



# Semantics & Pragmatics SoSe 2021

## Lecture 2: Information Theory I

29/04/2021, Christian Bentz



---

# Overview

## Section 1: Historical Overview

## Section 2: Introduction

The Box Game

Information and Language

## Section 3: Measuring Entropy

Definition

Application to Languages

## Section 4: Two Problems and Solutions

## Summary

## References

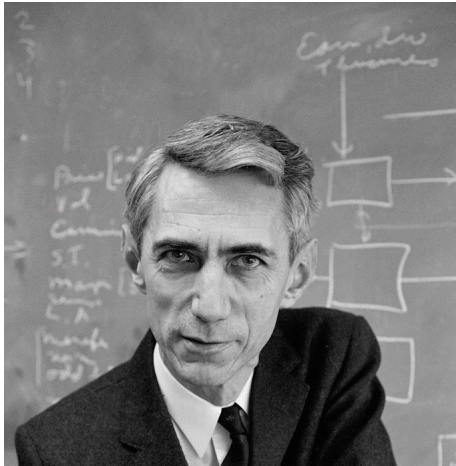


---

# Section 1: Historical Overview



# A Brief History of Information and Language



*The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. [...] **semantic aspects of communication are irrelevant to the engineering problem.** The significant aspect is that the actual message is one selected from a set of possible messages.*

Shannon, Claude E. (1948). A mathematical theory of communication, p. 1.

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References



## Example

### Article 1

All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

### Universal Declaration of Human Rights (UDHR) in English

### Raeiclt 1

Rll humrn btings rat boan fatt and tqurl in digniey rnd aighes. Ehty rat tndowtd wieh atrson rnd conscitnct rnd should rce eowrads ont rnoehta in r spiaie of baoehtahood.

### Universal Declaration of Human Rights (UDHR) in ???

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

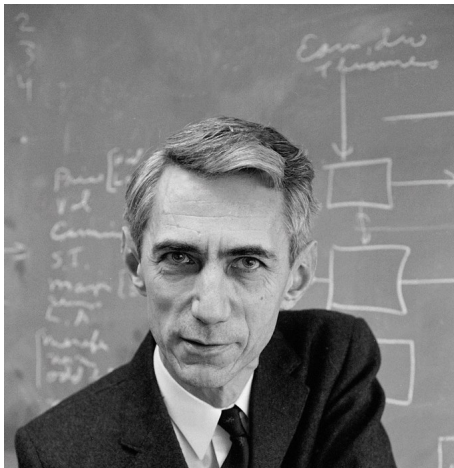
Section 4: Two  
Problems and  
Solutions

Summary

References



# A Brief History of Information and Language



*[...] two messages, one of which is heavily loaded with meaning and the other which is pure nonsense, can be exactly equivalent, from the present viewpoint, as regards information. It is this, undoubtedly, that Shannon means when he says that “the semantic aspects of communication are irrelevant to the engineering aspects.” **But this does not mean that the engineering aspects are necessarily irrelevant to the semantic aspects.***

Shannon & Weaver (1949). The mathematical theory of communication, p. 8.

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

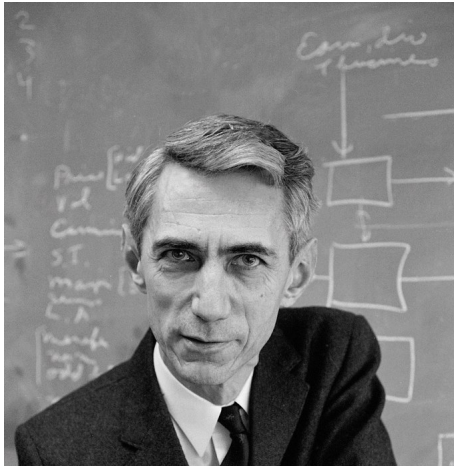
Section 4: Two  
Problems and  
Solutions

Summary

References



# Three Levels of Communication Problems



- ▶ **Level A:** How accurately can the symbols of communication be transmitted? (The technical problem.)
- ▶ **Level B:** How precisely do the transmitted symbols convey the desired meaning? (The semantic problem.)
- ▶ **Level C:** How effectively does the received meaning affect conduct in the desired way? (The effectiveness problem.)

Shannon & Weaver (1949). The mathematical theory of communication, p. 4.

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

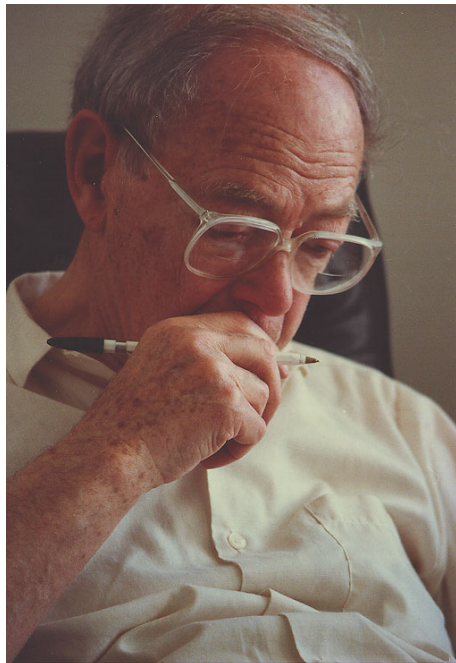
Section 4: Two  
Problems and  
Solutions

Summary

References



# A Brief History of Information and Language



ZELIG HARRIS  
A THEORY OF  
LANGUAGE AND  
INFORMATION  
A Mathematical Approach

*The theory of syntax is stated in terms related to mathematical Information Theory: as constraints on word combination, each later constraint being defined on the resultants of a prior one. This structure not only permits a finitary description of the unbounded set of sentences, but also admits comparison of language with other notational systems, [...]*

Harris, Zellig (1991). A theory of language and information. A mathematical approach.

