EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Semantics & Pragmatics SoSe 2022

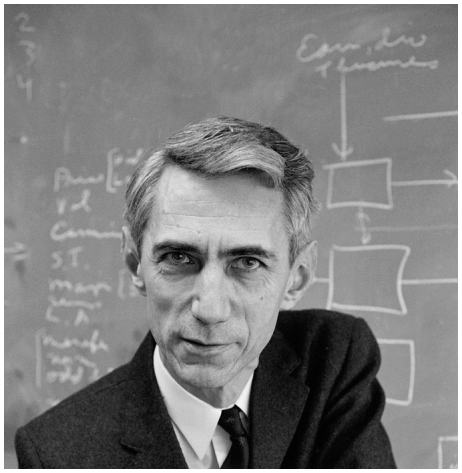
Lecture 2: Information Theory I

28/04/2022, Christian Bentz

Overview

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.

Example

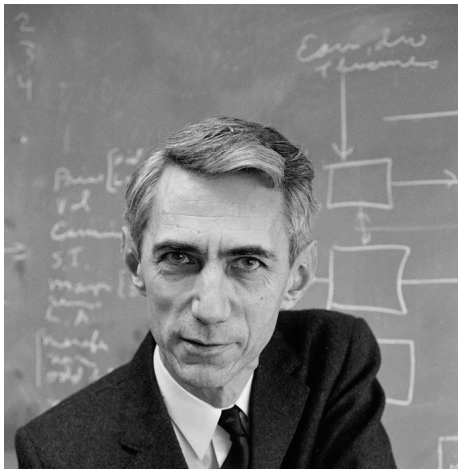
```
Article 1
All human beings are born free and equal in dignity
and rights.  They are endowed with reason and
conscience and should act towards one another in a
spirit of brotherhood.
```

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 ???

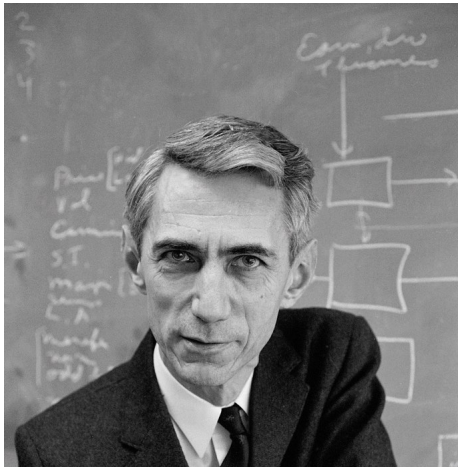
A Brief History of Information and Language



[...] 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.

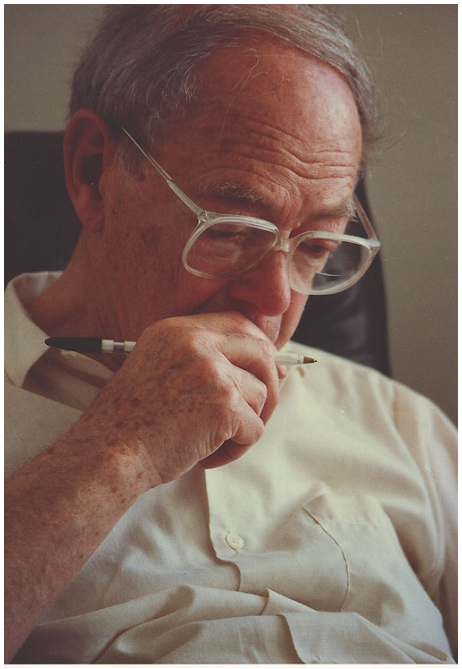
Three Levels of Communication Problems

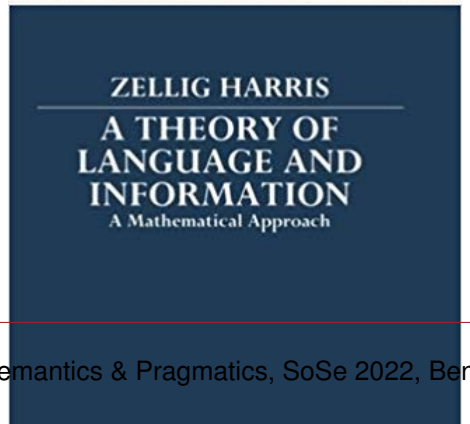


- **Level A**: How accurately can the symbols of communication be transmitted? (The technical problem.)
- **Level B**: How precisely do the transmitted symbols convey the desired meaning? (The semantic problem.)
- **Level C**: How effectively does the received meaning affect conduct in the desired way?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

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

A Brief History of Information and Language

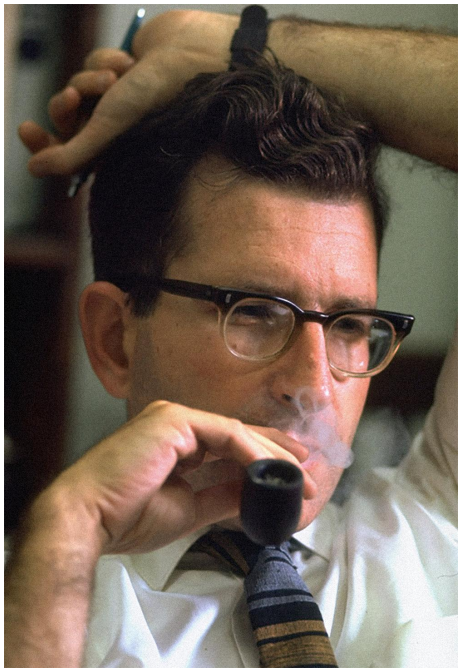




***The theory of syntax is stated in terms related to mathematical Information Theory: as constraints on word combination**, each later constraint being defined on the resultants of a prior one. This structure not only permits a finitary description of the unbounded set of sentences, but also admits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

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

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.

A Brief History of Information and Language

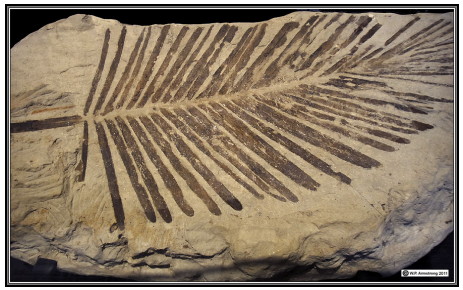
In the last 40 years, research on models of spoken and written language has been split between two seemingly irreconcilable traditions: formal linguistics in the Chomsky tradition, and information theory in the Shannon tradition. [...]

