



# **Semantics & Pragmatics SoSe 2021**

## Lecture 3: Information Theory II

04/05/2021, Christian Bentz



---

# Overview

## Section 1: Recap of Lecture 2

## Section 2: Relative Entropy

- Definition

- Relation to Natural Language

- Summary

## Section 3: Mutual Information

- Conditional Entropy

- Definition of Mutual Information

## Section 4: Relation to Meaning

- Implications for Natural Language

## Summary

## References



---

## **Section 1: Recap of Lecture 2**



## 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: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

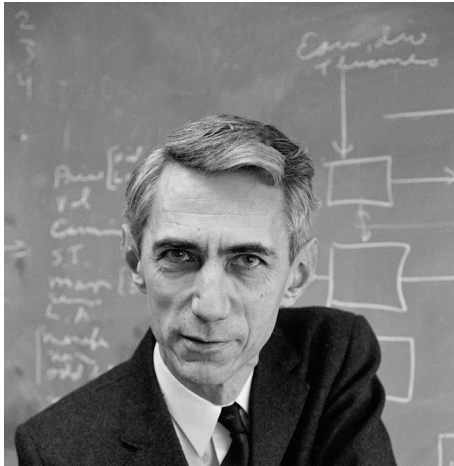
Section 4:  
Relation to  
Meaning

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: Recap of Lecture 2

Section 2: Relative Entropy

Section 3: Mutual Information

Section 4: Relation to Meaning

Summary

References



## Some Intuitive Terminology

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

Entropy

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

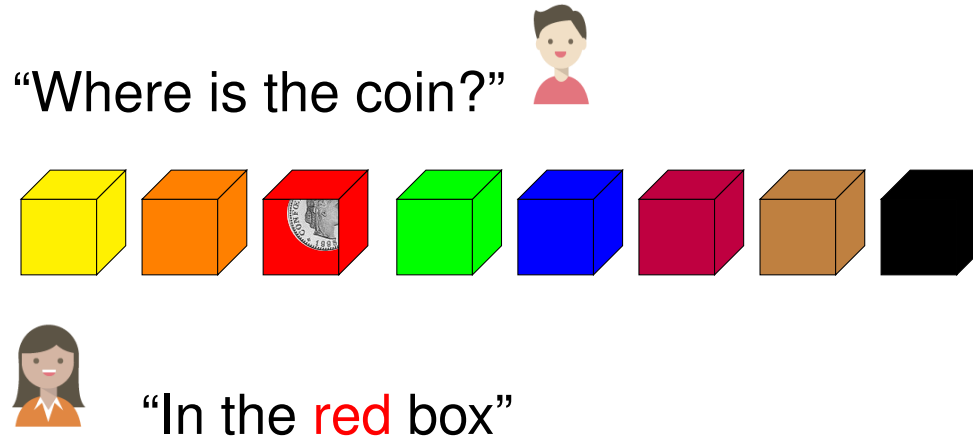
Section 4:  
Relation to  
Meaning

Summary

References



## How does this relate to language?



Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References

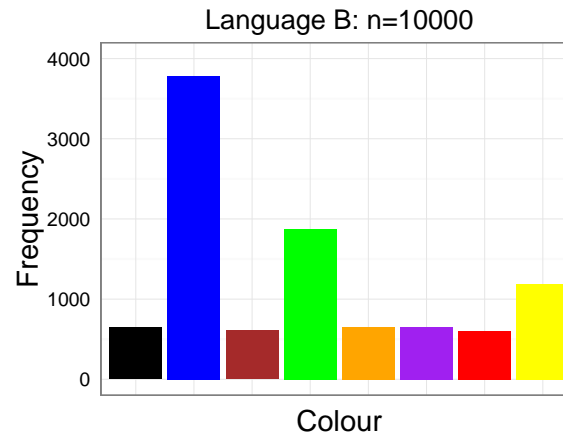
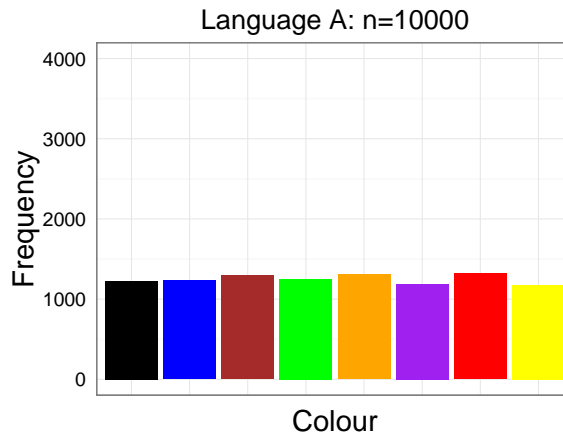
- ▶ The “alphabet” (here words) of the “language” they use does not need more than 8 colour adjectives to disambiguate:

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



## 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: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

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: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

