Entropy, Signs, Symbols and Natural Languages

Christian Bentz
University of Tübingen

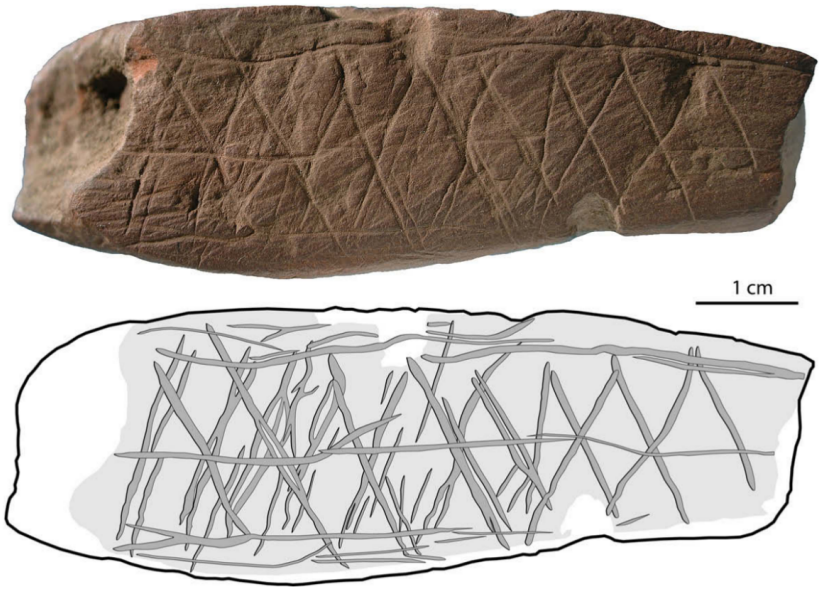
September 27, 2016

WORDS BONES GENES TOOLS
Tracking Linguistic, Cultural, and Biological Trajectories of the Human Past

EVOLAEMP
LANGUAGE EVOLUTION: THE EMPIRICAL TURN

Overview

What's the difference?



Some Intuitive Terminology

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

} Entropy

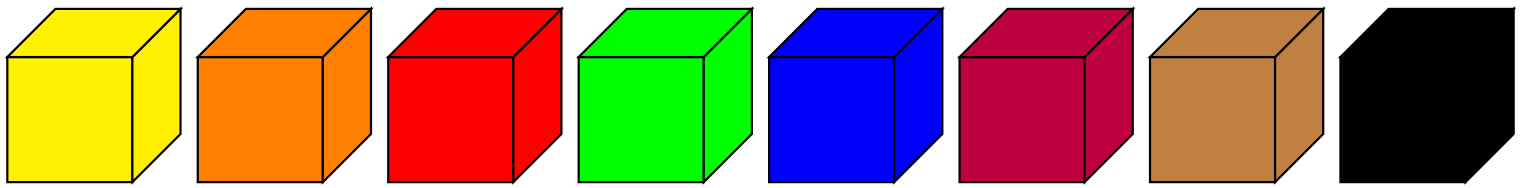
A more optimistic way of putting it

"*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."

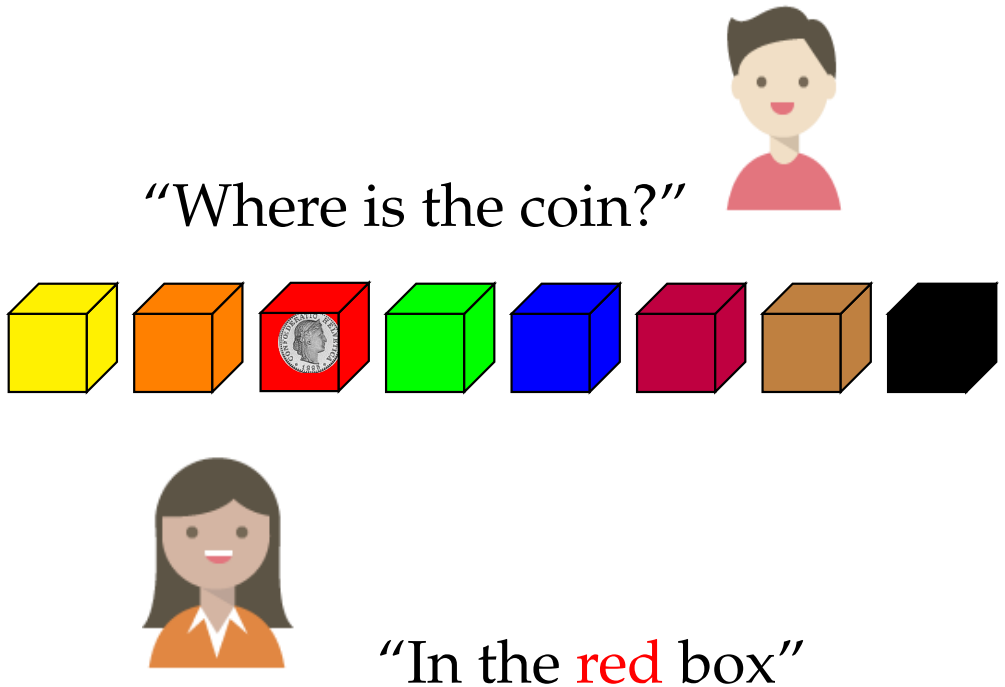
[Lemons, 2013]

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?



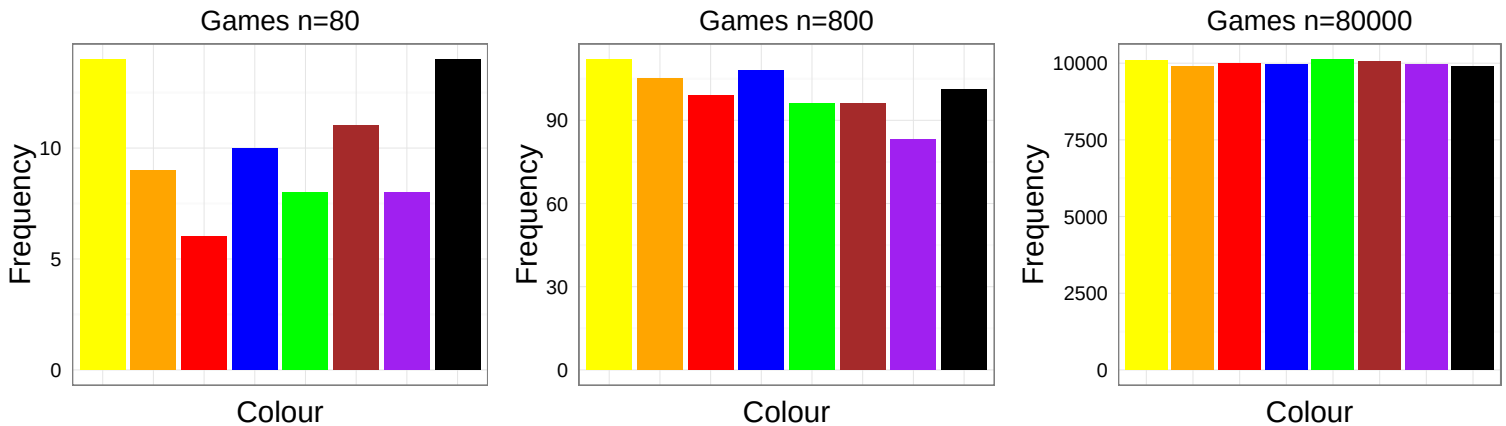
- The "alphabet" (here words) of the "language" they use does not need more than 8 colour adjectives to disambiguate:

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

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.

