EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

photo by Abdullah Mohiuddin

CC BY-NC-SA Hajime Yamauchi

Language Evolution WiSe 2023/2024

Lecture 10: Information Theory

23/11/2023, Christian Bentz

Overview

Recap

Section 1: Information-Theoretic Measures
- Surprisal
- Entropy
- Entropy Rate
- Further Entropic Measures

Section 2: Estimation Problems
- Probablities
- Encoding Units
- Sample Size
- Interdependence
- Extrapolation

Section 3: Estimation Methods
- Frequency-Based
- Language Models
- Experiments with Humans

Section 4: Information and Meaning

Summary

References

photo by Abdullah Mohiuddin

CC BY-NC-SA Hajime Yamauchi

Recap

What is Language?

Recap

Section 1: Information-Theoretic Measures

Section 2: Estimation Problems

Section 3: Estimation Methods

Section 4: Information and Meaning

Summary

References

Definition
(Usage-Based)

From the **usage-based** perspective **language** is ultimately a **mapping** from phonetic shapes (or hand shapes in sign language, or graphemes in writing) to semantic or pragmatic context. The strength of this mapping is determined by the frequency of co-occurrence.

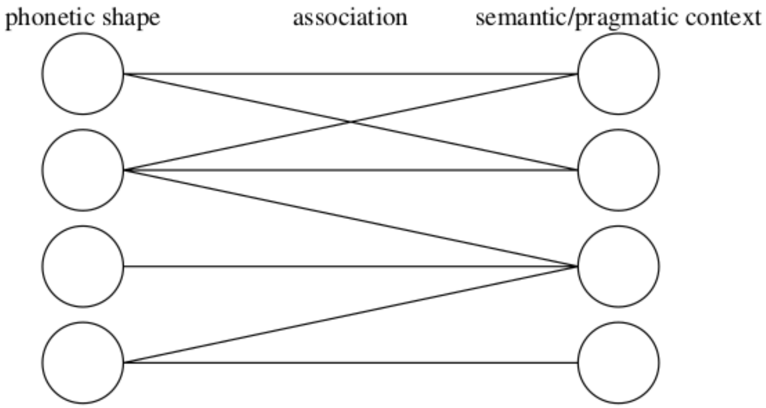
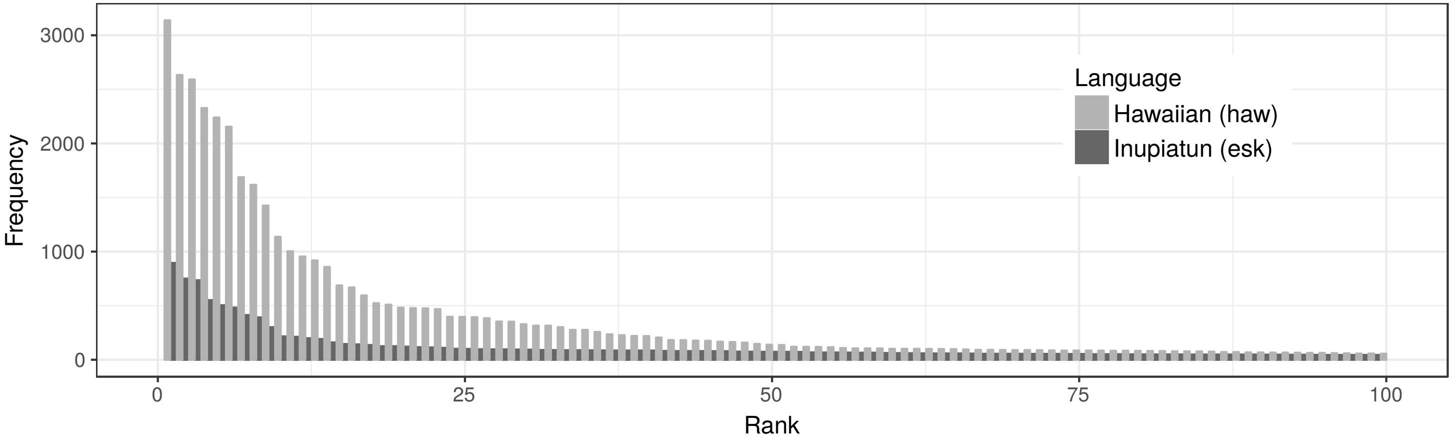


FIGURE 3. Variable associations of form and meaning in a linguistic sign.

Bybee (2006). From usage to grammar: The mind's response to repetition.

Recap
Section 1: Information-Theoretic Measures
Section 2: Estimation Problems
Section 3: Estimation Methods
Section 4: Information and Meaning
Summary
References

Word Frequency Distributions



Hawaiian (haw)

40001001 O ke kuauhau na ka hanauna o Iesu Kristo , ka mamo a Davida , ka mamo a Aberahama.

40001002 Na Aberahama o Isaaka ; na Isaaka o Iakoba ; na Iakoba o Iuda a me kona poe hoahanau;

[...]

Iñupiatun (esk)

40001001 Uvva ukua aglang ich sivulliang iñ Jesus Christ-ng um , kinguviang upluni David-miñ Abraham-miñḷu .

40001002 Abraham aapagigaa Isaac-ng um , Isaac-li aapagigaa Jacob-ng um , Jacob-li aapagigaa Judah-ng um aniqataiñḷu .

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

Real World Salience ↔ Word Frequencies

.... Matterhorn ... Matterhorn ... Bietschhorn ... Jungfrau ... Matterhorn ... Pilatus ... Matterhorn ... Finsteraarhorn ... Bietschhorn ... Matterhorn ... Finsteraarhorn ... Matterhorn ... Matterhorn ... Jungfrau ... Bietschhorn ... Matterhorn ... Matterhorn ... Bietschhorn

Summary

Factors influencing Word Frequency Distributions:

- Lexicon
- Morphology
- Writing Systems
- Translation/Content
- Real World Salience
- etc.