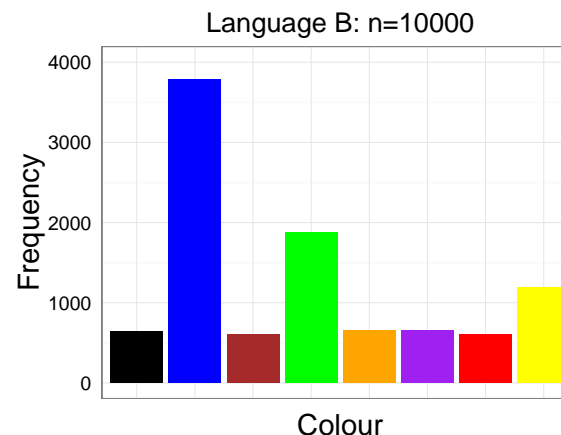
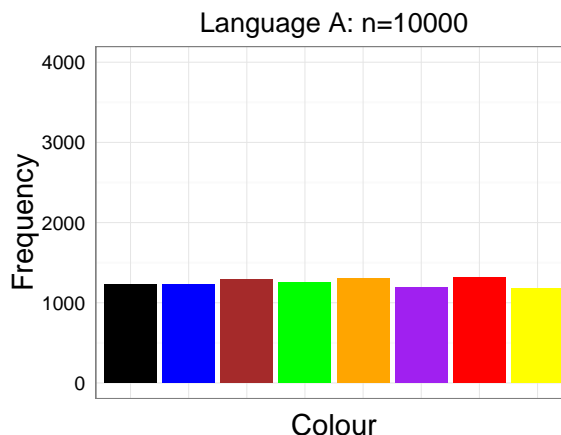
References

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

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

$$H(L_A) = -\left(\frac{1}{8} \times \log_2\left(\frac{1}{8}\right) + \frac{1}{8} \times \log_2\left(\frac{1}{8}\right) + \dots + \frac{1}{8} \times \log_2\left(\frac{1}{8}\right)\right) = 3^1 \quad (2)$$

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



<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: Recap of Lecture 2

Section 2: Relative Entropy

Section 3: Mutual Information

Section 4: Relation to Meaning

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: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References



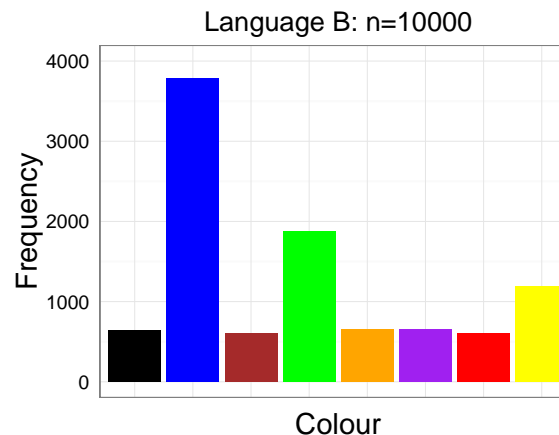
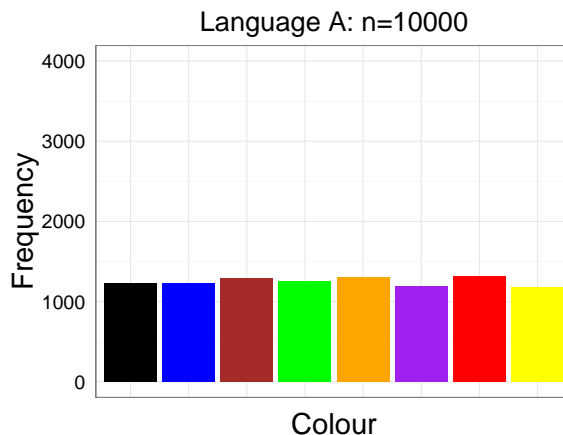
---

## **Section 2: Relative Entropy**

# Probability Distributions for Languages A and B

Language A :  $p(x) = \frac{1}{8}$ , with  $x \in \{black, blue, \dots, yellow\}$ . (4)

Language B :  $q(x) = \{\langle black, \frac{1}{16} \rangle, \langle blue, \frac{6}{16} \rangle, \langle brown, \frac{1}{16} \rangle, \langle green, \frac{3}{16} \rangle, \langle orange, \frac{1}{16} \rangle, \langle purple, \frac{1}{16} \rangle, \langle red, \frac{1}{16} \rangle, \langle yellow, \frac{2}{16} \rangle\}$ . (5)



Note: There is a difference in the probabilities of occurrences of colour adjectives between Language A and Language B. How big is this difference?

Section 1: Recap of Lecture 2

Section 2: Relative Entropy

Section 3: Mutual Information

Section 4: Relation to Meaning

Summary

References



## Relative Entropy (Kullback-Leibler distance/divergence)

“The *relative entropy* is a measure of the distance<sup>2</sup> between two distributions. [...] The relative entropy  $D(p||q)$  is a measure of the inefficiency of assuming that the distribution is  $q$  when the true distribution is  $p$ .”

Cover & Thomas (2006), p. 19.

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References

The relative entropy between two probability mass functions  $p(x)$  and  $q(x)$  is defined as

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)}. \quad (6)$$

It can take values between 0 and  $\infty$ . It is 0 if  $p = q$ .