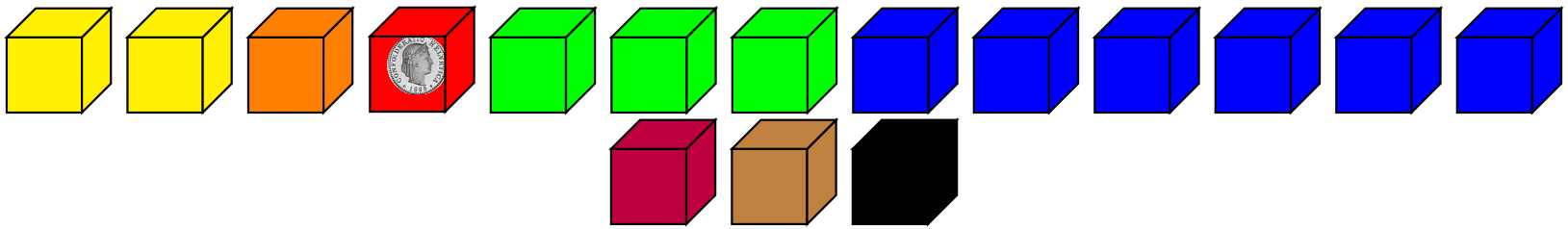
"In the red, green, blue, yellow, purple,... box"

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"

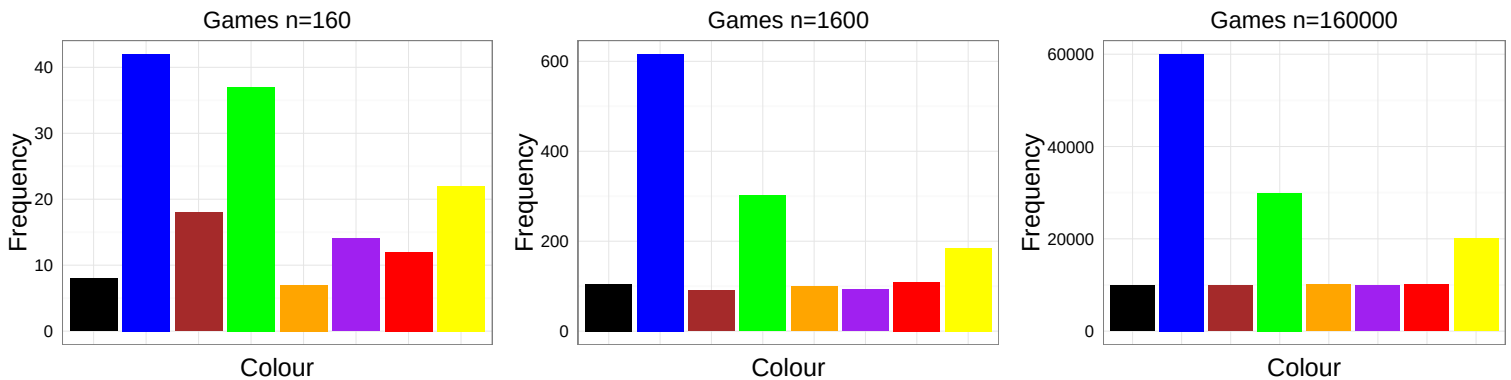
- The "alphabet" has **not** changed:

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

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

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

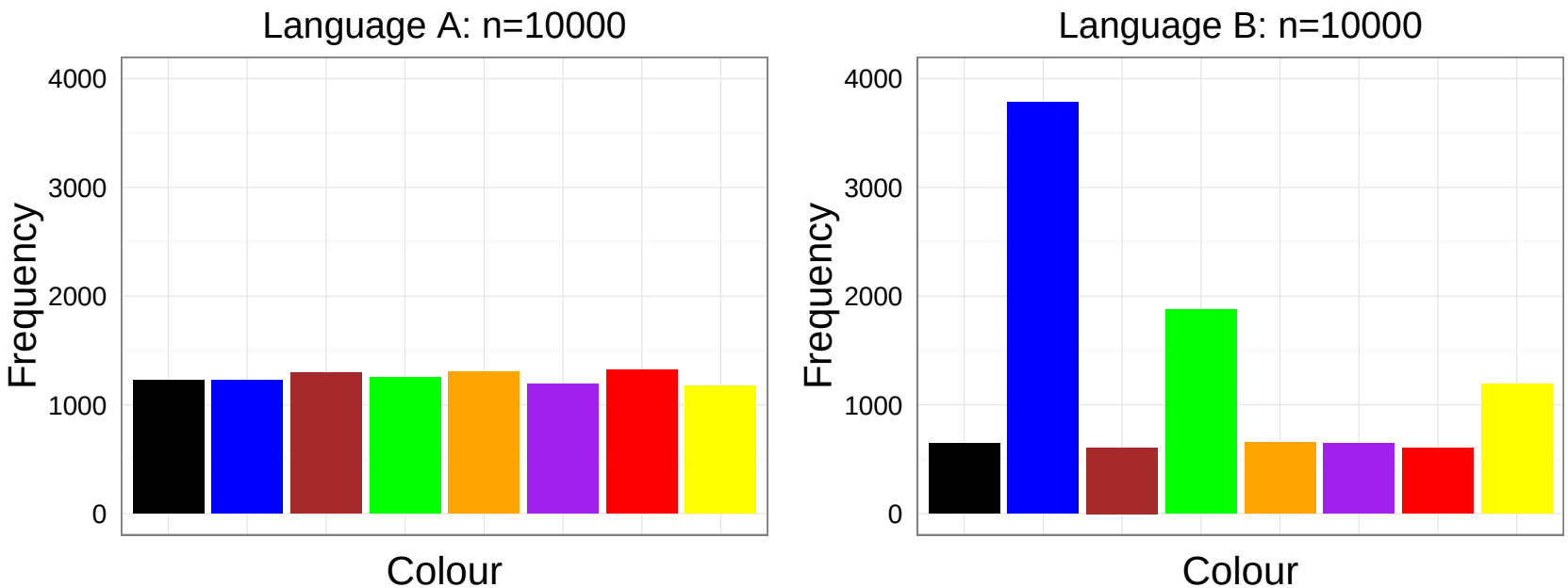
Again, this will be reflected in the **girl's language production**.