Methodological Question

How can we measure the differences in Frequency Distributions?

Type-Token Ratio (TTR)

$$TTR = \frac{V}{\sum_{i=1}^{V} f_i},$$

- $\mathcal{V}$: set of unique types (*vocabulary*), e.g. $\mathcal{V} = \{A, a, b, \ldots\}$, with $|\mathcal{V}| = V$,
- V: number of character types,
- f_i: Token frequency of given type x_i.

Example

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

char.types	freq
a	5
A	1
b	2
d	3
e	5
f	1
g	3
...	...

word.types	freq
All	1
human	1
beings	1
are	1
born	1
free	1
and	2
...	...

$TTR = \frac{19}{51} = 0.37$

$TTR = \frac{11}{12} = 0.92$

Recap
Section 1: Information-Theoretic Measures
Section 2: Estimation Problems
Section 3: Estimation Methods
Section 4: Information and Meaning
Summary
References

Zipf’s Law (of Word Frequencies)

Word	Rank	Freq	Char
the	1	12539	3
and	2	9964	3
of	3	7459	2
to	4	7317	2
in	5	3985	2
you	6	3747	3
for	7	3014	3
is	8	2957	2
he	9	2925	2
a	10	2862	1
...	...	...	...
work–then	2742	1	10
world-rulers	2743	1	12
worm	2744	1	4
wormwood	2745	1	8
wounding	2746	1	8
writer	2747	1	6
writers	2748	1	7
zarephath	2749	1	9
zenas	2750	1	5

Another (more common) formulation of the law:

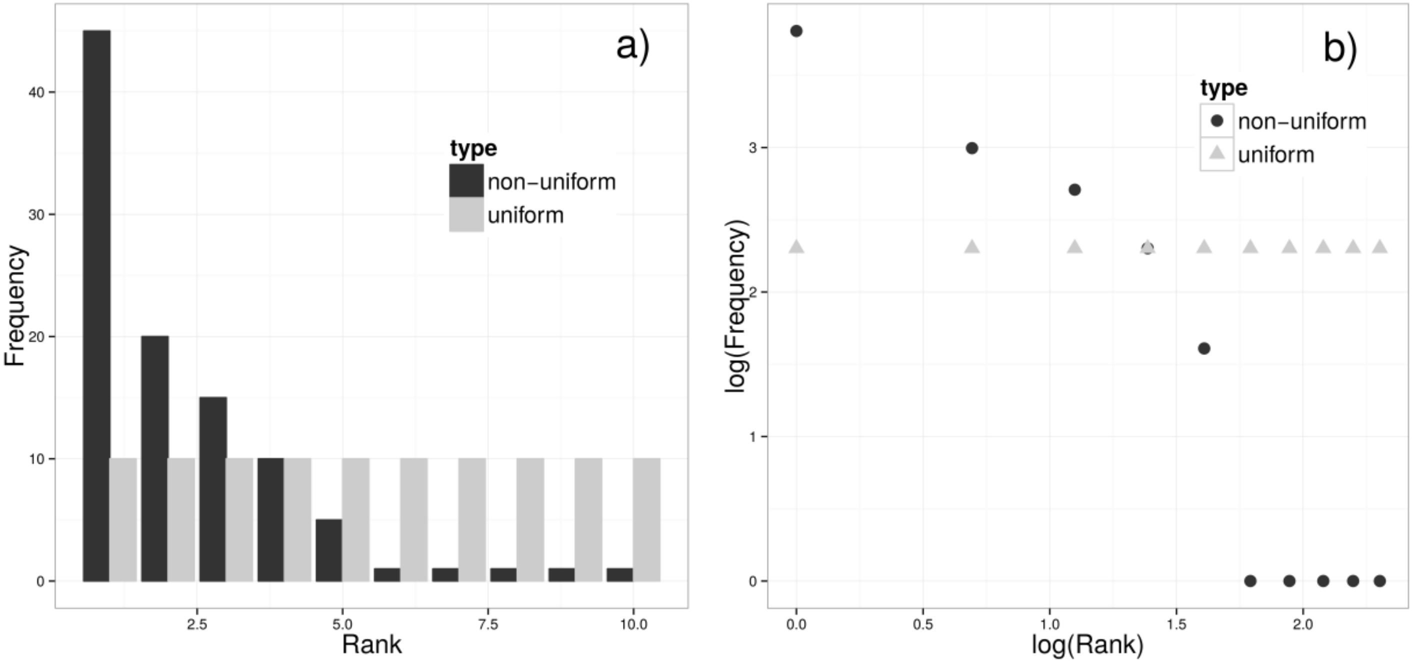
$$f(w) \propto \frac{1}{r^{\alpha}} \tag{1}$$

The α-paramter is the slope in log-log space (i.e. when both the ranks and frequencies are log transformed). Zipf assumed that $\alpha \sim 1$ holds across languages.

photo by Abdullah Mohiuddin
CC BY-NC-SA Hajime Yamauchi

Section 1: Information-Theoretic Measures

Frequency Distributions



Bentz (2018), p. 51.

