EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

DGfS Workshop 2022

Information-Theoretic Analyses of Natural Languages

22/02/2022, Christian Bentz & Ximena Gutierrez-Vasques

Table of Contents

10:00 – 11:00
Introduction &
General Setup

11:00 – 12:00
Information-
Theory: Basics

12:00 – 13:00
Lunch Break

13:00 – 14:00
Estimation
Methods

14:00 – 15:00
Application 1:
Voynich
Manuscript

15:00 – 16:00
Application 2:
Morphology &
Syntax

16:00 – 17:00
Summary & Final
Discussion

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Table of Contents

10:00 – 11:00 Introduction & General Setup

11:00 – 12:00 Information-Theory: Basics

12:00 – 13:00 Lunch Break

13:00 – 14:00 Estimation Methods

14:00 – 15:00 Application 1: Voynich Manuscript

15:00 – 16:00 Application 2: Morphology & Syntax

16:00 – 17:00 Summary & Final Discussion

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Introduction

Dr. Ximena Gutierrez-Vasques

University of Zurich[UZH]

10:00 – 11:00
Introduction & General Setup

11:00 – 12:00
Information-Theory: Basics

12:00 – 13:00
Lunch Break

13:00 – 14:00
Estimation Methods

14:00 – 15:00
Application 1: Voynich Manuscript

15:00 – 16:00
Application 2: Morphology & Syntax

16:00 – 17:00
Summary & Final Discussion

Dr. Christian Bentz

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

10:00 – 11:00
Introduction &
General Setup

11:00 – 12:00
Information-
Theory: Basics

12:00 – 13:00
Lunch Break

13:00 – 14:00
Estimation
Methods

14:00 – 15:00
Application 1:
Voynich
Manuscript

15:00 – 16:00
Application 2:
Morphology &
Syntax

16:00 – 17:00
Summary & Final
Discussion

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Acknowledgements

URPP Language and Space

University of Zurich[UZH]

10:00 – 11:00
Introduction & General Setup

11:00 – 12:00
Information-Theory: Basics

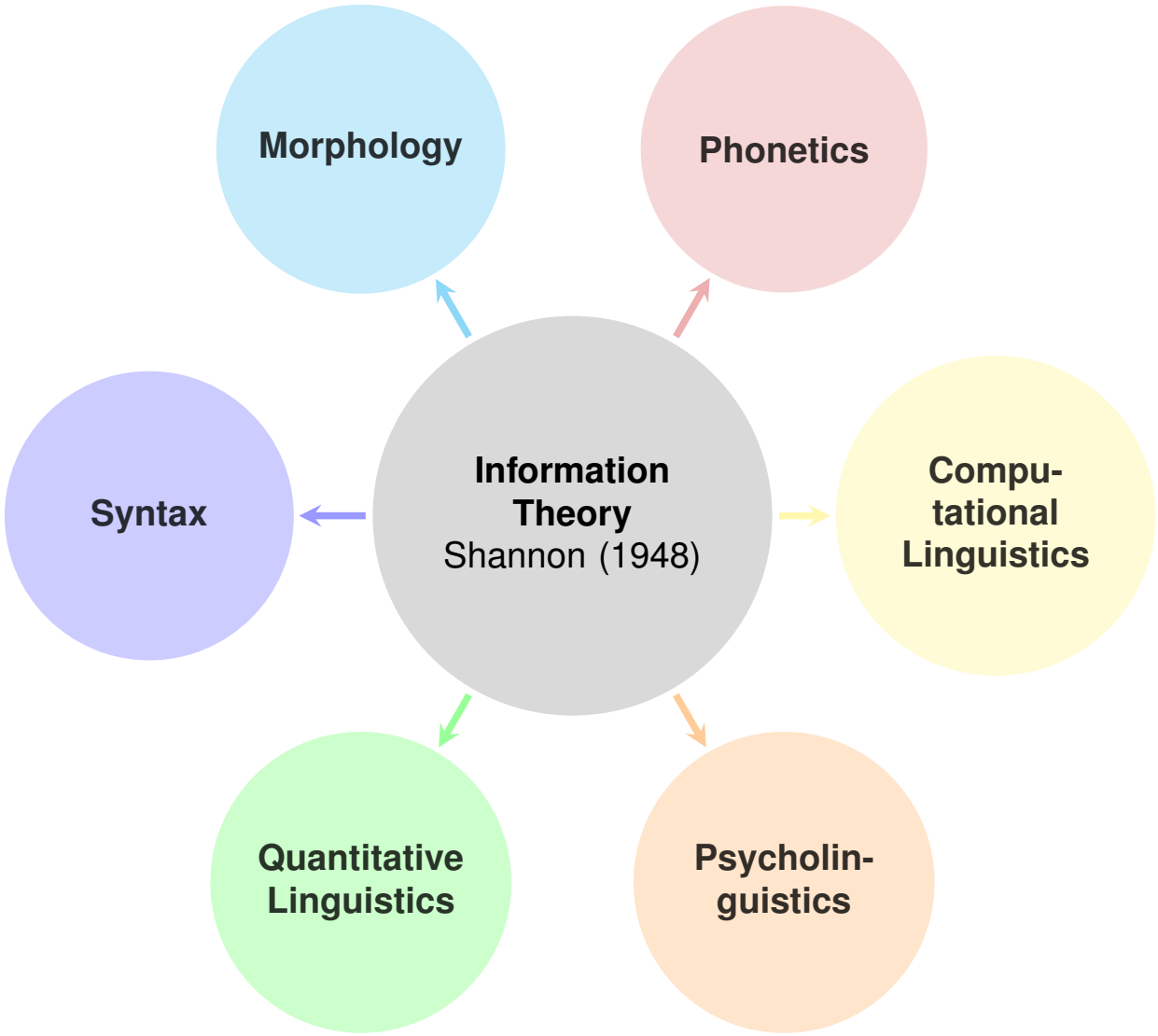
12:00 – 13:00
Lunch Break

13:00 – 14:00
Estimation Methods

14:00 – 15:00
Application 1: Voynich Manuscript

15:00 – 16:00
Application 2: Morphology & Syntax

16:00 – 17:00
Summary & Final Discussion



10:00 – 11:00 Introduction & General Setup

11:00 – 12:00 Information-Theory: Basics

12:00 – 13:00 Lunch Break

13:00 – 14:00 Estimation Methods

14:00 – 15:00 Application 1: Voynich Manuscript

15:00 – 16:00 Application 2: Morphology & Syntax

16:00 – 17:00 Summary & Final Discussion

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Quantitative Linguistics

The Voynich Manuscript
https://en.wikipedia.org/wiki/Voynich_manuscript

10:00 – 11:00
Introduction & General Setup

11:00 – 12:00
Information-Theory: Basics

12:00 – 13:00
Lunch Break

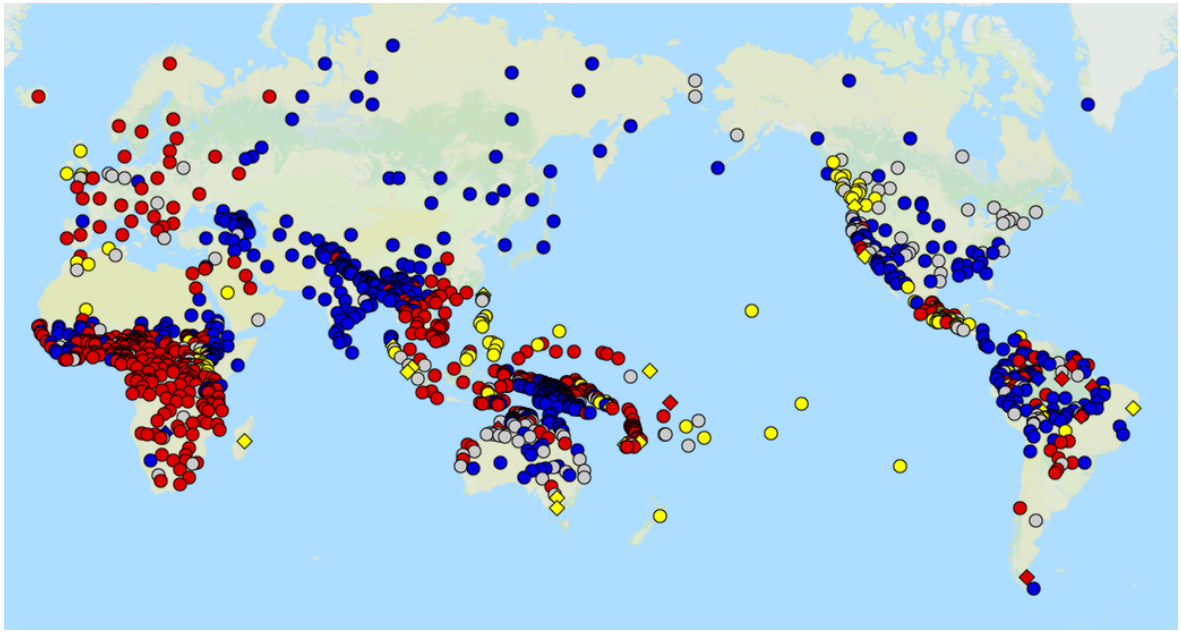
13:00 – 14:00
Estimation Methods

14:00 – 15:00
Application 1: Voynich Manuscript

15:00 – 16:00
Application 2: Morphology & Syntax

16:00 – 17:00
Summary & Final Discussion

Syntax & Morphology



Word Order ↔ Morphological Marking
https://wals.info/chapter/81

General Setup

Practical session

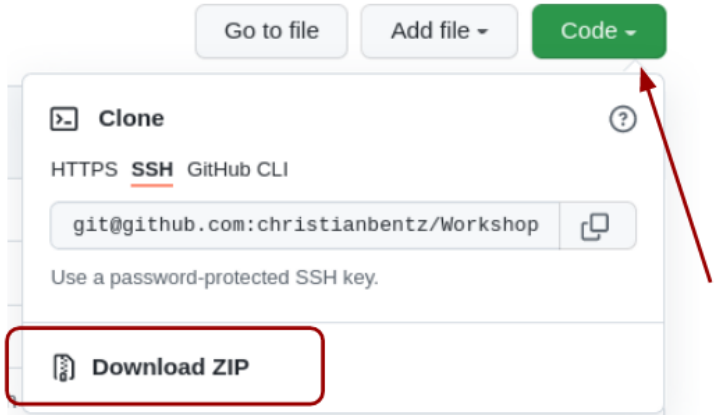
- **R** and **Python**
- No need of installation. We'll run everything through **Google Colab** `https://colab.research.google.com/`

- **Only requirement:** account linked to Google

General Setup

Practical session

Step 1: Download the material from Github and decompress it on your local machine (or use git clone [URL_Repository])



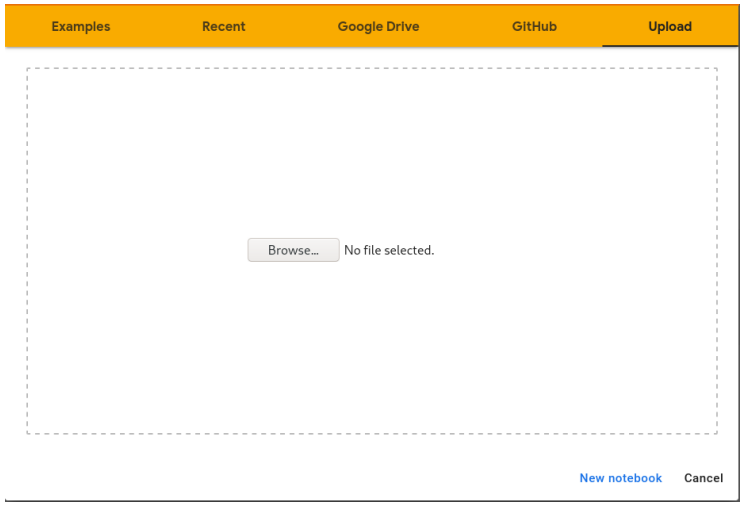
`https://github.com/christianbentz/Workshop_DGfS2022`

General Setup

Practical session

Step 2: Go to Google Colab, and upload the current notebook (.ipynb) that we are going to use

- For example:
  Workshop_DGfS2022-main/Code/Application1/TextPreprocessing.ipynb



Alternatively, you can upload notebooks using the Github tab (you must specify the url of the .ipynb notebook)

Practical Part

Run the following code files (in this order):

- `Application1/TextPreprocessing.ipynb`

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Table of Contents

10:00 – 11:00 Introduction & General Setup

11:00 – 12:00 Information-Theory: Basics

12:00 – 13:00 Lunch Break

13:00 – 14:00 Estimation Methods

14:00 – 15:00 Application 1: Voynich Manuscript

15:00 – 16:00 Application 2: Morphology & Syntax

16:00 – 17:00 Summary & Final Discussion

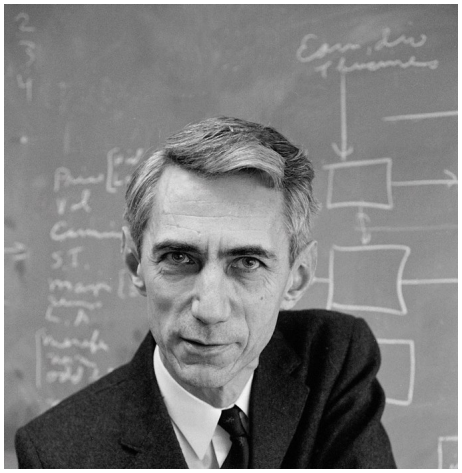
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Information-Theory: Basics

Section 1: A Brief History

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

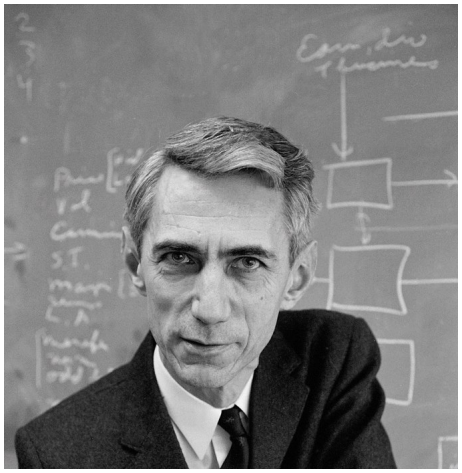
```
Article 1
All human beings are born free and equal in dignity
and rights.  They are endowed with reason and
conscience and should act towards one another in a
spirit of brotherhood.
```

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 ???

