polished ?

(To whom do I present this lovely new little book now polished with a dry pumice?)

- Gender (and case) agreement links adjectives to nouns

Marking of Relationships to Verb: Case

12



- German:

| | | | |
|-----------|-------|-----------------|-----------|
| Die Frau | gibt | dem Mann | den Apfel |
| The woman | gives | the man | the apple |
| subject | | indirect object | object |

| | | | |
|-----------------|-------|----------|-----------|
| Der Frau | gibt | der Mann | den Apfel |
| The woman | gives | the man | the apple |
| indirect object | | subject | object |

- Case inflection indicates role of noun phrases

Case Morphology vs. Prepositions

- Two different word orderings for English:

- The woman gives the man the apple
- The woman gives the apple **to** the man

- Japanese:

女の人 は 男の人に リンゴを あげます
woman TOPIC man OBJ2 apple OBJ1 give

- Is there a real difference between prepositions and noun phrase case inflection?

Words

14



This is a simple sentence **WORDS**

This is a simple sentence

be
3sg
present

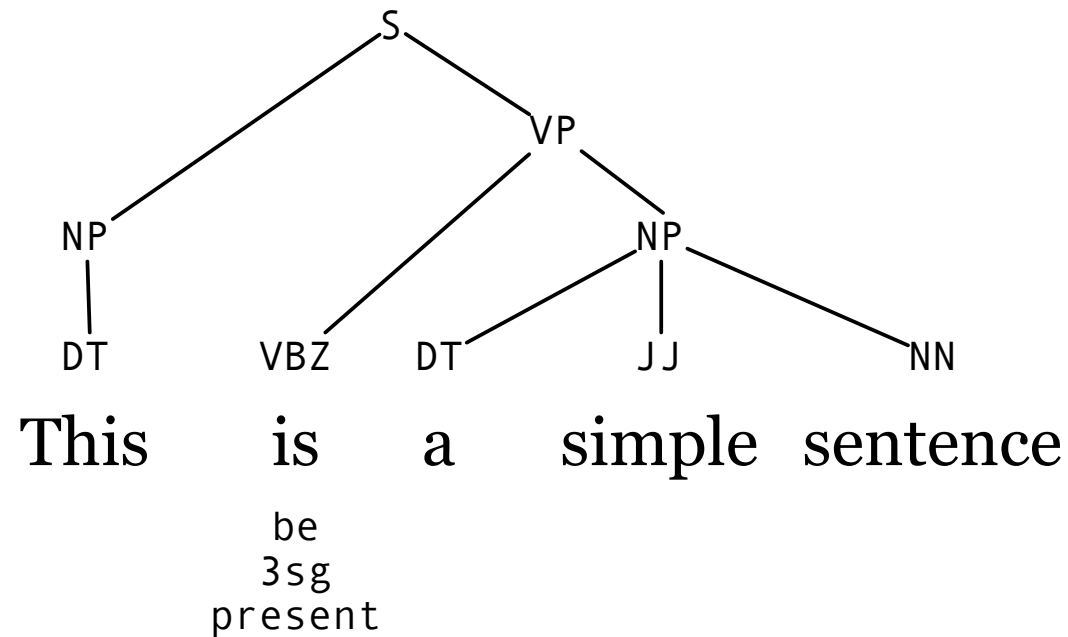
WORDS
MORPHOLOGY

Parts of Speech

16



| DT | VBZ | DT | JJ | NN | PART OF SPEECH |
|------|----------------------|----|--------|----------|----------------|
| This | is | a | simple | sentence | WORDS |
| | be 3sg present | | | | MORPHOLOGY |