Comparing language production

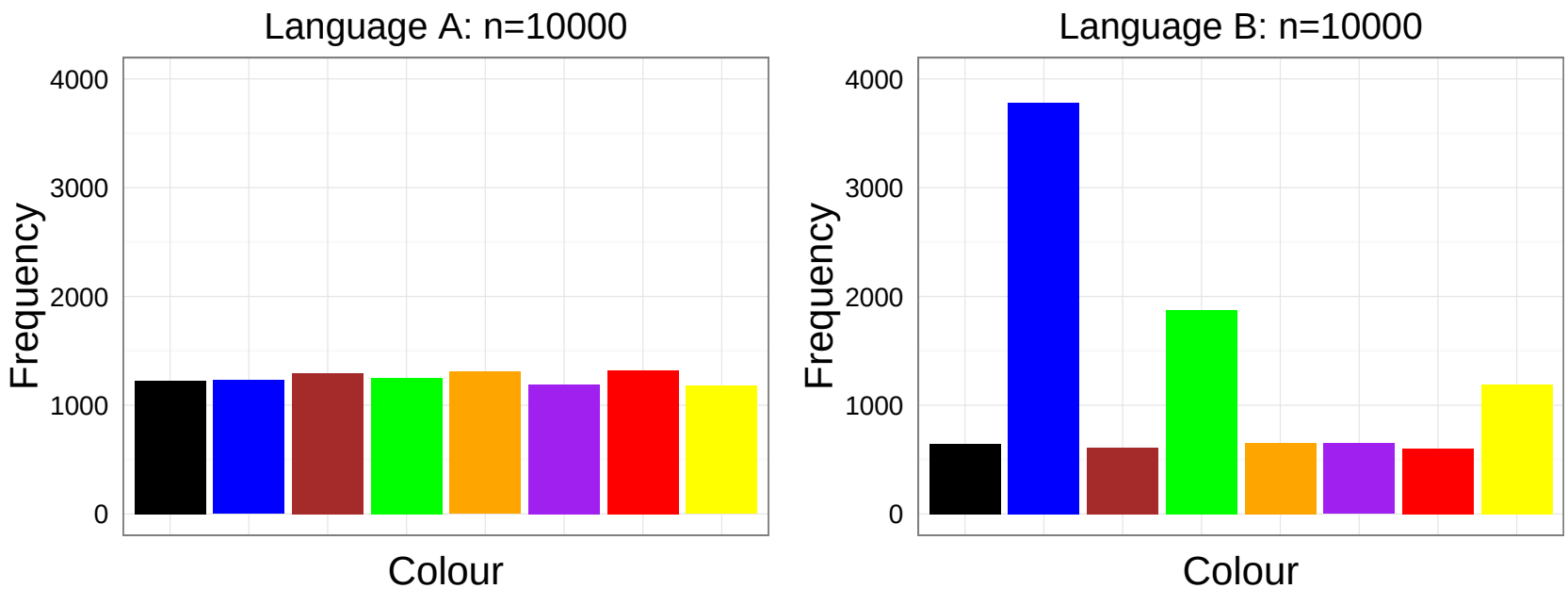
If we play the two games the same number of times n, we will get the same two languages L_A and L_B in terms of **word types** (8 in this case), and the number of **word tokens** (10K in this case).

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



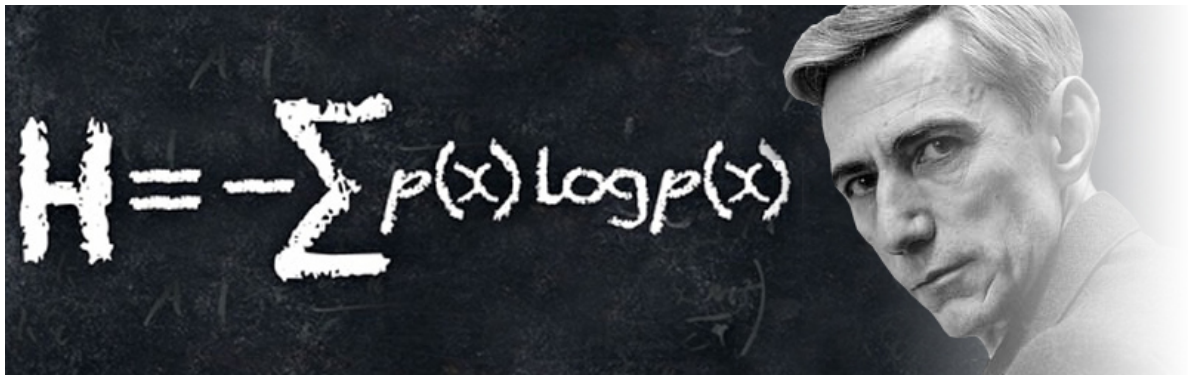
Comparing language production

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".



How can we measure this difference in the distributions?

Claude Shannon came up with a measure for this difference in "The 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

Then the entropy is defined as

$$H(X) = -K \sum_{i=1}^{N} p(x_i) \log_2 p(x_i) \tag{1}$$

Notes:

- K is a positive constant that was introduced by Shannon, but mostly it is assumed to be 1 and hence dropped.
- $H(X)$ does not mean that H is a function of X! In fact, the correct way of writing it is $H(p_1, p_2, ..., p_N)$, which is mostly shortened to $H(X)$.

Let's look at the component parts

$$H(X) = -\sum_{i=1}^{N} p(x_i) \log_2 p(x_i) \tag{2}$$

- $-\log_2 p(x_i)$ is the **information content** of a unit x_i (word type in our case). 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 ($0 < p < 1$) is negative (except for $p = 1$, 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

$$H(X) = -\sum_{i=1}^{N} p(x_i) \log_2 p(x_i) \tag{3}$$

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

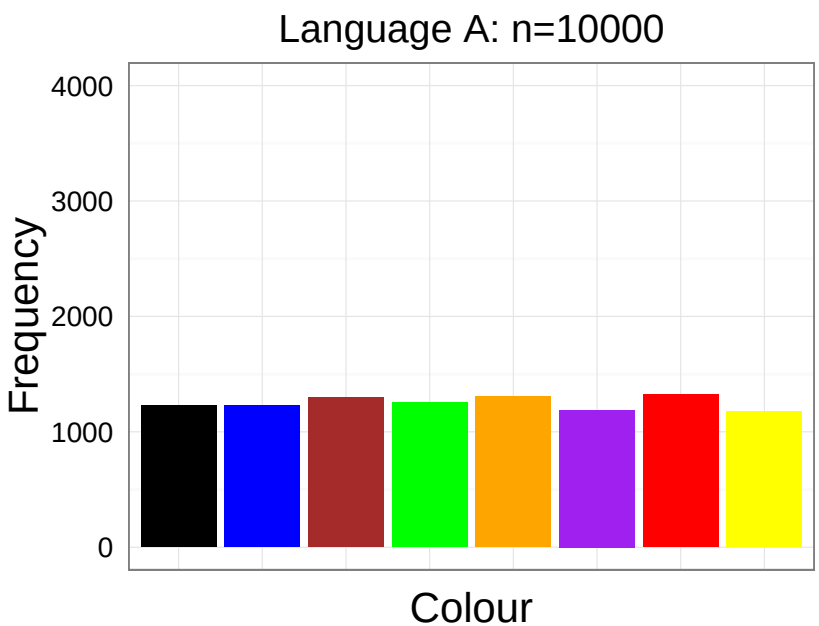
Hence, the entropy $H(X)$ can be seen as the average information content of information encoding units, i.e. word types in our case.