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

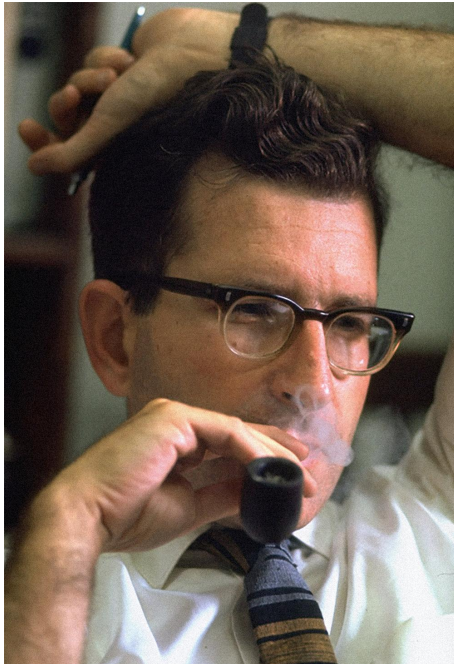
Summary

References





# A Brief History of Information and Language



*[...] To complete this elementary communication theoretic model for language, we assign a probability to each transition from state to state. We can then calculate the "uncertainty" associated with each state and we can define the "information content" of the language as the average uncertainty, weighted by the probability of being in the associated states. **Since we are studying grammatical, not statistical structure of language here, this generalization does not concern us.***

Chomsky, Noam (1957). Syntactic Structures, p. 20.

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References

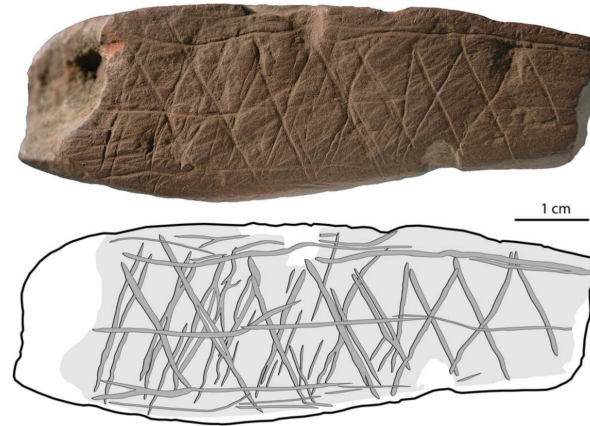


---

## Section 2: Introduction



# What's the difference?



Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References



## Some Intuitive Terminology

- ▶ order  $\leftrightarrow$  disorder
- ▶ regularity  $\leftrightarrow$  irregularity
- ▶ predictability  $\leftrightarrow$  unpredictability
- ▶ certainty  $\leftrightarrow$  uncertainty
- ▶ choice  $\leftrightarrow$  restriction

# Entropy

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References



## Entropy as Possibility

“*Entropy as possibility* is my favorite short description of entropy because possibility is an apt word and, unlike *uncertainty* and *missing information*, has positive connotation.”

**“Entropy is an additive measure of the number of possibilities available to a system.”**

Lemons (2013). A student’s guide to entropy, p. 160.

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

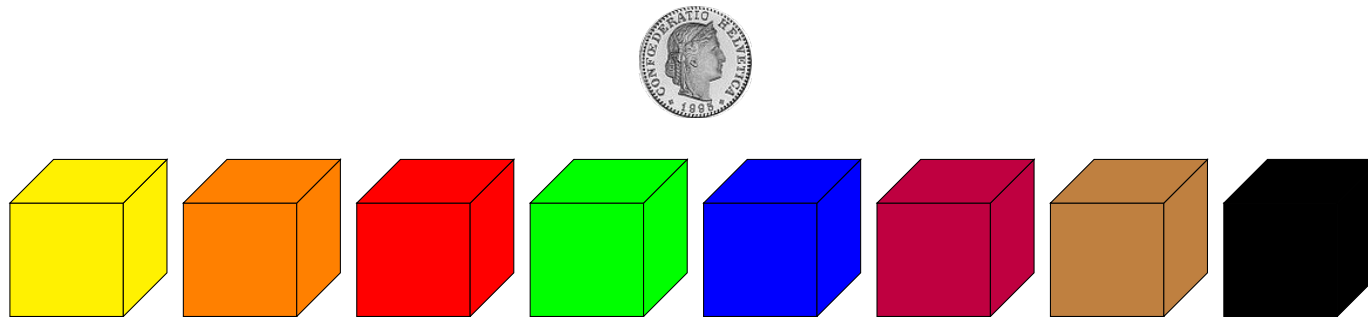
Section 4: Two  
Problems and  
Solutions

Summary

References



# How can you measure possibility? Let's play the box game!



Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

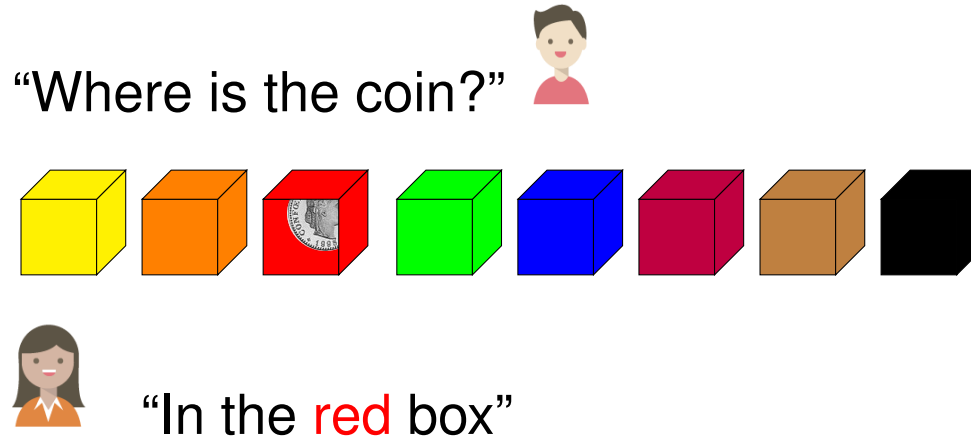
Summary

References

- ▶ How many choices do you have? – Well, 8.
- ▶ Just to make it more complicated: in **bits** this is  $\log_2(8) = 3$
- ▶ Translated into binary code:  
000 001 010 100 011 110 101 111