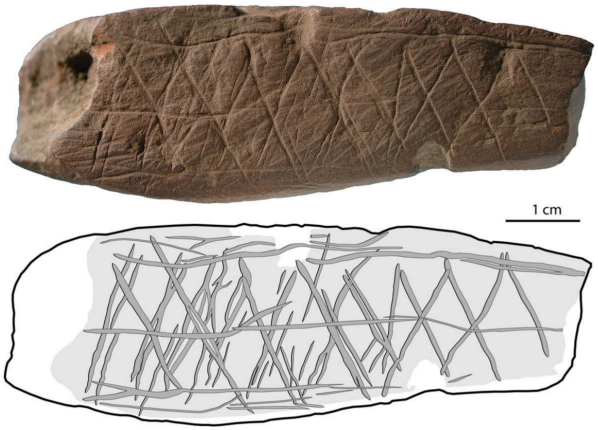
[...] information-theoretic and computational ideas are [...] playing an increasing role in the scientific understanding of language, and will help bring together formal-linguistic and information-theoretic perspectives.

Pereira, Fernando (2000). Formal grammar and information theory: together again?

Section 2: Introduction

What's the difference?





Some Intuitive Terminology

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

} Entropy

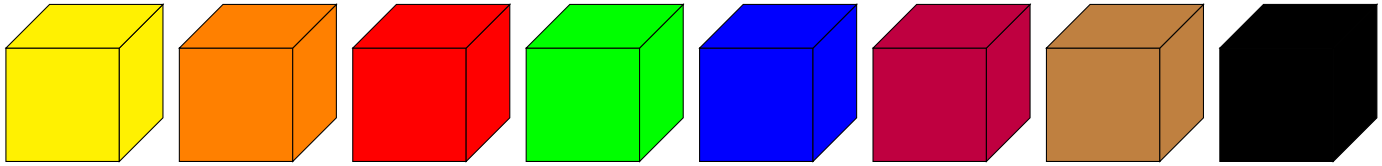
Entropy as Possibility

"*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."

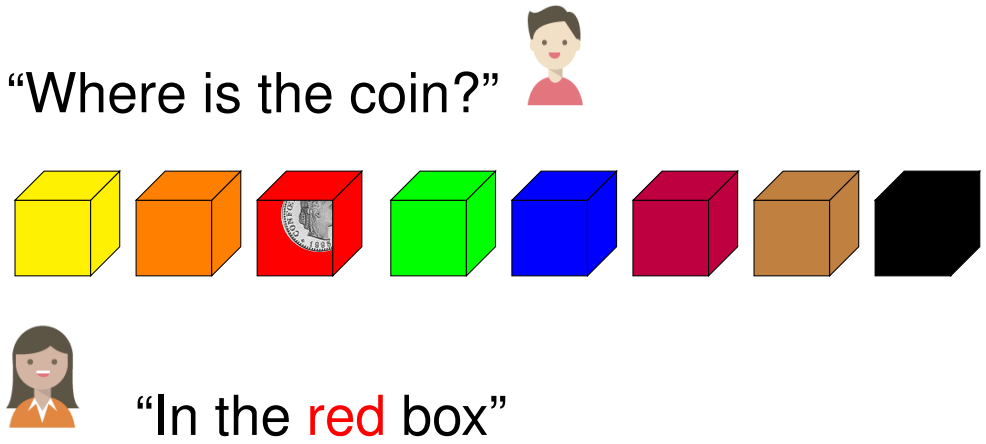
Lemons (2013). A student's guide to entropy, p. 160.

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



- 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?



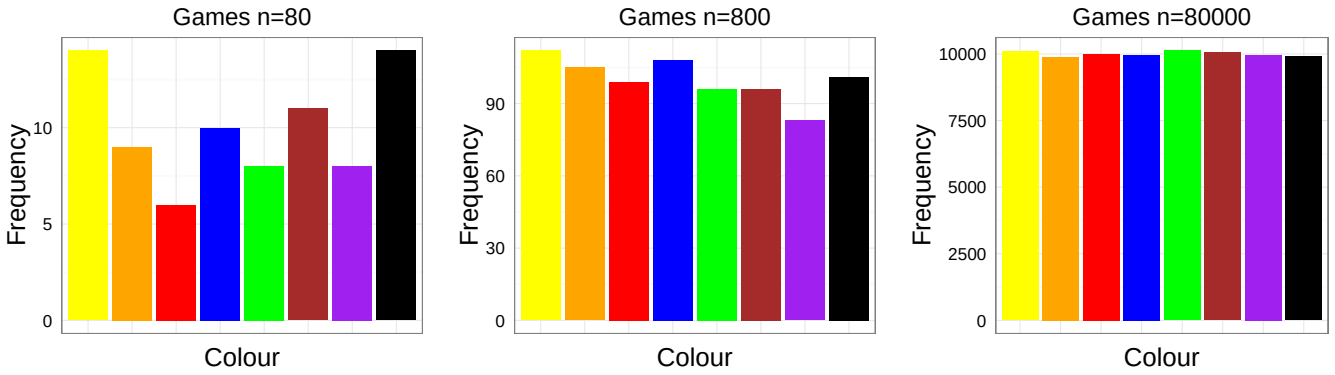
- The “alphabet” (here words) of the “language” they use does not need more than 8 colour adjectives to disambiguate:

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

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 $n \to \infty$.