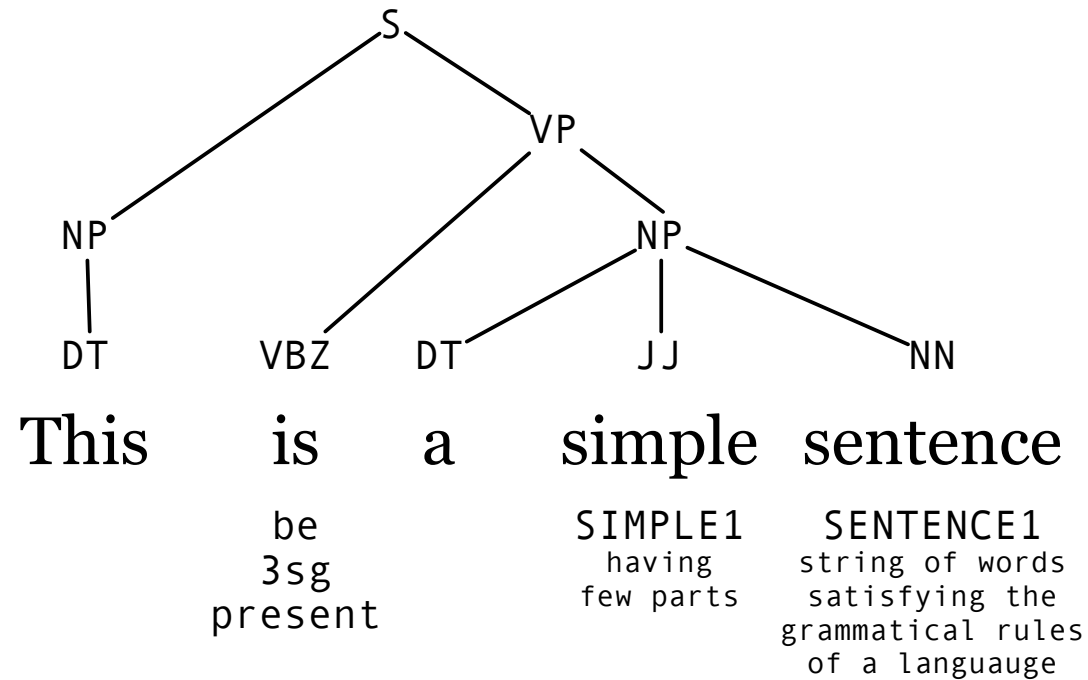


SYNTAX

PART OF SPEECH

WORDS

MORPHOLOGY



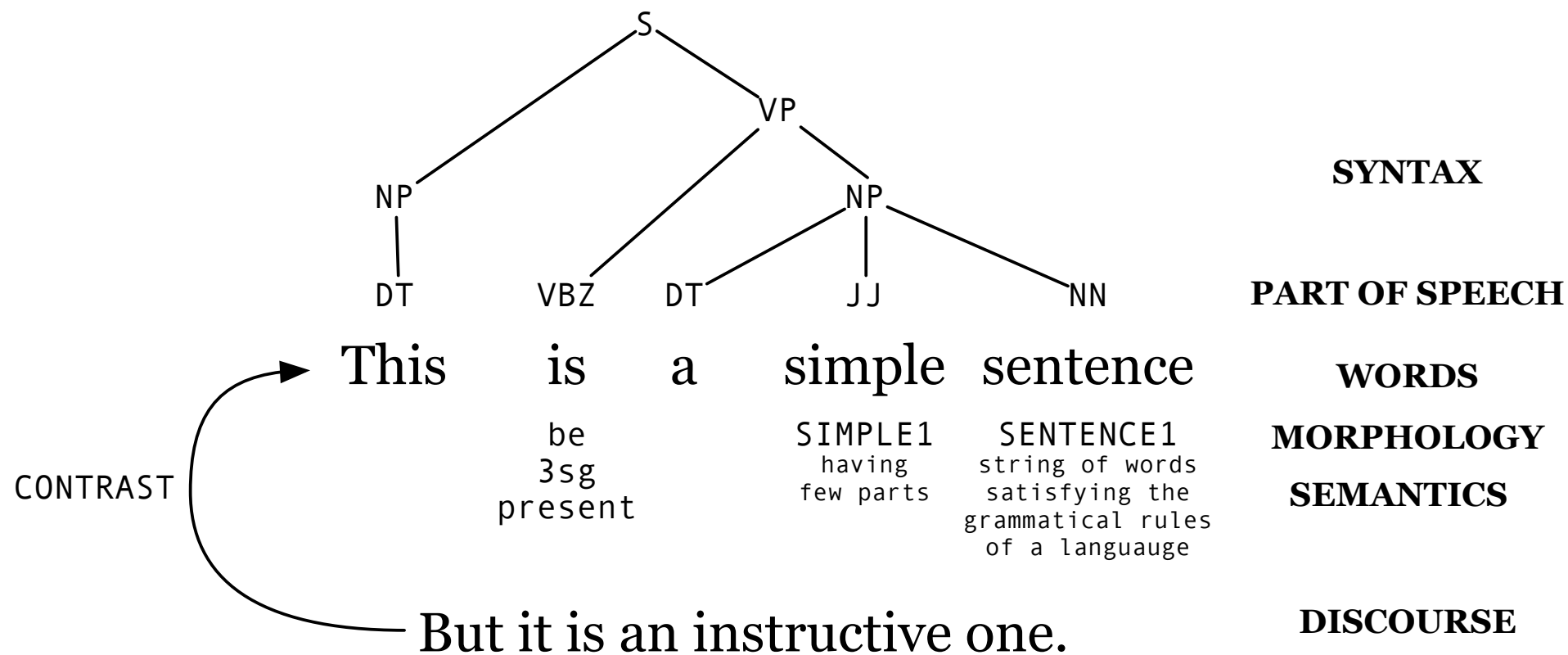
SYNTAX

PART OF SPEECH

WORDS

MORPHOLOGY

SEMANTICS



Why is Language Hard?

- Ambiguities on many levels
- Rules, but many exceptions
- No clear understand how humans process language
- Can we learn everything about language by automatic data analysis?



data

- Definition: strings of letters separated by spaces
- But how about:
 - punctuation: commas, periods, etc. typically separated (tokenization)
 - hyphens: **high-risk**
 - clitics: **Joe's**
 - compounds: **website, Computerlinguistikvorlesung**
- And what if there are no spaces:

伦敦每日快报指出,两台记载黛安娜王妃一九九七年巴黎死亡车祸调查资料的手提电脑,被从前大都会警察总长的办公室里偷走.

Word Counts

Most frequent words in the English Europarl corpus

any word

| Frequency in text | Token |
|-------------------|-------|
| 1,929,379 | the |
| 1,297,736 | , |
| 956,902 | . |
| 901,174 | of |
| 841,661 | to |
| 684,869 | and |
| 582,592 | in |
| 452,491 | that |
| 424,895 | is |
| 424,552 | a |

nouns

| Frequency in text | Content word |
|-------------------|--------------|
| 129,851 | European |
| 110,072 | Mr |
| 98,073 | commission |
| 71,111 | president |
| 67,518 | parliament |
| 64,620 | union |
| 58,506 | report |
| 57,490 | council |
| 54,079 | states |
| 49,965 | member |

But also:

There is a large tail of words that occur only once.