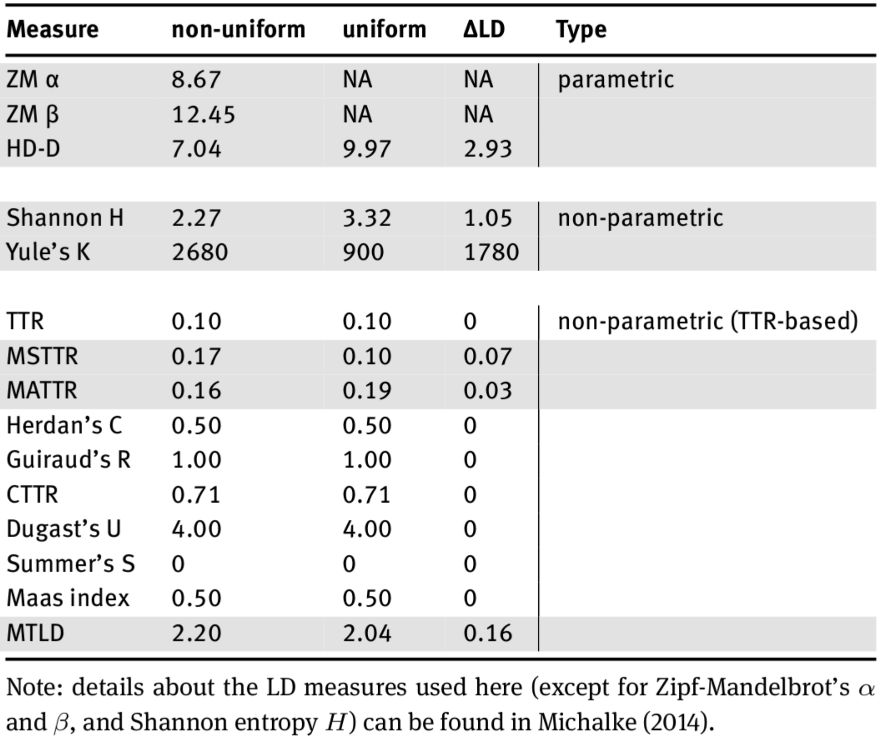
Methodological Question

How can we measure the differences in Frequency Distributions?

Problems

Parametric models such as Zipf-Mandelbrot (ZM) require complicated fitting procedures which can fail for particular kinds of data (e.g. uniform distribution).

Some **non-parametric** methods (e.g. TTR-based) fail to distinguish certain types of distributions (e.g. uniform vs. non-uniform in this example).