## How does this relate to language?



- ▶ The “alphabet” (here words) of the “language” they use does not need more than 8 colour adjectives to disambiguate:

$$\mathcal{A} = \{\text{yellow, orange, red, green, blue, purple, brown, black}\}$$

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References

Assume we play this game  $n$  times. The probability of a coin being put into any of the boxes is  $p(col) = \frac{1}{8}$ . This is a *random* and *uniform* distribution of probabilities.



“In the **red/green/blue/ yellow/purple/brown/black ... box**”

The probabilities of words occurring in the **girl’s language** will match this distribution in the limit, i.e. as  $n \rightarrow \infty$ .

Section 1:  
Historical  
Overview

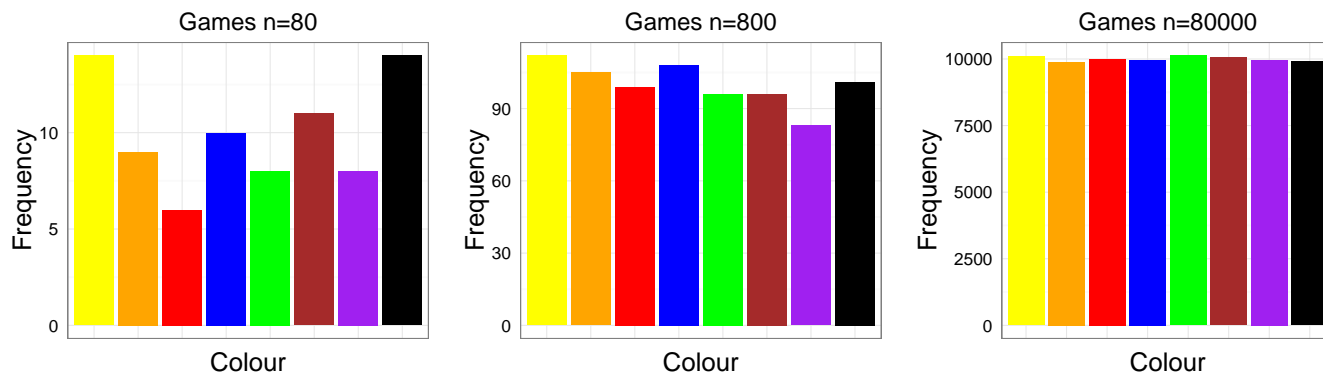
Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References







## The bottom line is:

Notice how in this simple communication game, the probabilities of occurrences of **words** (color adjectives) start to reflect the probabilities of occurrences of **situations** in the “real world” (coins in boxes) – **if communication is truthful.**

However, is this relevant to “real” natural language?

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

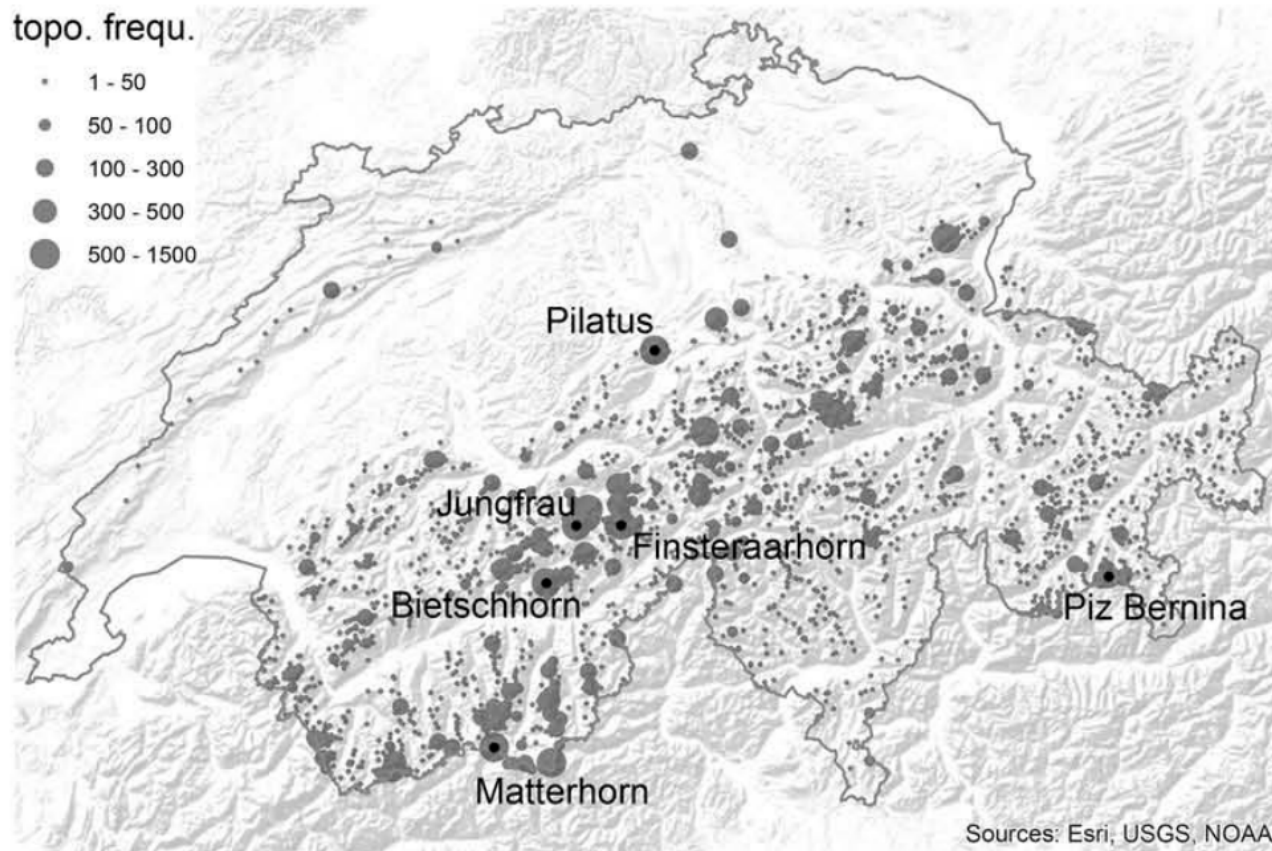
Section 4: Two  
Problems and  
Solutions