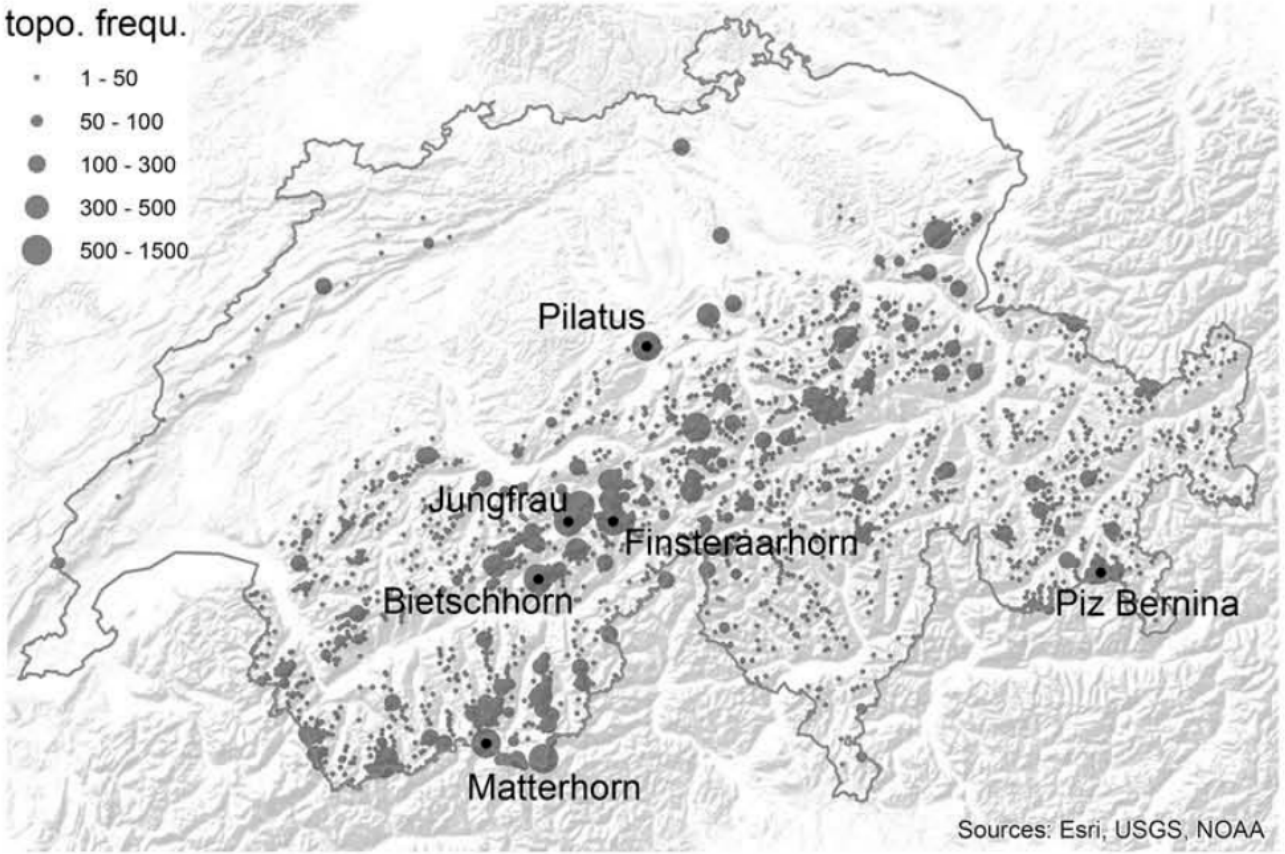


The bottom line is:

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?

Example: Frequencies of Mountain Names



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

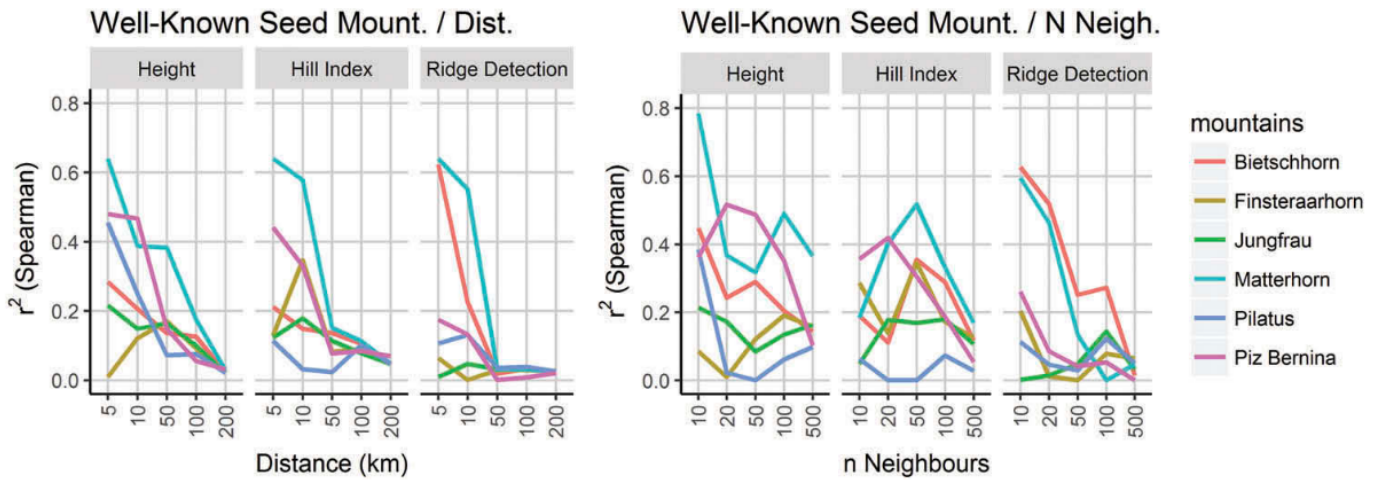


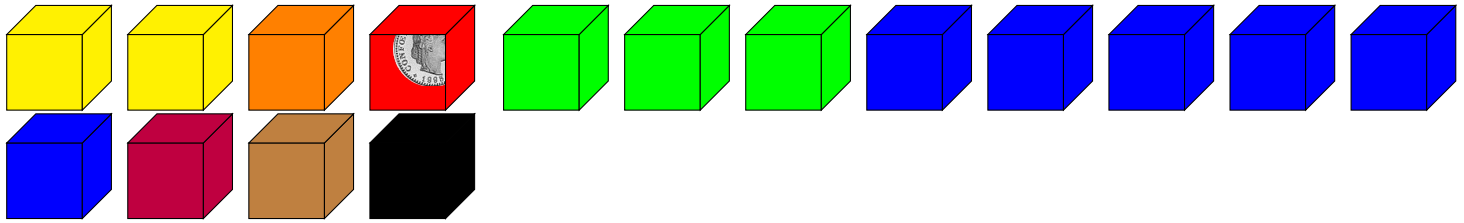
Figure 5. The relation toponym frequency: spatial measure tested for different spatial extents and a set of well-known seed mountains.

