Quantitative Comparison of Natural Languages and Other Sequences

Christian Bentz

November 1, 2023

Department of General Linguistics, University of Tübingen

Acknowledgements

URPP Language and Space

University of Zurich[UZH]

DFG Center for Advanced Studies “Words, Bones, Genes, Tools”

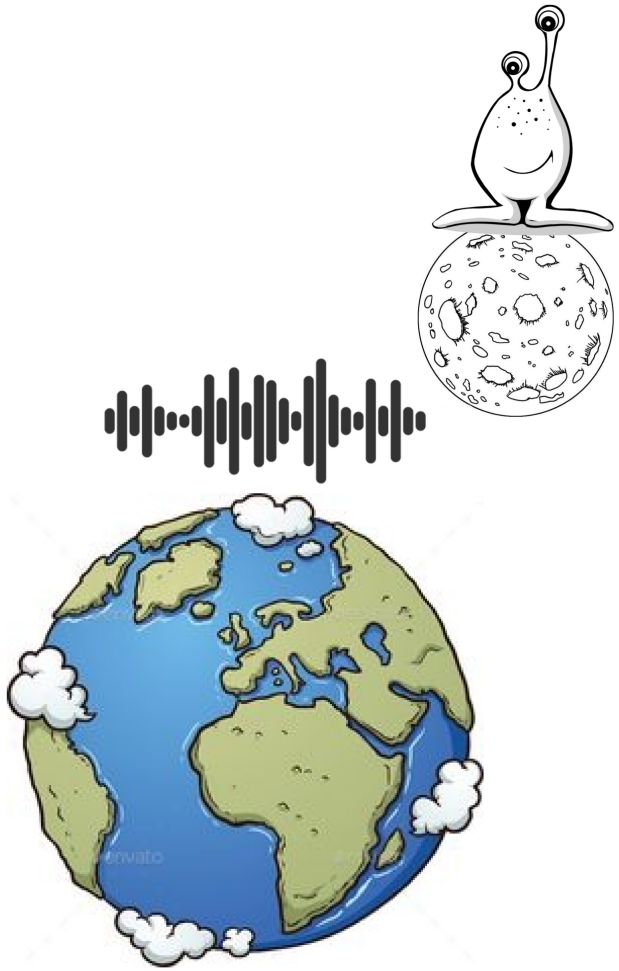
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Introduction

The Martian Scientist (Zipfian View)

If a Martian scientist [...] received from Earth the broadcast of an extensive speech [...] what criteria would [...] determine whether the reception represented the effect of an animate process on Earth, or merely the latest thunderstorm on Earth? It seems that the only criteria would be the ***arrangement of occurrences of the elements*** *[...]: the arrangement of the occurrences would be neither of* ***rigidly fixed regularity*** *[...] nor yet a completely* ***random scattering*** *of the same.*

Zipf (1936). The psycho-biology of language, p. 187.



The Martian Scientist Challenge

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.

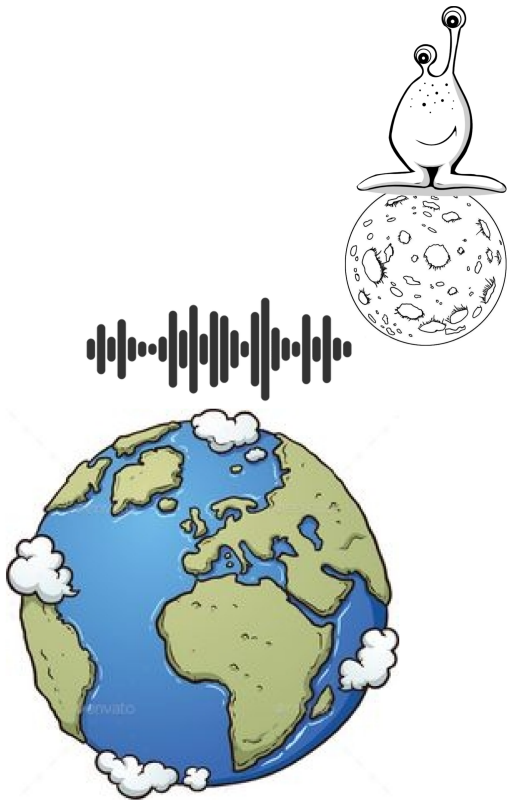
ހުރިހާ ބަނޭ ރަހުމަތްތެރިކަމާއި ދިވެހިންނަކީ، ހަމަހަމަ ގަދަރާއި ހައްގުތަކުގެ ގޮތުން، ފަށާ ފަށާ ވަގުތުން ހިންގާ ފަރާތެއް ކަމަށް ވާ.