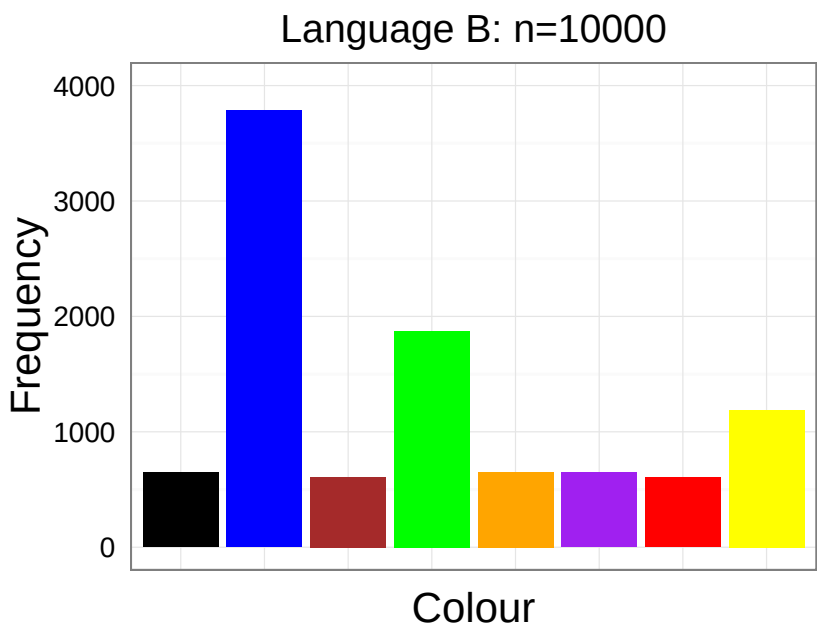
Let's apply this to Languages A and B

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

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

$$H(L_B) = -(\frac{6}{16} \times \log_2(\frac{6}{16}) + \frac{3}{16} \times \log_2(\frac{3}{16}) + ... + \frac{1}{8} \times \log_2(\frac{1}{8})) = \mathbf{2.61} \quad (5)$$

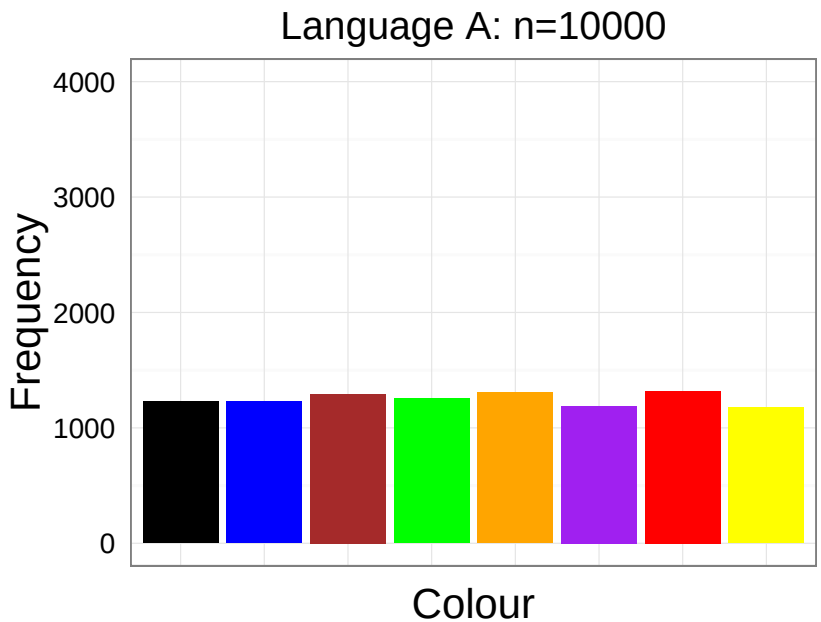


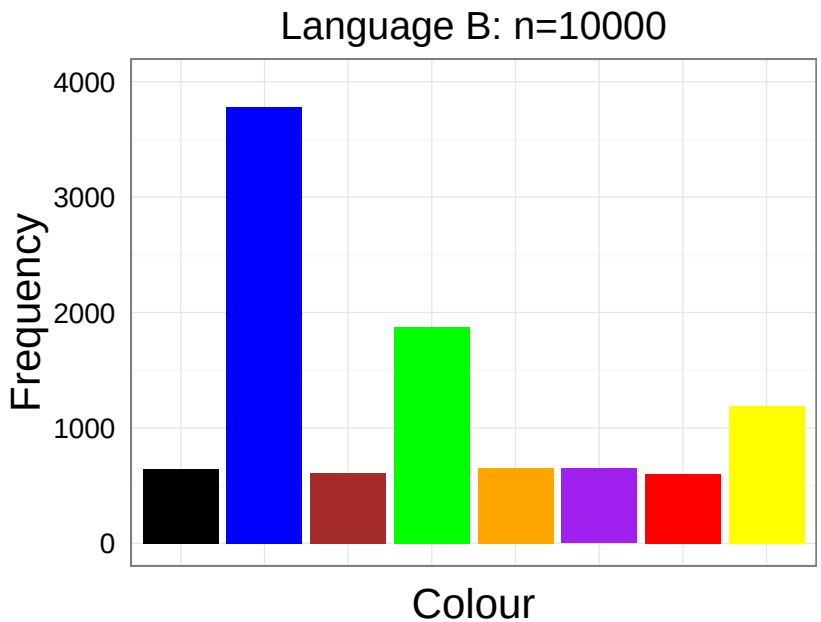


Let's apply this to Languages A and B

Word types in Language *A* carry **3 bits** of information on average, whereas word types in Language *B* carry only **2.61 bits**.

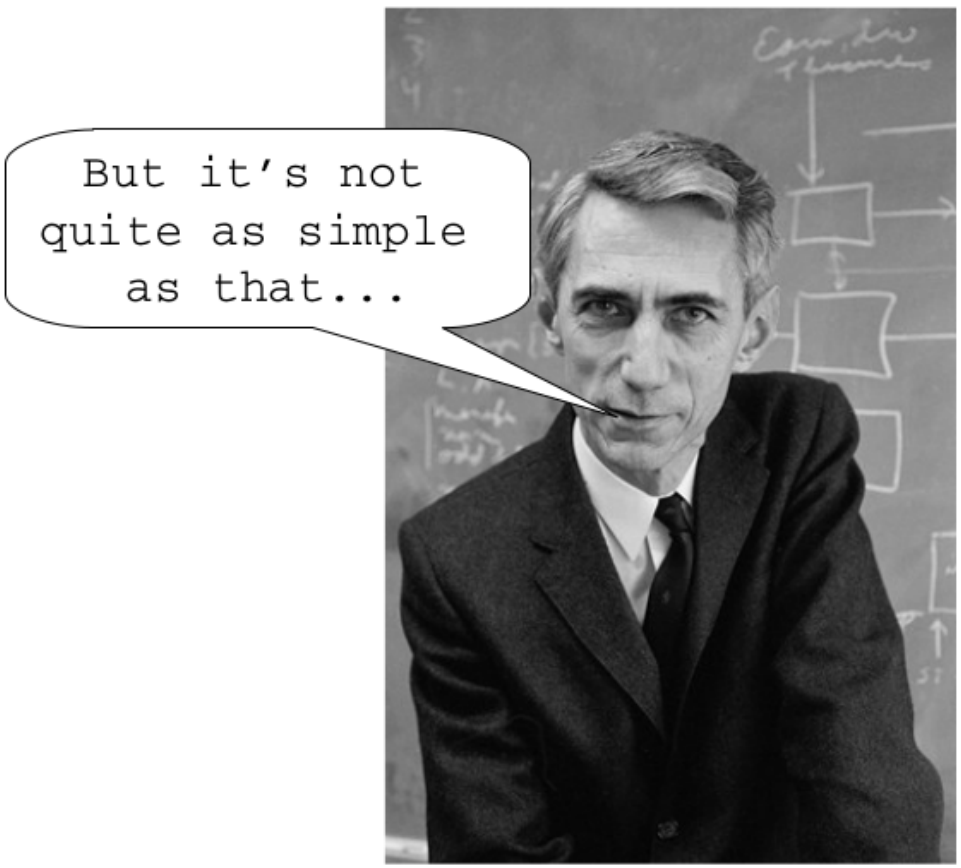
Note that 3 bits is actually the **maximum entropy** possible for a language with 8 word types, since this is the case with uniform probabilities $\frac{1}{8}$. The **minimum entropy** would be 0, namely in the case where only 1 word type is used, since $\log_2(\frac{8}{8}) = 0$.