Summary

References



# Example: Frequencies of Mountain Names



Section 1:  
Historical  
Overview

Section 2:  
Introduction

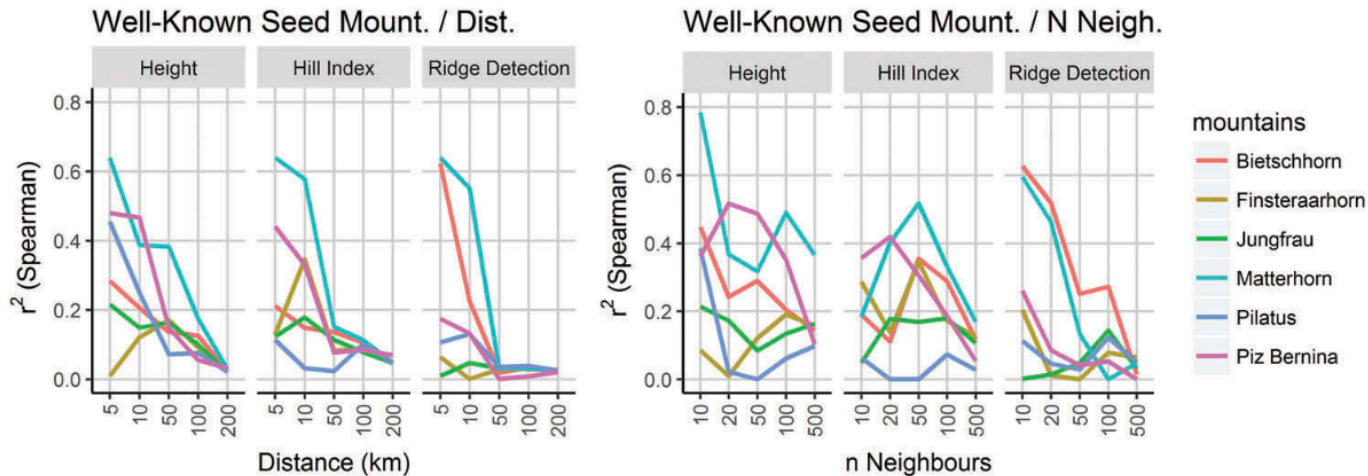
Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References

Derungs & Samardžić (2017). Are prominent mountains frequently mentioned in text?



Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References

**Figure 5.** The relation toponym frequency: spatial measure tested for different spatial extents and a set of well-known seed mountains.

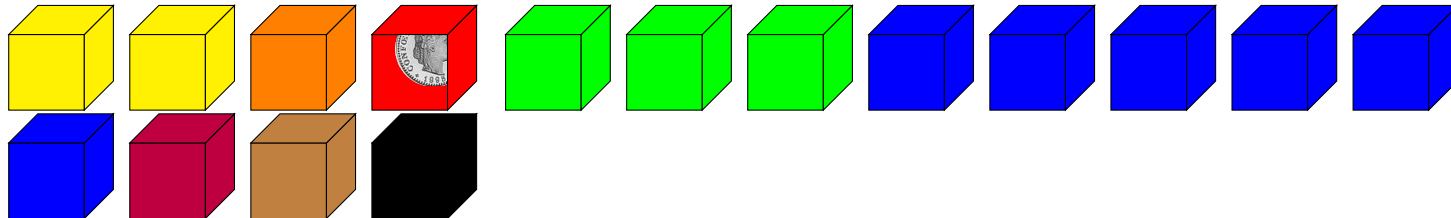
The frequency of occurrence of so-called toponyms (in this case names of famous mountains) in texts is significantly correlated with measures of spatial salience (e.g. height), especially if a text is written in a location close-by.

Hence, this is an example of how **real world salience** is reflected in **probabilities of occurrence in language**.



# What if we change the game?

“Where is the coin?”



“In the **red** box”

- ▶ The “alphabet” has **not** changed:

$$\mathcal{A} = \{\text{yellow, orange, red, green, blue, purple, brown, black}\}$$

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

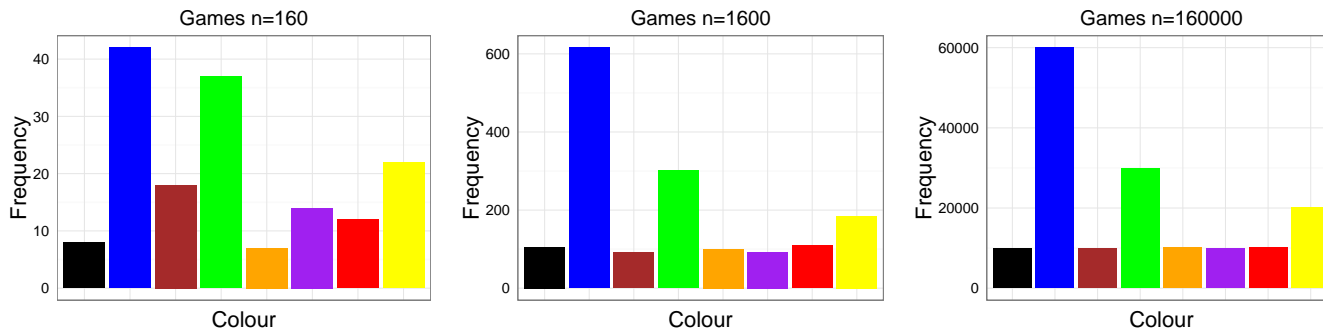
References

However, the probabilities of boxes/colours has changed:  $p(\text{blue}) = \frac{6}{16}$ ,  
 $p(\text{green}) = \frac{3}{16}$ ,  $p(\text{yellow}) = \frac{2}{16}$ ,  $p(\text{purple}) = \frac{1}{16}$ , etc.



