EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

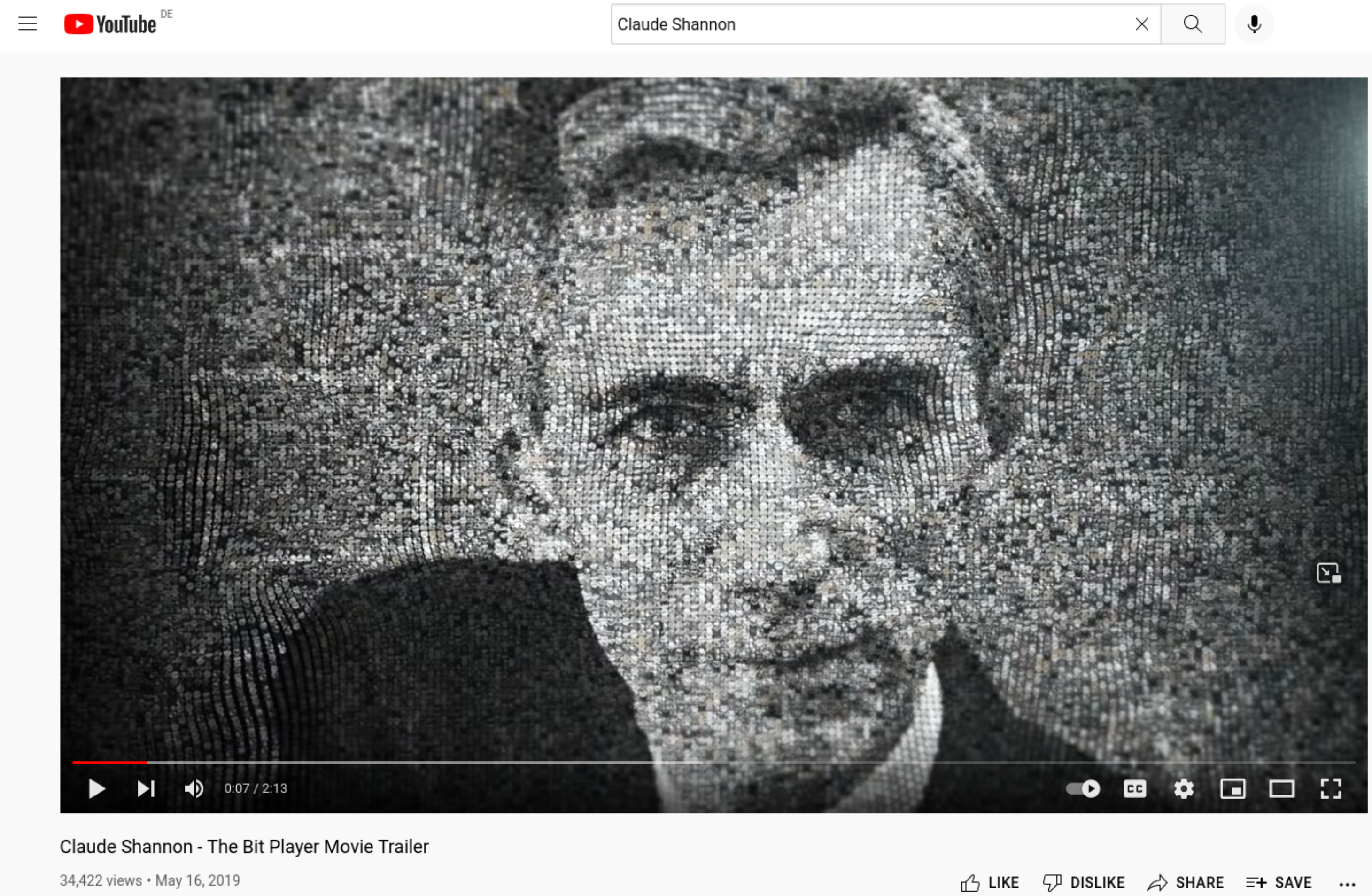
Faculty of Philosophy
General Linguistics