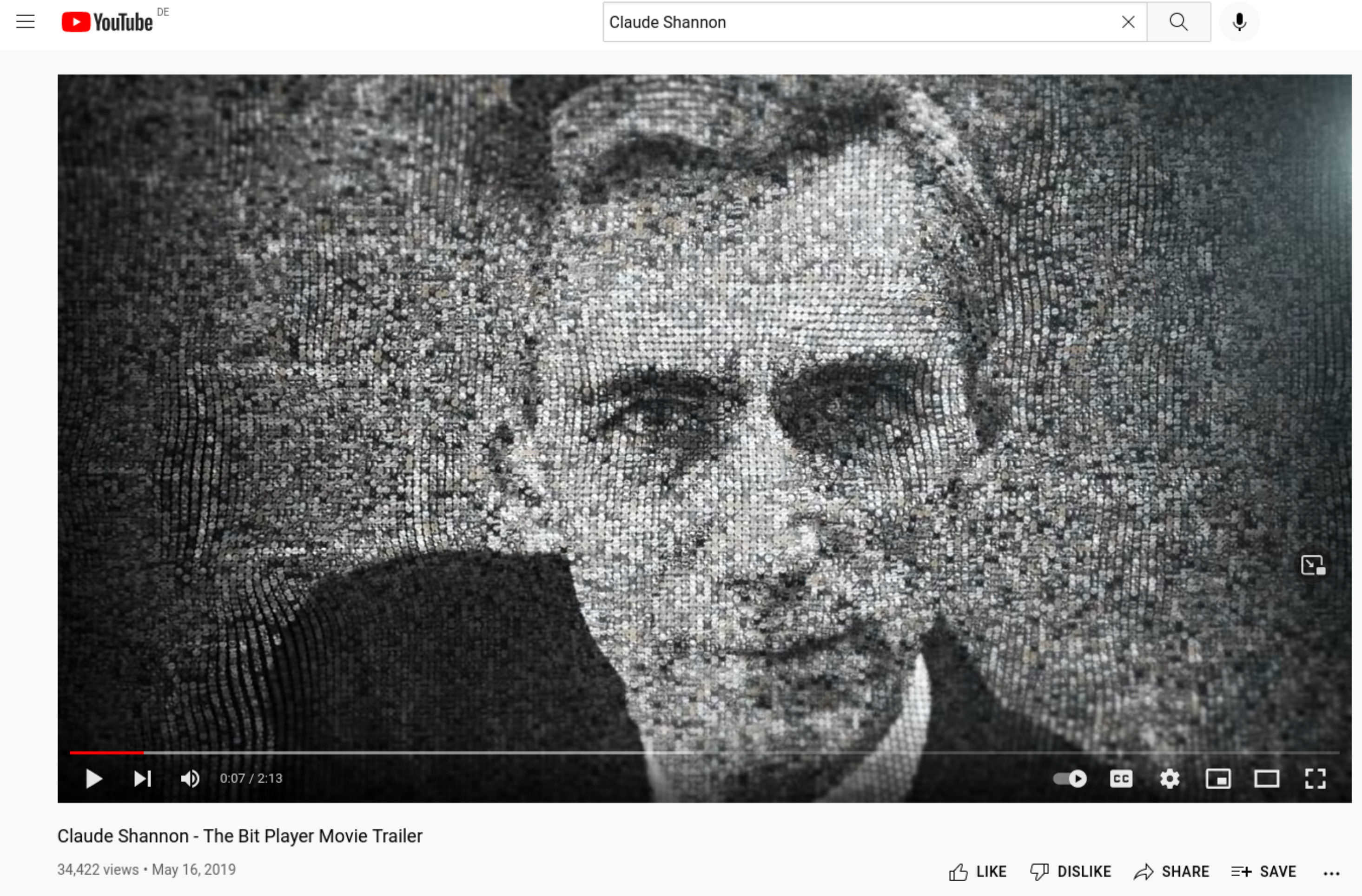


Measure	non-uniform	uniform	ΔLD	Type
ZM α	8.67	NA	NA	parametric
ZM β	12.45	NA	NA	
HD-D	7.04	9.97	2.93	
Shannon H	2.27	3.32	1.05	non-parametric
Yule's K	2680	900	1780	
TTR	0.10	0.10	0	non-parametric (TTR-based)
MSTTR	0.17	0.10	0.07	
MATTR	0.16	0.19	0.03	
Herdan's C	0.50	0.50	0	
Guiraud's R	1.00	1.00	0	
CTTR	0.71	0.71	0	
Dugast's U	4.00	4.00	0	
Summer's S	0	0	0	
Maas index	0.50	0.50	0	
MTLD	2.20	2.04	0.16	

Note: details about the LD measures used here (except for Zipf-Mandelbrot's α and β, and Shannon entropy H) can be found in Michalke (2014).

Bentz (2018), p. 52.

Recap
Section 1: Information-Theoretic Measures
Section 2: Estimation Problems
Section 3: Estimation Methods
Section 4: Information and Meaning
Summary
References



Claude Shannon - The Bit Player Movie Trailer

34,422 views • May 16, 2019

https://www.youtube.com/watch?v=CCrpgUM_rYc (5:30)

Information Content (Surprisal)

The *information content* or *surprisal* measures how “suprised” we are to encounter a certain character/word. If its probability is low we are more surprised to encounter it.

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

- x: one particular type,
- $p(x)$: probability of x,
- f_i: token frequency of a given type x_i.

Example

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

char.types	freq
a	5
A	1
b	2
d	3
e	5
f	1
g	3
...	...

word.types	freq
All	1
human	1
beings	1
are	1
born	1
free	1
and	2
...	...

$$\hat{I}(a) = -\log_2 \hat{p}(a) = -\log_2 \frac{5}{51} = 3.35\,\text{bits}$$

$$\hat{I}(and) = -\log_2 \hat{p}(and) = -\log_2 \frac{2}{12} = 2.58\,\text{bits}$$

Note: This example uses the so-called *maximum likelihood* (ML) estimator for probabilities. This gives the estimated $\hat{p}$ and $\hat{I}$.

Unigram Entropy

The unigram entropy is the **average information content** of all types.

$$H(X) = -\sum_{i=1}^{V} p(x_i) \log_2 p(x_i),$$

- X: random variable drawn from the set of types (i.e. $\mathcal{V}$),
- V: number of types (as before).

Shannon, Claude E. (1948). A mathematical theory of communication.

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

Example (Characters)

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

unit	char.freq
a	5
A	1
b	2
d	3
e	5
f	1
...	...

$\widehat{H}(X) = -(\frac{5}{51}\log_2(\frac{5}{51}) + \frac{1}{51}\log_2(\frac{1}{51}) + \dots) \sim$ 3.97 bits/char

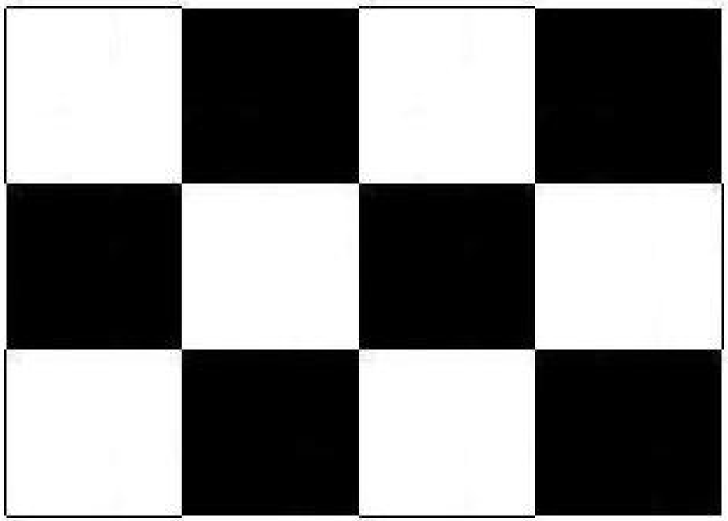
Note: This example uses the so-called *maximum likelihood* (ML) estimator for probabilities. This gives the estimated $\hat{p}$ and $\widehat{H}$.

Recap
Section 1: Information-Theoretic Measures
Section 2: Estimation Problems
Section 3: Estimation Methods
Section 4: Information and Meaning
Summary
References

Exercise

Take the picture below and calculate its entropy (assuming that *white square* = 0 and *black square* = 1). Do the same for the word “square”. Now go to a text to binary converter and convert “square” into binary (`https://cryptii.com/pipes/text-to-binary`). What is the difference between the word “square” and this picture of squares from an information theoretic perspective?

square



Further Entropic Measures

There is a whole range of “entropic” measures derived within *Standard Information Theory*. Some of the most well-known ones are here given for completeness.