<sup>2</sup>Note that it is not a ‘true’ distance measure, since it does not satisfy the triangle inequality.



## Example

Assume the probabilities of Language A constitute the probability mass function  $p(x)$ , and the probabilities of Language B the function  $q(x)$ . We then have:

$$\begin{aligned}
 D(p||q) = & 1/8 \log_2 \frac{1/8}{1/16} + 1/8 \log_2 \frac{1/8}{6/16} + 1/8 \log_2 \frac{1/8}{1/16} + \\
 & 1/8 \log_2 \frac{1/8}{3/16} + 1/8 \log_2 \frac{1/8}{1/16} + 1/8 \log_2 \frac{1/8}{1/16} + 1/8 \log_2 \frac{1/8}{1/16} + \quad (7) \\
 & 1/8 \log_2 \frac{1/8}{2/16} \sim \mathbf{0.35} \text{ bits per word.}
 \end{aligned}$$

Thus, there is relatively little difference in the probability distributions of colour adjectives in these two languages.

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References



## How does this relate to language?

In the example above, we compared two artificial languages of the box game by using the **relative entropy**. Another way of looking at it – which is arguably closer to a denotational semantics point of view – is to consider the difference between the “real” world situations (i.e. colored boxes), and the language that transmits information about them (i.e. the colour adjectives).

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

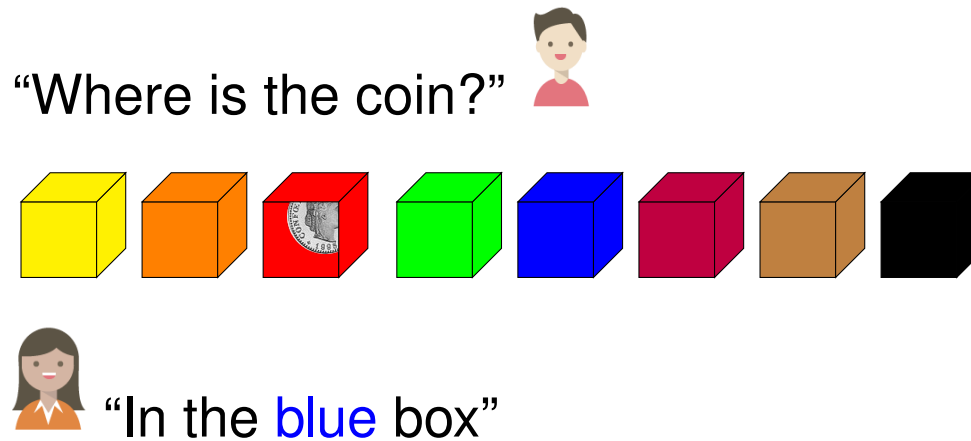
Summary

References





# The Cost of Miss-Information



Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References

Assume the “alphabet” of the “language” is still the same:

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

Assume they play the game 16 times. The probability of the coin being in any of the boxes is still  $1/8$ . However, half of the time the coin is in the **red** box, the girl actually says it is in the **blue** box. Otherwise she is faithful.



## Example

In the box game **with miss-information** we thus have a discrepancy between the probability distribution of real colours of boxes ( $p(x)$ ) and the probabilities of colour adjectives denoting these boxes ( $q(x)$ ). This discrepancy can be measured by the relative entropy:

$$\text{World} : p(x) = \frac{1}{8}, \text{ with } x \in \{black, blue, \dots, yellow\}. \quad (8)$$

$$\text{Language} : q(x) = \left\{ \left\langle black, \frac{2}{16} \right\rangle, \left\langle blue, \frac{3}{16} \right\rangle, \left\langle brown, \frac{2}{16} \right\rangle, \right. \\ \left. \left\langle green, \frac{2}{16} \right\rangle, \left\langle orange, \frac{2}{16} \right\rangle, \left\langle purple, \frac{2}{16} \right\rangle, \left\langle red, \frac{1}{16} \right\rangle, \left\langle yellow, \frac{2}{16} \right\rangle \right\}. \quad (9)$$

$$D(p||q) \sim \mathbf{0.05} \text{ bits per word} \quad (10)$$

Conclusion: The **cost of miss-information** is here 0.05 bits per word (on average).

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References



## Summary: Relative Entropy

- ▶ The **relative entropy** measures the discrepancy in probability distributions *over the same variable*  $x$ , e.g.  $p(x)$  and  $q(x)$ .
- ▶ **First possible application:** calculate the discrepancy between the probability distributions of elements of the “alphabet” in two instances of language usage (Language A and Language B above).
- ▶ **Second possible application:** calculate the discrepancy between the probability distributions of real world situations (coloured boxes), and the language used to communicate about them (colour adjectives). This is the communicative **cost of miss-information**.