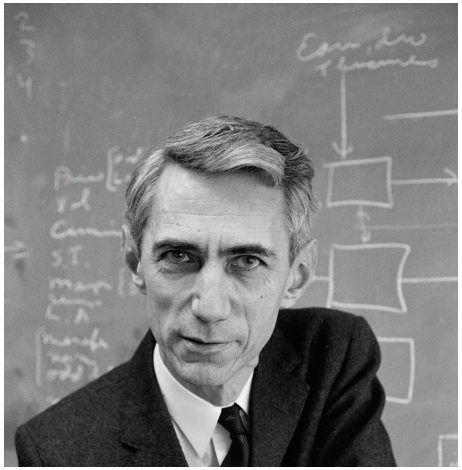
A Brief History of Information and Language



[...] 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.)

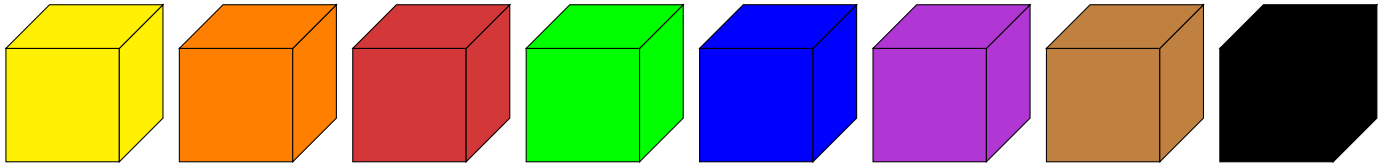
Shannon & Weaver (1949). The mathematical theory of communication, p. 4.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Section 2: Information Content (Surprisal)

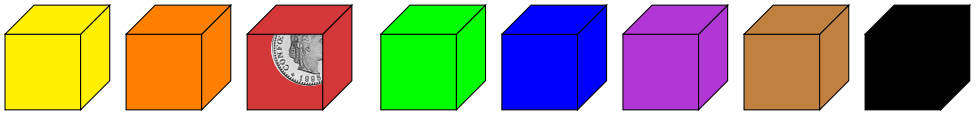
Let's play the Box Game!



- 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

Language in The Box Game

"Where is the coin?"



"In the red box"

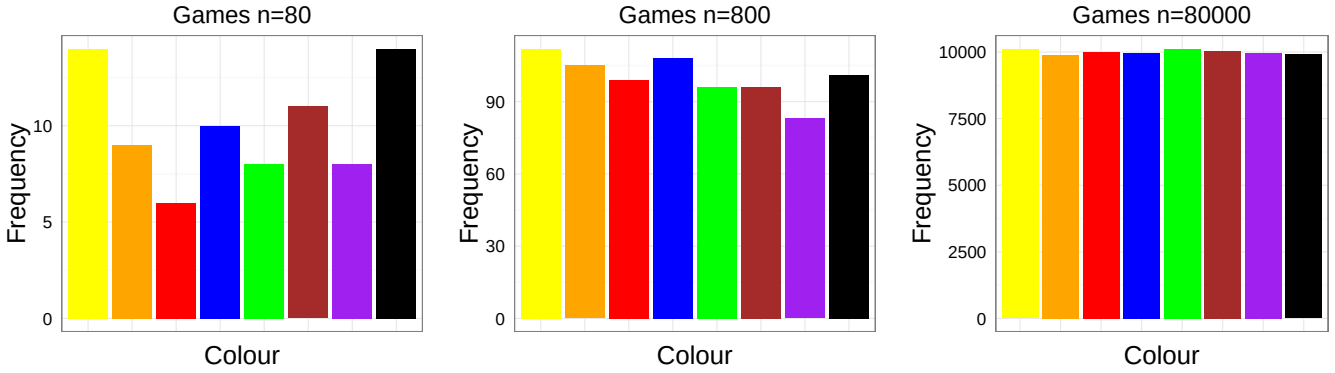
- The "alphabet" (here words) of the "language" they use does not need more than 8 colour adjectives to disambiguate:

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

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.

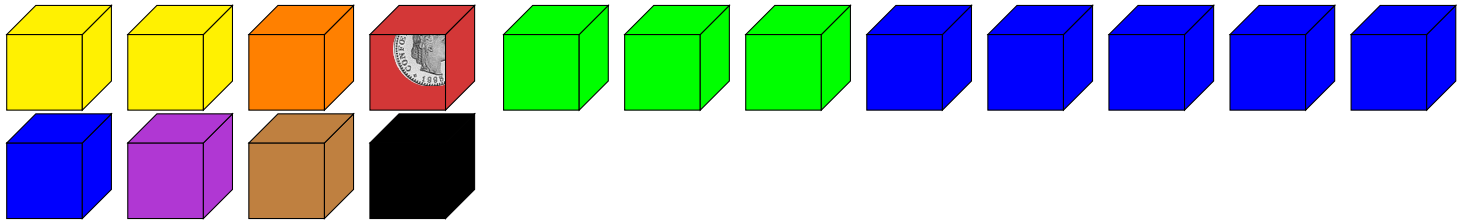
"In the red/green/blue/ yellow/purple/brown/black ... box"

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



What if we change the game?

"Where is the coin?"



"In the red box"

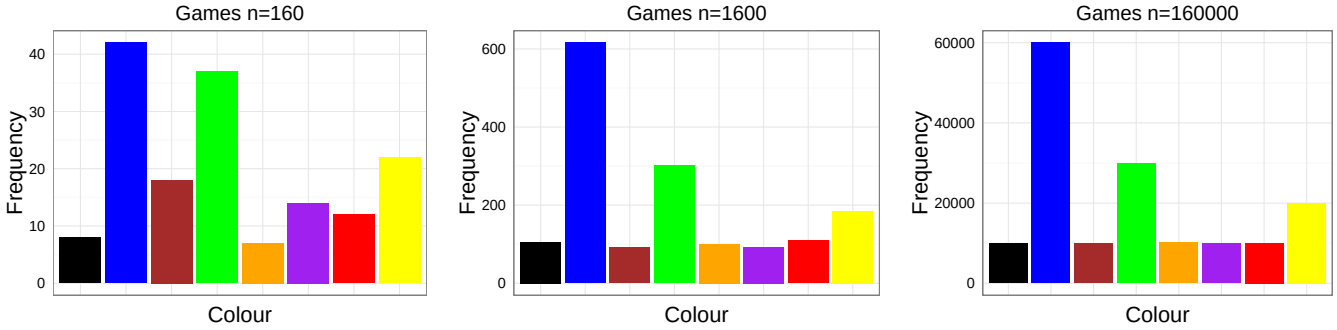
- The "alphabet" has **not** changed:

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

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.

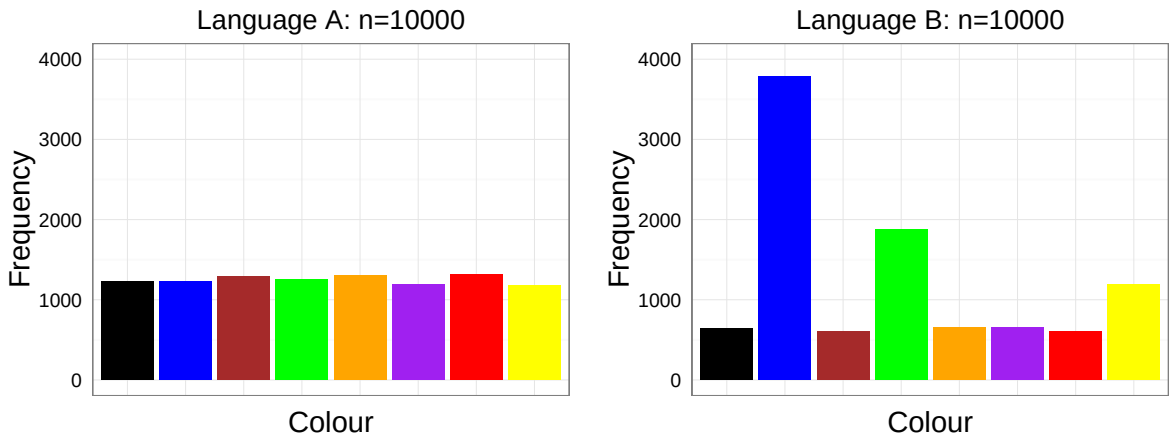
"In the red, green, blue, blue yellow, purple, blue,... box"

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



Information in the Game

How “surprised” should we be to hear the girl say “blue” in the first version of the game, and how “surprised” in the second version of the game? In other words, how much information do we gain by hearing “blue”?



A Precise Formulation

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.

Cover & Thomas (2006). Elements of information theory, p. 13-14.

Information Content (Surprisal)

The **information content**[1] of a given outcome x is then defined as

$$I(x) = -\log_2 p(x) = \log_2 \frac{1}{p(x)} \tag{1}$$

Example: The information content of the word "blue" in language A is $-\log_2 p(\frac{1}{8}) = 3$ bits, while for language B it is $-\log_2 p(\frac{6}{16}) = 1.4$ bits.

Borda (2011). Fundamentals in information theory and coding, p. 9-11.

[1] This is also variously called *self-information*, *Shannon information*, or *surprisal* in the literature.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Section 3: Entropy

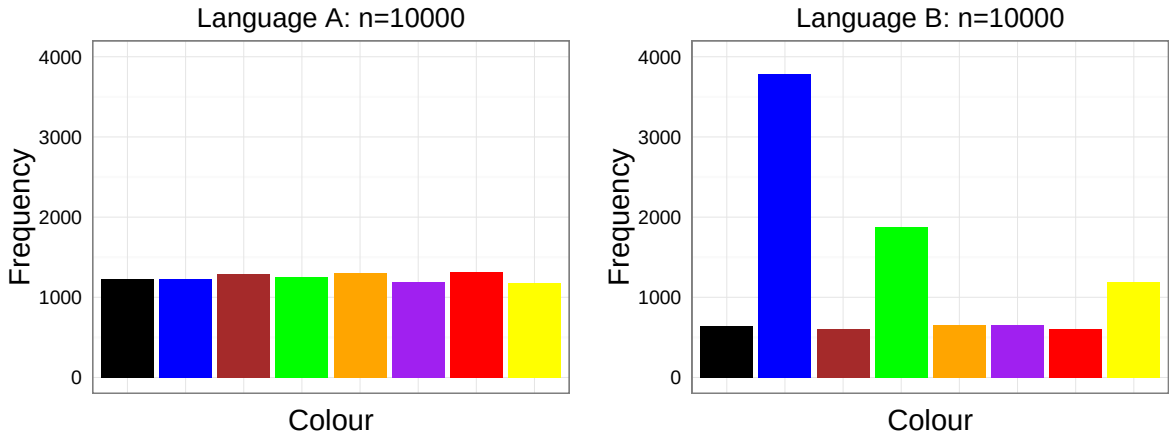
Some Intuitive Terminology

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

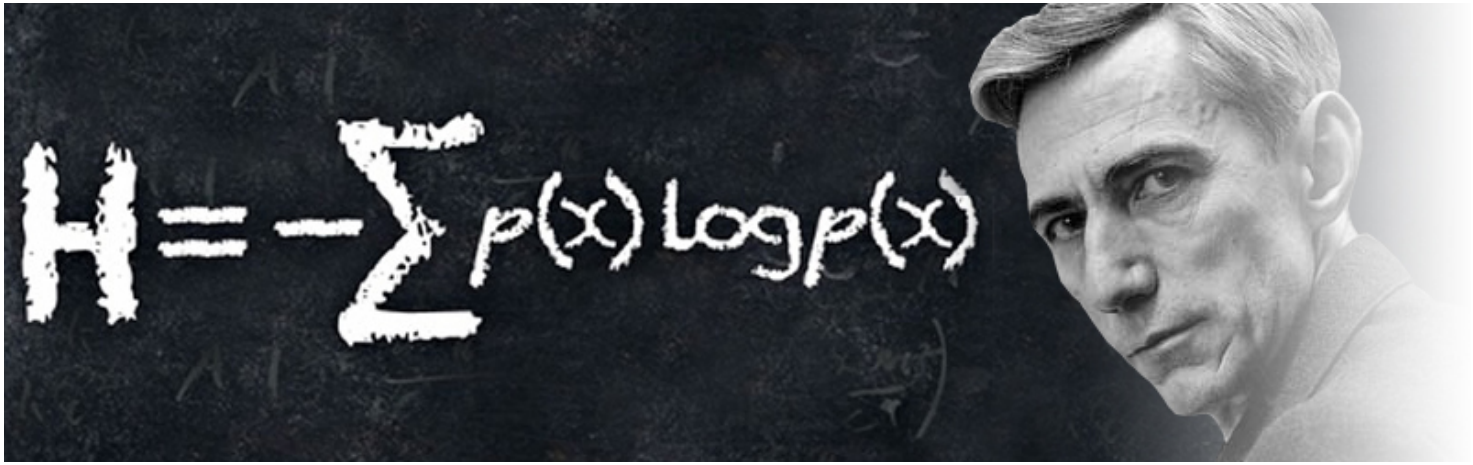
} Entropy

Certainty/Uncertainty in the Game

Note that overall 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”.



How can we measure the difference?



Shannon (1948). A mathematical theory of communication.

A more precise formulation

Given these definitions, the entropy is then defined as

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

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

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

$-\log_2 p(x)$ is the **information content** of a unit x (word type in the case of the box game).