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

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



Two Major Caveats

1. What is an information encoding "unit" in the first place?
2. What is the "real" probability of letters, words, sentences, or symbols more generally?

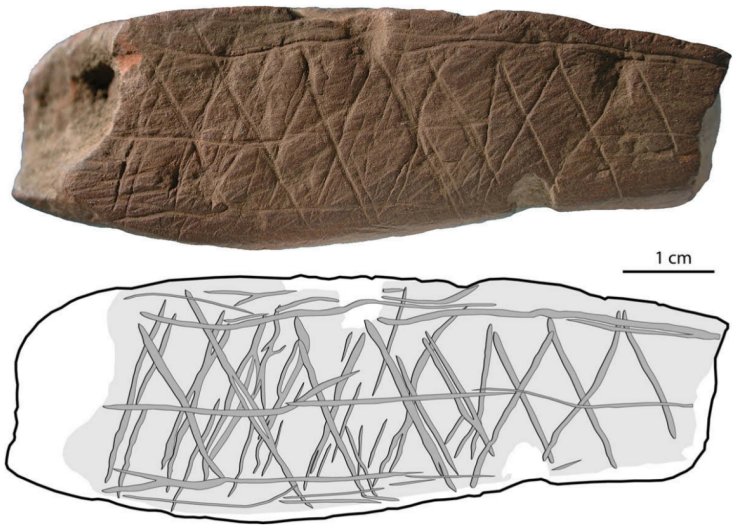
Problem 1

Exercise
Define what you think are the information encoding units in the three "corpora", and calculate the entropy by using the given R code:

1. English_UDHR.csv
2. Amharic_UDHR.csv
3. paleoSign1.png, paleoSign2.png, paleoSign3.png, paleoSign4.png, paleoSign5.png, paleoSign6.png

Problem 1

Example

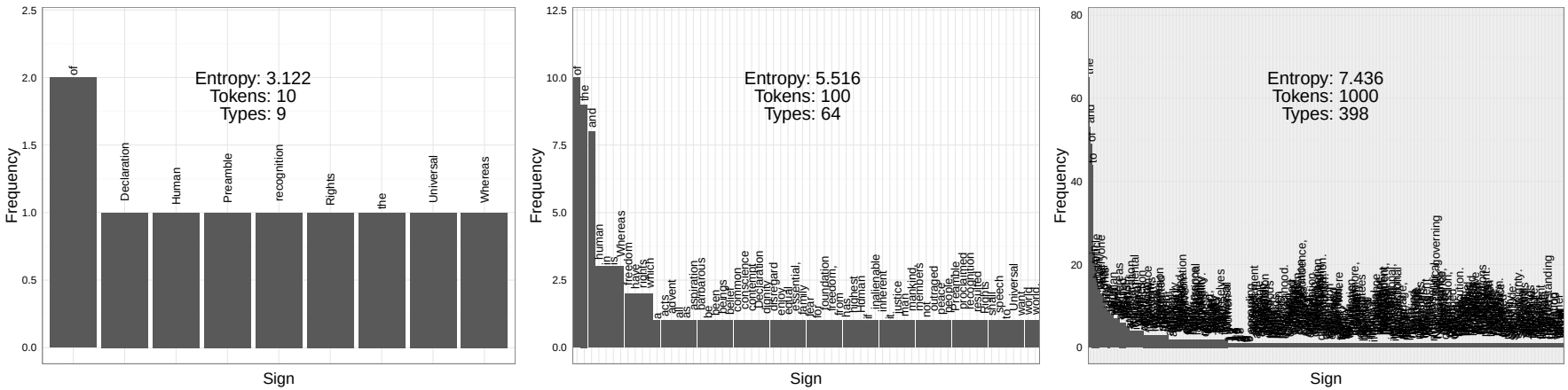


cross cross cross cross cross cross cross cross longline longline

Create a simple .txt or .csv file (e.g. with notepad) where you define names for (potential) signs and put white spaces between them. All "paleo signs" should be in one file, just like the English and Amharic UDHR. You can use line breaks to delimit different paleo sign files, but you don't have to.

PROBLEM 2: WHAT'S THE "REAL" PROBABILITY?

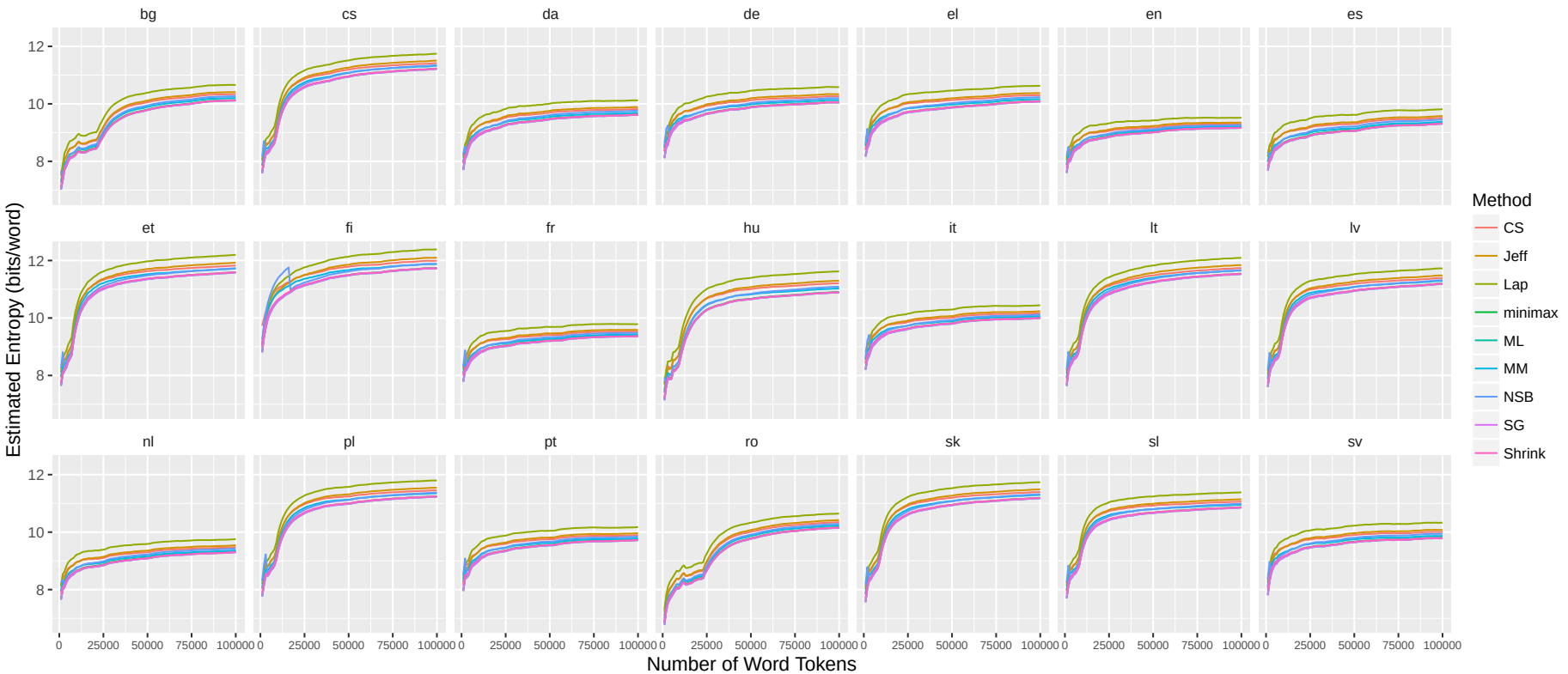
a) The probabilities of letters, words, phrases, etc. depend on the **corpus size**, and so does the entropy $H(p_1, p_2, ...p_N)$.