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References



## Drawback

Since the **relative entropy is defined over the same variable**  $x$ , this means that the “alphabet” between the systems compared has to be exactly the same. If we had, for example,  $\mathcal{X} = \{\text{black, blue, brown}\}$  and  $\mathcal{Y} = \{\text{black, blue}\}$ , then the relative entropy between the probability distributions over these alphabets would be defined as

$$D(p||q) = \infty. \quad (11)$$

This is because we have to assume

$$q(\text{brown}) = 0. \quad (12)$$

Cover & Thomas (2006), p. 19.

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References



---

## **Section 3: Mutual Information**



## Another Version of the Box Game

Imagine a version of the box game in which the girl consistently uses the colour adjective **blue** instead of **red**, such that the latter is actually not in her alphabet anymore. Otherwise she names the correct colours.

Section 1: Recap of Lecture 2

Section 2: Relative Entropy

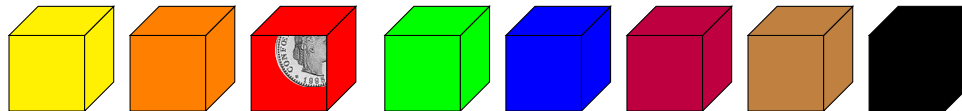
Section 3: Mutual Information

Section 4: Relation to Meaning

Summary

References

“Where is the coin?” 



 “In the **blue** box”

Assume the “alphabet” of the “language” is then:

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



We now have the situation described above, namely, the alphabet of the “language” (here  $\mathcal{Y}$ ) does not fit the alphabet of the “real world” (here  $\mathcal{X}$ ) anymore. The relative entropy would give us  $D(p||q) = \infty$  for  $p(x)$  and  $q(y)$ .

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

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

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

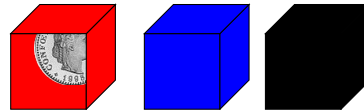
Summary

References



For ease of exposition, and for making later calculations easier, let's simplify this game to three boxes.

“Where is the coin?” 



 “In the **blue** box.”

Such that we have the alphabets

$$\mathcal{Y} = \{ \text{blue}, \text{black} \},$$

$$\mathcal{X} = \{ \text{red}, \text{blue}, \text{black} \}.$$

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References





Assume we play the box game with the probability of the coin being in any of the three boxes being uniform, i.e.  $\frac{1}{3}$ . We thus get the probability mass function for the “**real world**” **variable**  $x$  as

$$p(x) = \{ \langle \text{red}, \frac{1}{3} \rangle, \langle \text{blue}, \frac{1}{3} \rangle, \langle \text{black}, \frac{1}{3} \rangle \}. \quad (13)$$

Since the girl consistently replaces “red” for “blue”, and is otherwise faithful, we furthermore get the following **conditional probability function** for a colour in the language ( $y$ )<sup>3</sup> conditioned on a colour in the real world ( $x$ ):

$$p(y|x) = \{ \langle (\text{red}|\text{red}), 0 \rangle, \langle (\text{red}|\text{blue}), 0 \rangle, \langle (\text{red}|\text{black}), 0 \rangle, \langle (\text{blue}|\text{red}), 1 \rangle, \langle (\text{blue}|\text{blue}), 1 \rangle, \langle (\text{blue}|\text{black}), 0 \rangle, \langle (\text{black}|\text{red}), 0 \rangle, \langle (\text{black}|\text{blue}), 0 \rangle, \langle (\text{black}|\text{black}), 1 \rangle \}. \quad (14)$$

<sup>3</sup>For reasons of symmetry, we assume that for the variable  $y$ :  $p(\text{red}) = 0$ . In other words, rather than not having a probability value at all, “red” is assigned 0 probability.

Section 1: Recap of Lecture 2

Section 2: Relative Entropy

Section 3: Mutual Information

Section 4: Relation to Meaning

Summary

References



## Conditional Entropy

Given  $p(x)$  and  $p(y|x)$ , we can define the so-called **conditional entropy** of the random variable  $Y$  given the random variable  $X$  as:

$$H(Y|X) = - \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log_2 p(y|x) \quad (15)$$

This gives the amount of information (in bits) which is needed to describe the random variable  $Y$  (our language production in the box game), conditioned on another random variable  $X$  (the real world outcomes of where the coin goes in the box game).

Cover & Thomas (2006), p. 17.

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References



## Example: Calculating $H(Y|X)$

Given  $p(x)$  and  $p(y|x)$  defined for the box game above, we thus get the conditional entropy as:

$$\begin{aligned}
 H(Y|X) = & -(p(\text{red}) \times (p(\text{red}|\text{red}) \log_2 p(\text{red}|\text{red}) + \\
 & p(\text{blue}|\text{red}) \log_2 p(\text{blue}|\text{red}) + \\
 & p(\text{black}|\text{red}) \log_2 p(\text{black}|\text{red})) + \\
 & p(\text{blue}) \times (p(\text{red}|\text{blue}) \log_2 p(\text{red}|\text{blue}) + \\
 & p(\text{blue}|\text{blue}) \log_2 p(\text{blue}|\text{blue}) + \\
 & p(\text{black}|\text{blue}) \log_2 p(\text{black}|\text{blue})) + \\
 & p(\text{black}) \times (p(\text{red}|\text{black}) \log_2 p(\text{red}|\text{black}) + \\
 & p(\text{blue}|\text{black}) \log_2 p(\text{blue}|\text{black}) + \\
 & p(\text{black}|\text{black}) \log_2 p(\text{black}|\text{black})))
 \end{aligned} \tag{16}$$