Example: in L_B the word type "blue" has a probability of $\frac{3}{8}$, and its information content is hence $-\log_2(\frac{3}{8}) \sim 1.42$ bits. The word type "orange", on the other hand, has a probability of $\frac{1}{8}$, its information content is $-\log_2(\frac{1}{8}) = 3$ bits. Hence, the word type "orange" has higher information content.

Let's look at the component parts

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \tag{4}$$

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**.

The Entropy

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \tag{5}$$

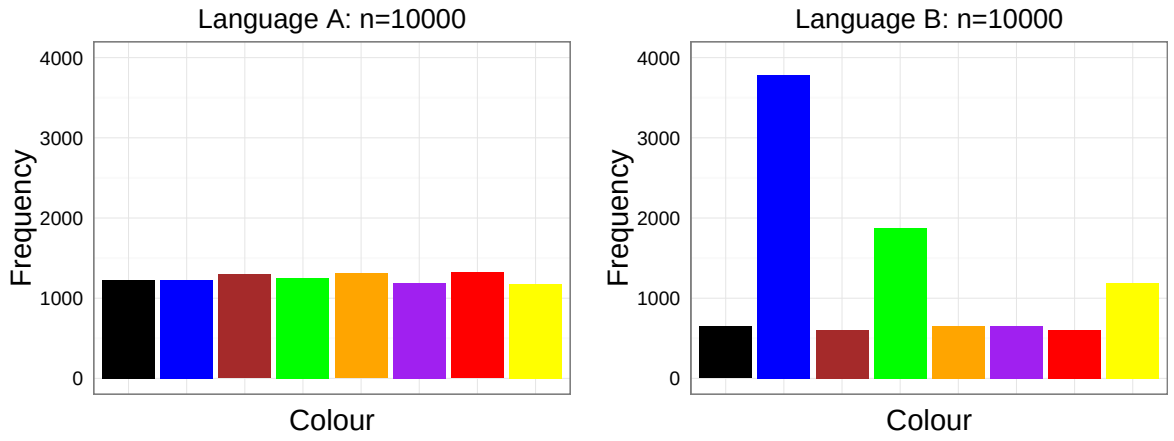
Hence, the entropy $H(X)$ can be seen as the **average information content** of information encoding units, i.e. adjective word types in the case of the box game.

Let's apply this to Languages A and B

Plugging the predefined probabilities of words into the formula gives us:

$$H(L_A) = -(\frac{1}{8} \times \log_2(\frac{1}{8}) + \frac{1}{8} \times \log_2(\frac{1}{8}) + ... + \frac{1}{8} \times \log_2(\frac{1}{8})) = \mathbf{3}^2 \quad (6)$$

$$H(L_B) = -(\frac{6}{16} \times \log_2(\frac{6}{16}) + \frac{3}{16} \times \log_2(\frac{3}{16}) + ... + \frac{1}{16} \times \log_2(\frac{1}{16})) = \mathbf{2.61} \quad (7)$$



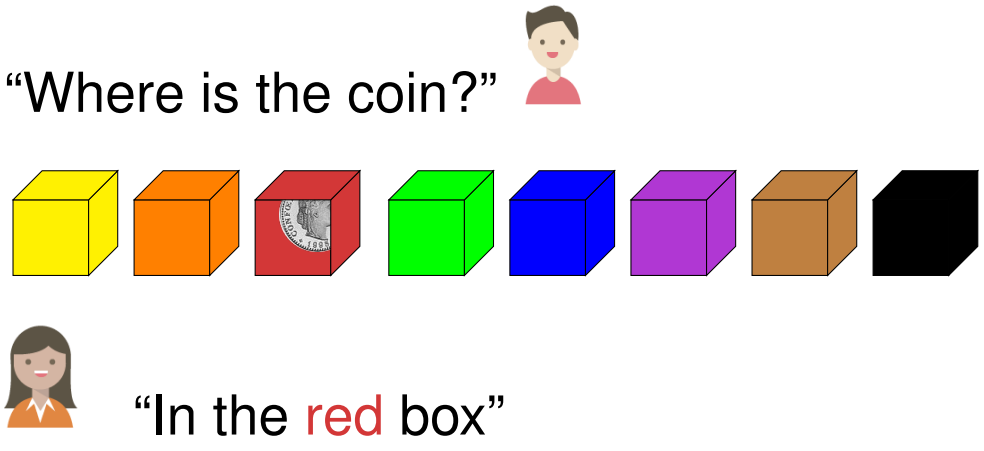
[2]Note: the case where we have a uniform distribution of probabilities, i.e. all events (adjectives here) are exactly equally likely, is the *maximum entropy* case. In this case, the equation simplifies to $log_2(N)$. Such that here we have $\log_2(8)=3$.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Section 4: Joint Entropy

Two Random Variables

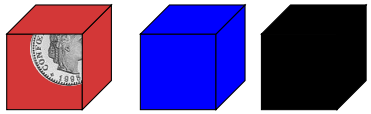


So far, we have assumed that the girl’s language production is a perfect mirror of the world (box game outcome). Now, imagine we rather see the **world as one random variable** X, and the **language of the girl as another random variable** Y. In a sense, this means that we allow for discreptancies between the two.

Simplified Version of the Box Game

Let's assume a simplified version with only three boxes.

"Where is the coin?" 



 "In the red box."

Such that we have the alphabets

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

Two Random Variables

Assume the probability mass function for the **"real world" variable** X is

$$p(x) = \{\langle \text{red}, \frac{1}{3}\rangle, \langle \text{blue}, \frac{1}{3}\rangle, \langle \text{black}, \frac{1}{3}\rangle\}. \tag{8}$$

And the one for the **girl's language variable** Y is

$$p(y) = \{\langle \text{red}, \frac{1}{2}\rangle, \langle \text{blue}, \frac{1}{4}\rangle, \langle \text{black}, \frac{1}{4}\rangle\}. \tag{9}$$

If these are **independent** from one another, i.e. the girl is not actually influenced in her choice of words by the real world outcome, then the **joint probability** is given by

$$p(x, y) = p(x) \times p(y) \tag{10}$$

Joint Probability Matrix

In this case we have a joint probability matrix as below.[3]

		World X: red	World X: blue	World X: black
Girl Y	red	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
	blue	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
	black	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

[3] In case of independence the cells are filled by $p(x, y) = p(x) \times p(y)$. In case the outcomes are fully dependent (as in the examples above, i.e. when the girl's language is a perfect mirror of the world) then $p(x, y) = p(x) = p(y)$. If there is some statistical dependence of Y on X, then we would have to use conditional probabilities, i.e. $p(x, y) = p(x) \times p(y|x)$.

Joint Entropy

Given the joint probabilities $p(x, y)$ in the matrix, the so-called **joint entropy** of the two random variables is defined as:

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x, y) \tag{11}$$

This gives the amount of information (in bits) which both variables together carry.

Cover & Thomas (2006), p. 16.

Example: Calculating $H(X, Y)$

Given the joint probabilities $p(x, y)$ of the matrix defined for the box game above, we thus get the joint entropy as:

$$\begin{aligned} H(X, Y) = -(&p(red, red) \log_2 p(red, red) + \\ &p(red, blue) \log_2 p(red, blue) + \\ &p(red, black) \log_2 p(red, black) + \\ \\ &p(blue, red) \log_2 p(blue, red) + \\ &p(blue, blue) \log_2 p(blue, blue) + \\ &p(blue, black) \log_2 p(blue, black) + \\ \\ &p(black, red) \log_2 p(black, red) + \\ &p(black, blue) \log_2 p(black, blue) + \\ &p(black, black) \log_2 p(black, black)) \end{aligned} \tag{12}$$

Example: Calculating $H(X, Y)$

Which gives us:

$$\begin{aligned} H(X,Y) = -(&\frac{1}{6}\log_2\frac{1}{6} + \frac{1}{12}\log_2\frac{1}{12} + \frac{1}{12}\log_2\frac{1}{12} + \\ &\frac{1}{6}\log_2\frac{1}{6} + \frac{1}{12}\log_2\frac{1}{12} + \frac{1}{12}\log_2\frac{1}{12} + \\ &\frac{1}{6}\log_2\frac{1}{6} + \frac{1}{12}\log_2\frac{1}{12} + \frac{1}{12}\log_2\frac{1}{12}) \\ &\sim \mathbf{3.08}\,\text{bits} \end{aligned} \tag{13}$$

Note that the *maximum entropy* in the case above, i.e. with uniform distributions of probabilities would be

$$H(X,Y) = -(9 \times (\frac{1}{9}\log_2(\frac{1}{9}))) = \log_2(9) \sim \mathbf{3.17}\,\text{bits}. \tag{14}$$

We are thus already close to the maximum joint entropy in this scenario.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

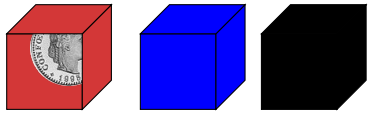
Faculty of Philosophy
General Linguistics

Section 5: Conditional Entropy

Yet Another Version of the Box Game

Let's assume the 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

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

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

$$p(x) = \{\langle \text{red}, \frac{1}{3}\rangle, \langle \text{blue}, \frac{1}{3}\rangle, \langle \text{black}, \frac{1}{3}\rangle\}. \tag{15}$$

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

$$\begin{aligned} p(y|x) = \{&\langle(\text{red}|\text{red}), 0\rangle, \langle(\text{red}|\text{blue}), 0\rangle, \langle(\text{red}|\text{black}), 0\rangle, \\ &\langle(\text{blue}|\text{red}), 1\rangle, \langle(\text{blue}|\text{blue}), 1\rangle, \langle(\text{blue}|\text{black}), 0\rangle, \\ &\langle(\text{black}|\text{red}), 0\rangle, \langle(\text{black}|\text{blue}), 0\rangle, \langle(\text{black}|\text{black}), 1\rangle\}. \end{aligned} \tag{16}$$

[4]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:

$$H(Y|X) = -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log_2 p(y|x) \tag{17}$$

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:

$$\begin{aligned} H(Y|X) = -(&p(red) \times (p(red|red) \log_2 p(red|red) + \\ &p(blue|red) \log_2 p(blue|red) + \\ &p(black|red) \log_2 p(black|red)) + \\ \\ &p(blue) \times (p(red|blue) \log_2 p(red|blue) + \\ &p(blue|blue) \log_2 p(blue|blue) + \\ &p(black|blue) \log_2 p(black|blue)) + \\ \\ &p(black) \times (p(red|black) \log_2 p(red|black) + \\ &p(blue|black) \log_2 p(blue|black) + \\ &p(black|black) \log_2 p(black|black))) \end{aligned} \tag{18}$$

Further plugging the conditional probabilities of (16) into Equation (18) gives us:

$$\begin{aligned} H(Y|X) = -(&\frac{1}{3} \times (0 \times \log_2(0) + 1 \times \log_2(1) + 0 \times \log_2(0)) + \\ &\frac{1}{3} \times (0 \times \log_2(0) + 1 \times \log_2(1) + 0 \times \log_2(0)) + \\ &\frac{1}{3} \times (0 \times \log_2(0) + 0 \times \log_2(0) + 1 \times \log_2(1))) \end{aligned} \tag{19}$$

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

$$H(Y|X) = 0. \tag{20}$$

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

$$p(y) = \{\langle \text{red}, 0\rangle, \langle \text{blue}, \frac{2}{3}\rangle, \langle \text{black}, \frac{1}{3}\rangle\}. \tag{21}$$

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):

$$\begin{aligned} p(x|y) = \{&\langle (\text{red}|\text{blue}), \frac{1}{2}\rangle, \langle (\text{red}|\text{black}), 0\rangle, \\ &\langle (\text{blue}|\text{blue}), \frac{1}{2}\rangle, \langle (\text{blue}|\text{black}), 0\rangle, \\ &\langle (\text{black}|\text{blue}), 0\rangle, \langle (\text{black}|\text{black}), 1\rangle\}. \end{aligned} \tag{22}$$

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:

$$H(X|Y) = -\sum_{y \in \mathcal{Y}} p(y) \sum_{x \in \mathcal{X}} p(x|y) \log_2 p(x|y). \tag{23}$$

And thus we have

$$\begin{aligned} H(X|Y) = -(&p(blue) \times (p(red|blue) \log_2 p(red|blue) + \\ &p(blue|blue) \log_2 p(blue|blue) + \\ &p(black|blue) \log_2 p(black|blue)) + \\ &p(black) \times (p(red|black) \log_2 p(red|black) + \\ &p(blue|black) \log_2 p(blue|black) + \\ &p(black|black) \log_2 p(black|black))). \end{aligned} \tag{24}$$

Further plugging the conditional probabilities of (22) into Equation (24) gives us:

$$H(X|Y) = -(\frac{2}{3} \times (\frac{1}{2} \times \log_2(\frac{1}{2}) + \frac{1}{2} \times \log_2(\frac{1}{2}) + 0 \times \log_2(0)) + \frac{1}{3} \times (0 \times \log_2(0) + 0 \times \log_2(0) + 1 \times \log_2(1))). \quad (25)$$

We thus get

$$H(X|Y) = \frac{2}{3} \sim \mathbf{0.67}\,\text{bits}. \quad (26)$$

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).

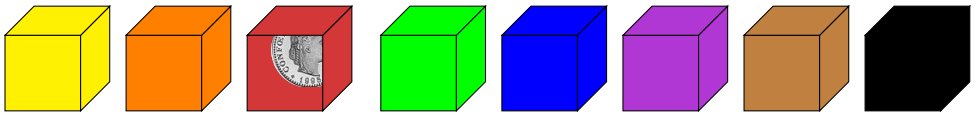
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Section 6: 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

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n), \tag{27}$$

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:

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2}\ldots, X_1), \tag{28}$$

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*[5] processes these two definitions are equivalent, i.e.

$$H(\mathcal{X}) = H'(\mathcal{X}). \tag{29}$$

Cover & Thomas (2006), p. 75.

[5] "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.