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

Solution for Problem 2a

Get better entropy estimators (e.g. Hausser & Strimmer 2014 via R package *entropy*), and estimate when the entropy converges to stable values.



Bentz et al. (forthcoming)

Problem 2: What's the "real" probability?

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

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

Example for the first 100 tokens of the English UDHR:

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

While the simple unigram probability of "the" is

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

Solution for Problem 2b

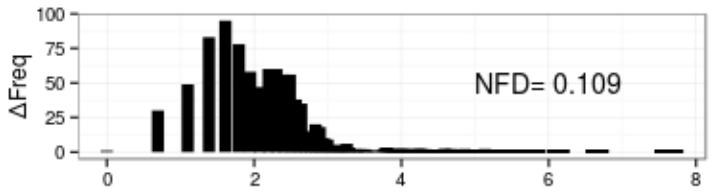
- estimate **n-gram** (bigram, trigram, etc.) entropies instead of unigram entropies. However, this soon requires very big corpora as n increases
- estimate the **entropy rate** h using the results from Gao et al. (2008) on minimum match lengths within a string, and how these relate to $H(X)$ (see Bentz et al. forthcoming, for an application to natural languages)

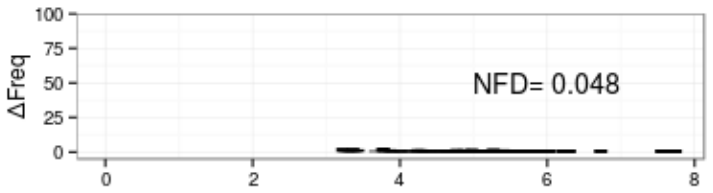
Some Applications

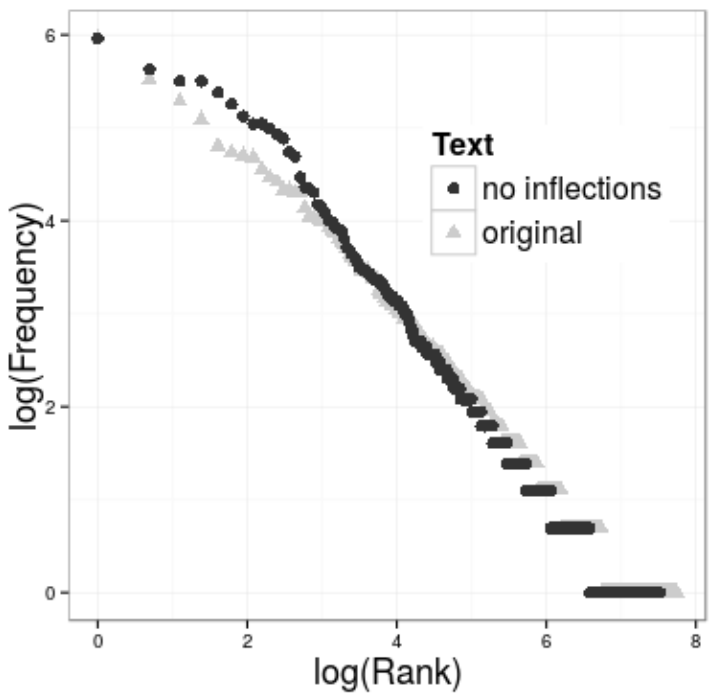
1. Measuring the **information encoding potential** of different levels of structure within the same language
2. **Corpus-based typology**: Measuring and comparing the word entropy across languages

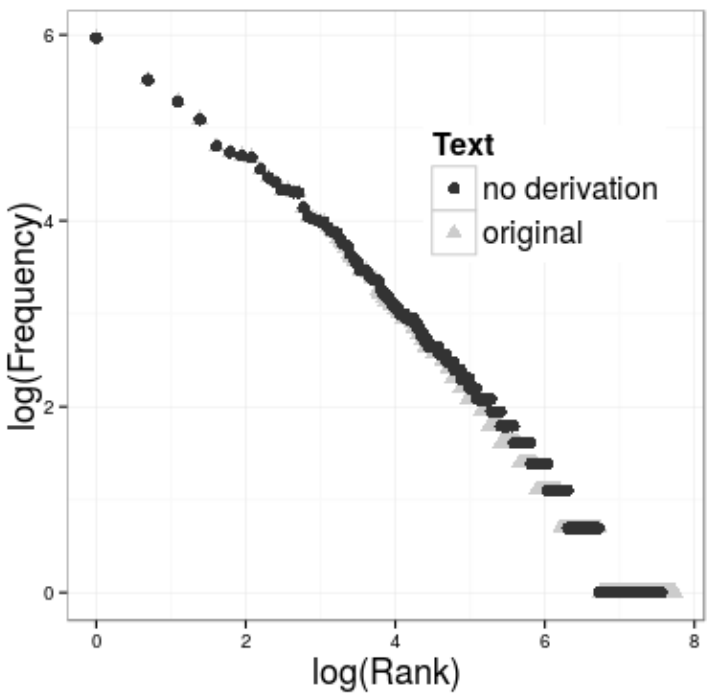
Synchronic Analyses

What happens if we remove **inflection** and **derivation** from a German corpus? Bentz, Alikaniotis, Samardžić & Buttery (accepted)