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References



Further plugging the conditional probabilities of (14) into Equation (16) gives us:

$$\begin{aligned}
 H(Y|X) = & -\left(\frac{1}{3} \times (0 \times \log_2(0) + 1 \times \log_2(1) + 0 \times \log_2(0))\right) + \\
 & \frac{1}{3} \times (0 \times \log_2(0) + 1 \times \log_2(1) + 0 \times \log_2(0)) + \\
 & \frac{1}{3} \times (0 \times \log_2(0) + 0 \times \log_2(0) + 1 \times \log_2(1))
 \end{aligned} \tag{17}$$

Note that we define  $0 \times \log_2(0) = 0$  (Cover & Thomas, 2006, p. 14). Furthermore, it generally holds that  $1 \times \log_2(1) = 0$ . We thus actually get

$$H(Y|X) = 0. \tag{18}$$

Why is this?

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References



In words: the conditional entropy (i.e. uncertainty or choice) of the language variable ( $Y$ ) given the real world variable ( $X$ ) is 0 in our current version of the box game, meaning that we know everything about  $Y$  by knowing  $X$ .

This is true, since we know:

- ▶ If the coin is in the **red** box, the girl will **always** say “**blue**”.
- ▶ If the coin is in the **blue** box, the girl will **always** say “**blue**”.
- ▶ If the coin is in the black box, the girl will **always** say “black”.

Hence, for every possible value of  $X$  we know exactly, i.e. with probability 1, what the outcome is going to be in  $Y$ .

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References



## Example: Calculating $H(X|Y)$

What if we calculate the conditional entropy for the real world outcomes based on knowing the language production? The probability mass function for the “**language**” variable  $y$  is

$$p(y) = \{ \langle \text{red}, 0 \rangle, \langle \text{blue}, \frac{2}{3} \rangle, \langle \text{black}, \frac{1}{3} \rangle \}. \quad (19)$$

Since the girl consistently replaces “red” for “blue”, and is otherwise faithful. We furthermore get the following **conditional probability function** for a colour in the the real world scenario ( $x$ ) conditioned on a colour in language ( $y$ ):

$$p(x|y) = \{ \langle (\text{red}|\text{blue}), \frac{1}{2} \rangle, \langle (\text{red}|\text{black}), 0 \rangle, \\ \langle (\text{blue}|\text{blue}), \frac{1}{2} \rangle, \langle (\text{blue}|\text{black}), 0 \rangle, \\ \langle (\text{black}|\text{blue}), 0 \rangle, \langle (\text{black}|\text{black}), 1 \rangle \}. \quad (20)$$

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References



## Example: Calculating $H(X|Y)$

Given  $p(y)$  and  $p(x|y)$  defined for the box game above, we thus get the conditional entropy as:

$$H(X|Y) = - \sum_{y \in \mathcal{Y}} p(y) \sum_{x \in \mathcal{X}} p(x|y) \log_2 p(x|y). \quad (21)$$

And thus we have

$$\begin{aligned} H(X|Y) = & -(p(\text{blue}) \times (p(\text{red}|\text{blue}) \log_2 p(\text{red}|\text{blue}) + \\ & p(\text{blue}|\text{blue}) \log_2 p(\text{blue}|\text{blue}) + \\ & p(\text{black}|\text{blue}) \log_2 p(\text{black}|\text{blue}))) + \\ & (p(\text{black}) \times (p(\text{red}|\text{black}) \log_2 p(\text{red}|\text{black}) + \\ & p(\text{blue}|\text{black}) \log_2 p(\text{blue}|\text{black}) + \\ & p(\text{black}|\text{black}) \log_2 p(\text{black}|\text{black}))). \end{aligned} \quad (22)$$

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References



Further plugging the conditional probabilities of (20) into Equation (22) gives us:

$$H(X|Y) = -\left(\frac{2}{3} \times \left(\frac{1}{2} \times \log_2\left(\frac{1}{2}\right) + \frac{1}{2} \times \log_2\left(\frac{1}{2}\right) + 0 \times \log_2(0)\right) + \frac{1}{3} \times (0 \times \log_2(0) + 0 \times \log_2(0) + 1 \times \log_2(1))\right). \quad (23)$$

We thus get

$$H(X|Y) = \frac{2}{3} \sim \mathbf{0.67} \text{ bits.} \quad (24)$$

**Conclusion:** This means that there is some conditional entropy (uncertainty or choice) in the real world outcome (X) given we know the language production (Y). Again, this makes sense given that there is an **ambiguity** in the girls language: when she says “blue”, the coin could either be in the blue or the red box (with equal probability).