모든 인간은 태어날 때부터 자유로우며 그 존엄과 권리에 있어 동등하다.

يولد جميع الناس أحرارًا متساوين في الكرامة والحقوق.

የሰው፡ልጅ፡ሁሉ፡ሲወለድ፡ነጻና፡በክብርና፡በመብት፡ም፡እኩልነት፡ያለው፡ነው።

人人生而自由,在尊严和权利上一律平等。



Research Questions

- Is there a measure/algorithm that can distinguish between natural languages and other sequences?

Research Questions

- Is there a measure/algorithm that can distinguish between natural languages and other sequences?
- If yes - why? What makes languages different?

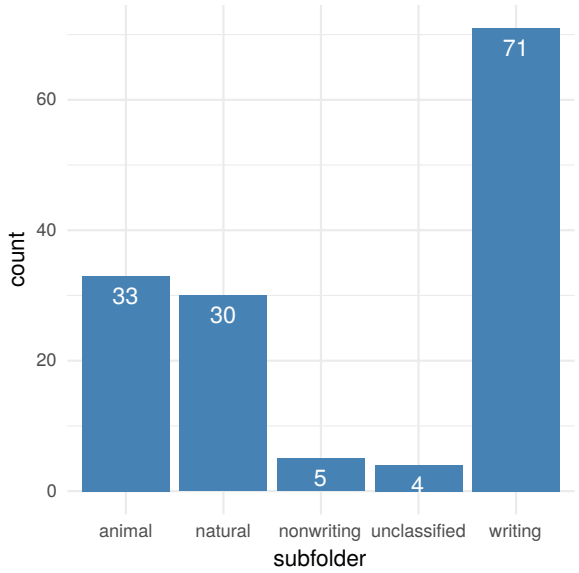
Data

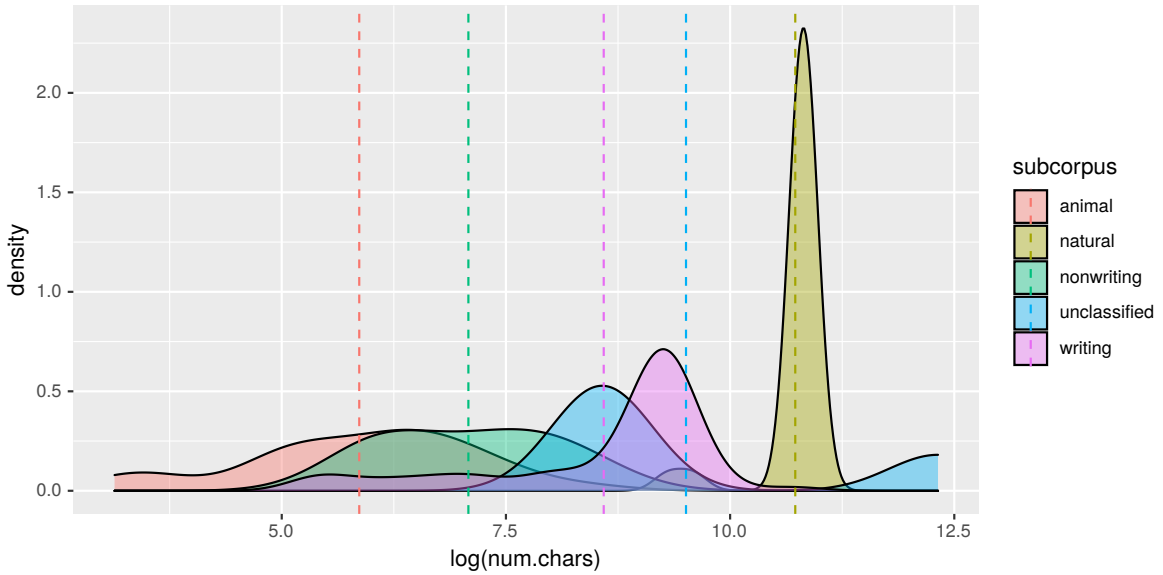
Corpus of Sequences (StringBase)