Example: Calculating $H(\mathcal{X})$

Let us use the same example of the box game as for the *joint entropy* above (with two independent variables), such that $X = X_1$ and $Y = X_2$, then – according to the first entropy rate definition – we simply have

$$H(\mathcal{X}) = \frac{1}{2}H(X_1, X_2) = \frac{3.08\,bits}{2} = \mathbf{1.54}\,bits \tag{30}$$

Note: in the case of these two *independent* random variables, the indivdual entropies are

$$H(X_1) = -(\frac{1}{3}\log_2\frac{1}{3} + \frac{1}{3}\log_2\frac{1}{3} + \frac{1}{3}\log_2\frac{1}{3}) = \log_2(3) \sim 1.58\,bits, \tag{31}$$

and

$$H(X_2) = -(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4}) = \log_2(4) = 1.5\,bits. \tag{32}$$

Hence, the entropy rate with finite n is here just the average entropy of the individual variables. However, this would be different for variables which are are dependent, i.e. have conditional probabilities, since conditioning would further decrease the entropy rate.

Summary

- Shannon & Weaver identified three levels to the communication problem. **Classic information theory** deals mainly with Level A (the technical problem of information transmission), but is also relevant to Level B and C (semantics and pragmatics).
- There is a range of (interrelated) **information-theoretic measures**: information content (surprisal), entropy, joint entropy, conditional entropy, relative entropy, etc.
- All of these build on the fundamental principle of **probability**.

References

References

Borda, Monica (2011). *Fundamentals in information theory and coding*. Berlin/Heidelberg: Springer.

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

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

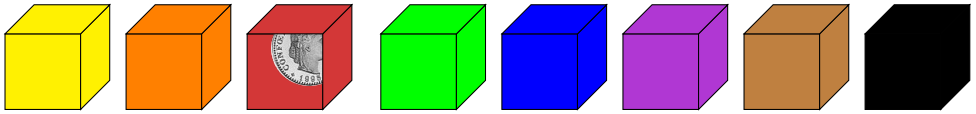
Shannon, Claude E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, Vol. 27, pp. 379–423.

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

Appendix: Relative Entropy

What is the Cost of Miss-Information?

"Where is the coin?"



"In the blue box"

Assume the "alphabet" of the "language" is still the same:

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

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.

Relative Entropy (Kullback-Leibler distance/divergence)

"The *relative entropy* is a measure of the distance[6] 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.

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

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)}. \tag{33}$$

It can take values between 0 and ∞. It is 0 if p = q.

[6] Note that it is not a 'true' distance measure, since it does not satisfy the triangle inequality.

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:

$$\textbf{World} : p(x) = \frac{1}{8}, \text{with } x \in \{black, blue, \ldots, yellow\}. \tag{34}$$

$$\begin{aligned}\textbf{Language} : q(x) = \{&\langle black, \frac{2}{16}\rangle, \langle blue, \frac{3}{16}\rangle, \langle brown, \frac{2}{16}\rangle, \\ &\langle green, \frac{2}{16}\rangle, \langle orange, \frac{2}{16}\rangle, \langle purple, \frac{2}{16}\rangle, \langle red, \frac{1}{16}\rangle, \langle yellow, \frac{2}{16}\rangle\}.\end{aligned} \tag{35}$$

$$D(p||q) \sim \textbf{0.05} \text{ bits per word} \tag{36}$$

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

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, $\mathcal{X} = \{\text{black}, \text{blue}, \text{brown}\}$ and $\mathcal{Y} = \{\text{black}, \text{blue}\}$, then the relative entropy between the probability distributions over these alphabets would be defined as

$$D(p||q) = \infty. \tag{37}$$

This is because we have to assume

$$q(\textit{brown}) = 0. \tag{38}$$

Cover & Thomas (2006), p. 19.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Table of Contents

10:00 – 11:00 Introduction & General Setup
11:00 – 12:00 Information-Theory: Basics
12:00 – 13:00 Lunch Break
13:00 – 14:00 Estimation Methods
14:00 – 15:00 Application 1: Voynich Manuscript
15:00 – 16:00 Application 2: Morphology & Syntax
16:00 – 17:00 Summary & Final Discussion

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Table of Contents

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Estimation Methods

Probabilities

For all of the information-theoretic measures we have discussed in the previous session, a crucial ingredient are the **probabilities** of information encoding units:

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

Information Content (Surprisal)

$$I(x) = -\log_2 p(x) \tag{39}$$

Entropy

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \tag{40}$$

Joint Entropy

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x, y) \tag{41}$$

Conditional Entropy

$$H(Y|X) = -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log_2 p(y|x) \tag{42}$$

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

$$\hat{p}(x_i) = \frac{f_i}{\sum_i^N f_i}, \tag{43}$$

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

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

$$\hat{p}(blue) = \frac{3}{6} \tag{44}$$

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.

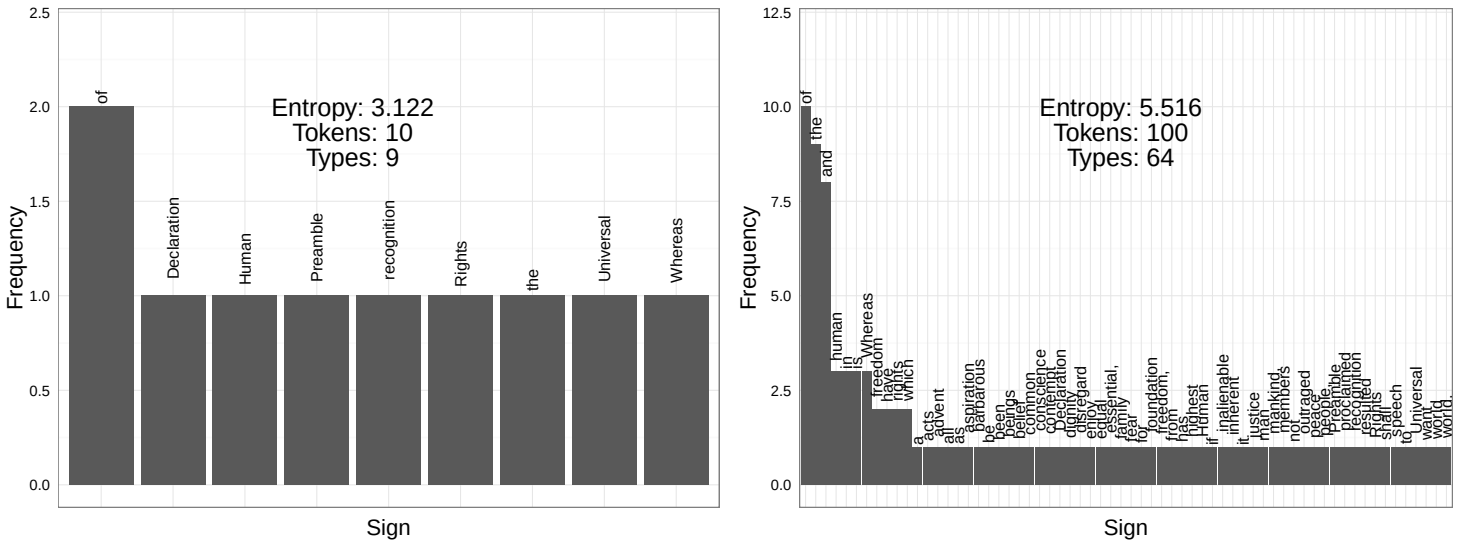


Figure. Frequency distributions and word type entropies for the English UDHR according to the first 10 and 100 word tokens.

Possible Solution for Problem 2

Use **less biased estimators**, and estimate the text size (in number of tokens) for which the information-theoretic measures (e.g. entropy) converge. For instance, Hausser & Strimmer (2009) show that the ML estimator has an *underestimation bias*, while other estimators (e.g. James-Stein Shrinkage) converge more quickly to the true entropy (if known). See also Figure on next slide.

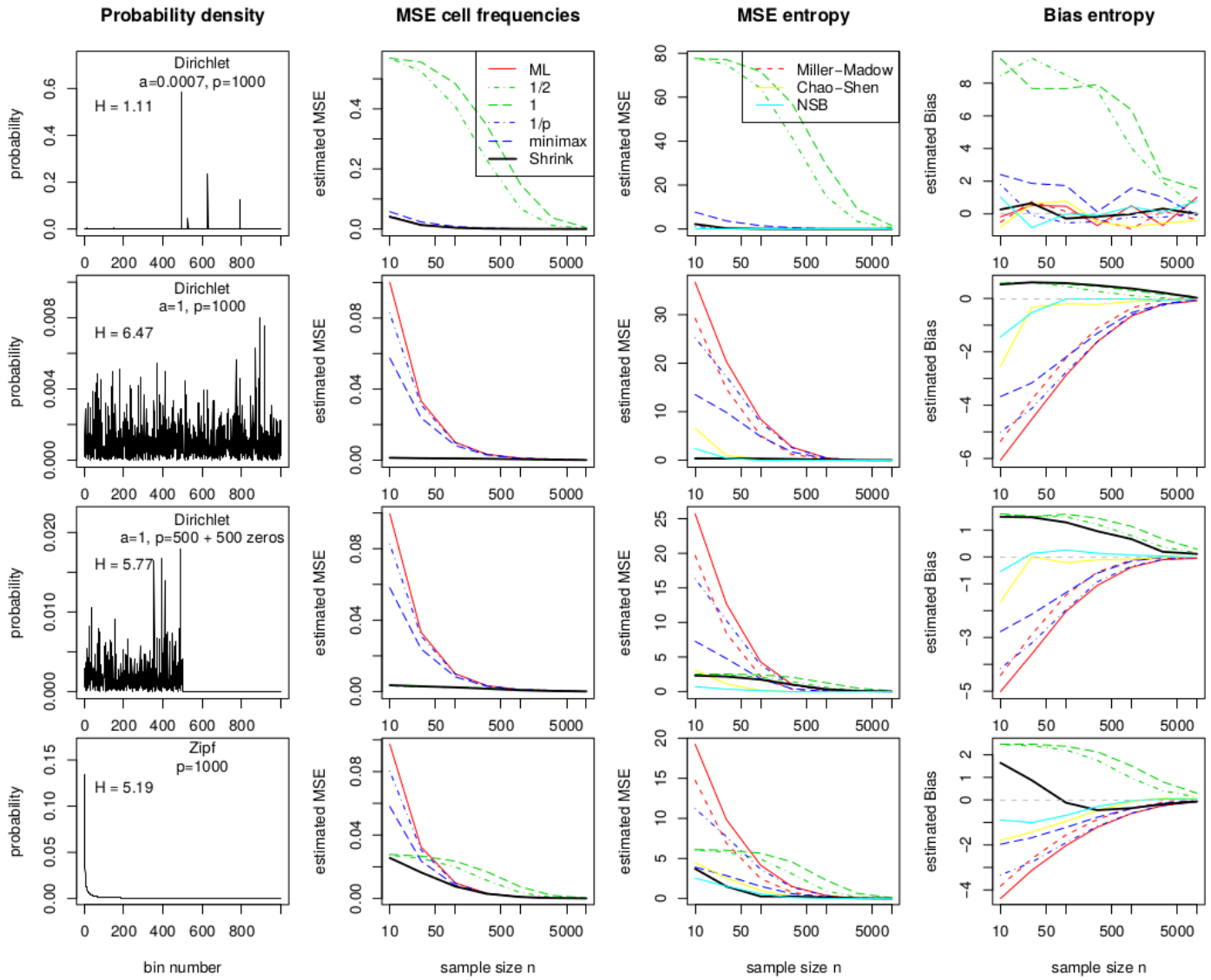
Hausser & Strimmer (2009). Entropy inference and the James-Stein estimator, with application to nonlinear Gene Association Networks.

Hausser & Strimmer (2015). R package entropy.

Lozano, Casas, Bentz, & Ferrer-i-Cancho (2017). Fast calculation of entropy with Zhang's estimator.

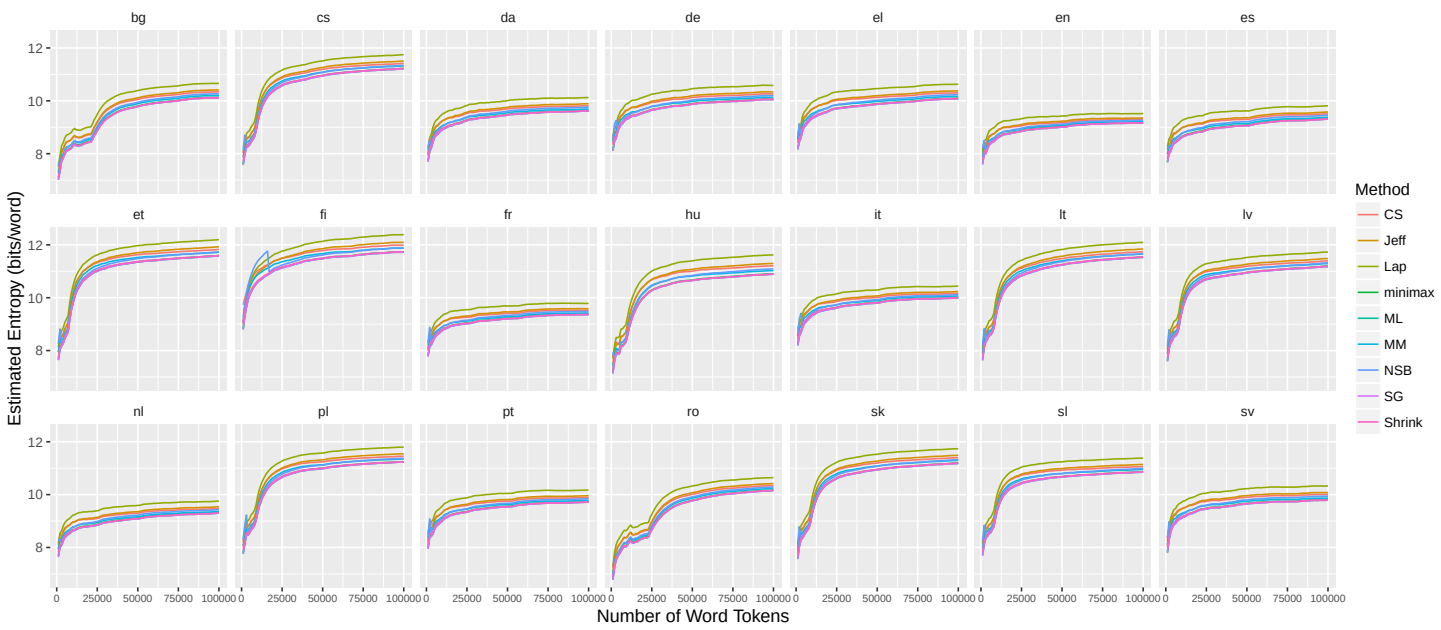
Back & Wiles (2021). Entropy estimation using a linguistic Zipf-Mandelbrot-Li model for natural sequences.