Semantics & Pragmatics SoSe 2022

Lecture 3: Information Theory II

03/05/2022, Christian Bentz

Overview



https://www.youtube.com/watch?v=JP1Ljp8X6hg

Section 1: Recap of Lecture 2

Example

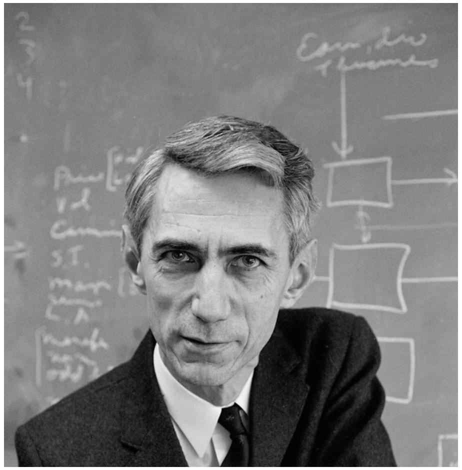

```
Article 1
All human beings are born free and equal in dignity
and rights.  They are endowed with reason and
conscience and should act towards one another in a
spirit of brotherhood.
```

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 ???

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.

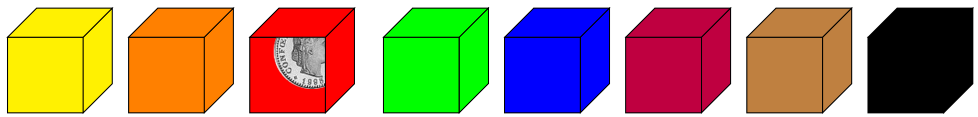
Some Intuitive Terminology

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

Entropy

How does this relate to language?

“Where is the coin?”



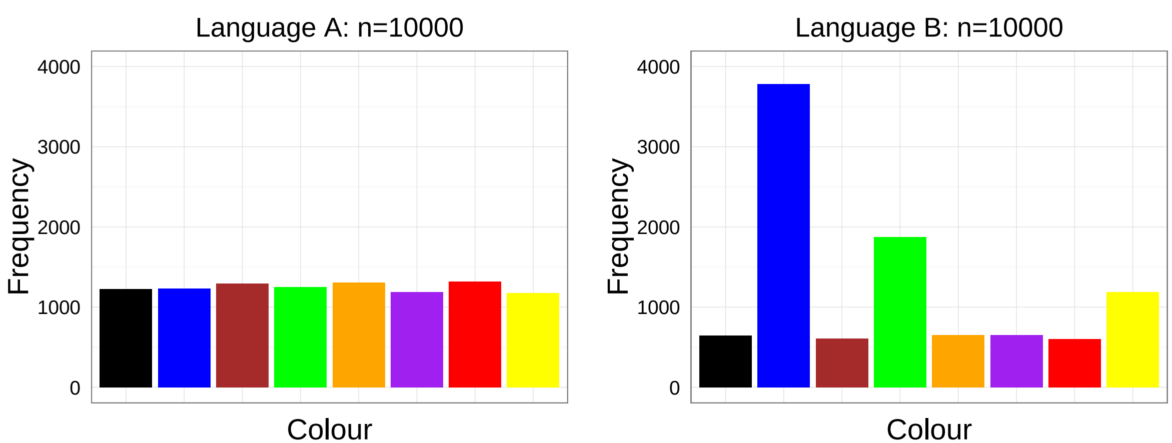
“In the red box”

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

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

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".



A more precise formulation

Given these definitions, the entropy is then defined as

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

Notes:

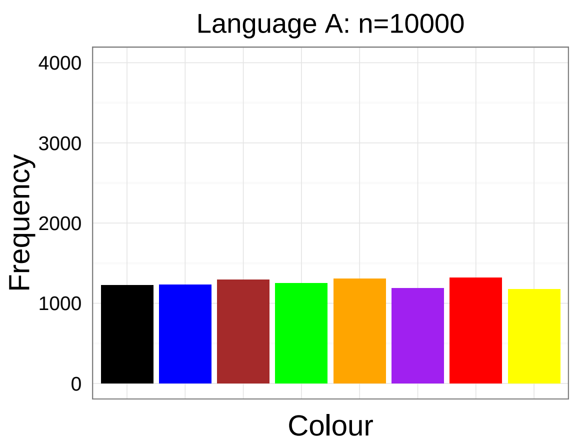
- The logarithm is typically taken to the base 2, i.e. giving bits of information. We will henceforth indicate this explicitly.
- In the original article by Shannon, there was also a positive constant K before the summation sign, but henceforth it was mostly assumed to be 1, and hence dropped.
- There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