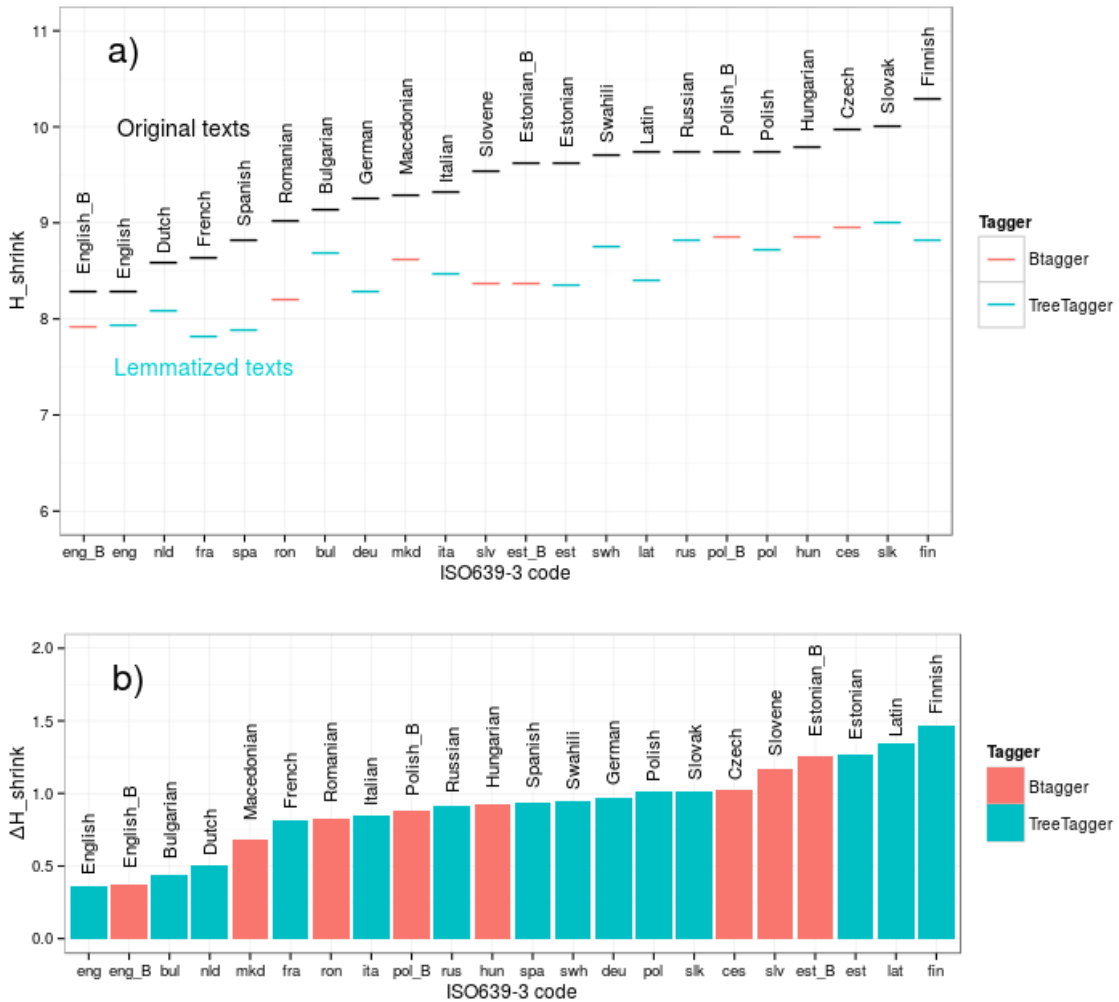






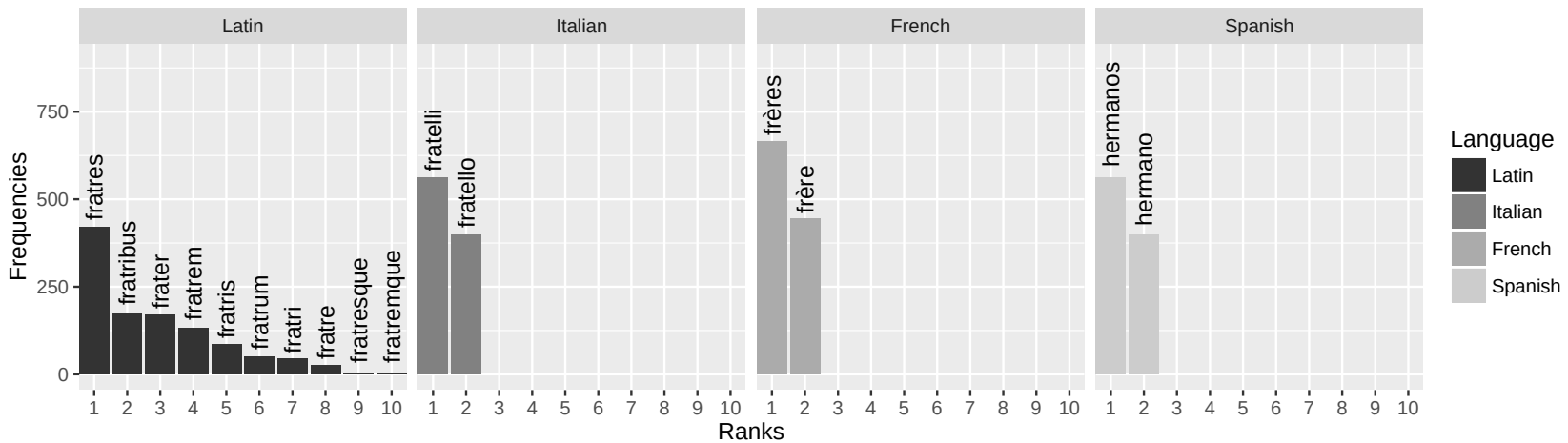
Synchronic Analyses

The information encoding potential of inflection across 21 languages



Diachronic Analyses

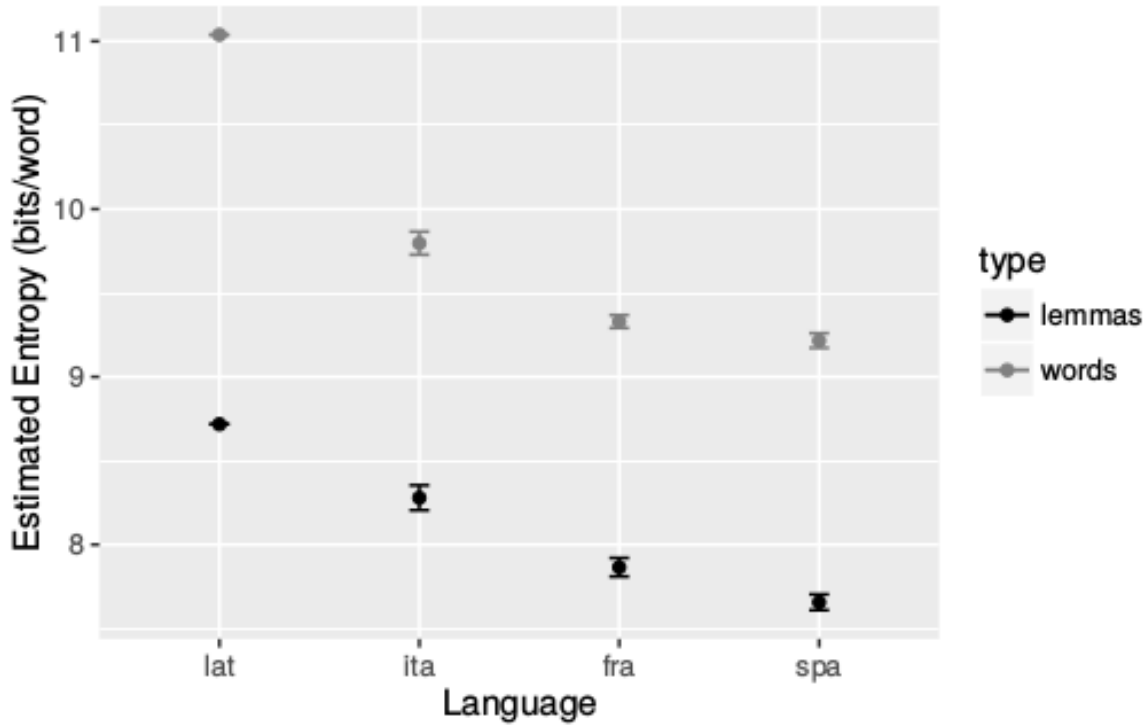
The drop of word entropy from **Latin towards the Romance languages**