“In the red, green, blue, blue yellow, purple, blue,... box”

Again, this will be reflected in the **girl’s language production**.



Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

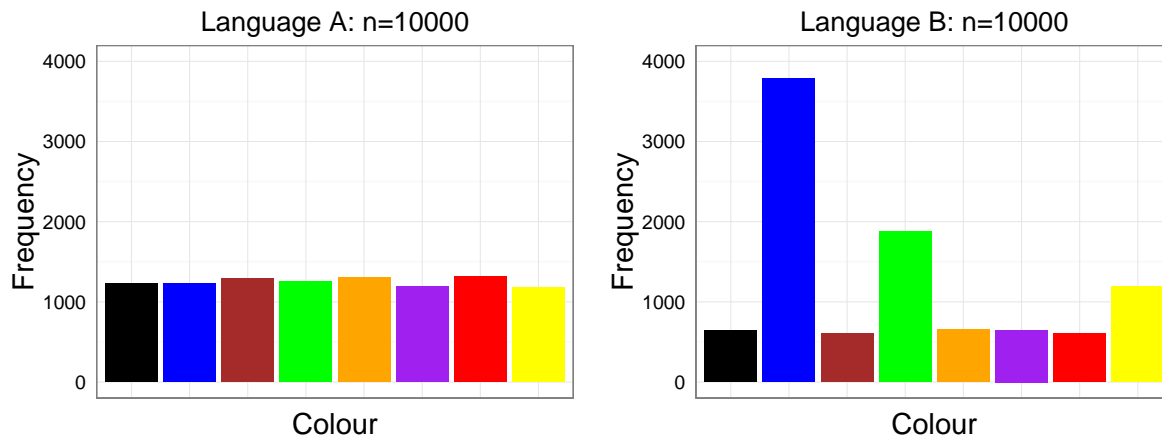
References



# Comparing language production

If we play the two games the same number of times  $n$ , we will get the same two languages  $L_A$  and  $L_B$  in terms of **word types** (8 in this case), and the number of **word tokens** (10K in this case).

However, the **distributions of word token counts** differ!



Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

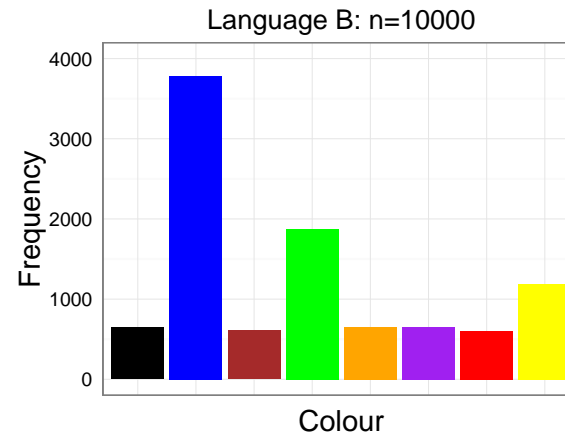
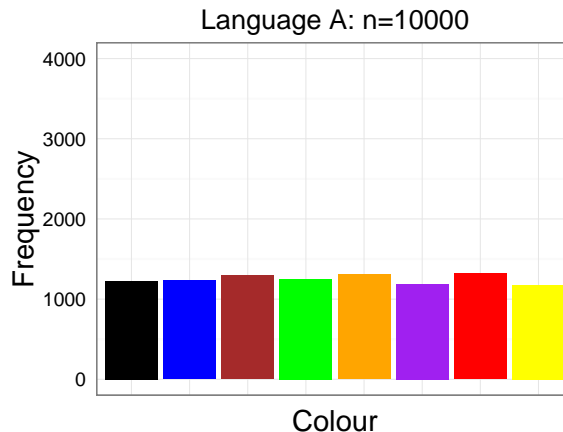
Summary

References



# Crucially: Certainty and Uncertainty in the Game

Note that in  $L_A$  there is **more uncertainty, more choice/possibility** than in  $L_B$ . If we had to take a guess what the girl says next, then in  $L_A$  we have a uniform chance of  $\frac{1}{8} = 0.125$  of being right, whereas in  $L_B$  we have a better chance of  $\frac{6}{16} = \frac{3}{8} = 0.375$  if we guess “blue”.



Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References



---

## **Section 3: Measuring Entropy**





## How can we measure this difference in the distributions?

Claude Shannon came up with a measure for this difference in "A mathematical theory of communication" (1948). He called it the **entropy H**, after the concept known from thermodynamics.

$$H = -\sum p(x) \log p(x)$$

Note that there can be different notations and versions of that formula, which is confusing at times.

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References



# A more precise formulation

(See also Cover & Thomas, 2006)

Assume that

- ▶  $X$  is a *discrete random variable*, drawn from an alphabet of possible values  $\mathcal{X} = \{x_1, x_2, \dots, x_N\}$ , where  $N = |\mathcal{X}|$

Example: The “alphabet” or set of colour adjectives, e.g.