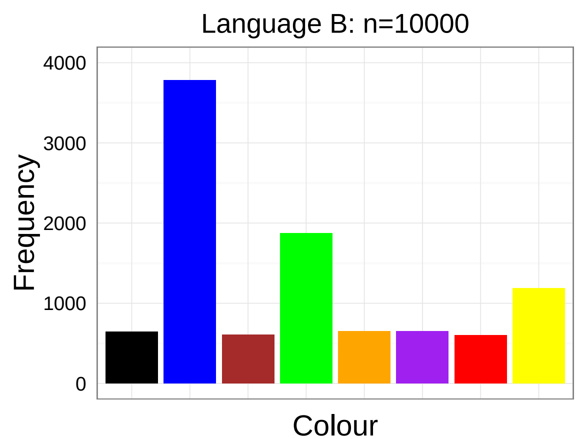
Let's apply this to Languages A and B

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

$$H(L_A) = -(\frac{1}{8} \times \log_2(\frac{1}{8}) + \frac{1}{8} \times \log_2(\frac{1}{8}) + \ldots + \frac{1}{8} \times \log_2(\frac{1}{8})) = \mathbf{3} \tag{2}$$ [1]

$$H(L_B) = -(\frac{6}{16} \times \log_2(\frac{6}{16}) + \frac{3}{16} \times \log_2(\frac{3}{16}) + \ldots + \frac{1}{16} \times \log_2(\frac{1}{16})) = \mathbf{2.61} \tag{3}$$





[1] 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 $\log_2(8)=3$.

Estimation Problems in Natural Languages

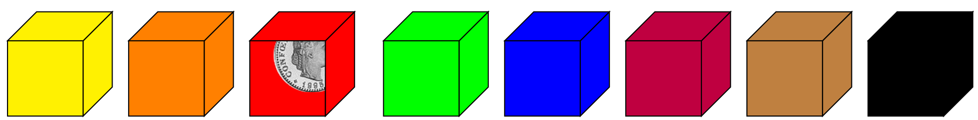
1. **Unit Problem**
   What is an information encoding "unit" in the first place – and how does the choice effect the results?
2. **Sample Size Problem**
   How do estimations change with sample sizes?
3. **Interdependence Problem**
   What is the "real" probability of "units" in natural language, given that they are interdependent?
4. **Extrapolation Problem**
   Do estimations extrapolate across different texts, and corpora?

Section 2: Conditional Entropy

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.

“Where is the coin?” 



 “In the blue box”

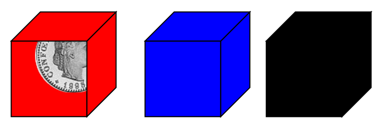
Assume the “alphabet” of the “language” is then:

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

Simplified Version of the Box Game

Let's assume a simplified version with only three boxes. The girl is generally faithful, however, she never uses the color word red, but systematically replaces it by blue.

"Where is the coin?" 



 "In the blue box."

Such that we have the alphabets

$$\mathcal{X} = \{\textit{red}, \textit{blue}, \textit{black}\},$$
$$\mathcal{Y} = \{(\textit{red}), \textit{blue}, \textit{black}\}.$$

Assume, again, that 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

$$p(x) = \{\langle \text{red}, \frac{1}{3} \rangle, \langle \text{blue}, \frac{1}{3} \rangle, \langle \text{black}, \frac{1}{3} \rangle\}. \tag{4}$$

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)[2] conditioned on a colour in the real world (x):

$$\begin{aligned} p(y|x) = \{&\langle(\text{red}|\text{red}), 0\rangle, \langle(\text{red}|\text{blue}), 0\rangle, \langle(\text{red}|\text{black}), 0\rangle, \\ &\langle(\text{blue}|\text{red}), 1\rangle, \langle(\text{blue}|\text{blue}), 1\rangle, \langle(\text{blue}|\text{black}), 0\rangle, \\ &\langle(\text{black}|\text{red}), 0\rangle, \langle(\text{black}|\text{blue}), 0\rangle, \langle(\text{black}|\text{black}), 1\rangle\}. \end{aligned} \tag{5}$$

[2]For reasons of symmetry, we assume that for the variable *y*: p(red) = 0. In other words, rather than not having a probability value at all, "red" is assigned 0 probability.

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:

$$H(Y|X) = -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log_2 p(y|x) \tag{6}$$

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.