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"

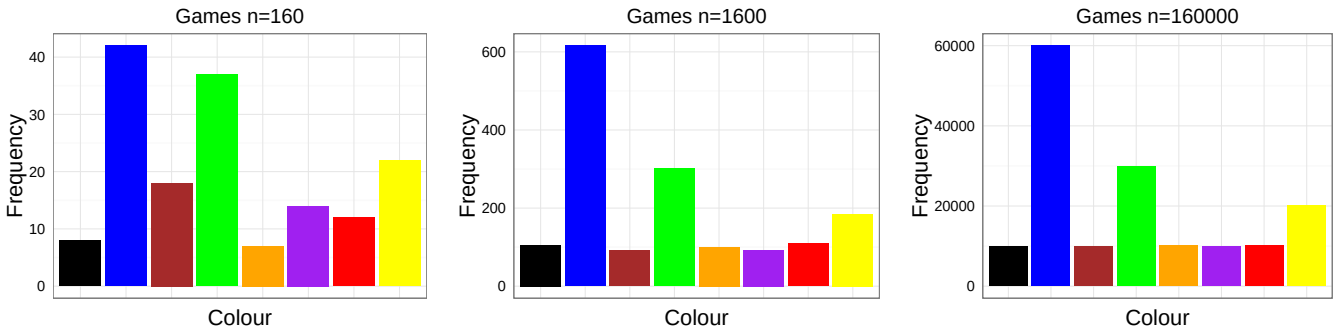
- The "alphabet" has **not** changed:

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

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

"In the red, green, blue, blue yellow, purple, blue,... box"

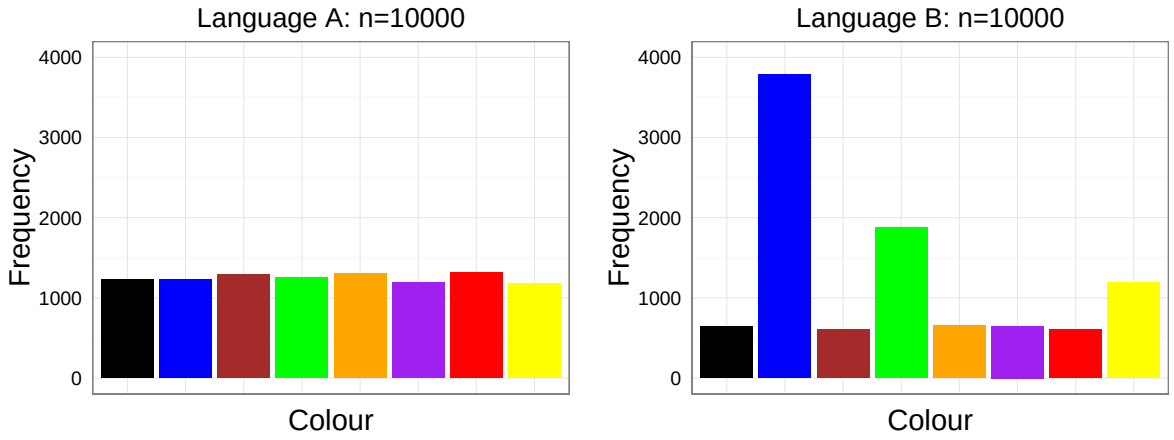
Again, this will be reflected in the **girl's language production**.



Comparing language production

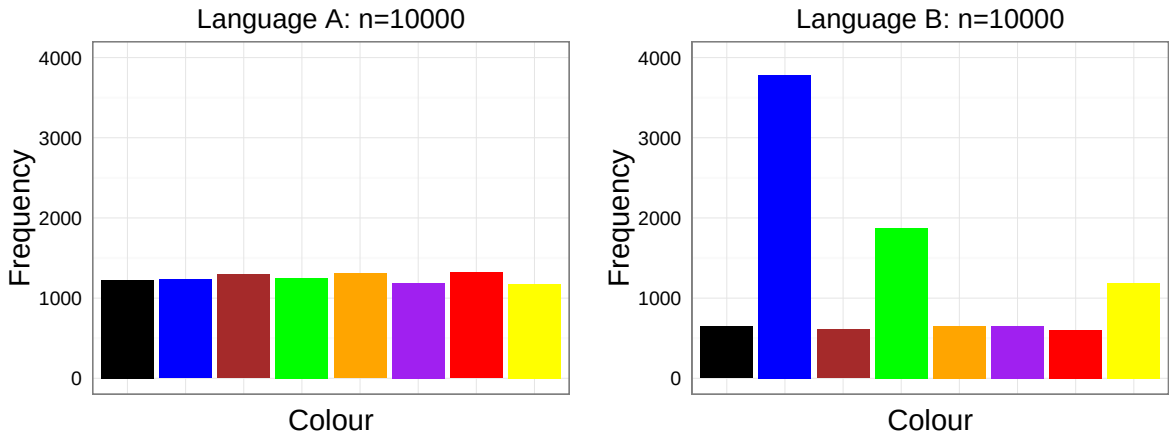
If we play the two games the same number of times n, we will get the same two languages L_A and L_B in terms of **word types** (8 in this case), and the number of **word tokens** (10K in this case).

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



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”.



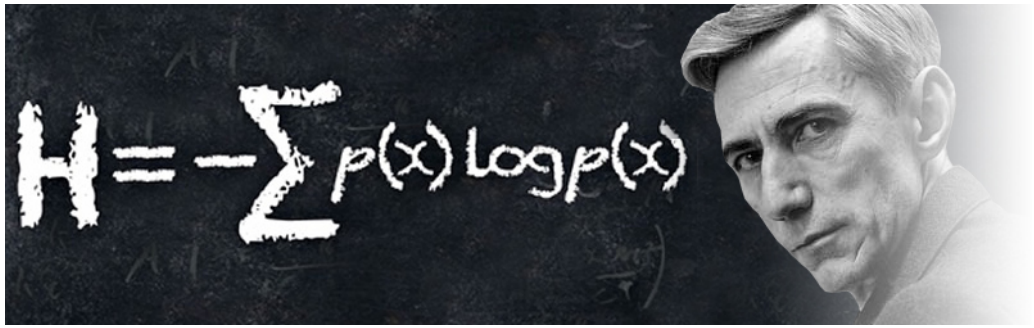
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

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.