Information Content (Surprisal):

$$I(x) = -\log_2 p(x) \tag{2}$$

Entropy:

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

Joint Entropy:

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

Conditional Entropy:

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

Entropy Rate:

$$H(\mathcal{X}) = \lim_{N \to \infty} \frac{1}{N} H(X_1, X_2, \ldots, X_N), \tag{6}$$

Mutual Information:

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{7}$$

Relative Entropy (Kullback-Leibler Divergence):

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)}. \tag{8}$$

Recap

Section 1: Information-Theoretic Measures

Section 2: Estimation Problems

Section 3: Estimation Methods

Section 4: Information and Meaning

Summary

References

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

photo by Abdullah Mohiuddin
CC BY-NC-SA Hajime Yamauchi

Section 2: Estimation Problems

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)

$$I(x) = -\log_2 p(x) \tag{9}$$

Entropy

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \tag{10}$$

Joint Entropy

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

Conditional Entropy

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

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

$$\hat{p}(x_i) = \frac{f_i}{\sum_i^V f_i}, \qquad (13)$$

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

.... Matterhorn ... Matterhorn ... Bietschhorn ... Jungfrau ... Matterhorn ... Pilatus ... Matterhorn ... Finsteraarhorn ... Bietschhorn ... Matterhorn ... Finsteraarhorn ... Matterhorn ... Matterhorn ... Jungfrau ...

$$\hat{p}(Matterhorn) = \frac{7}{14} = 0.5 \qquad (14)$$

Note: The hat above the probability symbol $\hat{p}$ indicates that we are *estimating* the probability, rather than *pre-defining* it.

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.

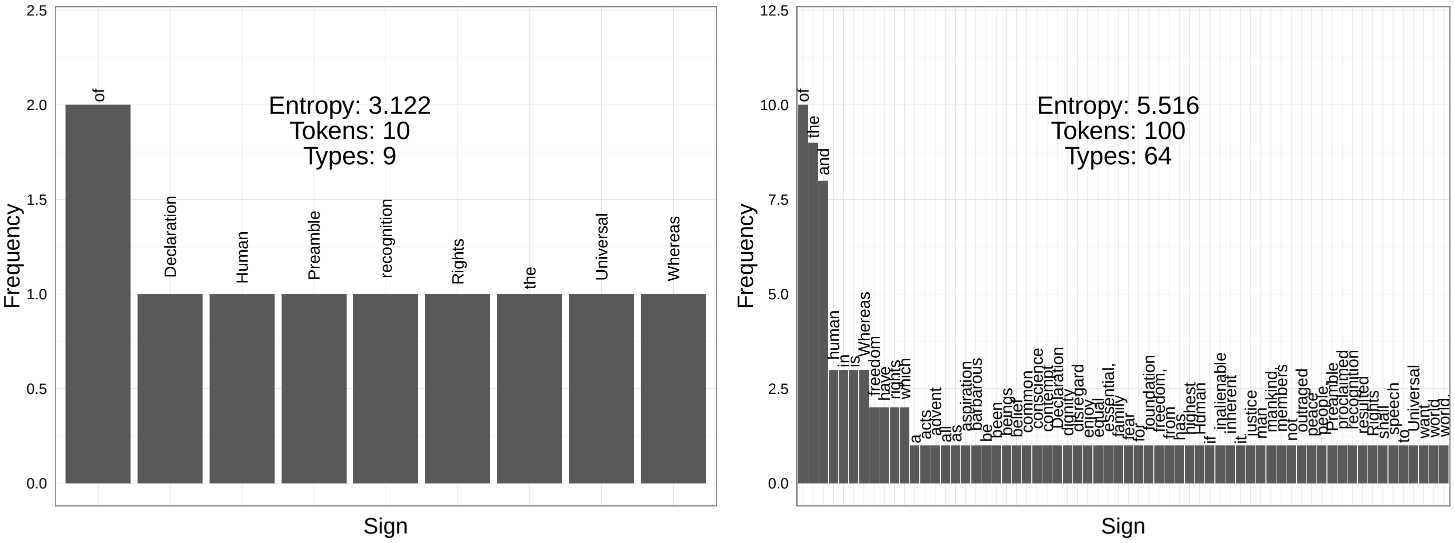
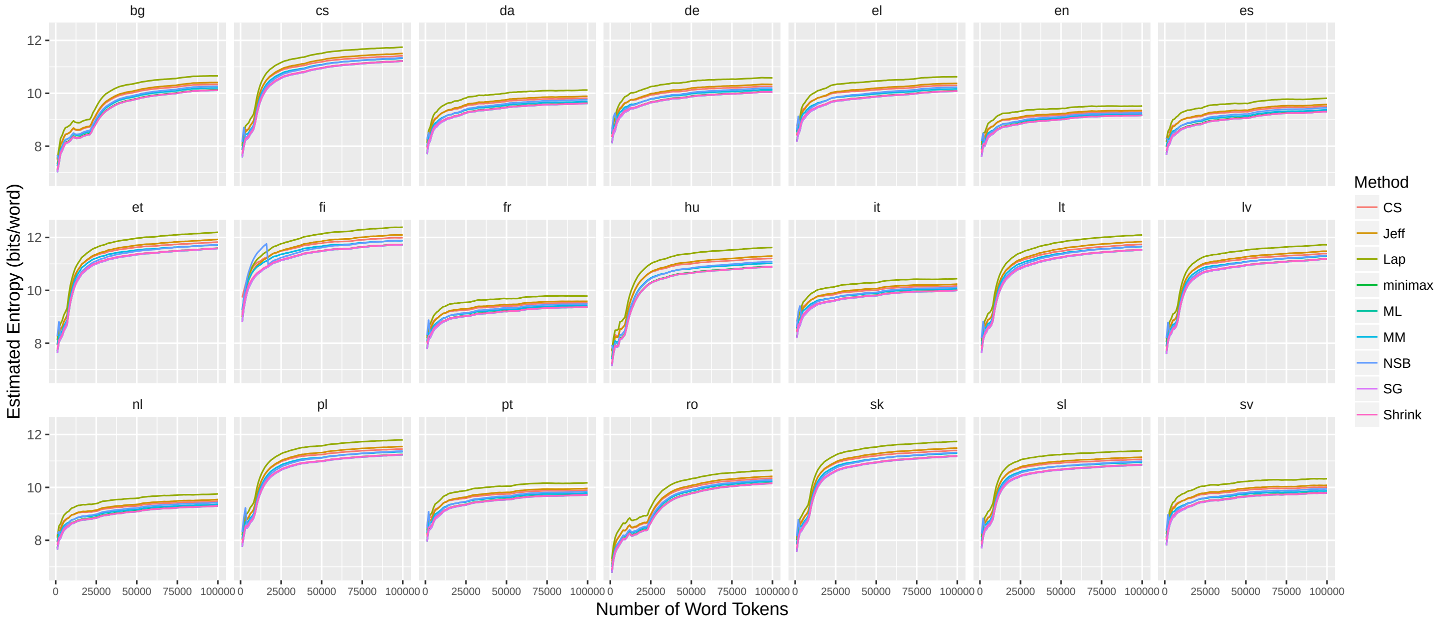