Shi & Lei (2022). Lexical richness and text length: An entropy-based perspective.



Possible Solution for Problem 2

Recent analyses suggest that for orthographic words, unigram entropies "stabilize" around 50K tokens. Shi & Lei (2022) argue that the "boundary point" lies even lower, at around 1K tokens.



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

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:

$$p(y|x) = \frac{p(x,y)}{p(x)} \tag{45}$$

`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.

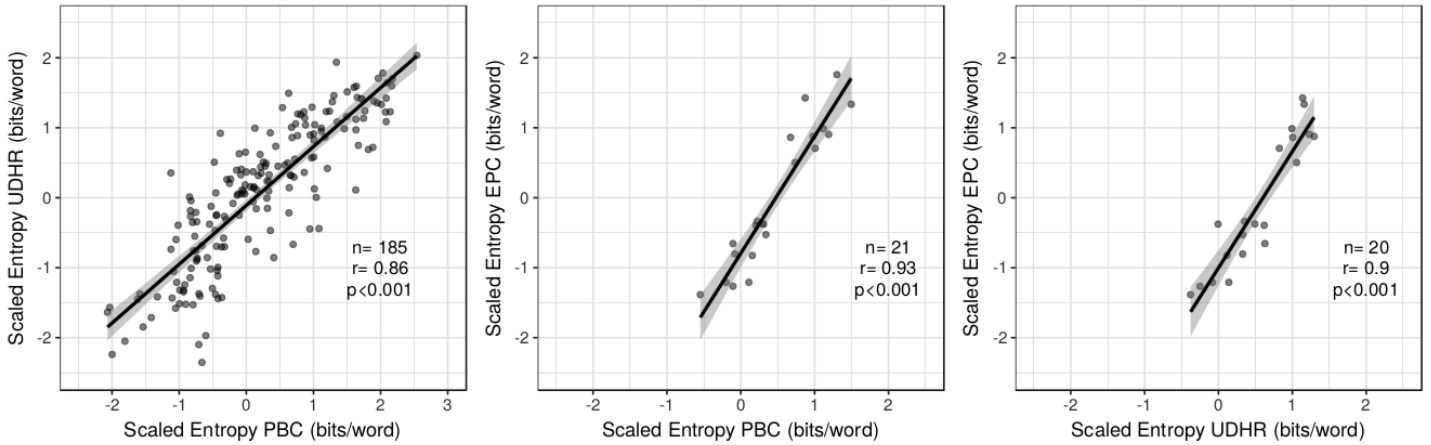
Gao, Kontoyiannis, & Bienenstock (2008). Estimating the entropy of binary time series: Methodology, some theory, and a simulation study.

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.



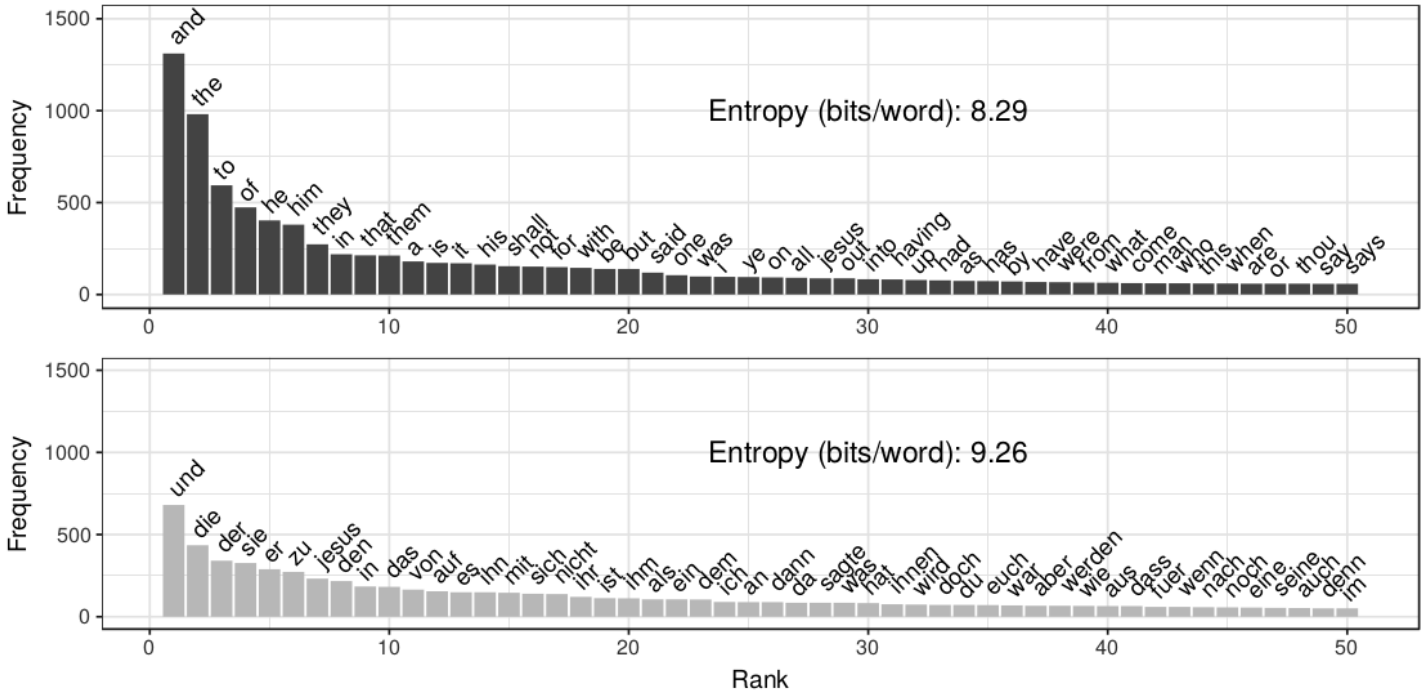
Bentz (2018). Adaptive languages: An information-theoretic account of linguistic diversity, p. 108.

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).

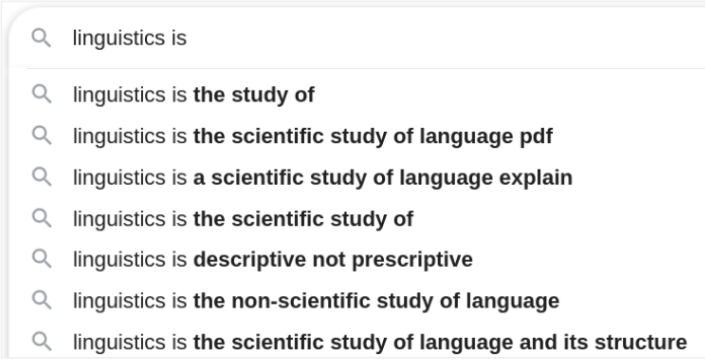


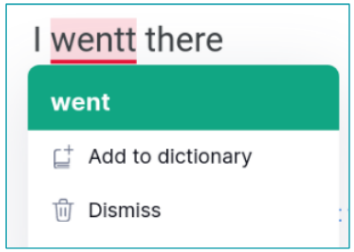
Bentz (2018). Adaptive languages, p. 88.

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: $P(w_1, w_2, \ldots, w_n)$

Language Models

The probability of this sequence of tokens can be estimated as:

$$p(w_1, w_2, \ldots, w_n) = p(w_1) \cdot p(w_2|w_1) \cdot p(w_3|w_1, w_2) \ldots p(w_n|w_1, \ldots, w_{n-1}) = \prod_{k=1}^{n} p(w_k|w_{<k})$$

Chain rule

How to calculate these conditional probabilities? Instead of computing the probability of a token given its entire history, we can approximate the history by just using the last few tokens (Markov property):

- Trigram model $p(w_k|w_1, \ldots, w_{k-1}) = p(w_k|w_{k-2}, w_{k-1})$
- Bigram model $p(w_k|w_1, \ldots, w_{k-1}) = p(w_k|w_{k-1})$
- Unigram model $p(w_k|w_1, \ldots, w_{k-1}) = p(w_k)$

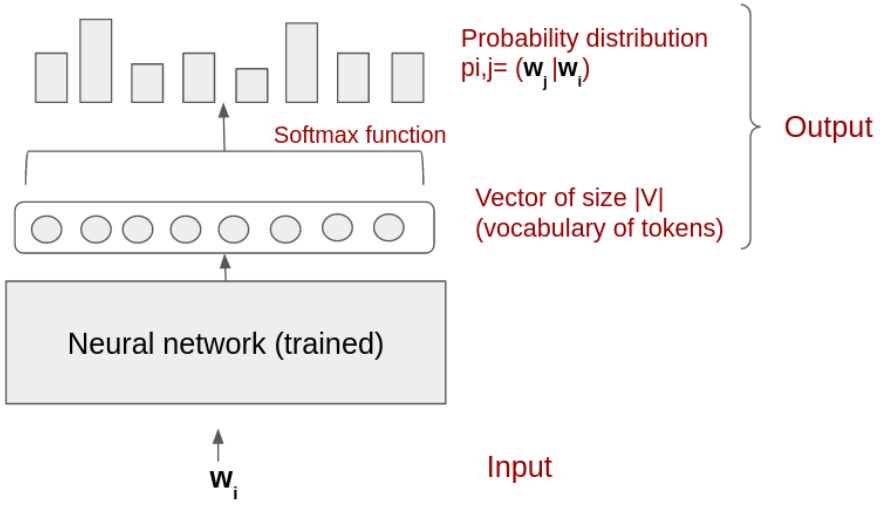
Speech and Language Processing. Jurafsky & Martin. Chapter 3.
NLP Course For You. Lena Voita. https://lena-voita.github.io/nlp_course.html

Language Models

- Example bigram model

p(linguistics is the study of language)=
p(linguistics)·p(is|linguistics)·p(the|is)·p(study|the)·p(of|study)·p(language|of)

- Several ways to obtain these conditional probabilities $P(w_k|w_{k-1})$
- We can learn them using a Neural Network (NN) and a corpus:



Language Models

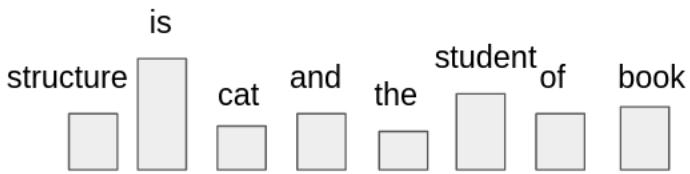
Instead of applying one formula based on the global corpus statistics, we teach a neural network to **predict** these conditional probabilities

- This is done by training the NN with many examples of units that occur together. Training instances for a bigram language model:

(linguistics, is) (linguistics, is) (linguistics,book) (linguistics, student)
(the, student) (book, structure) (student, book)
(linguistics, structure) (the, cat) (linguistics, is) (linguistics, is) (linguistics, student)

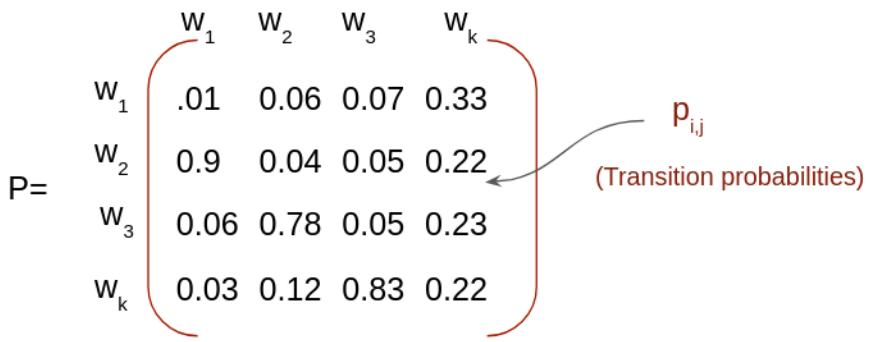
- Nowadays there are many types of neural language models, parameters and complex architectures
- Keep in mind: Underneath they are trying to predict what is the probability distribution of the tokens, given certain context

P(* |**linguistics**)



Language Models

The estimated probabilities of bigrams can also be interpreted as transition probabilities. This is equivalent to a Markov chain, which can be expressed with a transition (or stochastic) matrix:



If we assume this is a stationary Markov chain, we can calculate the entropy rate:

$$H(P) = -\sum_{i,j} \mu_i P_{i,j} log P_{i,j} \quad (46)$$

Cover & Thomas (2006), p. 77

Language Models

P=

	w_1	w_2	w_3	w_4
w_1	.01	0.06	0.07	0.33
w_2	0.9	0.04	0.05	0.22
w_3	0.06	0.78	0.05	0.23
w_4	0.03	0.12	0.83	0.22

$$H(P) = -\sum_{i,j} \mu_i P_{i,j} log P_{i,j}$$

$\mu_i = 1/4 = 0.25$ *(*if we assume the stationary distribution is uniform)*

$$H(P) = 0.25 * [0.01 * log_2(0.01) + 0.9 * log_2(0.9) + 0.06 * log_2(0.06) + 0.03 * log_2(0.03)]$$
$$+0.25 * [0.06 * log_2(0.06) + 0.04 * log_2(0.04) + 0.78 * log_2(0.78) + 0.12 * log_2(0.12)] +$$
$$0.25 * [0.07 * log_2(0.07) + 0.05 * log_2(0.05) + 0.05 * log_2(0.05) + 0.83 * log_2(0.83)] +$$
$$0.25 * [0.33 * log_2(0.33) + 0.22 * log_2(0.22) + 0.23 * log_2(0.23) + 0.22 * log_2(0.22)]$$

$H(P) = 1.1437$

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.