$\mathcal{A} = \{\text{yellow, orange, red, green, blue, purple, brown, black}\}$ , with  $N = 8$

- ▶ The *probability mass function* is defined by  
$$p(x) = \text{Pr}\{X = x\}, x \in \mathcal{X}$$

Example: each word type is assigned a probability, e.g. in  $L_B$

$p(\text{blue}) = \frac{6}{16}$ ,  $p(\text{green}) = \frac{3}{16}$  etc.

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References



## A more precise formulation

Given these definitions, the entropy is then defined as

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x). \quad (1)$$

Notes:

- ▶ The logarithm is typically taken to the base 2, i.e. giving bits of information. We will henceforth indicate this explicitly.
- ▶ In the original article by Shannon, there was also a positive constant  $K$  before the summation sign, but henceforth it was mostly assumed to be 1, and hence dropped.
- ▶ There are many alternative - notationally different, but conceptually equivalent - formulations of the entropy. Shannon, for instance, used  $H(p_1, p_2, \dots, p_N)$ , which is mostly shortened to  $H(X)$ .

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References



## Let's look at the component parts

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log_2 p(x). \quad (2)$$

- ▶  $-\log_2 p(x)$  is the **information content** of a unit  $x$  (word type in the case of the box game). In the case where units are independent of each other, the probability is essentially a normalized frequency. The frequency of a unit determines how much information it carries. The minus sign is just there to not get a negative value, since the logarithm of probabilities ( $0 < p(x) < 1$ ) is negative (except for  $p(x) = 1$ , for which it is 0).

For example, in  $L_B$  the word type “blue” occurs ca. 3750 times in 10000 tokens, and its information content is  $-\log_2\left(\frac{3750}{10000}\right) \sim 1.42$  bits. The word type “orange”, on the other hand, occurs ca. 625 times in 10000 tokens, its information content is  $-\log_2\left(\frac{625}{10000}\right) \sim 4$  bits. Hence, the word type “orange” has higher information content.

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References



## Let's look at the component parts

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \quad (3)$$

- ▶ The summation part of the equation means that we multiply the information content of each element  $x$  with its probability  $p(x)$ , and sum over all of them. Note that multiplying all elements with their probabilities just means that we take the average.

**Hence, the entropy  $H(X)$  can be seen as the average information content of information encoding units, i.e. adjective word types in the case of the box game.**

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

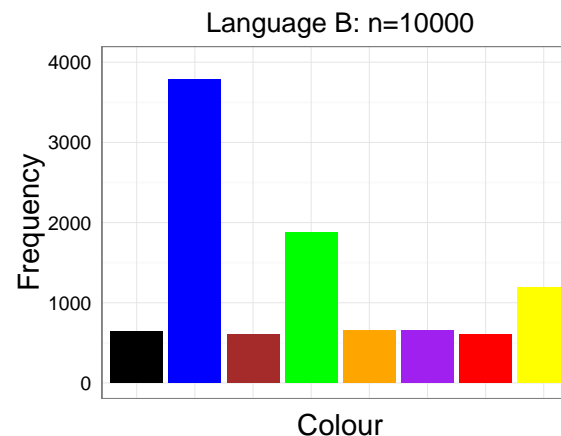
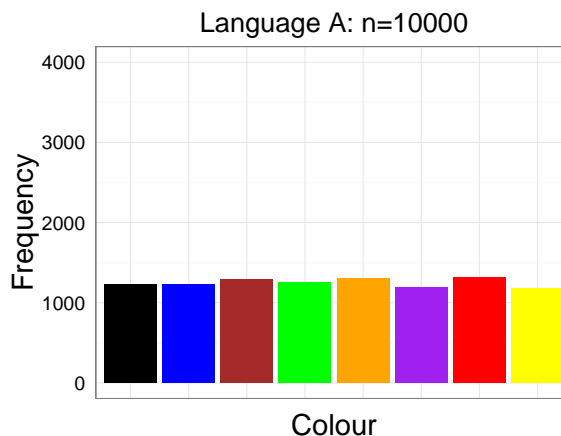
References

## Let's apply this to Languages A and B

For reasons of simplicity let's take the expected values and not actual counts:

$$H(L_A) = -\left(\frac{1}{8} \times \log_2\left(\frac{1}{8}\right) + \frac{1}{8} \times \log_2\left(\frac{1}{8}\right) + \dots + \frac{1}{8} \times \log_2\left(\frac{1}{8}\right)\right) = 3^1 \quad (4)$$

$$H(L_B) = -\left(\frac{6}{16} \times \log_2\left(\frac{6}{16}\right) + \frac{3}{16} \times \log_2\left(\frac{3}{16}\right) + \dots + \frac{1}{16} \times \log_2\left(\frac{1}{16}\right)\right) = 2.61 \quad (5)$$



<sup>1</sup>Note: the case where we have a uniform distribution of probabilities, i.e. all events (adjectives here) are exactly equally likely, is the **maximum entropy** case. In this case, the equation simplifies to  $\log_2(N)$ . Such that here we have  $\log_2(8)=3$ .

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References



- ▶ Word types in Language *A* carry **3 bits** of information on average, whereas word types in Language *B* carry only **2.61 bits**.
- ▶ Note that 3 bits is actually the **maximum entropy** possible for a language with 8 word types, since this is the case with uniform probabilities  $\frac{1}{8}$ .
- ▶ The **minimum entropy** would be 0, namely in the case where only 1 word type is used, since  $\log_2\left(\frac{8}{8}\right) = 0$ .