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

A more precise formulation

(See also Cover & Thomas, 2006)

Assume that

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

  Example: The "alphabet" or set of colour adjectives, e.g. $\mathcal{A} = \{yellow, orange, red, green, blue, purple, brown, black\}$, with $N = 8$
- The *probability mass function* is defined by $p(x) = Pr\{X = x\}, x \in \mathcal{X}$

  Example: each word type is assigned a probability, e.g. in L_B $p(blue) = \frac{6}{16}$, $p(green) = \frac{3}{16}$ etc.

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 explicitely.
- 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 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 $-\log_2(\frac{3750}{10000}) \sim 1.42$ bits. The word type "orange", on the other hand, occurs ca. 625 times in 10000 tokens, its information content is $-\log_2(\frac{625}{10000}) \sim 4$ bits. Hence, the word type "orange" has higher information content.

Let's look at the component parts

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

- 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.

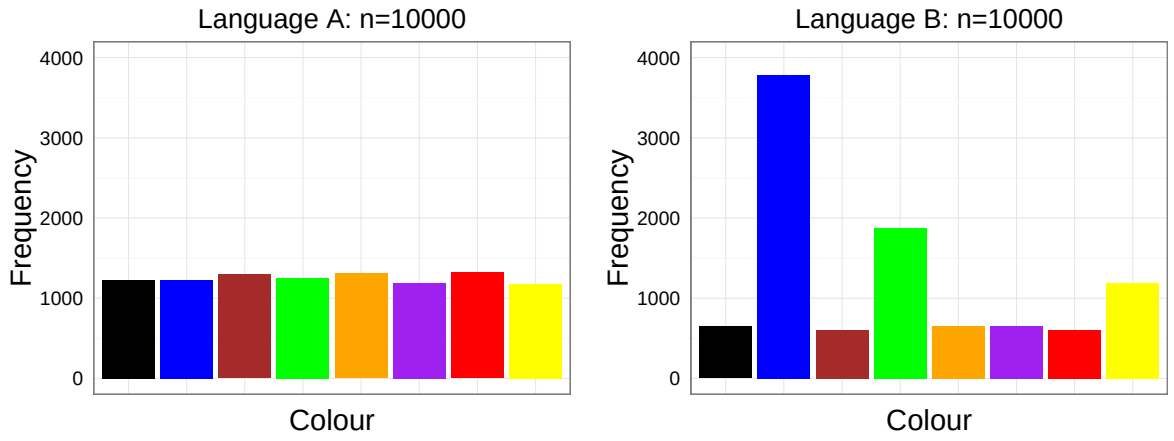
  Hence, the entropy $H(X)$ can be seen as the average information content of information encoding units, i.e. adjective word types in the case of the box game.

Let's apply this to Languages A and B

With the pre-defined probabilities we get:

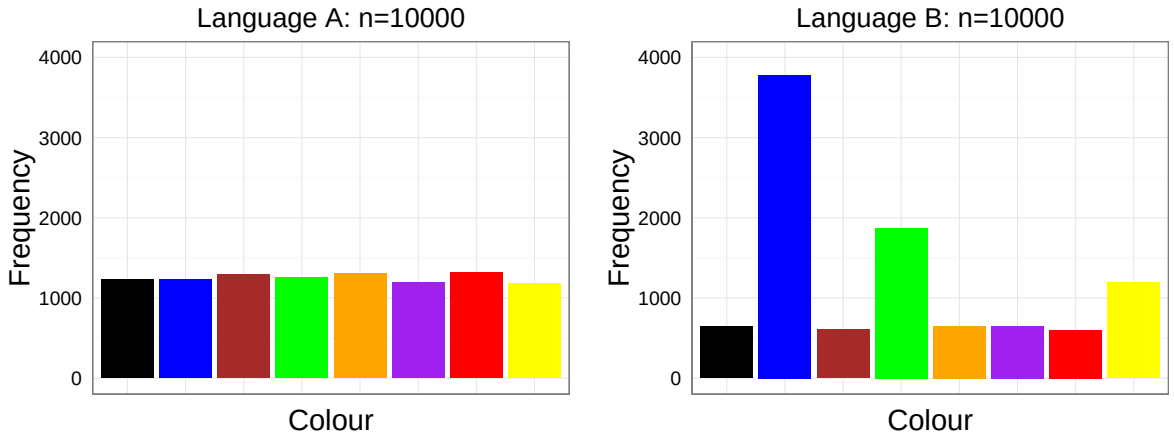
$$H(L_A) = -(\frac{1}{8} \times \log_2(\frac{1}{8}) + \frac{1}{8} \times \log_2(\frac{1}{8}) + ... + \frac{1}{8} \times \log_2(\frac{1}{8})) = \mathbf{3}^1 \quad (4)$$

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



[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 $\log_2(8)=3$.

- 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 $\log_2(1) = 0$.

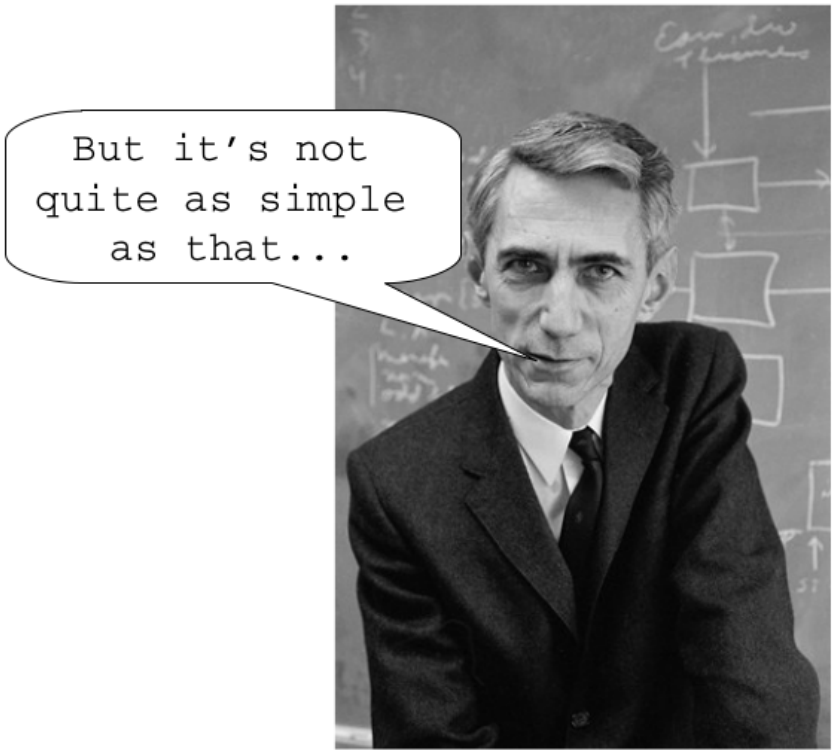


EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Section 4: Entropy Estimation