Figure. Frequency distributions and word type entropies for the English UDHR according to the first 10 and 100 word tokens.

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.

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:

$$p(y|x) = \frac{p(x,y)}{p(x)} \tag{15}$$

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 [...]

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

$\hat{p}(the|of) = \frac{p(of,the)}{p(of)} = \frac{\frac{3}{31}}{\frac{5}{32}} \sim \mathbf{0.6}$.

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.

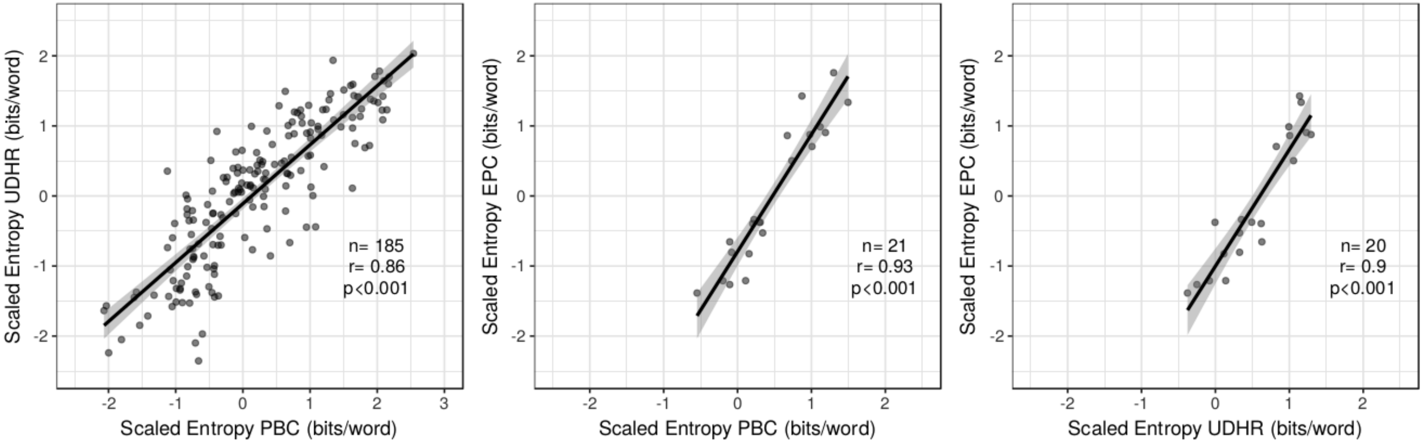
Gao, Kontoyiannis, & Bienenstock (2008). Estimating the entropy of binary time series: Methodology, some theory, and a simulation study.

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.