Experiments with Humans

"Shannon's experiment, however, used only one subject, bringing into question the statistical validity of his value of $h = 1.3$ bits per character for the English language entropy rate. [...] Our final entropy estimate was $h \sim 1.22$ bits per character."

Table 1. Comparison of the scales of cognitive experiments undertaken in previous works for the entropy rate estimation in English [1,9–11] and that of the present work.

	Total Number of Samples	Number of Subjects	Number of Phrases	Max n for a Session	Number of Sample Per n
Shannon [1]	1600	1	100	100	100
Jamison and Jamison [9]	360	2	50 and 40	100	50 and 40
Cover and King [10] No.1	440	2	1	220	2
Cover and King [10] No.2	900	12	1	75	12
Moradi et al. [11] No.1	6400	1	100	64	100
Moradi et al. [11] No.2	3200	8	400	32	100
Our Experiment	172,954	683	225	87.51	1954.86

Ren, Takahasi, & Tanaka-Ishii (2019). Entropy rate estimation for English via a large cognitive experiment using Mechanical Turk.

Summary

- The **probabilities of units** are a fundamental ingredient to any estimation of information-theoretic measures.
- There are **fundamental problems** with estimations of probabilties relating to: the *choice of units*, *sample sizes*, *interdependencies* between units, and *extrapolation* of results.
- There are at least three different strains of **estimation methods**: *frequency-based*, *language models*, *experiments* with human participants.

Practical Part

Run the following code files (in this order):

- `Application1/EstimationML.ipynb`
- `Application1/EstimationNeuralNet.ipynb`

References

References

Back, Andrew, & Wiles, Janet (2021). Entropy estimation using a linguistic Zipf-Mandelbrot-Li model for natural sequences. *Entropy* (23), 1100.

Bentz, Christian, Alikaniotis, Dimitrios, Cysouw, Michael, & Ferrer-i-Cancho, Ramon (2017). The entropy of words – learnability and expressivity across more than 1000 languages. *Entropy*, 19.

Bentz, Christian (2018). *Adaptive languages: An information-theoretic account of linguistic diversity*. Trends in Linguistics. Studies and Monographs (TiLSM), volume 316. Berlin/Boston, De Gruyter Mouton.

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

Gao, Yun, Kontoyiannis, Ioannis, & Bienenstock, Elie (2008). Estimating the entropy of binary time series: Methodology, some theory and a simulation study. *Entropy*, 10, p. 71–99.

Gutierrez-Vasques, Ximena, & Mijangos, Victor (2020). Productivity and predictability for measuring morphological complexity. *Entropy* 22.

Hausser, Jean, & Strimmer, Korbinian (2009). Entropy inference and the James-Stein Estimator, with application to Nonlinear Gene Association Networks. *Journal of Machine Learning Research*, 10, p. 1469–1484.

Hausser, Jean, & Strimmer, Korbinian (2015). R package entropy.
`http://strimmerlab.org/software/entropy/`

Kontoyiannis, I., Algoet, P. H., Suhov, Y. M., & Wyner, A. J. (1998). Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Trans. Inform. Theory*, 44, p. 1319–1327.

Lesne, Annick, Blanc, Jean-Luc, & Pezard, Laurent (2009). Entropy estimation for very short symbolic sequences. *Physical Review E*, 79.

Lozano, A. Casas, B. Bentz, C., & Ferrer-i-Cancho, R. (2017). Fast calculation of entropy with Zhang's estimator. In: *Studies in Quantitative Linguistics 23*, ed. by E. Kelih, R. Knight, J. Mačutek, and A. Wilson. Lüdenscheid: RAM Verlag. p. 273–285.
Ren, Geng, Takahashi, Shuntaro, & Tanaka-Ishii, Kumiko (2019). Entropy rate estimation for English via a large cognitive experiment using Mechanical Turk. *Entropy*, 21, 1201.

Shannon, Claude E. (1951). Prediction and entropy of printed English. *The Bell System Technical Journal*, 30(1), p. 50–64.

Shi, Yiqian & Lei, Lei (2022). Lexical richness and text length: An entropy-based perspective. *Journal of Quantitative Linguistics*, 29(1).

Table of Contents

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Application 1: Voynich Manuscript

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

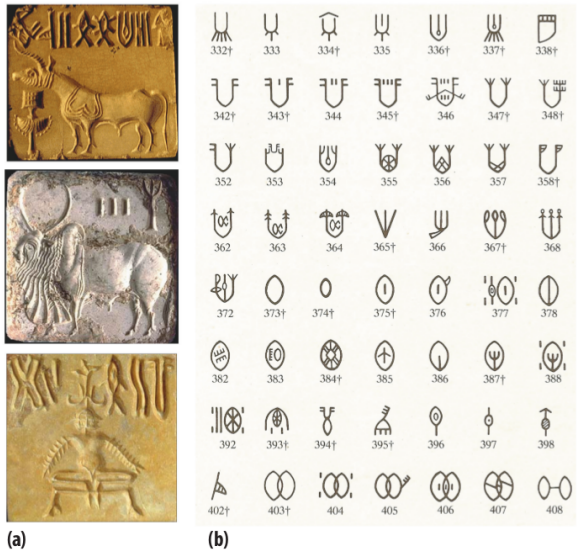
Faculty of Philosophy
General Linguistics

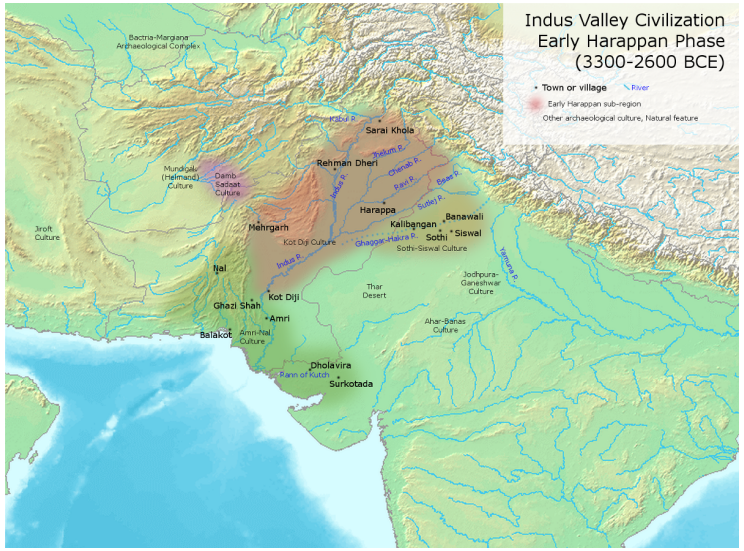
Introduction

Quantitative Linguistics

The Voynich Manuscript
https://en.wikipedia.org/wiki/Voynich_manuscript

Entropic Analyses of Undeciphered Scripts





Rao et al. (2009). Entropic evidence for linguistic structure in the Indus script.
Rao (2010). Probabilistic analysis of an ancient undeciphered script.
Rao et al. (2010). Entropy, the Indus script, and language.

Entropic Analyses: Block Entropy

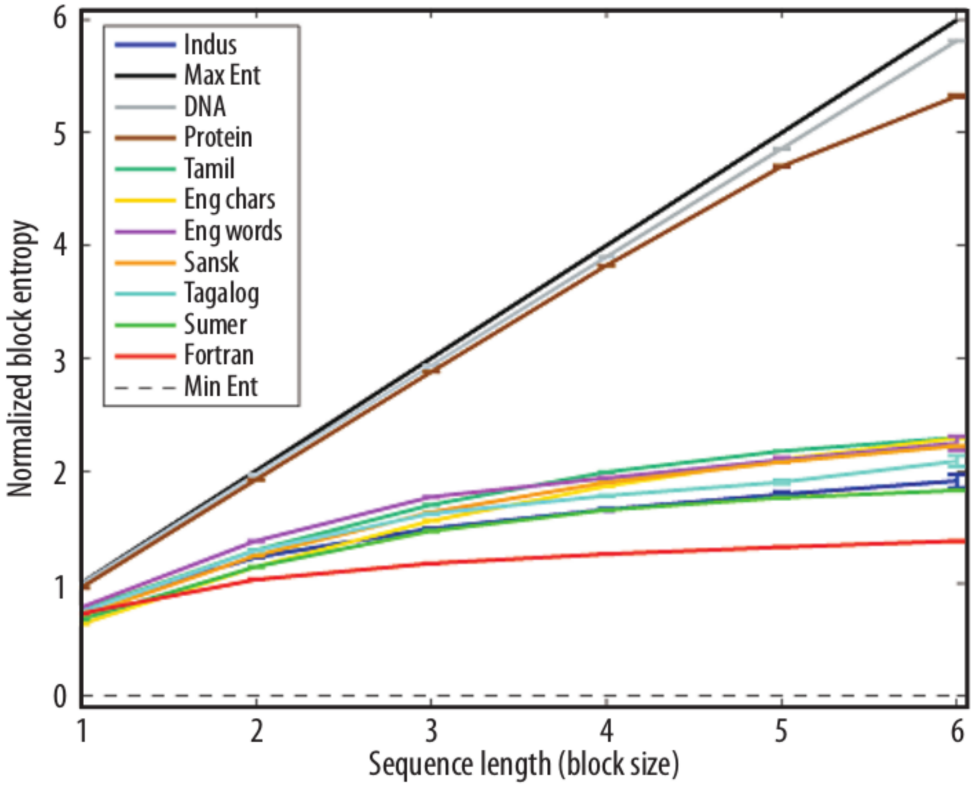
"Block entropy for block size N is defined as:

$$H_N = -\sum_i p_i^{(N)} \log p_i^{(N)} \tag{47}$$

where $p_i^{(N)}$ are the probabilities of sequences (blocks) of N symbols. Thus for $N = 1$, block entropy is simply the standard unigram entropy and for $N = 2$, it is the entropy of bigrams."

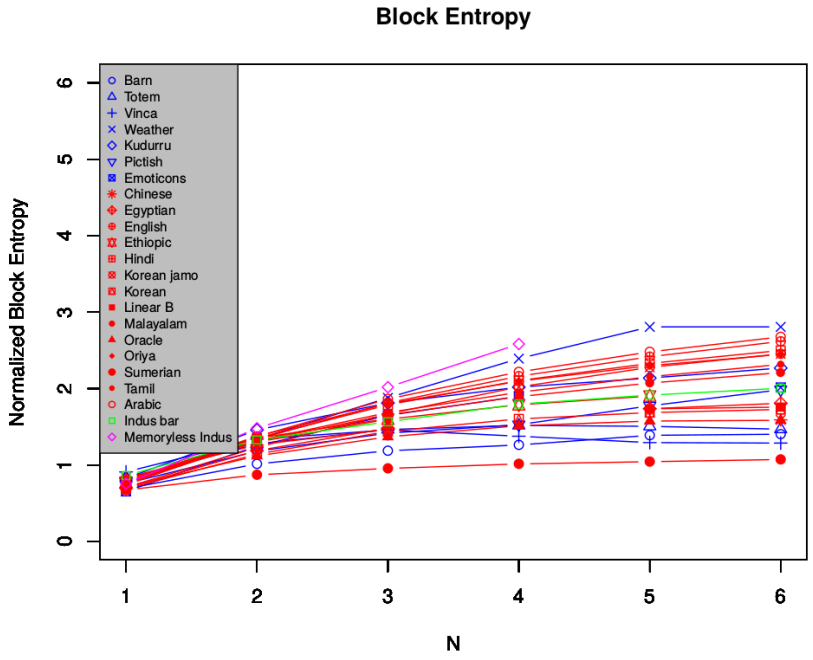
Rao et al. (2010). Entropy, the Indus script, and language, p. 4

Written language or not?



Rao (2010). Probabilistic analysis of an ancient undeciphered script.

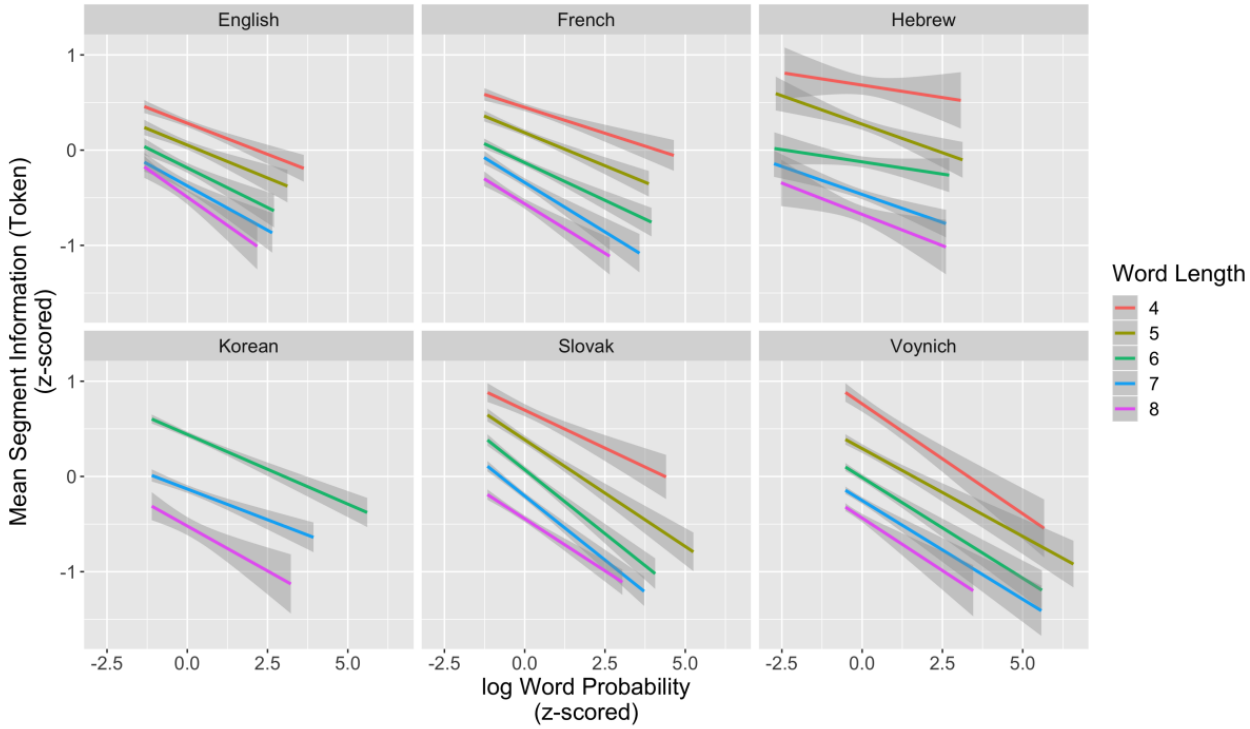
However...