Bentz & Berdicevskis (forthcoming)

Diachronic Analyses

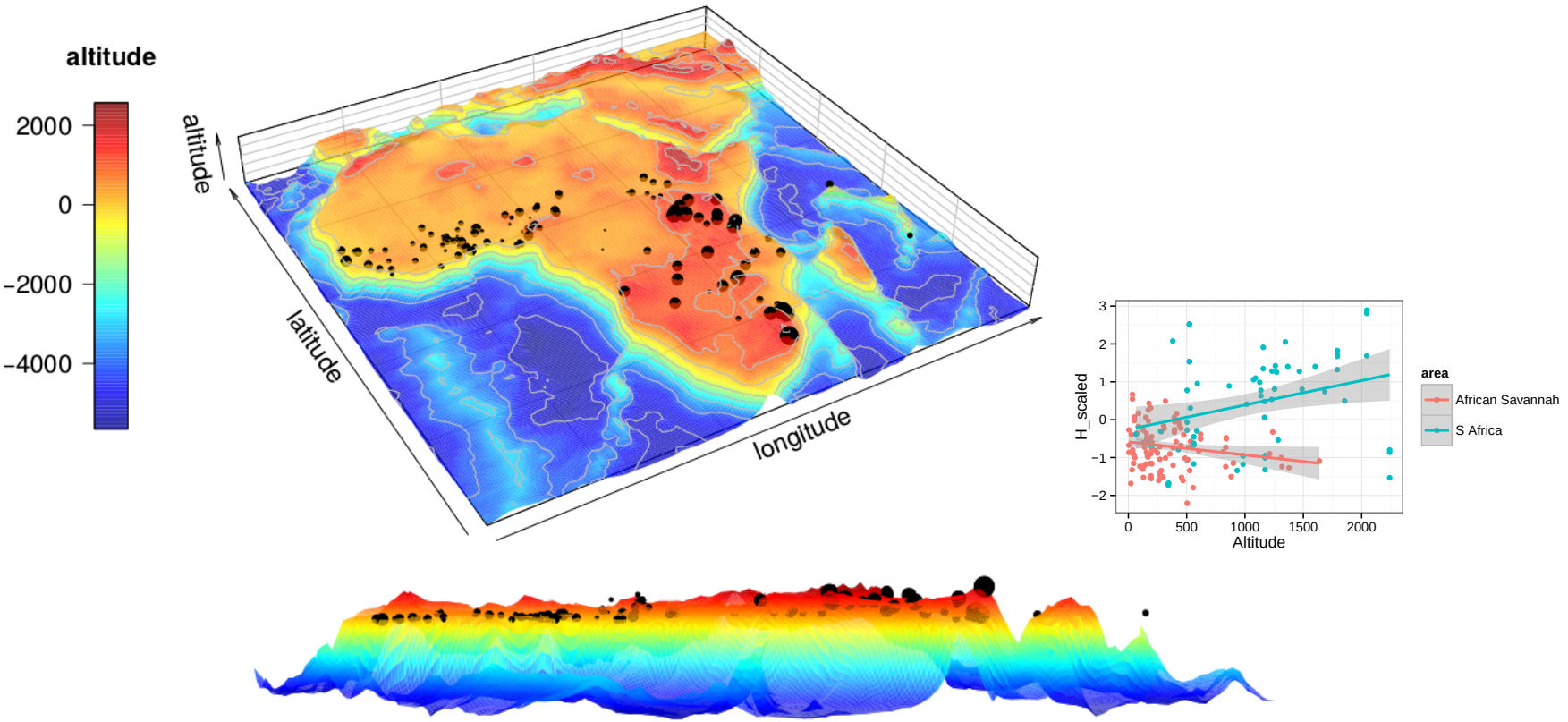
The drop of word entropy from **Latin towards the Romance languages**



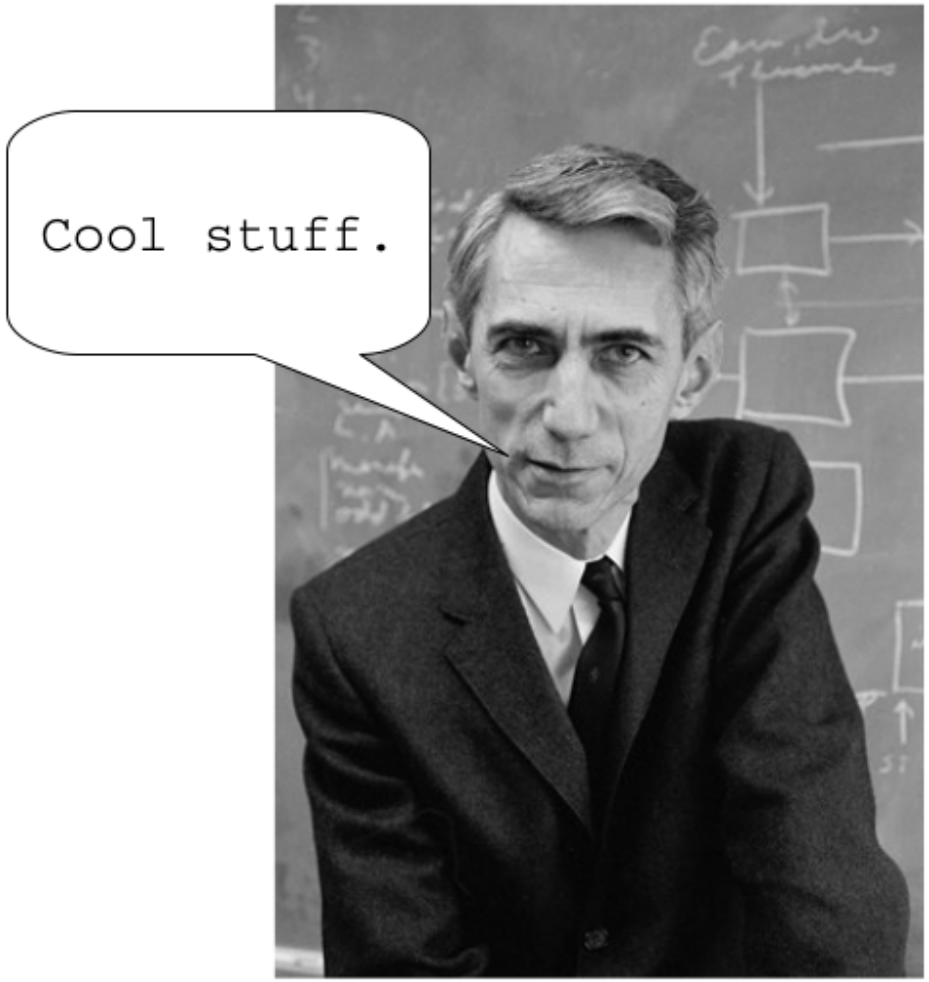
Bentz & Berdicevskis (forthcoming)

Entropy and External Factors?

127 languages, 21 families, 2 areas



It's useful after all!



Thank you