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:

$$
\begin{aligned}
H(Y|X) = -(p(red) \times (p(red|red)\log_2 p(red|red)+ \\
p(blue|red)\log_2 p(blue|red)+ \\
p(black|red)\log_2 p(black|red))+ \\
\\
p(blue) \times (p(red|blue)\log_2 p(red|blue)+ \\
p(blue|blue)\log_2 p(blue|blue)+ \\
p(black|blue)\log_2 p(black|blue))+ \\
\\
p(black) \times (p(red|black)\log_2 p(red|black)+ \\
p(blue|black)\log_2 p(blue|black)+ \\
p(black|black)\log_2 p(black|black)))
\end{aligned} \tag{7}
$$

Further plugging the conditional probabilities of (5) into Equation (7) gives us:

$$
\begin{aligned}
H(Y|X) = -(\frac{1}{3} \times (0 \times \log_2(0) + 1 \times \log_2(1) + 0 \times \log_2(0))+ \\
\frac{1}{3} \times (0 \times \log_2(0) + 1 \times \log_2(1) + 0 \times \log_2(0))+ \\
\frac{1}{3} \times (0 \times \log_2(0) + 0 \times \log_2(0) + 1 \times \log_2(1)))
\end{aligned} \tag{8}
$$

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

$$H(Y|X) = 0. \tag{9}$$

Why is this?

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.

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

$$p(y) = \{\langle \text{red}, 0\rangle, \langle \text{blue}, \frac{2}{3}\rangle, \langle \text{black}, \frac{1}{3}\rangle\}. \tag{10}$$

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):

$$\begin{aligned} p(x|y) = \{&\langle(\text{red}|\text{blue}), \frac{1}{2}\rangle, \langle(\text{red}|\text{black}), 0\rangle, \\ &\langle(\text{blue}|\text{blue}), \frac{1}{2}\rangle, \langle(\text{blue}|\text{black}), 0\rangle, \\ &\langle(\text{black}|\text{blue}), 0\rangle, \langle(\text{black}|\text{black}), 1\rangle\}. \end{aligned} \tag{11}$$

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:

$$H(X|Y) = -\sum_{y \in \mathcal{Y}} p(y) \sum_{x \in \mathcal{X}} p(x|y) \log_2 p(x|y). \tag{12}$$

And thus we have

$$\begin{aligned} H(X|Y) = -(p(blue) \times (p(red|blue) \log_2 p(red|blue) + \\ p(blue|blue) \log_2 p(blue|blue) + \\ p(black|blue) \log_2 p(black|blue)) + \\ p(black) \times (p(red|black) \log_2 p(red|black) + \\ p(blue|black) \log_2 p(blue|black) + \\ p(black|black) \log_2 p(black|black))). \end{aligned} \tag{13}$$

Further plugging the conditional probabilities of (11) into Equation (13) gives us:

$$\begin{aligned} H(X|Y) = -(\frac{2}{3} \times (\frac{1}{2} \times \log_2(\frac{1}{2}) + \frac{1}{2} \times \log_2(\frac{1}{2}) + 0 \times \log_2(0)) + \\ \frac{1}{3} \times (0 \times \log_2(0) + 0 \times \log_2(0) + 1 \times \log_2(1))). \end{aligned} \tag{14}$$

We thus get

$$H(X|Y) = \frac{2}{3} \sim \mathbf{0.67}\,\text{bits}. \tag{15}$$

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).

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Faculty of Philosophy
General Linguistics

Section 3: Mutual Information

Mutual Information

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

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{16}$$

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**.[3]

Cover & Thomas (2006), p. 21.

[3]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

$$H(X) = -(\frac{1}{3}\log_2(\frac{1}{3}) + \frac{1}{3}\log_2(\frac{1}{3}) + \frac{1}{3}\log_2(\frac{1}{3})) \sim 1.58\,\text{bits}, \quad (17)$$

as well as

$$H(Y) = -(0\log_2(0) + \frac{2}{3}\log_2(\frac{2}{3}) + \frac{1}{3}\log_2(\frac{1}{3})) \sim 0.92\,\text{bits}. \quad (18)$$

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

$$I(X;Y) = 1.58 - 0.67 \sim \mathbf{0.92}\,\text{bits}. \tag{19}$$

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

Section 4: Relation to Meaning

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:

$$I(X;Y) = H(X) - H(X|Y) \tag{20}$$

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. $H(X) \geq 0$.

Maximum Mutual Information

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

$$I(X;Y) \leq H(X). \tag{21}$$

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 **0.67** bits given the language, but there are **0.92** bits of uncertainty left.