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



Probabilities

For all information-theoretic measures (not only the entropy) a crucial ingredient are the **probabilities** of information encoding units:

$$p(x),\ p(x, y),\ p(y|x)$$

Information Content (Surprisal)

$$I(x) = -\log_2 p(x) \tag{6}$$

Entropy

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \tag{7}$$

Joint Entropy

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x, y) \tag{8}$$

Conditional Entropy

$$H(Y|X) = -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log_2 p(y|x) \tag{9}$$

Probability Estimation

The simplest, most straightforward, but also most naive estimator for probabilities is the so-called **Maximum Likelihood (ML)** or plug-in estimator, i.e. taking the *relative frequency* f_i of a unit x_i as its probability such that

$$\hat{p}(x_i) = \frac{f_i}{\sum_i^N f_i}, \tag{10}$$

where i is a running index, and N is the alphabet size.

"blue ... blue ... red ... blue ... orange ... green"

$$\hat{p}(blue) = \frac{3}{6} \tag{11}$$

Note: The hat above the probability symbol $\hat{p}$ indicates that we are *estimating* the probability, rather than *pre-defining* it, as we did in the session on Information Theory Basics.

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?

Problem 1: Information Encoding Units

In the case of natural language writing, the "units" of information encoding could be characters, syllables, morphemes, orthographic words, phrases, sentences, etc. That is, the "alphabet" over which we estimate information-theoretic measures can differ vastly.

`All human beings are born free and equal in dignity and rights.`

UTF-8 characters: $\mathcal{A} = \{A, a, b, d, e, f, g, h, i, l, \dots\}$

Character bigrams: $\mathcal{A} = \{Al, ll, lh, hu, um, ma, an, nb, be, ei, in, ng, \dots\}$

Syllables: $\mathcal{A} = \{All, hu, man, be, ings, are, born, \dots\}$

Morphemes: $\mathcal{A} = \{All, human, be, ing, s, are, born, \dots\}$

Orthographic words: $\mathcal{A} = \{All, human, beings, are, born, \dots\}$

Word bigrams: $\mathcal{A} = \{All\ human, human\ beings, beings\ are, are\ born, \dots\}$

etc.

Problem 2: Sample Size

The probabilities of characters, syllables, words, etc. depend on the **corpus size**, and so do the estimations of information-theoretic measures.

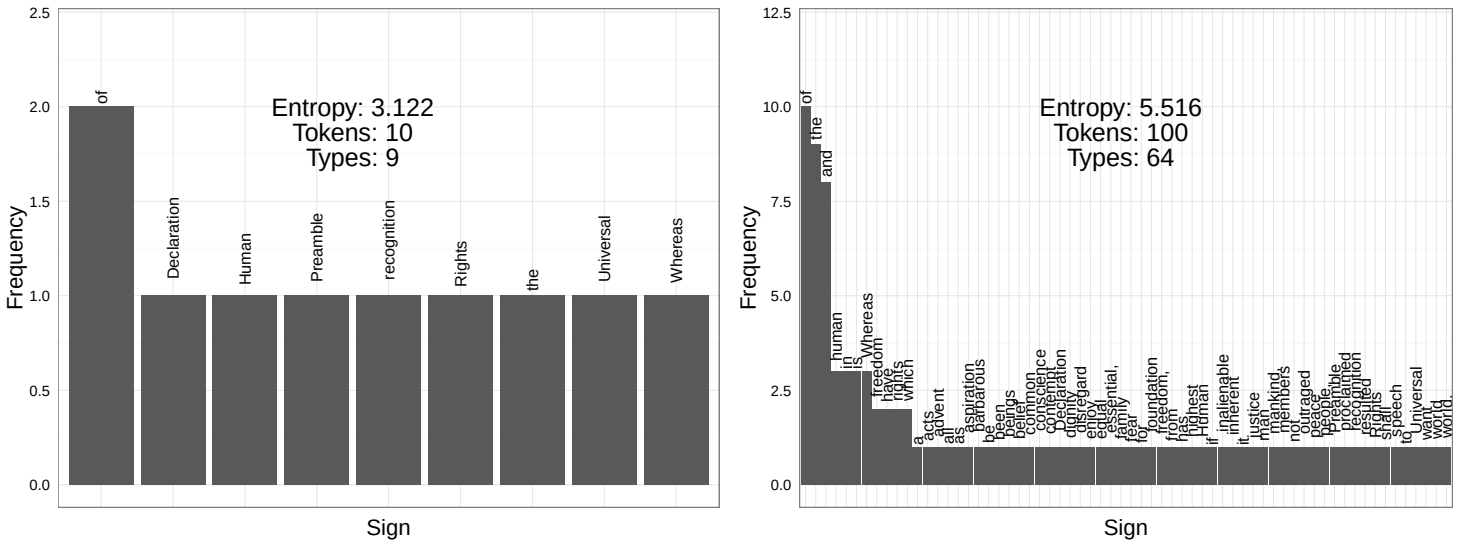
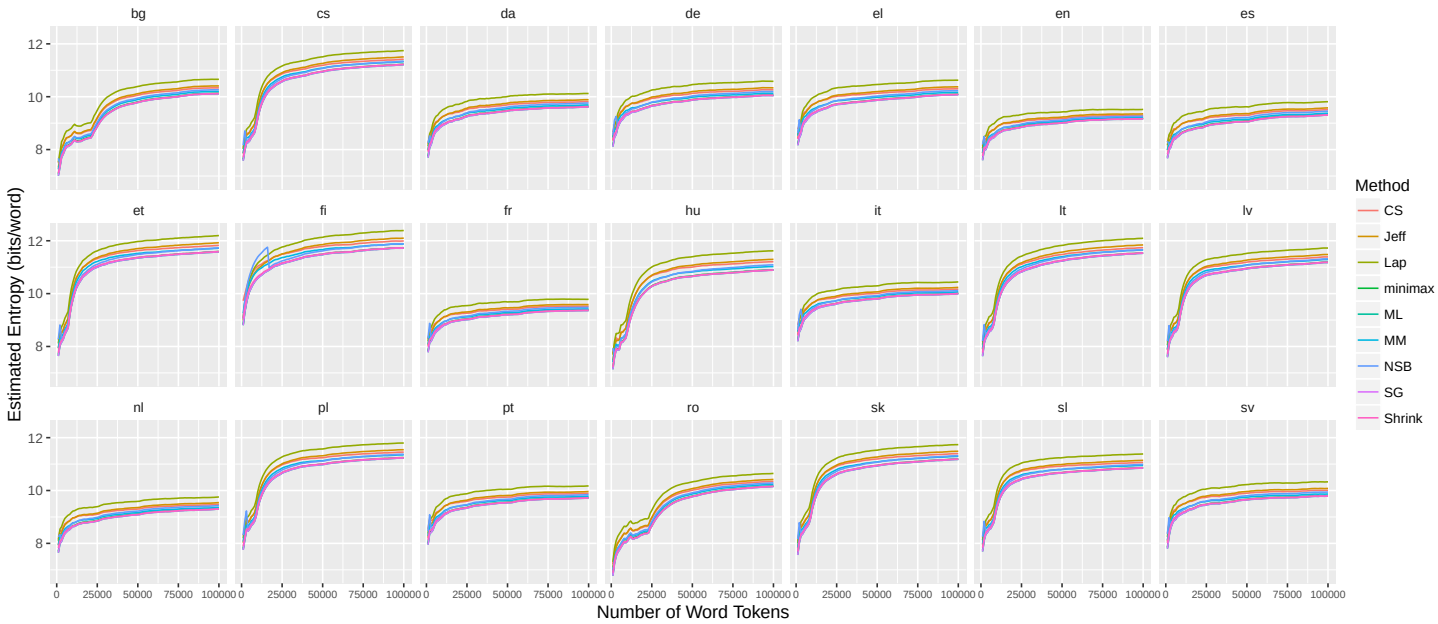