Section 1:  
Historical  
Overview

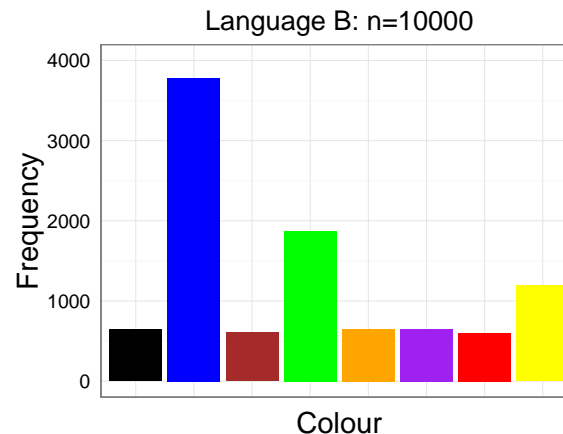
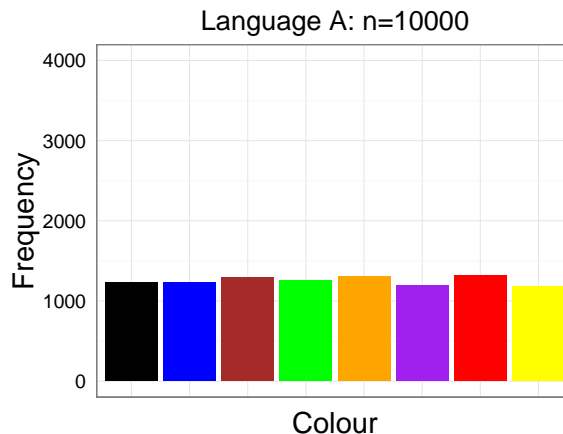
Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References





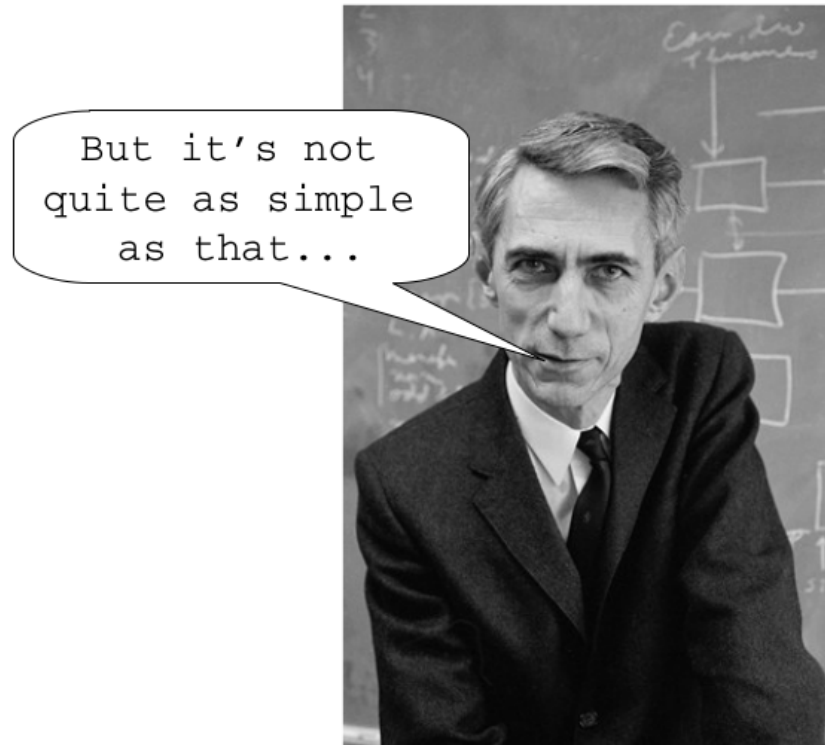
---

## **Section 4: Two Problems and Solutions**





That's great! We have a tool at hand to measure the information encoding potential of any communicative (and non-communicative) system!



Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References



## Two Major Problems

1. What is an information encoding “unit” in the first place - and how does this effect the results?
2. What is the “real” probability of letters, words, sentences, or symbols more generally?

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References



# Problem 1

We will deal with the problem of information encoding units and their impact on the results of entropy estimation in the Exercise Sheet for Tutorial Week 1.

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References

## Problem 2a: What's the “real” probability?

The probabilities of letters, words, phrases, etc. depend on the **corpus size**, and so does the entropy  $H(X)$ .

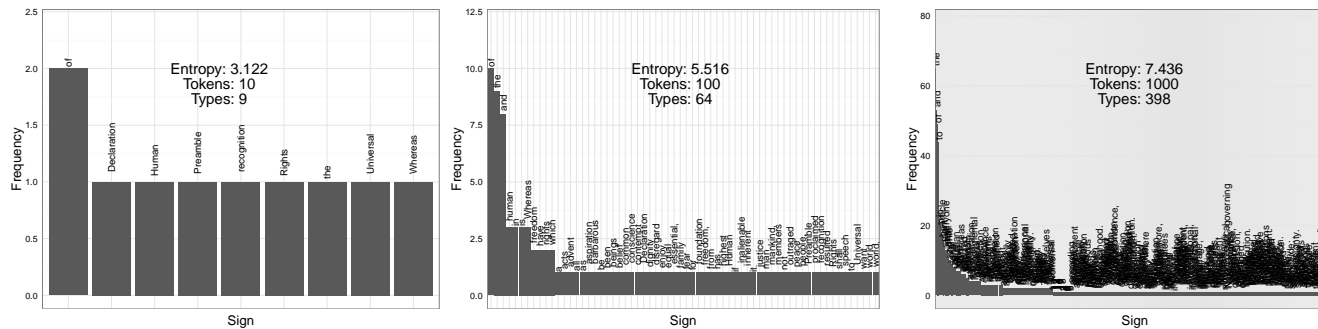


Figure. Frequency distributions and word type entropies for the English UDHR according to the first 10, 100, 1000 word tokens.

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References



## Possible Solution for Problem 2a

Get better entropy estimators (e.g. Hausser & Strimmer 2014 via R package *entropy*), and estimate the text size for which the entropy stabilizes.