photo by Abdullah Mohiuddin

CC BY-NC-SA Hajime Yamauchi

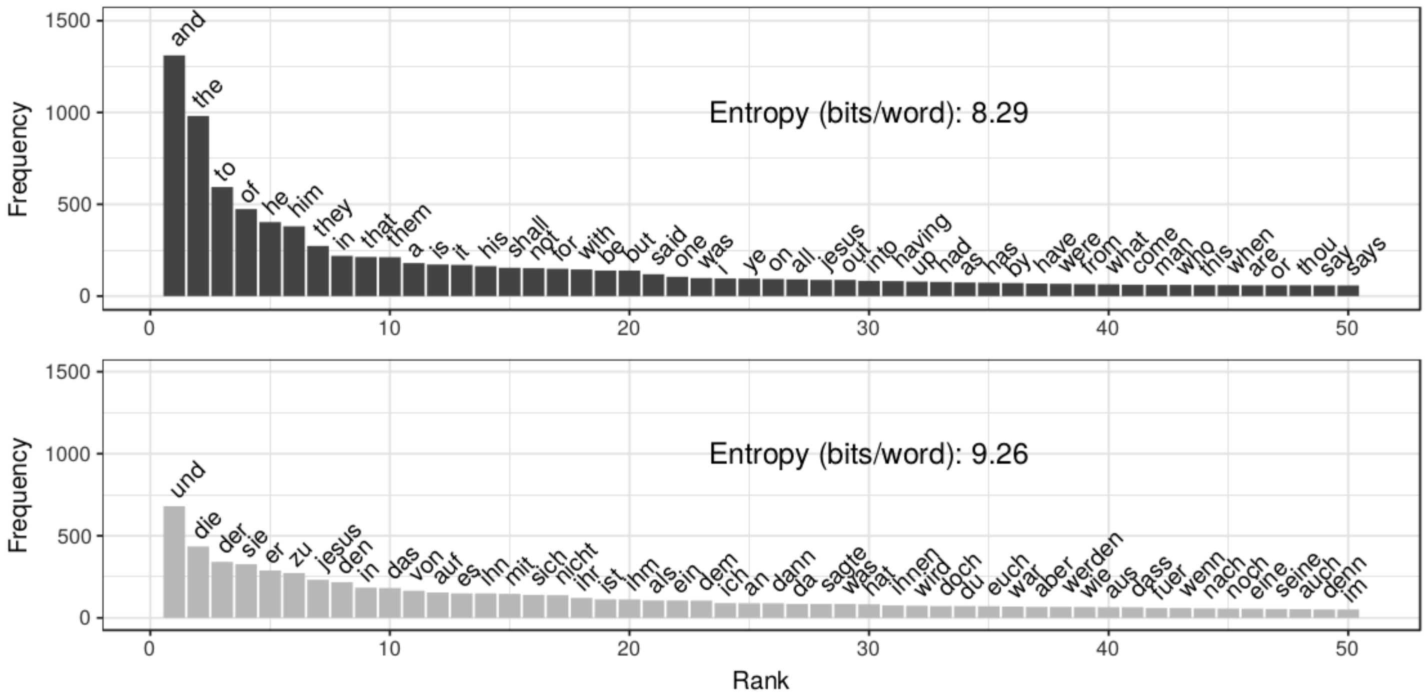
Section 3: Estimation Methods

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).



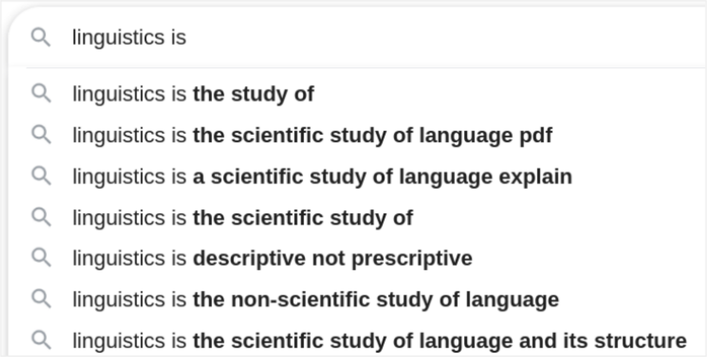
Bentz (2018). Adaptive languages, p. 88.

Language Models

Useful tool in NLP for estimating the probability of sequences

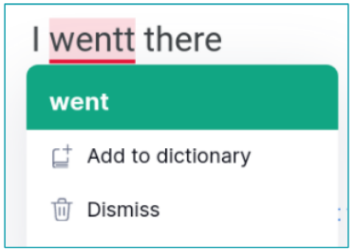
- For example, we can use them for calculating the probability of a sentence in a language (based on a text corpus)
- Many applications in NLP

Recap
Section 1: Information-Theoretic Measures
Section 2: Estimation Problems
Section 3: Estimation Methods
Section 4: Information and Meaning
Summary
References



linguistics is

linguistics is **the study of**
linguistics is **the scientific study of language pdf**
linguistics is **a scientific study of language explain**
linguistics is **the scientific study of**
linguistics is **descriptive not prescriptive**
linguistics is **the non-scientific study of language**
linguistics is **the scientific study of language and its structure**



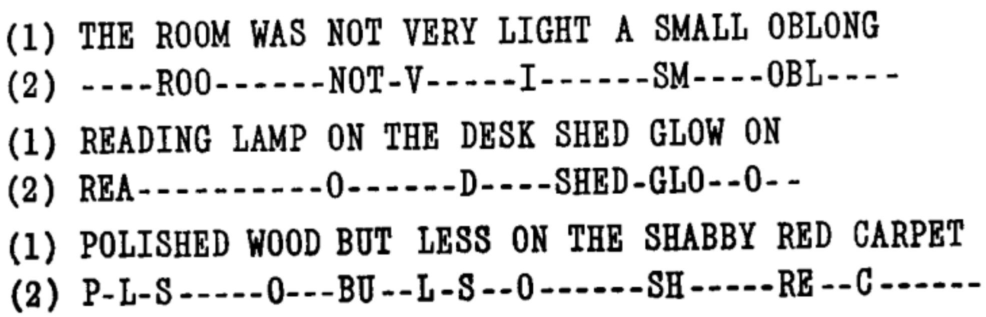
I wentt there

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

See also `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.

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.”

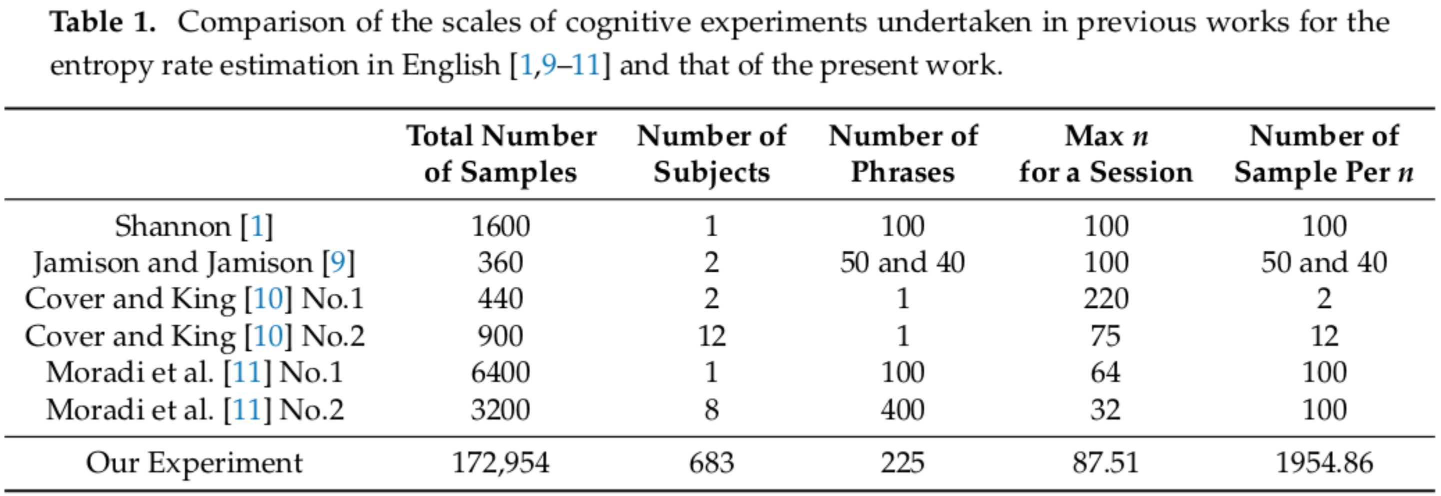


Table 1. Comparison of the scales of cognitive experiments undertaken in previous works for the entropy rate estimation in English [1,9–11] and that of the present work.