Figure. Frequency distributions and word type entropies for the English UDHR according to the first 10 and 100 word tokens.

Possible Solution for Problem 2

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



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

Problem 3: Interdependence of Units

In the case of natural language writing, characters, words, phrases etc. are **not identically** and **independently** distributed variables (i.i.d). Instead, the **co-text** and **context** results in systematic **conditional probabilities** between units:

$$p(y|x) = \frac{p(x,y)}{p(x)} \tag{12}$$

`Preamble Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world [...]`

$\hat{p}(the) = \frac{5}{32} \sim \mathbf{0.16}$,

$\hat{p}(the|of) = \frac{p(of,the)}{p(of)} = \frac{\frac{3}{32}}{\frac{5}{32}} \sim \mathbf{0.6}$.

Note: There are 32 orthographic word tokens, and 31 orthographic word bigram tokens in this example. We here take a simple ML estimate of unigram and bigram probabilities.

Possible Solution for Problem 3

- 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.

Kontoyiannis et al. (1998). Nonparametric entropy estimation for stationary processes and random fields, with applications to English text.

Cover & Thomas (2006). Elements of information theory, p. 74.

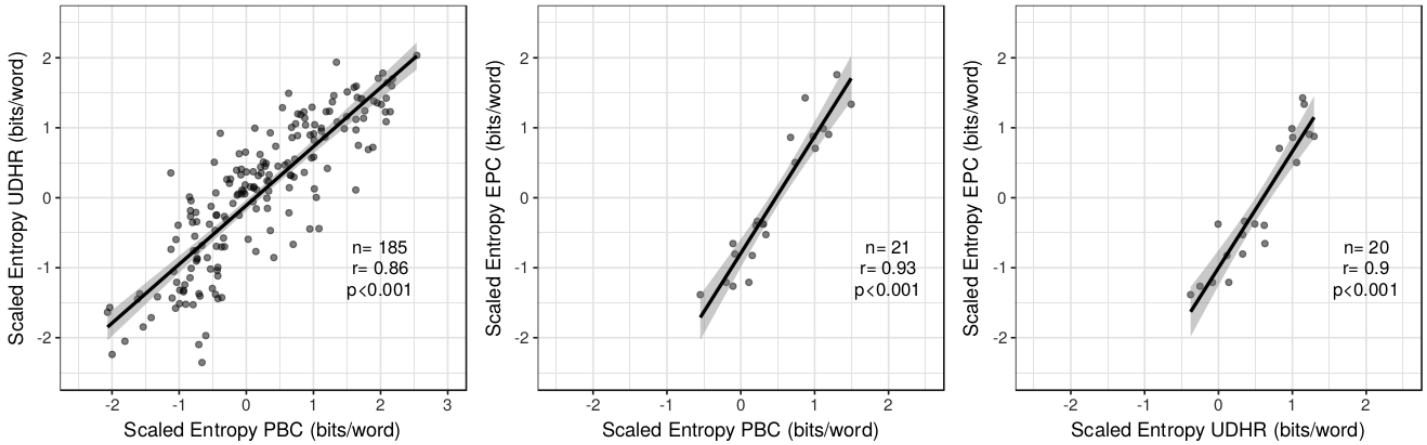
Gao, Kontoyiannis, & Bienenstock (2008). Estimating the entropy of binary time series: Methodology, some theory, and a simulation study.

Lesne et al. (2009). Entropy estimation for very short symbolic sequences.

Gutierrez-Vasques & Mijangos (2020). Productivity and predictability for measuring morphological complexity.

Problem 4: Extrapolation

When estimating information-theoretic measures for natural languages, we can only use a snapshot of the overall language production (of all speakers and writers). The question then is to what extend our results **extrapolate** beyond our limited sample. A possible solution to this problem is to compare estimations between different corpora.