Section 1:  
Historical  
Overview

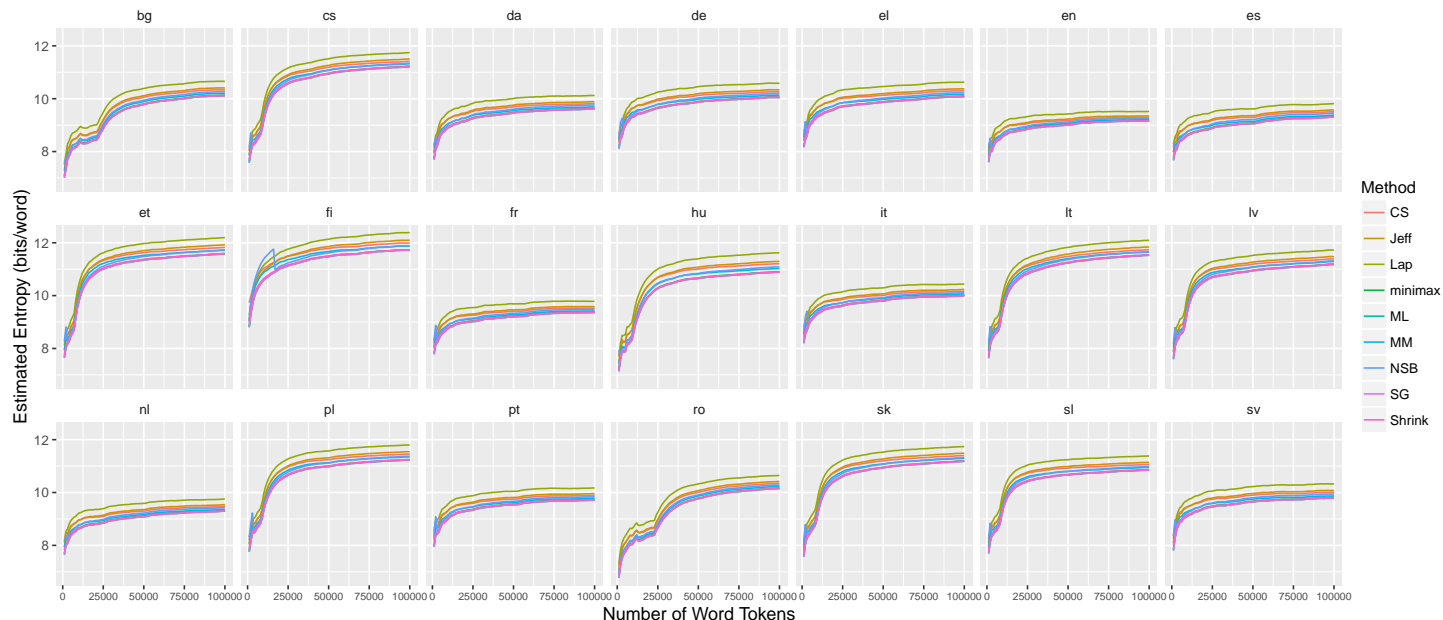
Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References



Bentz et al. (2017). The entropy of words - learnability and expressivity across more than 1000 languages. *Entropy*.



## Problem 2b: What's the “real” probability?

Letters, words, phrases etc. are **not** drawn randomly and **independently** from one another. Instead, the **co-text** and **context** results in **conditional probabilities and entropies**.

Conditional probability:  $p(y|x) = \frac{p(x,y)}{p(x)}$

Example for the first 100 tokens of the English UDHR:

While the simple unigram probability of “the” is

$$p(\text{the}) = \frac{9}{100} = \mathbf{0.09},$$

the probability of “the” conditioned on “of” is

$$p(\text{the}|\text{of}) = \frac{p(\text{of}, \text{the})}{p(\text{of})} = \frac{\frac{4}{100}}{\frac{10}{100}} = \frac{4}{10} = \mathbf{0.4}.$$

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References



## Possible Solution for Problem 2b

- ▶ Estimate **n-gram** (bigram, trigram, etc.) entropies instead of unigram entropies. However, this soon requires very big corpora as  $n$  increases. This is a fundamental problem often referred to as *data sparsity*.
- ▶ Estimate the **entropy rate**  $h$ , which reflects the growth of the entropy with the length of a string, i.e.  $n$  in our case (Cover & Thomas, 2006, p. 74).

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References



---

# Summary





# Summary

- ▶ Information theory gives us an understanding of the fundamentals of **information encoding and decoding**. Communication also harnesses these processes.
- ▶ The information contained in a string of symbols can be **defined mathematically**, and **measured empirically**.
- ▶ Information contained in a communication system might reflect information contained in the real world.
- ▶ Entropy is a measure of the **information encoding potential** of a symbol system.

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References



---

## References



## References

Bentz, Christian, Alikaniotis, Dimitrios, Ferrer-i-Cancho, Ramon & Cysouw, Michael (2017). The entropy of words - learnability and expressivity across more than 1000 languages. *Entropy*.

Cover, Thomas M. & Thomas, Joy A. (2006). *Elements of Information Theory*. New Jersey: Wiley & Sons.

Derungs, Curdin & Samardžić, Tanja (2017). Are prominent mountains frequently mentioned in text? Exploring the spatial expressiveness of text frequency. *International Journal of Geographical Information Science*.

Lemons, Don S. (2013). *A student's guide to entropy*. Cambridge: Cambridge University Press.

Shannon, Claude E. & Weaver, Warren (1949). *The mathematical theory of communication*. Chicago: University of Illinois Press.

Section 1:  
Historical  
Overview

Section 2:  
Introduction

Section 3:  
Measuring  
Entropy

Section 4: Two  
Problems and  
Solutions

Summary

References



# Thank You.

Contact:

**Faculty of Philosophy**

General Linguistics

Dr. Christian Bentz

SFS Wihlemstraße 19-23, Room 1.24

[chris@christianbentz.de](mailto:chris@christianbentz.de)

Office hours:

During term: Wednesdays 10-11am

Out of term: arrange via e-mail