	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.

photo by Abdullah Mohiuddin
CC BY-NC-SA Hajime Yamauchi

Section 4: Information and Meaning

Information $\neq$ Meaning

Recap
Section 1: Information-Theoretic Measures
Section 2: Estimation Problems
Section 3: Estimation Methods
Section 4: Information and Meaning
Summary
References

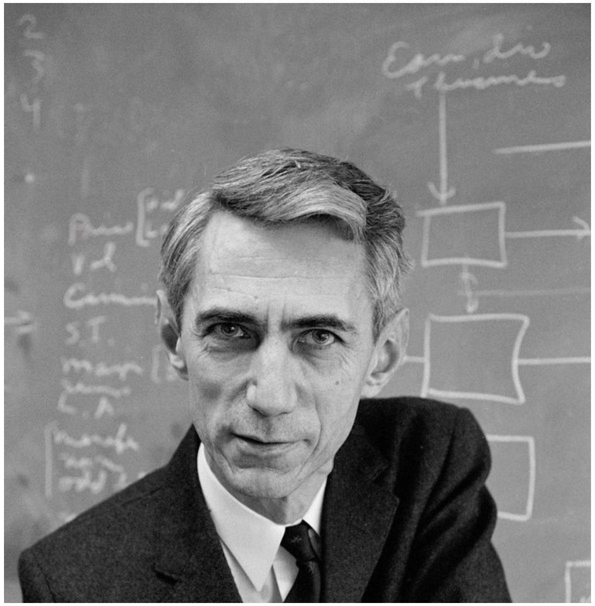
Article 1
All human beings are born free and equal in dignity
and rights. They are endowed with reason and
conscience and should act towards one another in a
spirit of brotherhood.

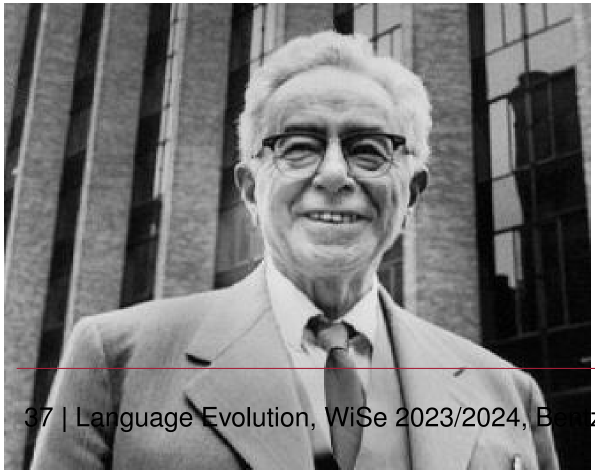
Universal Declaration of Human Rights (UDHR) in English

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 ???

Information and Meaning





[...] 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.

Entropy and Mutual Information

The entropy could be seen as a **necessary** but **not sufficient** condition for meaning encoding. That is, the entropy of signals is an upper bound on the mutual information between signals (S) and referents/meanings (R), i.e.

$$H(S) \geq I(S, R) \tag{16}$$

Ferrer-i-Cancho & Diaz-Guilera (2007). The global minima of the communicative energy of natural communication systems.

Example: Bird Song and Human Language

```
rn rn kd rq rp km jx km rn rn kd rq rp ro as rr rs rt
ls as am rn rn kd rq rp ro ro lo rn rn kd rq rp as rr
rs rt rh rn rn tw nn ir rh tx rn lo rs rt rh
```

$$\widehat{H}(X) \sim 3.1 \text{ bits/char}$$
$$\widehat{H}(X) \sim 3.9 \text{ bits/char.string}$$

```
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
```

$$\widehat{H}(X) \sim 4.1 \text{ bits/char}$$
$$\widehat{H}(X) \sim 4.5 \text{ bits/char.string}$$

Implication: Bird Song and Human Language

Human language (English UDHR) has a higher entropy, i.e. average information content, for both single characters and strings of characters (delimited by white spaces) than bird song (of this particular example).

While we do not strictly know the meaning(s) this bird song encodes, we know that it cannot encode more meanings (unambiguously) than the English UDHR passage.

photo by Abdullah Mohiuddin

CC BY-NC-SA Hajime Yamauchi

Summary

Summary

- There is a range of (interrelated) **information-theoretic measures**: information content (surprisal), entropy, joint entropy, conditional entropy, relative entropy, etc.
- 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.
- While it is true that **information** $\neq$ **meaning**, the entropy of a signal system can be seen as the **upper bound** on how much mutual information there can be between signals and the meanings they encode.

photo by Abdullah Mohiuddin

CC BY-NC-SA Hajime Yamauchi

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

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).

Thank You.

Contact:

Faculty of Philosophy
General Linguistics
Dr. Christian Bentz
SFS Keplerstraße 2, Room 168
chris@christianbentz.de

Office hours:
During term: Wednesdays 10-11am
Out of term: arrange via e-mail