"Using a larger set of nonlinguistic and comparison linguistic corpora than were used in these and other studies, I show that none of the previously proposed methods are useful as published. However, one of the measures proposed by Lee and colleagues (2010a) (with a different cut-off value) and a novel measure based on repetition turn out to be good measures for classifying symbol systems into the two categories."

Sproat (2014). A statistical comparison of written language and nonlinguistic symbol systems.

Information-Theoretic Analyses of Voynichese



Layfield et al. (2020). Word probability findings in the Voynich manuscript.
Montemurro & Zanette (2011). Keywords and co-occurrence patterns in the Voynich manuscript: An information-theoretic analysis.

Methods

Methods

- We estimate unigram, bigram and trigram **character** entropies.
- We use the **Maximum likelihood method** (ML) as well as a **Neural Network** to estimate probabilities.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Data

Universal Declaration of Human Rights (UDHR)

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

Все люди рождаются свободными и равными в своем достоинстве и правах.

כל בני אדם נולדו בני חורין ושווים בערכם ובזכויותיהם.

.ᏧᎾᏓᏂᎸᎬᏁᏍᎨᏍᏗ ᎨᏒᎢᏴᏛᏁᎵᏍᏗᏍᎩᏂ ᎤᏂᏣᏘᏂᏙᎯᏳᎲᏍᎩᏂ ᎠᏂᏴᏫᏂᎨᏒᎢ ᎤᏂᏰᎸᎭᏂᏙᎯᏳᎲᏍᎩᏂ ᎠᎴ ᎤᏂᏣᏘᏂᏙᎯᏳᎲᏍᎩᏂ

모든 인간은 태어날 때부터 자유로우며 그 존엄과 권리에 있어 동등하다.

يولد جميع الناس أحرارًا متساوين في الكرامة والحقوق.

የሰው፡ልጅ፡ሁሉ፡ሲወለድ፡ነጻና፡ክብር፡እና፡መብት፦እኩል፡ነት፡ያለው፡ነው።

人人生而自由,在尊严和权利上一律平等。

Latin Transliteration of Voynichese

```
#type:          unclassified
#specification: Voynich manuscript (voy)
#scriptcode:    Latn
#source:        http://www.voynich.com/pages/PagesH.txt (last accessed 05.02.2019)
#encoding:      utf-8
#copyright:     NA
#comments:      This is a transcription of the Voynich manuscript by Takeshi Takahashi in the so-called EVA alphabet. Details on the
transcription and line identifiers can be found at http://www.voynich.nu/transcr.html#n15. "." as word separators have been replaced by blank
spaces. The character "*", which is used in the original EVA transcription system for unreadable characters, is here replaced by "?". Note
that "?" has the meaning of "missing word" in the original EVA transcription system, but there are actually no "?" found in the transcription
we used.

<f1r>
<f1r.P1.1;H>    fachys ykal ar ataiin shol shory cthres y kor sholdy
<f1r.P1.2;H>    sory ckhar or y kair chtaiin shar are cthar cthar dan
<f1r.P1.3;H>    syaiir sheky or ykaiin shod cthoary cthes daraiin sa
<f1r.P1.4;H>    ooiin oteey oteos roloty cth?ar daiin otaiin or okan
```

Research Question

Does the language of the Voynich manuscript fall in the range of character entropies of written natural languages?

Practical Part

Run the following code files (in this order):

- `Application1/Visualization.ipynb`

Table of Contents

10:00 – 11:00 Introduction & General Setup
11:00 – 12:00 Information-Theory: Basics
12:00 – 13:00 Lunch Break
13:00 – 14:00 Estimation Methods
14:00 – 15:00 Application 1: Voynich Manuscript
15:00 – 16:00 Application 2: Morphology & Syntax
16:00 – 17:00 Summary & Final Discussion

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

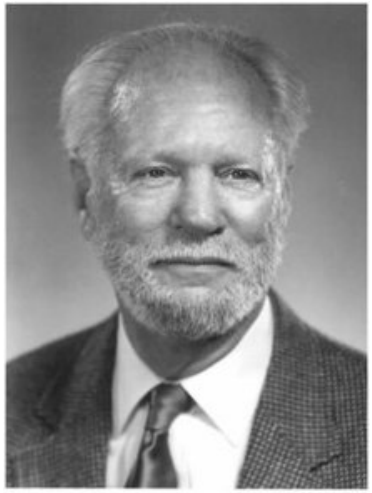
Applications 2:
Word Entropy and Word Order Entropy

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Introduction

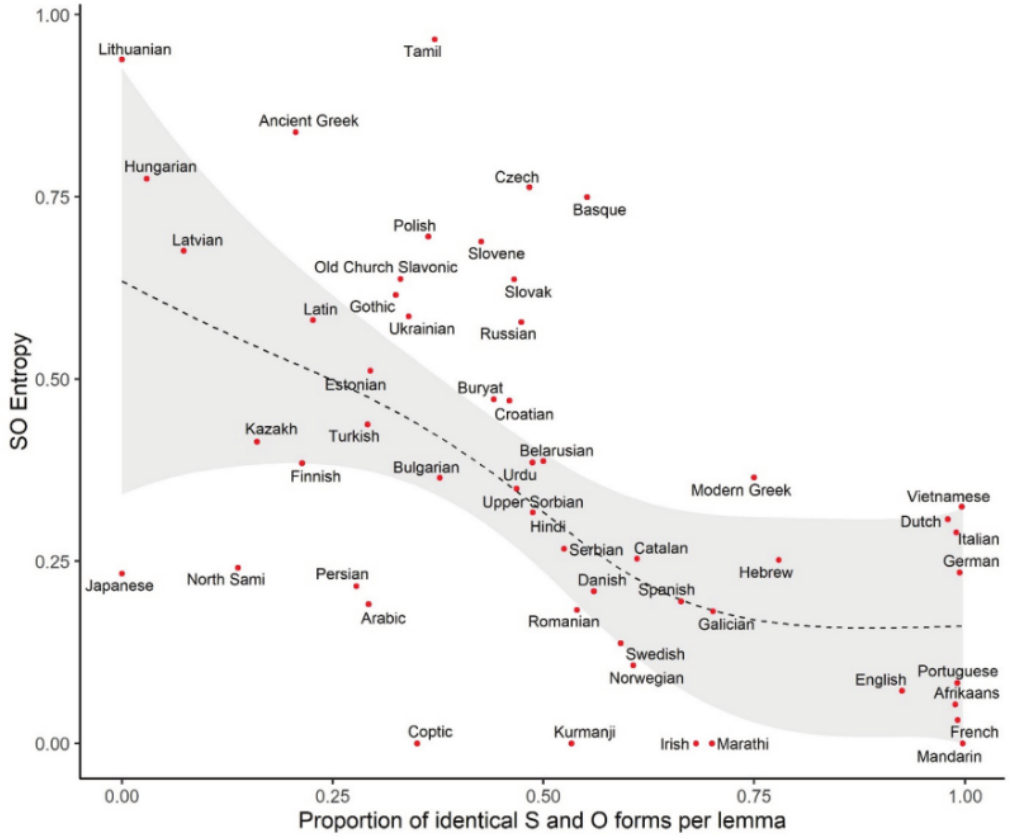
Morphology and Syntax



"[...] all languages have about equally complex jobs to do, and what is not done morphologically has to be done syntactically."

Hockett (1958). A course in modern linguistics, p. 180-181.

Word Order Entropy



Levshina, N. (2019). Token-based typology and word order entropy: A study based on Universal Dependencies.

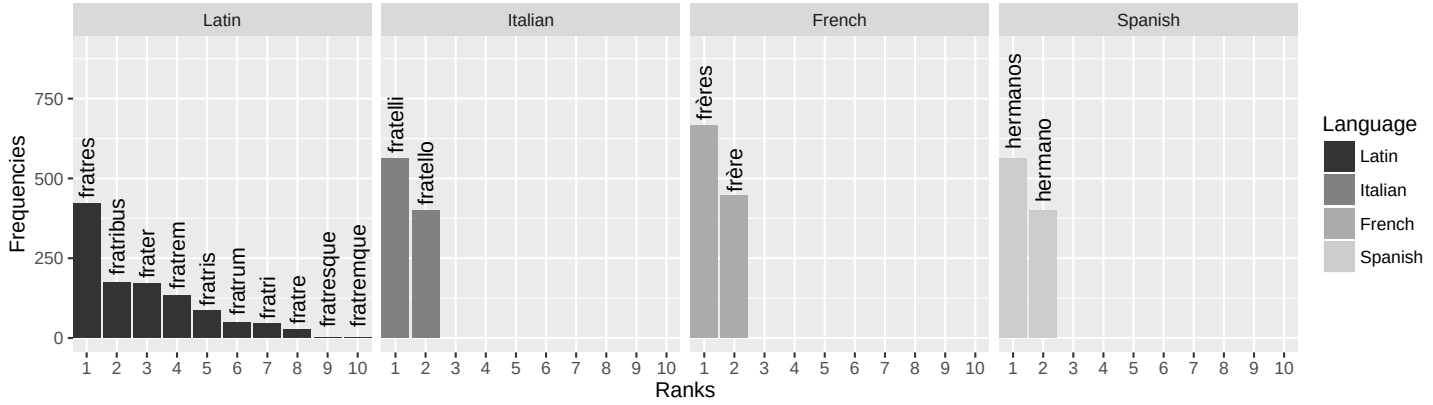
Research Question

Can we replicate these findings? – If it holds true that languages with free word order tend to have complex morphology, and languages with fixed word order tend to have simpler morphology, then we expect a **positive correlation** between word order entropy and orthographic word entropy.

Methods

Orthographic Word Distributions

Languages differ with regards to their distributions of different word types and their token frequencies. This is related (among other factors) to the productiveness of inflectional morphology.



Bentz & Berdicevskis (2016). Learning pressures reduce morphological complexity: linking corpus, computational and experimental evidence.

Unigram Entropy

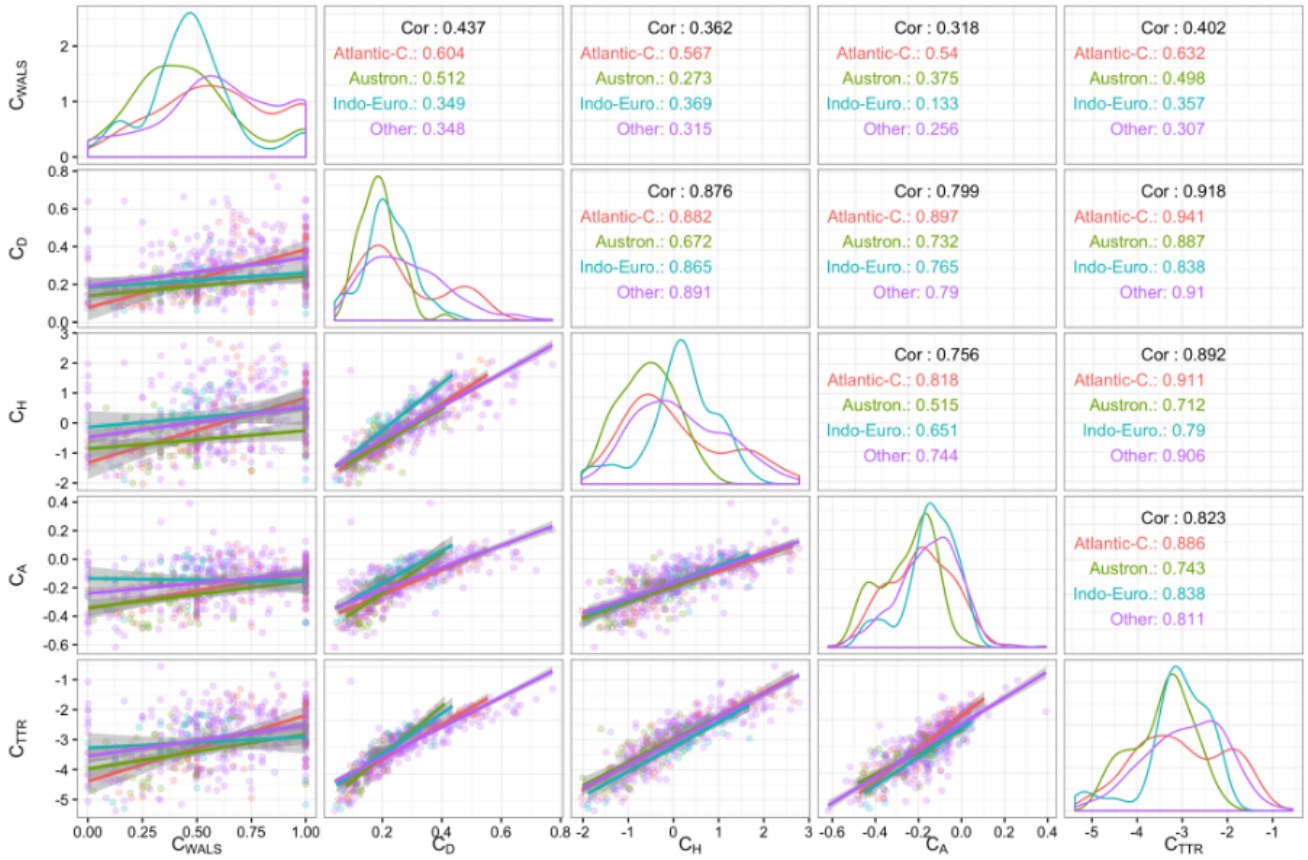
- A text can be regarded as a sequence of symbols. Each symbol is generated with a certain probability, and hence carries a certain information content.
- The average information content of a text can be estimated by Shannon entropy,

A text T with a vocabulary of types $V = \{t_1, t_2, ..., t_V\}$ of size $|V|$

$$H(T) = -\sum_{i=1}^{V} p(t_i) \log_2 p(t_i)$$

The vocabulary of types V correspond to any information encoding unit that we choose, e.g., UTF-8 characters, character ngrams, syllables, morphemes, orthographic words,...