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References





## Mutual Information

In the last step, we can now define the **mutual information** between  $X$  and  $Y$  as

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (25)$$

Note that while the conditional entropies  $H(X|Y)$  and  $H(Y|X)$  are asymmetrical, i.e. can give different values (as we have seen above), the mutual information is symmetrical. The mutual information is the **reduction in the uncertainty of  $X$  given  $Y$** .<sup>4</sup>

Cover & Thomas (2006), p. 21.

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References

<sup>4</sup>There is an alternative – but equivalent – way of defining mutual information with reference to *joint probabilities* of  $X$  and  $Y$  rather than conditional probabilities.



## Example: Calculating $I(X; Y)$

In the last lecture we have seen how to calculate the entropy of variables  $X$  and  $Y$  based on the probabilities of their possible outcomes. For our current version of the box game,  $p(x)$  and  $p(y)$  were defined above. This yields

Section 1: Recap of Lecture 2

Section 2: Relative Entropy

Section 3: Mutual Information

Section 4: Relation to Meaning

Summary

References

$$H(X) = -\left(\frac{1}{3} \log_2\left(\frac{1}{3}\right) + \frac{1}{3} \log_2\left(\frac{1}{3}\right) + \frac{1}{3} \log_2\left(\frac{1}{3}\right)\right) \sim 1.58 \text{ bits}, \quad (26)$$

as well as

$$H(Y) = -\left(0 \log_2(0) + \frac{2}{3} \log_2\left(\frac{2}{3}\right) + \frac{1}{3} \log_2\left(\frac{1}{3}\right)\right) \sim 0.92 \text{ bits}. \quad (27)$$

While above we have established that  $H(X|Y) = 0.67$  bits, and  $H(Y|X) = 0$  bits.



If we plug these results into the mutual information formula, we get

$$I(X; Y) = 1.58 - 0.67 \sim \mathbf{0.92} \text{ bits.} \quad (28)$$

We come to the conclusion that there is almost **one bit of uncertainty reduction** in the language given the real world outcomes of the box game, and the other way around.

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References



---

## **Section 4: Meaning and Mutual Information**



## Interpretation of Mutual Information

Let's look at the **mutual information** equation again from the perspective of  $X$ , i.e. the **real world outcomes** of the box game:

$$I(X; Y) = H(X) - H(X|Y) \quad (29)$$

There are several points to be noted:

- ▶ Note that the **conditional entropy** is strictly positive or zero, i.e.  $H(X|Y) \geq 0$ .
- ▶ The entropy is itself also strictly positive or zero, i.e.  $H(X) \geq 0$ .

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References



## Maximum Mutual Information

From this it follows that the **maximum of mutual information** is the entropy  $H(X)$ , i.e.

$$I(X; Y) \leq H(X). \quad (30)$$

This would be the case if the language of the box game was so precise that there is *no conditional entropy* left, i.e.  $H(X|Y) = 0$ .

However, as we have seen in our box game example, this is not the case. There is some ambiguity of the colour term “blue” in the language. Hence, the uncertainty about the real world outcomes is reduced by **0.92** bits given the language, but there are **0.67** bits of uncertainty left.

Section 1: Recap of Lecture 2

Section 2: Relative Entropy

Section 3: Mutual Information

Section 4: Relation to Meaning

Summary

References



## Minimum Mutual Information

The **minimal mutual information** is defined as 0. When is this the case? – When it holds that

$$H(X) = H(X|Y). \quad (31)$$

This would be the case if the language of the box game did not give us *any information at all* about the outcomes of the real world, meaning that the two variables  $X$  and  $Y$  are completely statistically independent.

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References



## Implications for Natural Language

Imagine a language that always maps exactly **one color adjective** with exactly **one box game outcome**. In this case, we have **maximum mutual information**  $I(X; Y)$ , since the conditional entropy is  $H(X|Y) = H(Y|X) = 0$ . However, as the number of colours increases, this would require a potentially infinite number of colour adjectives to cover all possible colours. In fact, the entropy  $H(Y)$  of the colour adjectives can be conceptualized as a **cost of learning**.

Section 1: Recap of Lecture 2

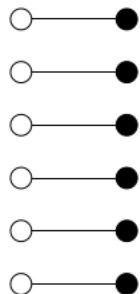
Section 2: Relative Entropy

Section 3: Mutual Information

Section 4: Relation to Meaning

Summary

References



**Figure 3.** A one-to-one mapping between  $n = 6$  signals (white circles) and  $m = 6$  stimuli (black circles). This configuration achieves maximum  $I(S, R)$ .

Ferrer-i-Cancho & Diaz-Guilera (2007).