Minimum Mutual Information

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

$$H(X) = H(X|Y). \tag{22}$$

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.

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**.

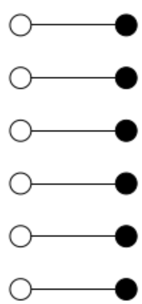


Figure 3. A one-to-one mapping between $n = 6$ signals (white circles) and $m = 6$ stimuli (black circles). This configuration achieves maximum $I(S,R)$.

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 expliciteness** $I(X;Y)$.

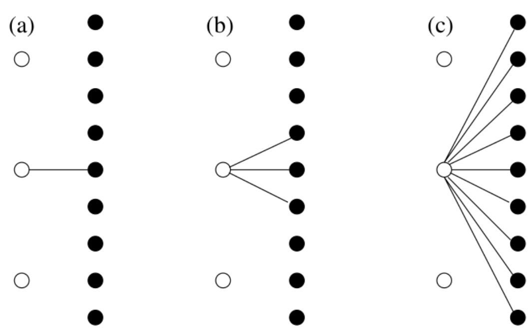
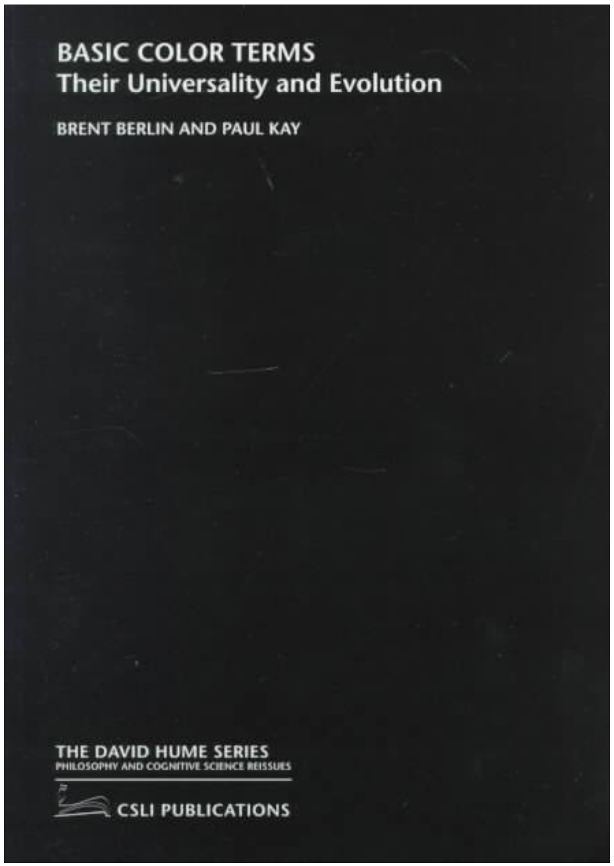


Figure 1. Some mappings between signals (white circles) and stimuli (black circles) that are minima of $H(S)$ and $H(S|R)$ with $n = 3$ signals and $m = 9$ stimuli. (a)–(c) are minima of model A while (c) is the only valid minima of model B.

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

Does this relate to Natural Language?



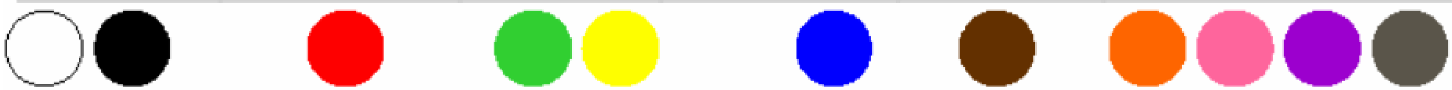
Two major hypotheses:

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.

Basic Color Terms: Implicational Hierarchy

WHITE BLACK	< RED <	GREEN YELLOW	< BLUE <	BROWN	PURPLE PINK ORANGE GREY
I	II	III/IV	V	VI	VII



Berlin & Kay (1969). Basic color terms.

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.