33,447 words occur once, for instance

- cornflakes
- mathematicians
- Tazhikistan

Zipf's law

$$f \times r = k$$

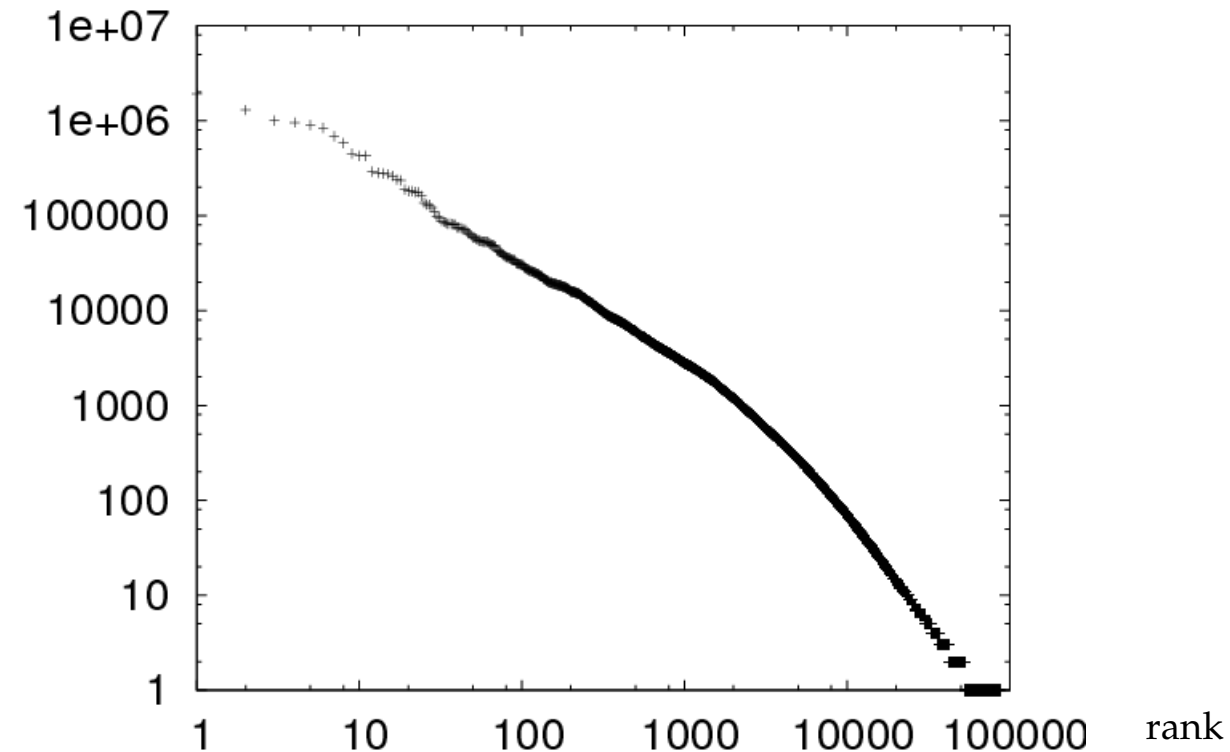
f = frequency of a word

r = rank of a word (if sorted by frequency)

k = a constant

Zipf's law as a graph

frequency



Why a line in log-scale?

$$fr = k \Rightarrow f = \frac{k}{r} \Rightarrow \log f = \log k - \log r$$

statistics

Probabilities

- Given word counts we can estimate a probability distribution:

$$P(w) = \frac{\text{count}(w)}{\sum_{w'} \text{count}(w')}$$

- Given word counts we can estimate a probability distribution:

$$P(w) = \frac{\text{count}(w)}{\sum_{w'} \text{count}(w')}$$

- This type of estimation is called *maximum likelihood estimation*. Why? We will get to that later.

Probabilities

- Given word counts we can estimate a probability distribution:

$$P(w) = \frac{\text{count}(w)}{\sum_{w'} \text{count}(w')}$$

- This type of estimation is called *maximum likelihood estimation*. Why? We will get to that later.
- Estimating probabilities based on frequencies is called the *frequentist approach* to probability.

- Given word counts we can estimate a probability distribution:

$$P(w) = \frac{\text{count}(w)}{\sum_{w'} \text{count}(w')}$$

- This type of estimation is called *maximum likelihood estimation*. Why? We will get to that later.
- Estimating probabilities based on frequencies is called the *frequentist approach* to probability.
- This probability distribution answers the question: If we randomly pick a word out of a text, how likely will it be word w ?

- We introduce a **random variable** W .
- We define a **probability distribution** p , that tells us how likely the variable W is the word w :

$$\text{prob}(W = w) = p(w)$$

- Sometimes, we want to deal with two random variables at the same time.
- Example: Words w_1 and w_2 that occur in sequence (a **bigram**)
We model this with the distribution: $p(w_1, w_2)$
- If the occurrence of words in bigrams is **independent**, we can reduce this to $p(w_1, w_2) = p(w_1)p(w_2)$. Intuitively, this not the case for word bigrams.
- We can estimate **joint probabilities** over two variables the same way we estimated the probability distribution over a single variable:

$$p(w_1, w_2) = \frac{\text{count}(w_1, w_2)}{\sum_{w_1', w_2'} \text{count}(w_1', w_2')}$$

- Another useful concept is **conditional probability**

$$p(w_2|w_1)$$

It answers the question: If the random variable $W_1 = w_1$, how what is the value for the second random variable W_2 ?