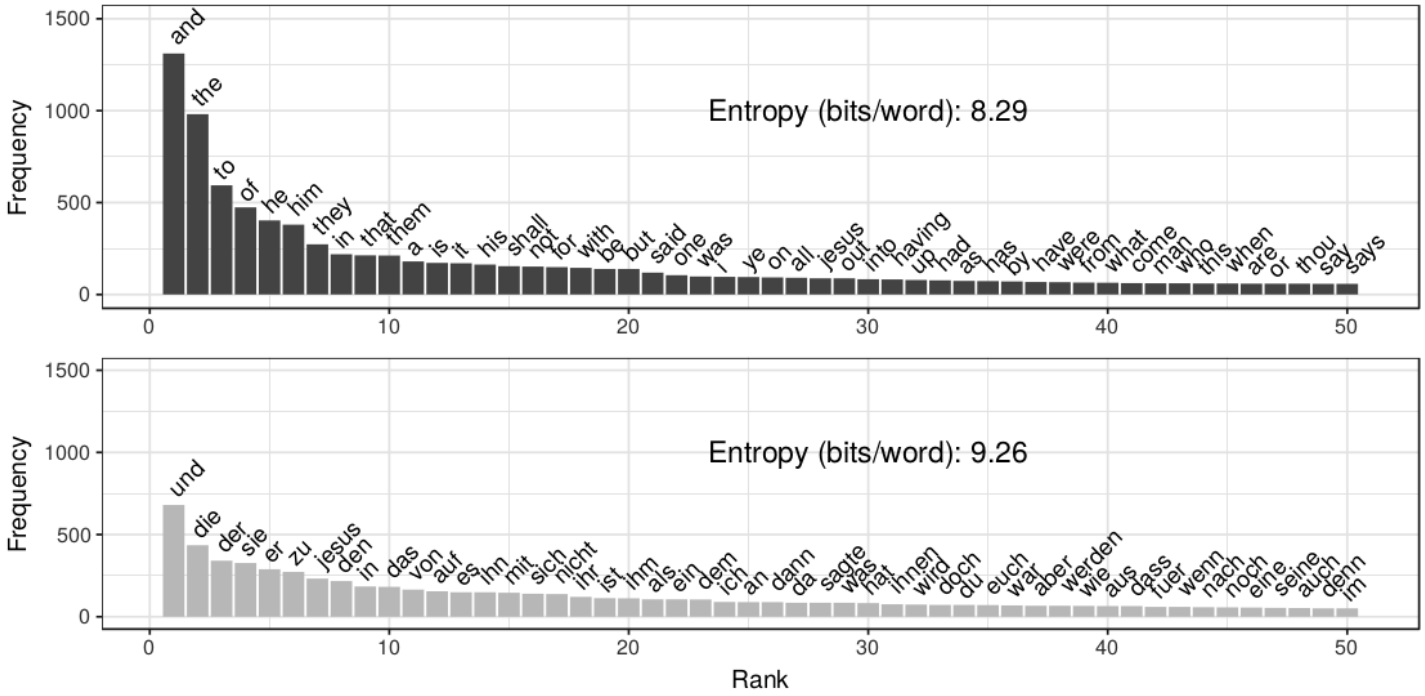
Bentz (2018). Adaptive languages: An information-theoretic account of linguistic diversity, p. 108.

Methods for Probability Estimation

- **Frequency-Based**: i.e. counting frequencies in corpora (and smoothing the counts with more advanced estimators).
- **Language Models**: train (neural) language models on texts, and get transition-probability estimates from these.
- **Experiments with Humans**: have humans predict the next character/word in a sentence, and calculate the probabilities from their precision.

Frequency-Based Estimation

We can estimate probabilities of units (here orthographic words) from written texts/corpora via the ML estimator (relative frequencies) or less biased estimators (here James-Stein Shrinkage estimator).

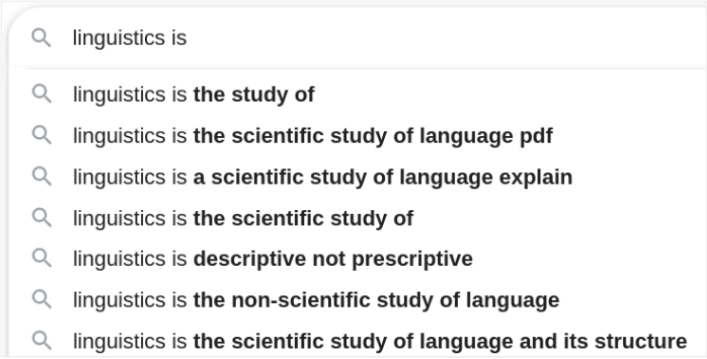


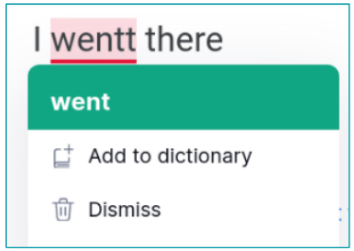
Bentz (2018). Adaptive languages, p. 88.

Language Models

Useful tool in NLP for estimating the probability of sequences

- For example, we can use them for calculating the probability of a sentence in a language (based on a text corpus)
- Many applications in NLP





We want to calculate: $P(w_1, w_2, \ldots, w_n)$

https://github.com/christianbentz/Workshop_DGfS2022

Experiments with Humans

"A new method of estimating the entropy and redundancy of a language is described. This method exploits the knowledge of the language statistics possessed by those who speak the language, and depends on experimental results in prediction of the next letter when the preceding text is known."


```
(1) THE ROOM WAS NOT VERY LIGHT A SMALL OBLONG
(2) ----ROO------NOT-V-----I------SM----OBL----
(1) READING LAMP ON THE DESK SHED GLOW ON
(2) REA----------O------D----SHED-GLO--O--
(1) POLISHED WOOD BUT LESS ON THE SHABBY RED CARPET
(2) P-L-S-----O---BU--L-S--O------SH-----RE--C------
```

Shannon (1951). Prediction and entropy of printed English.

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.

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.

Pereira, Fernando (2000). Formal grammar and information theory: together again? *Philosophical Transactions of the Royal Society A*. Volume 358, 1239-1253.

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