# Implications for Natural Language

Terms such as *ambiguity*, *vagueness*, *indeterminacy* are often associated with negative connotations. However, from an information-theoretic point of view these might be necessary aspects of human communication, in order to find a **compromise between minimum learning cost  $H(Y)$ , and maximum explicitness  $I(X; Y)$ .**

Section 1: Recap of Lecture 2

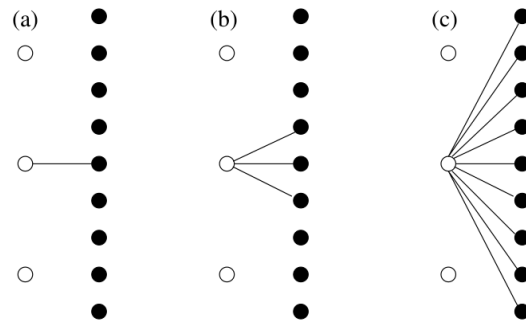
Section 2: Relative Entropy

Section 3: Mutual Information

Section 4: Relation to Meaning

Summary

References

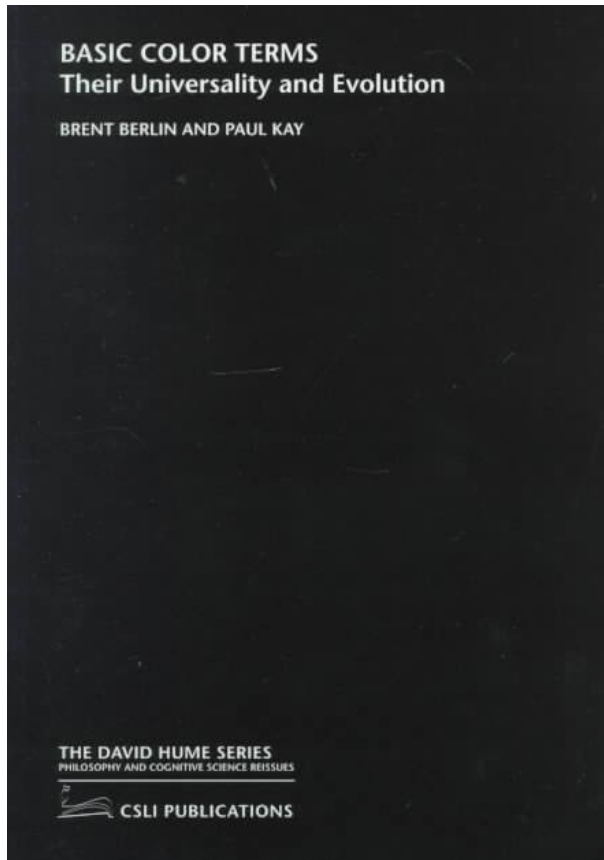


**Figure 1.** Some mappings between signals (white circles) and stimuli (black circles) that are minima of  $H(S)$  and  $H(S|R)$  with  $n = 3$  signals and  $m = 9$  stimuli. (a)–(c) are minima of model A while (c) is the only valid minima of model B.

Ferrer-i-Cancho & Diaz-Guilera (2007).  
Piantadosi et al. (2012).



# Does this relate to Natural Language?



## Two major hypotheses:

1. There is a finite inventory of 11 colors from which languages pick their basic terms.
2. While not all languages name the same set of colors, there are universal implicational hierarchies of which colors are picked.

Berlin & Kay (1969). Basic color terms.

Section 1: Recap of Lecture 2

Section 2: Relative Entropy

Section 3: Mutual Information

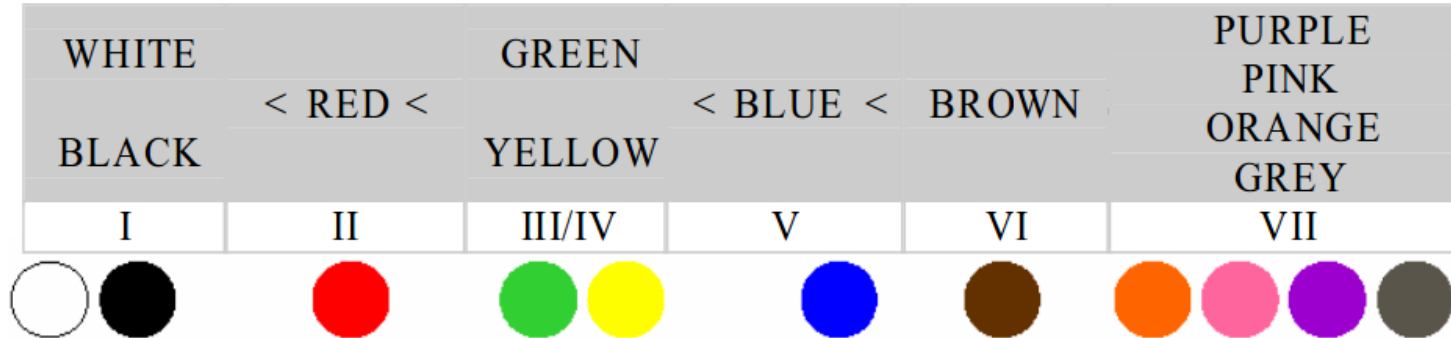
Section 4: Relation to Meaning

Summary

References



# Basic Color Terms: Implicational Hierarchy



Section 1: Recap of Lecture 2

Section 2: Relative Entropy

Section 3: Mutual Information

Section 4: Relation to Meaning

Summary

References

Berlin & Kay (1969). Basic color terms.



Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References

# The World Color Survey

The World Color Survey (WCS) was initiated in the late 1970's to test the hypotheses advanced by Berlin and Kay ([1969](#)) regarding

- (1) the existence of universal constraints on cross-language color naming, and
- (2) the existence of a partially fixed evolutionary progression according to which languages gain color terms over time.

[<http://www.icsi.berkeley.edu/wcs/>]



# Basic Color Terms: Implicational Hierarchy

BLACK, WHITE: Jalé (Papua New Guinea)

BLACK, WHITE, RED: Tiv (Nigeria)

BLACK, WHITE, RED, YELLOW: Ibo (Nigeria)

BLACK, WHITE, RED, GREEN: Ibibio (Nigeria)

BLACK, WHITE, RED, YELLOW, GREEN: Tzeltal (Mexico)

BLACK, WHITE, RED, YELLOW, GREEN, BLUE: Plains Tamil (India)

BLACK, WHITE, RED, YELLOW, GREEN, BLUE, BROWN: Nez Perce  
(State of Washington)

Moravcsik (2012). Introducing language typology, p. 57.

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

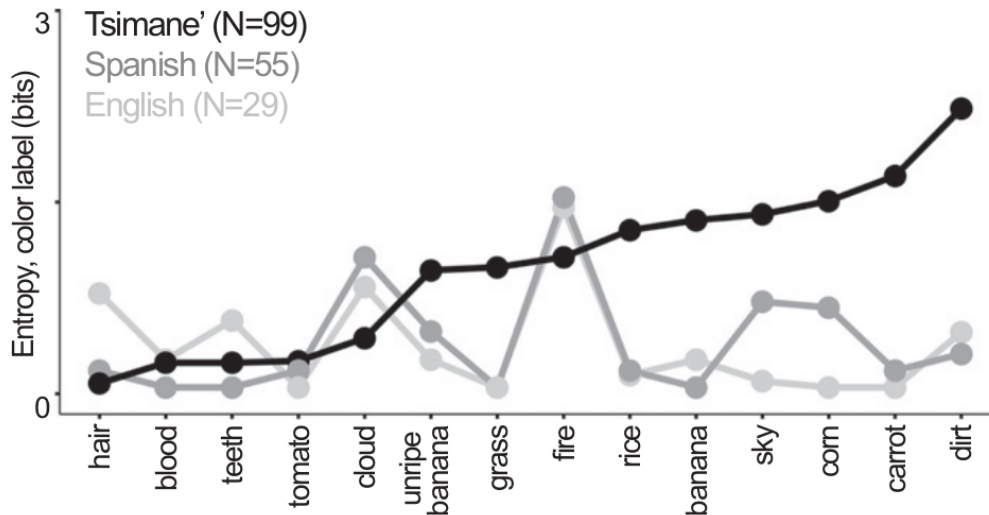
Section 4:  
Relation to  
Meaning

Summary

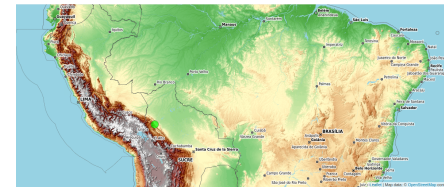
References



# Information-Theoretic Analyses



**Fig. 2.** Variability of color labels (entropy, Eq. 3) for familiar objects, ordered by Tsimane' results. On average, Tsimane' has higher entropy over color words for a particular object (1.06 bits, compared with English, 0.33 bits, and Bolivian-Spanish, 0.30 bits).



<https://glottolog.org>

Section 1: Recap of Lecture 2

Section 2: Relative Entropy

Section 3: Mutual Information

Section 4: Relation to Meaning

Summary

References

Gibson et al. (2017).



---

# Summary



# Summary

- ▶ **Mutual information** is defined as the reduction in entropy (uncertainty or choice) in a variable  $X$  given another variable  $Y$ .
- ▶ If  $X$  is conceptualized as events in the “real world”, and  $Y$  as language performance, then we can use mutual information to **measure how much language tells us about the world**.
- ▶ Since mutual information is calculated as the difference between the entropy of a variable  $X$  and the conditional entropy given another variable  $H(X|Y)$ , there is a **trade-off between minimizing the entropy**, while keeping the **mutual information high**.
- ▶ While entropy is not to be equated with meaning, it is the **upper bound on the mutual information between forms and meanings** – if we take a denotational view point on meaning.

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

Summary

References





---

## References



## References

Berlin, Brent & Kay, Paul (1969). *Basic Color Terms. Their Universality and Evolution*. CSLI Publication.

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

Ferrer-i-Cancho & Díaz-Guilera (2007). The global minima of the communicative energy of natural communication systems. *Journal of Statistical Mechanics: Theory and Experiment*.

Moravcsik, Edith A. (2012). *Introducing Language Typology*. Cambridge: Cambridge University Press.

Piantadosi, Steven, Tily, Hary & Gibson, Edward (2012). The communicative function of ambiguity in language. *Cognition*.

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

Section 1: Recap  
of Lecture 2

Section 2:  
Relative Entropy

Section 3: Mutual  
Information

Section 4:  
Relation to  
Meaning

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