Approximating Morphological Complexity



Bentz, C., Soldatova, T., Koplenig, A., Samardzic, T. (2016). A comparison between morphological complexity measures: typological data vs. language corpora.

Word Order Entropy

In a similar vein, Shannon's entropy can also be a tool to approach word order variation across languages.

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

Example: We have a random variable X with two possible values

x_1= word order SO x_2= word order OS

Language 1: $p(x_1) = 0.9$ $p(x_2) = 0.1 \rightarrow H(X) = 0.46$ bits

Language 2: $p(x_1) = 0.5$ $p(x_2) = 0.5 \rightarrow H(X) = 1$ bit

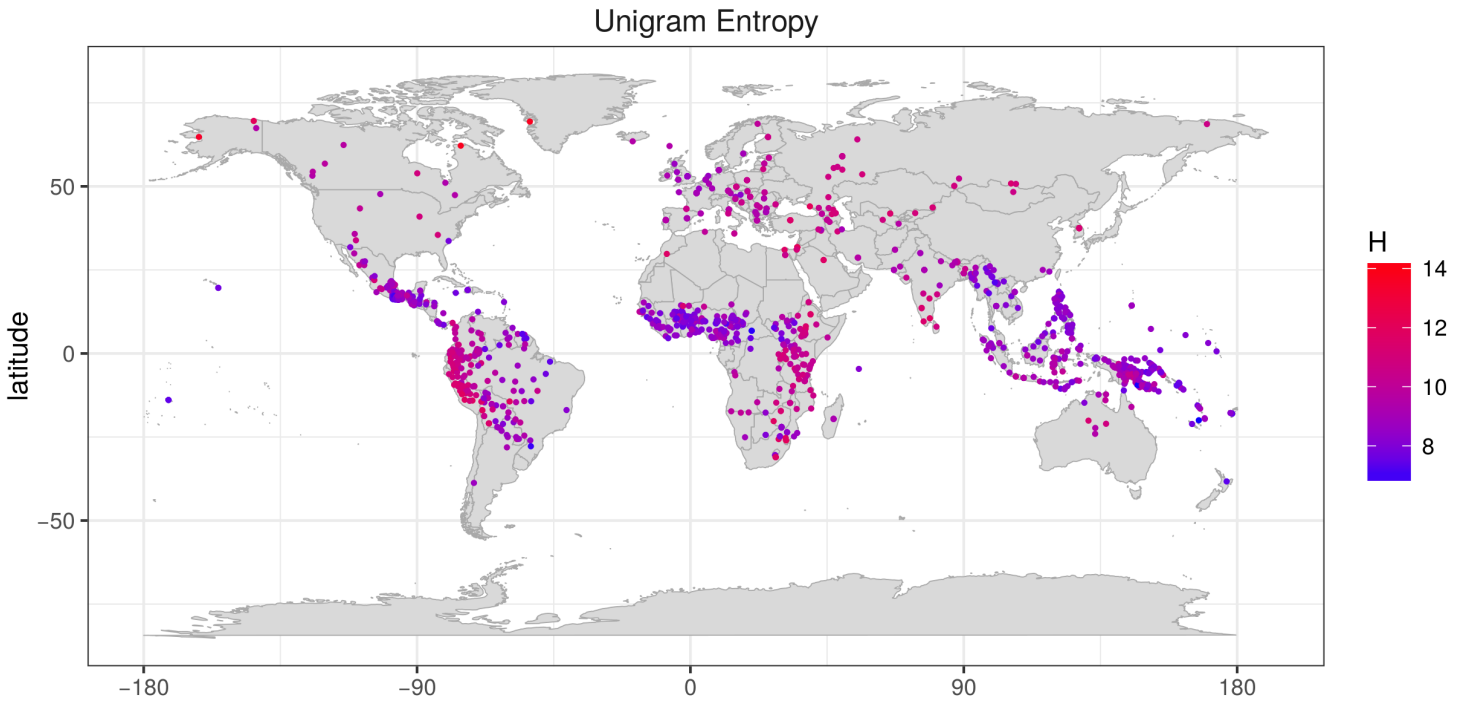
Language 2 has a word order which is harder to predict.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Data

Orthographic Word Entropy



Bentz et al. (2017). The entropy of words: Learnability and expressivity across more than 1000 languages.

Mayer & Cysouw (2014). A massively parallel Bible corpus.

Word Order Entropy

SOV	SVO	OSV	OVS	VSO	VOS
Polynesian (Hawaiian, Maori)					
3	31	2	2	70	3
6	26	5	4	76	18
Sinitic (Mandarin, Hakka)					
54	235	6	0	3	5
18	84	1	2	5	3
Turkic (Kara-Kalpak, Kumyk)					
114	2	8	7	0	0
89	1	12	11	4	1

Table 1: Number of transitive clauses with a given order of subject/object/verb, according to our algorithm, for six languages (from three families).

Östling. (2015). Word order typology through multilingual word alignment.

Practical Part

Run the following code files (in this order):

- `Application2/WordOrder.ipynb`
- `Application2/Visualization_WO.ipynb`

Table of Contents

10:00 – 11:00 Introduction & General Setup
11:00 – 12:00 Information-Theory: Basics
12:00 – 13:00 Lunch Break
13:00 – 14:00 Estimation Methods
14:00 – 15:00 Application 1: Voynich Manuscript
15:00 – 16:00 Application 2: Morphology & Syntax
16:00 – 17:00 Summary & Final Discussion

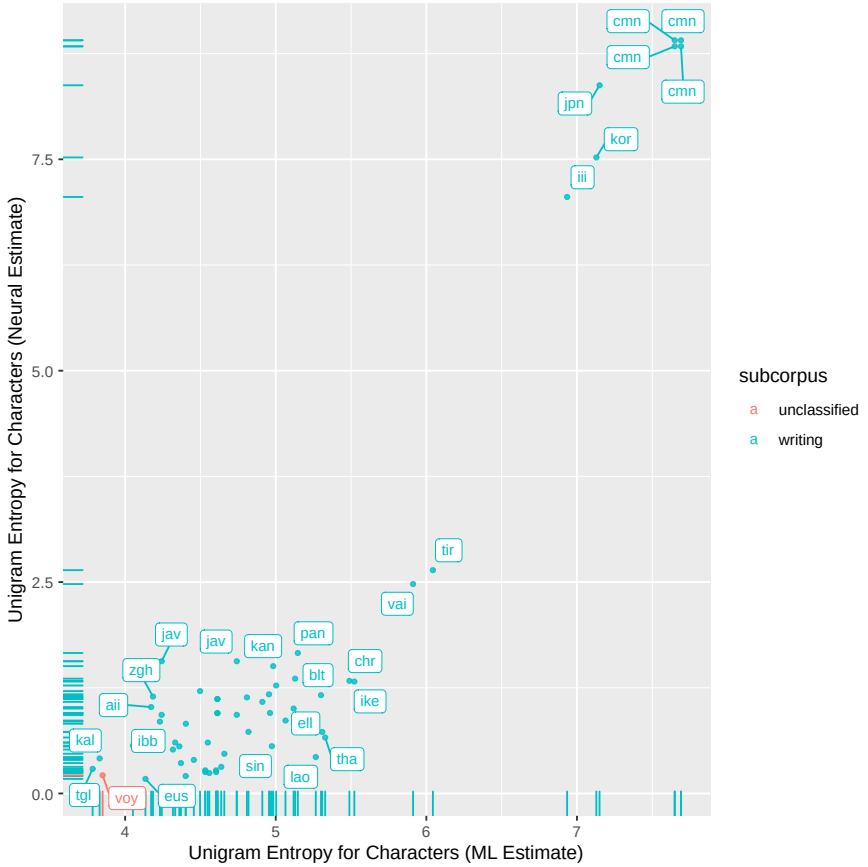
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Final Discussion

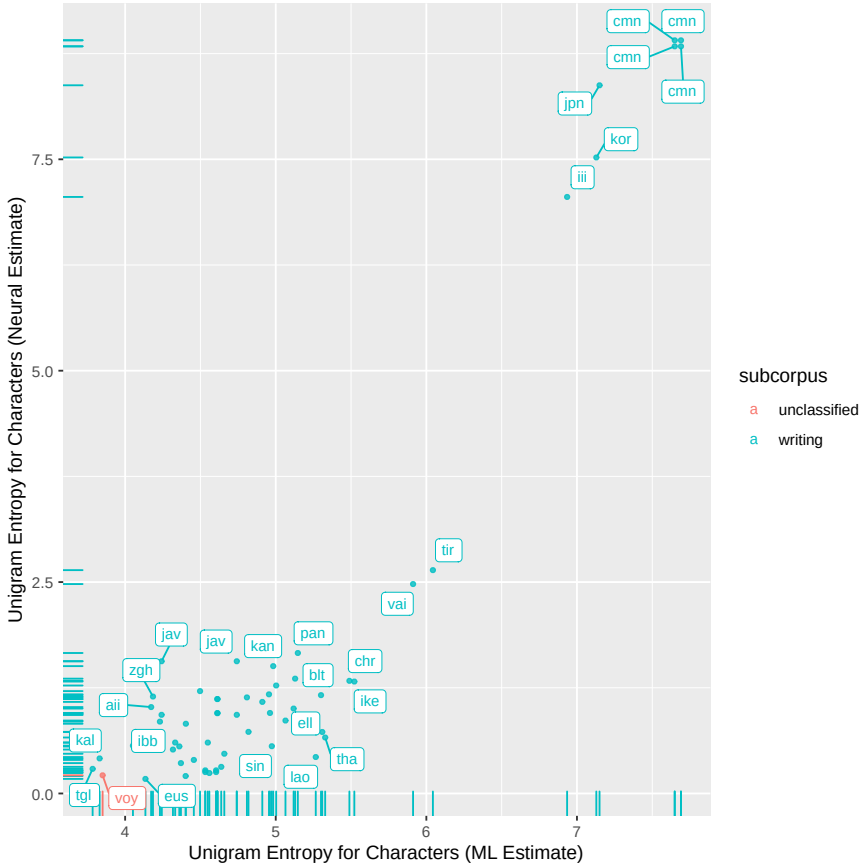
Application 1: Voynich Manuscript

- The unigram entropy of characters is generally high for logographic writing systems (e.g. Mandarin Chinese), in the middle range for syllabaries, and low for alphabets.
- The ML method overestimates entropies, while the neural network seems to underestimate them (at least for some alphabets).
- The *Voynich manuscript* just about falls in the range of human writing (close to Tagalog and Kalaallisut).



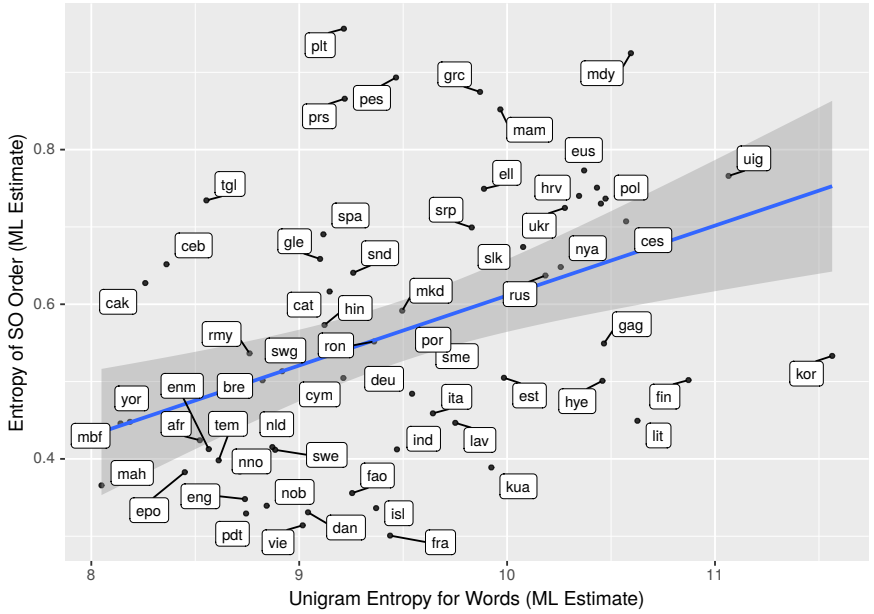
Application 1: Voynich Manuscript

- The trigram entropy of characters is more evenly distributed in the range of 4-11 bits per character trigram.
- The *Voynich manuscript* falls outside the range of human writing (except for Kalaallisut in the Neural Net estimation).



Application 2: Word Order Entropy

- There is a positive correlation between orthographic word entropy (morphology) and word order entropy (syntax), as we would expect.
- However, it is weak across the whole data set, and only becomes somewhat stronger when languages with few sentences are excluded.



Thank You.

Contact:

Dr. Christian Bentz
chris@christianbentz.de
Dr. Ximena Gutierrez-Vasques
ximena.gutierrezvasques@uzh.ch