Hiwis:
Tim Wientzek,
Clara Garcia
Baumgärtner

Number of Files:
138

Number of UTF-8 Characters:
ca. 10 to ca. 200 000 per file

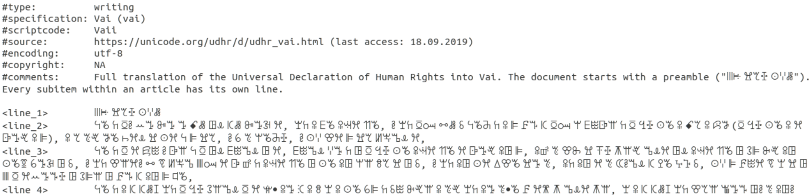




Universal Declaration of Human Rights

- 47 translations (different writing systems)

Eleanor Roosevelt



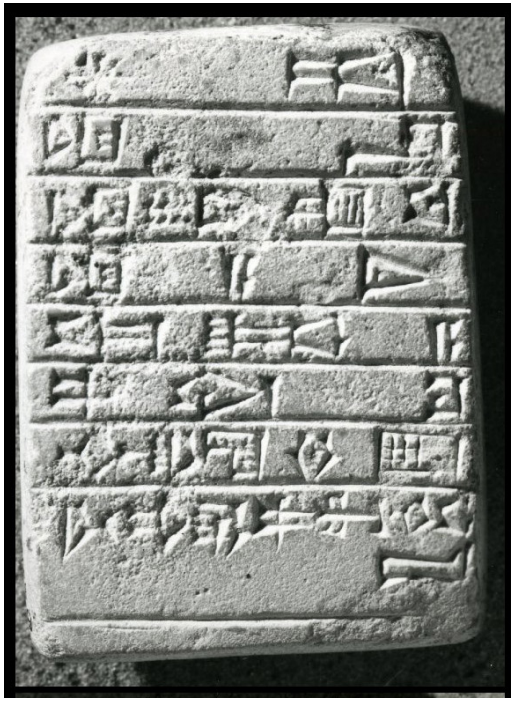
```
#type:          writing
#specification: Vai (vai)
#scriptcode:    Vaii
#source:        https://unicode.org/udhr/d/udhr_vai.html (last access: 18.09.2019)
#encoding:      utf-8
#copyright:     NA
#comments:      Full translation of the Universal Declaration of Human Rights into Vai. The document starts with a preamble ("[illegible]").
Every subitem within an article has its own line.

<line_1>        [illegible]
<line_2>        [illegible]
<line_3>        [illegible]
<line_4>        [illegible]
```

Writing (Ancient)

Sumerian, Akkadian, Elamite, Prakrit, Cretan Hieroglyphs

- 14 texts



```
#type:          writing
#specification: Cuneiform from the Lagash II period of the Sumerian language (sum)
#scriptcode:    Latn
#source:        https://cdli.ucla.edu/search/archival_view.php?ObjectID=P232529 (last access: 28.03.2019); Full sign list: https://
cdli.ucla.edu/tools/SignLists/Rosengarten.pdf (last access: 28.03.2019)
#encoding:      utf-8
#copyright:     "Text in the pages of CDLI may be freely copied, aggregated and re-used according to common academic practice; we request, in
the case of re-use of considerable textual data, that mention be made of the source of such material, with reference to CDLI and its URL
<http://cdli.ucla.edu>." From https://cdli.ucla.edu/?q=terms-of-use (last accessed 27.02.2020).
#comments:      Cuneiform of the Lagash II period (ca. 2200-2100 BC). A full sign list can be found on the above-mentioned link. Lines
beginning with '#' are comments. List of transliteration conventions: https://cdli.ucla.edu/methods/ednotes.html

<obverse>
<line_1>        {d}nansze
<line_2>        nin uru16
<line_3>        nin in-dub-ba
<line_4>        nin-a-ni
```

Nonwriting

Palaeolithic Signs of the Aurignacian (ca. 35 000 BP)