- Mathematically, we can define conditional probability as

$$p(w_2|w_1) = \frac{p(w_1, w_2)}{p(w_1)}$$

- If W_1 and W_2 are independent: $p(w_2|w_1) = p(w_2)$

- A bit of math gives us the chain rule:

$$p(w_2|w_1) = \frac{p(w_1, w_2)}{p(w_1)}$$

$$p(w_1) p(w_2|w_1) = p(w_1, w_2)$$

- What if we want to break down large joint probabilities like $p(w_1, w_2, w_3)$?

We can repeatedly apply the chain rule:

$$p(w_1, w_2, w_3) = p(w_1) p(w_2|w_1) p(w_3|w_1, w_2)$$

- Finally, another important rule: **Bayes rule**

$$p(x|y) = \frac{p(y|x) p(x)}{p(y)}$$

- It can easily derived from the chain rule:

$$p(x, y) = p(x, y)$$

$$p(x|y) p(y) = p(y|x) p(x)$$

$$p(x|y) = \frac{p(y|x) p(x)}{p(y)}$$

- We introduced the concept of a random variable X

$$\text{prob}(X = x) = p(x)$$

- Example: Roll of a dice. There is a $\frac{1}{6}$ chance that it will be 1, 2, 3, 4, 5, or 6.

- We define the **expectation** $E(X)$ of a random variable as:

$$E(X) = \sum_x p(x) x$$

- Roll of a dice:

$$E(X) = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 = 3.5$$

- **Variance** is defined as

$$\text{Var}(X) = E((X - E(X))^2) = E(X^2) - E^2(X)$$

$$\text{Var}(X) = \sum_x p(x) (x - E(X))^2$$

- Intuitively, this is a measure how far events diverge from the mean (expectation)
- Related to this is **standard deviation**, denoted as σ .

$$\text{Var}(X) = \sigma^2$$

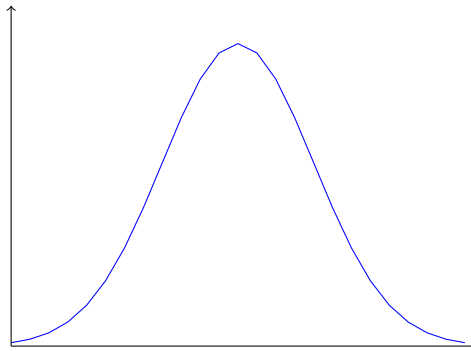
$$E(X) = \mu$$

- Roll of a dice:

$$\begin{aligned} \text{Var}(X) &= \frac{1}{6}(1 - 3.5)^2 + \frac{1}{6}(2 - 3.5)^2 + \frac{1}{6}(3 - 3.5)^2 \\ &\quad + \frac{1}{6}(4 - 3.5)^2 + \frac{1}{6}(5 - 3.5)^2 + \frac{1}{6}(6 - 3.5)^2 \\ &= \frac{1}{6}((-2.5)^2 + (-1.5)^2 + (-0.5)^2 + 0.5^2 + 1.5^2 + 2.5^2) \\ &= \frac{1}{6}(6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25) \\ &= 2.917 \end{aligned}$$

- **Uniform:** all events equally likely
 - $\forall x, y : p(x) = p(y)$
 - example: roll of one dice
- **Binomial:** a series of trials with only two outcomes
 - probability p for each trial, occurrence r out of n times:
$$b(r; n, p) = \binom{n}{r} p^r (1 - p)^{n-r}$$
 - a number of coin tosses

- **Normal**: common distribution for continuous values
 - value in the range $[-\infty, x]$, given expectation μ and standard deviation σ :
$$n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$$
 - also called **Bell curve**, or **Gaussian**
 - examples: heights of people, IQ of people, tree heights, ...



- We introduced previously an estimation of probabilities based on frequencies:

$$P(w) = \frac{\text{count}(w)}{\sum_{w'} \text{count}(w')}$$

- Alternative view: Bayesian: what is the most likely model given the data

$$p(M|D)$$

- Model and data are viewed as random variables
 - model M as random variable
 - data D as random variable

- Reformulation of $p(M|D)$ using Bayes rule:

$$p(M|D) = \frac{p(D|M) p(M)}{p(D)}$$

$$\operatorname{argmax}_M p(M|D) = \operatorname{argmax}_M p(D|M) p(M)$$

- $p(M|D)$ answers the question: What is the most likely model given the data
- $p(M)$ is a prior that prefers certain models (e.g. simple models)
- The frequentist estimation of word probabilities $p(w)$ is the same as Bayesian estimation with a uniform prior (no bias towards a specific model), hence it is also called the **maximum likelihood estimation**