Information-Theoretic Analyses

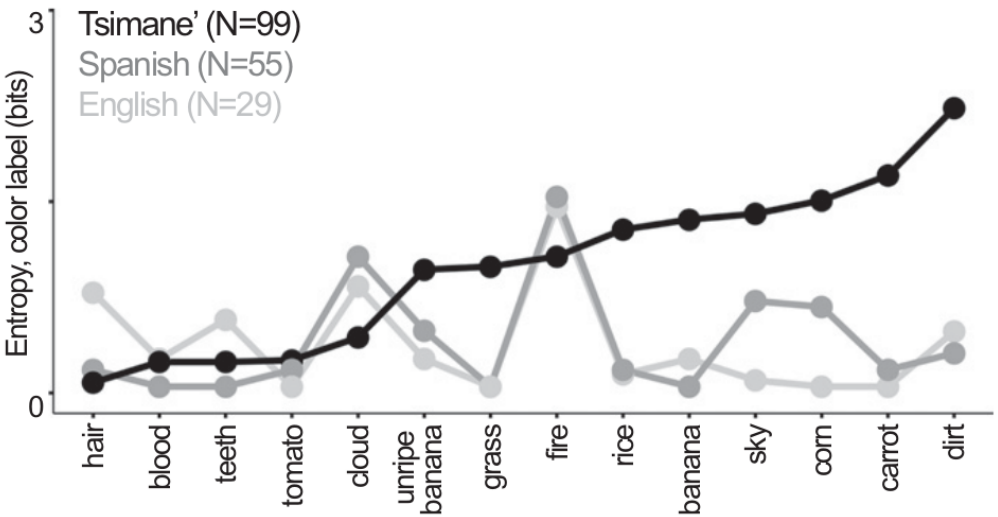
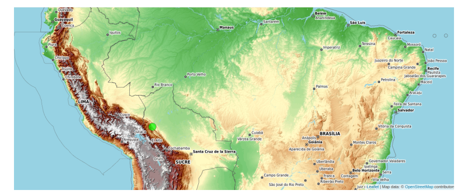


Fig. 2. Variability of color labels (entropy, Eq. 3) for familiar objects, ordered by Tsimane' results. On average, Tsimane' has higher entropy over color words for a particular object (1.06 bits, compared with English, 0.33 bits, and Bolivian-Spanish, 0.30 bits).



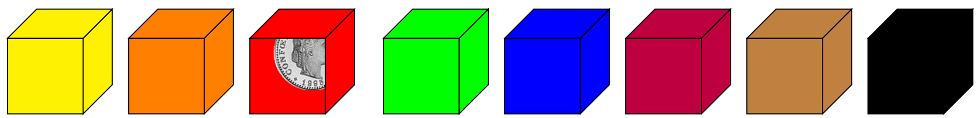
https://glottolog.org

Gibson et al. (2017).

Section 5: Entropy Rate

Many Random Variables (Stochastic Process)

"Where is the coin?"



"In the red box"

"In the blue box"

"In the red box"

"In the green box"

[...]

Finally, we might have many random variables concatenated.

Entropy Rate

Rather than giving the entropy for a single random variable X, we can also estimate the growth of the entropy with a sequence of random variables of length n, aka a *stochastic process* $\{X_i\}$. This is called the **entropy rate** and is defined as

$$H(\mathcal{X}) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n), \tag{23}$$

where $H(X_1, X_2, \ldots, X_n)$ is the *joint entropy* of the individual random variables (X_i). This quantity can be seen as the per symbol (unit) entropy for n random variables.

Cover & Thomas (2006), p. 74–75.

Beware notational confusion (!): Cover & Thomas (2006) use $H(\mathcal{X})$ here instead of $H(X)$, in order to indicate that the entropy is not taken over a single random variable. In many other publications, lower case h is used for the *entropy rate*, in order to distinguish it more clearly from the common definition of Shannon entropy above.

Entropy Rate (Alternative Formulation)

There is an **alternative formulation** of the entropy rate:

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2}\ldots, X_1), \tag{24}$$

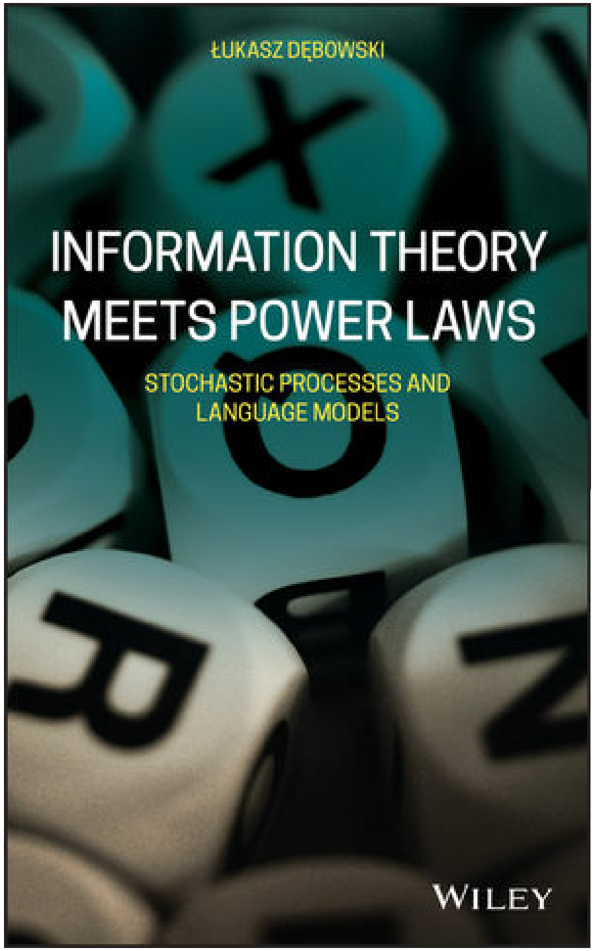
where $H(X_n|X_{n-1}, X_{n-2}\ldots, X_1)$ is the *conditional entropy* of the last random variable (X_n) conditioned on the entire past of random variables.

It can be proven that for *stationary*[4] processes these two definitions are equivalent, i.e.

$$H(\mathcal{X}) = H'(\mathcal{X}). \tag{25}$$

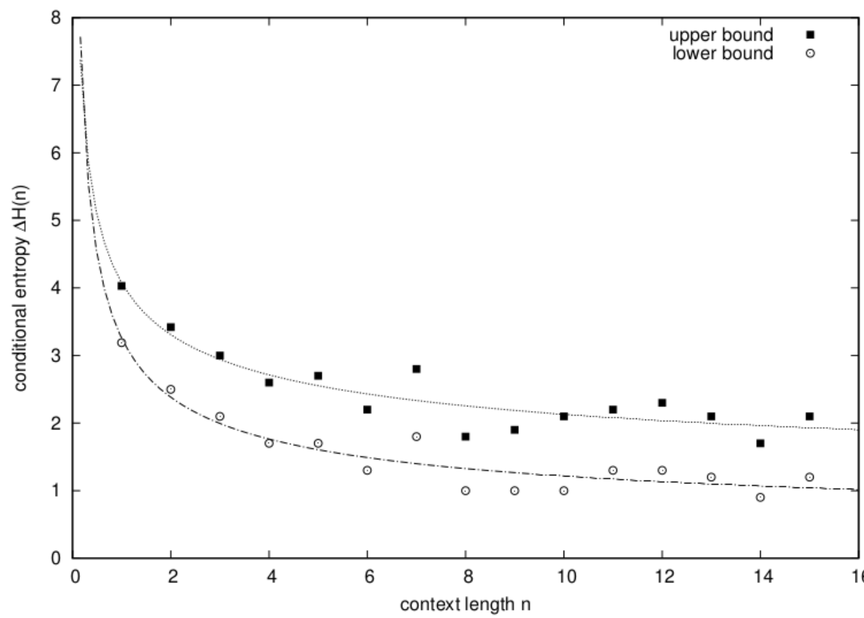
Cover & Thomas (2006), p. 75.

[4] "A distribution on the states such that the distribution at time $n+1$ is the same as the distribution at time n is called a *stationary* distribution." Cover & Thomas (2006), p. 72.



Is the entropy rate zero?

[...] four decades after Shannon, Wolfgang Hilberg, a German electric engineer, [...] supposed that conditional entropy [...] is inversely proportional to the square root of the context length n [...] As such, Hilberg's hypothesis implies that the entropy rate h equals zero. That is, Hilberg's hypothesis implies asymptotic determinism of human utterances.



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Summary

Summary

- There is a range of (interrelated) **information-theoretic measures**: information content (surprisal), entropy, joint entropy, conditional entropy, relative entropy, mutual information, entropy rate, etc.
- 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.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

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, Edtih 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.

Thank You.

Contact:

Faculty of Philosophy
General Linguistics
Dr. Christian Bentz
SFS Wilhelmstraße 19-23, Room 1.15
chris@christianbentz.de
Office hours:
During term: Wednesdays 10-11am
Out of term: arrange via e-mail