- An important concept is **entropy**:

$$H(X) = \sum_x -p(x) \log_2 p(x)$$

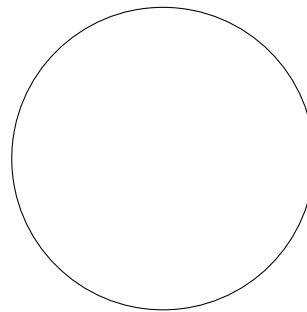
- A measure for the degree of disorder

Entropy Example

One event

$$p(a) = 1$$

$$\begin{aligned} H(X) &= -1 \log_2 1 \\ &= 0 \end{aligned}$$



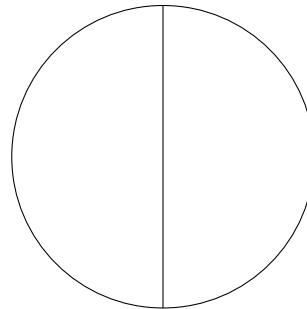
Entropy Example

2 equally likely events:

$$p(a) = 0.5$$

$$p(b) = 0.5$$

$$\begin{aligned} H(X) &= -0.5 \log_2 0.5 - 0.5 \log_2 0.5 \\ &= -\log_2 0.5 \\ &= 1 \end{aligned}$$



Entropy Example

4 equally likely events:

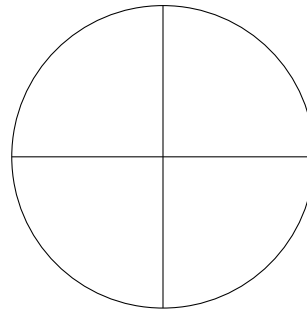
$$p(a) = 0.25$$

$$p(b) = 0.25$$

$$p(c) = 0.25$$

$$p(d) = 0.25$$

$$\begin{aligned} H(X) &= -0.25 \log_2 0.25 - 0.25 \log_2 0.25 \\ &\quad - 0.25 \log_2 0.25 - 0.25 \log_2 0.25 \\ &= -\log_2 0.25 \\ &= 2 \end{aligned}$$



Entropy Example

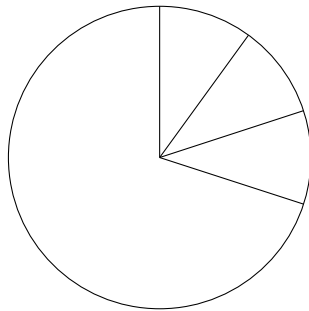
4 events, one more likely than the others:

$$p(a) = 0.7$$

$$p(b) = 0.1$$

$$p(c) = 0.1$$

$$p(d) = 0.1$$



$$\begin{aligned} H(X) &= -0.7 \log_2 0.7 - 0.1 \log_2 0.1 \\ &\quad - 0.1 \log_2 0.1 - 0.1 \log_2 0.1 \\ &= -0.7 \log_2 0.7 - 0.3 \log_2 0.1 \\ &= -0.7 \times -0.5146 - 0.3 \times -3.3219 \\ &= 0.36020 + 0.99658 \\ &= 1.35678 \end{aligned}$$

Entropy Example

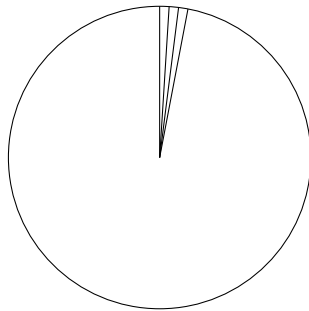
4 events, one much more likely than the others:

$$p(a) = 0.97$$

$$p(b) = 0.01$$

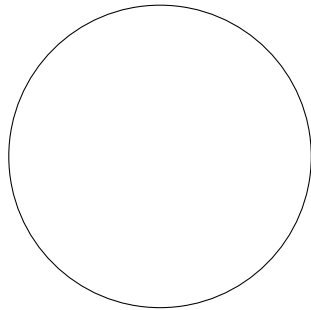
$$p(c) = 0.01$$

$$p(d) = 0.01$$

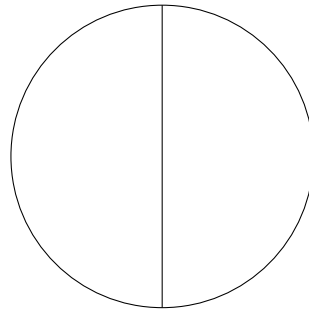


$$\begin{aligned} H(X) &= -0.97 \log_2 0.97 - 0.01 \log_2 0.01 \\ &\quad - 0.01 \log_2 0.01 - 0.01 \log_2 0.01 \\ &= -0.97 \log_2 0.97 - 0.03 \log_2 0.01 \\ &= -0.97 \times -0.04394 - 0.03 \times -6.6439 \\ &= 0.04262 + 0.19932 \\ &= 0.24194 \end{aligned}$$

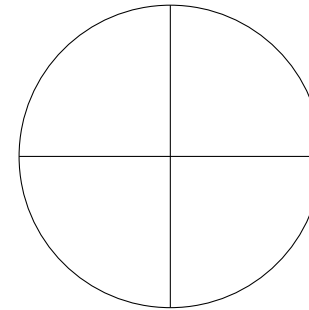
Examples



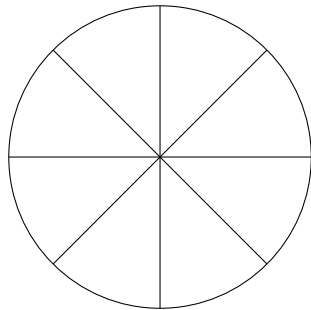
$$H(X) = 0$$



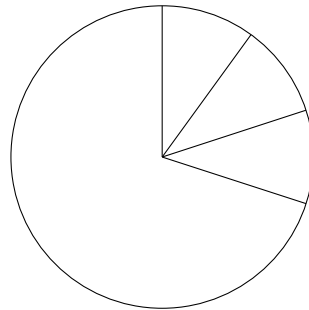
$$H(X) = 1$$



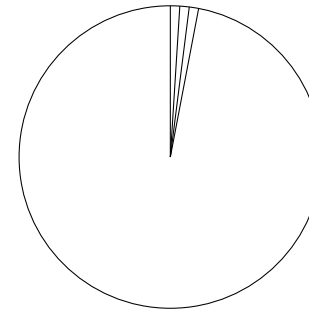
$$H(X) = 2$$



$$H(X) = 3$$



$$H(X) = 1.35678$$



$$H(X) = 0.24194$$

Intuition Behind Entropy

- A good model has low entropy

→ it is more certain about outcomes

- For instance a translation table

| e | f | $p(e f)$ |
|------|-----|----------|
| the | der | 0.8 |
| that | der | 0.2 |

is better than

| e | f | $p(e f)$ |
|------|-----|----------|
| the | der | 0.02 |
| that | der | 0.01 |
| ... | ... | ... |

- A lot of statistical estimation is about reducing entropy

- Assume that we want to encode a sequence of events X
- Each event is encoded by a sequence of bits
- For example
 - Coin flip: heads = 0, tails = 1
 - 4 equally likely events: $a = 00$, $b = 01$, $c = 10$, $d = 11$
 - 3 events, one more likely than others: $a = 0$, $b = 10$, $c = 11$
 - Morse code: e has shorter code than q
- Average number of bits needed to encode $X \geq$ entropy of X

The Entropy of English

- We already talked about the probability of a word $p(w)$
- But words come in sequence. Given a number of words in a text, can we guess the next word $p(w_n | w_1, \dots, w_{n-1})$?
- Assuming a model with a limited window size

| Model | Entropy |
|------------------|---------|
| 0th order | 4.76 |
| 1st order | 4.03 |
| 2nd order | 2.8 |
| human, unlimited | 1.3 |

Next Lecture: Language Models

- Next time, we will expand on the idea of a model of English in the form

$$p(w_n | w_1, \dots, w_{n-1})$$

- Despite its simplicity, a tremendously useful tool for NLP
- Nice machine learning challenge
 - sparse data
 - smoothing
 - back-off and interpolation