- 20 sequences

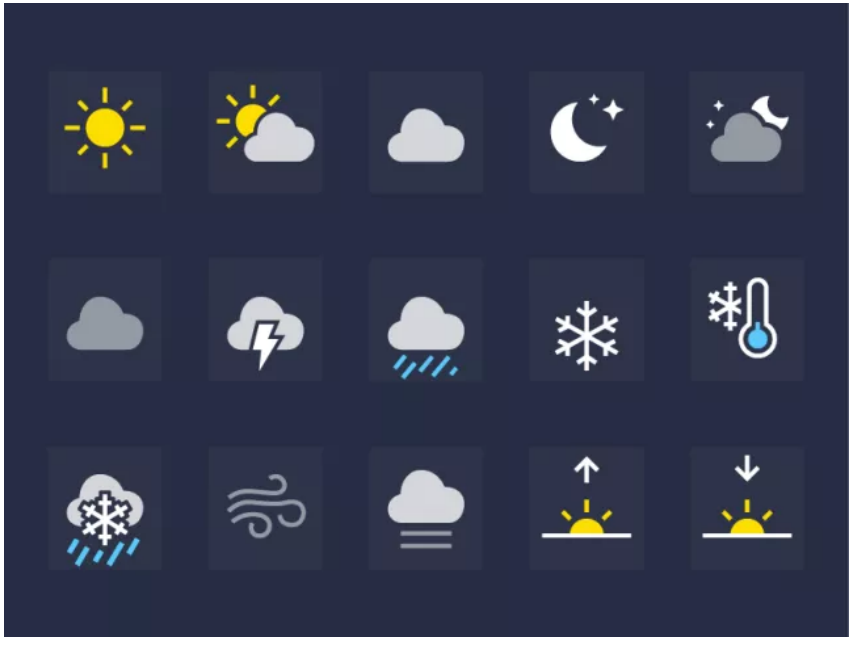

```
#type:          non-writing
#specification: palaeolithic signs (pal)
#scriptcode:    Latn
#source:        NA
#encoding:      utf-8
#copyright:     NA
#comments:      These are encodings of geometric signs found on mobile objects of the palaeolithic. Each line represents a separate object.
In "<>" the SignBase identifier for the object is given. White spaces indicate some visual distinction between rows and clusters of signs.
Alphabet of this particular transcription: "z": zoomorph; "d": dot; "V": vulva; "n": notch; "o": cupule; "N": obnotch; "r": rhombus; "h":
hatching; "l": line; "Z": zigzagrow; "C": circumnotch;

<laf0018>       z ddddd ddddd ddd ddddddd V dd
<bla0003>       nn nn nn nn nn
<laf0008>       nnnnn
<laf0007>       nnnnnnnnnn
<laf0015>       V
<laf0011>       V ooo V dd ddddddddddddddd ddd ddd d ddddd
```

Nonwriting

Weather Symbols

- 1 sequence



```
#type:          non-writing
#specification: weather symbols (wsy)
#scriptcode:    Latn
#source:        https://www.weather.com (last access: 03.03.2019)
#encoding:      utf-8
#copyright:     NA
#comments:      The data was collected by looking at the 10-day forecast (which actually is a 14-days forecast) for different cities around
the world. Following transcriptions of the weather symbols have been used: MostlyCloudy, Showers, Cloudy, PartlyCloudy, RainSnowShowers,
SnowShowers, Thunderstorms, MostlySunny, IsolatedThunderstorms, Sunny, Wind, ScatteredThunderstorms, Rain

<line_1>        Showers Wind Cloudy Cloudy Showers Showers Showers Showers Showers Showers Showers Showers Showers Showers
<line_2>        Cloudy Showers Sunny MostlySunny Showers PartlyCloudy PartlyCloudy PartlyCloudy MostlySunny PartlyCloudy PartlyCloudy
MostlySunny PartlyCloudy PartlyCloudy
<line_3>        Thunderstorms Thunderstorms MostlySunny MostlySunny MostlySunny IsolatedThunderstorms MostlySunny MostlySunny Thunderstorms
MostlySunny Thunderstorms Thunderstorms Thunderstorms Thunderstorms
<line_4>        Cloudy IsolatedThunderstorms Thunderstorms Mostlys MostlyCloudy MostlyCloudy PartlyCloudy IsolatedThunderstorms
```

The Voynich Manuscript

- 1 sequence (ca. 200 000 characters)

```
#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
```

Animal

Birdsong

- 5 species of birds
- 33 sequences

Arriaga et al. (2015). Bird-DB: A database for annotated bird song sequences.

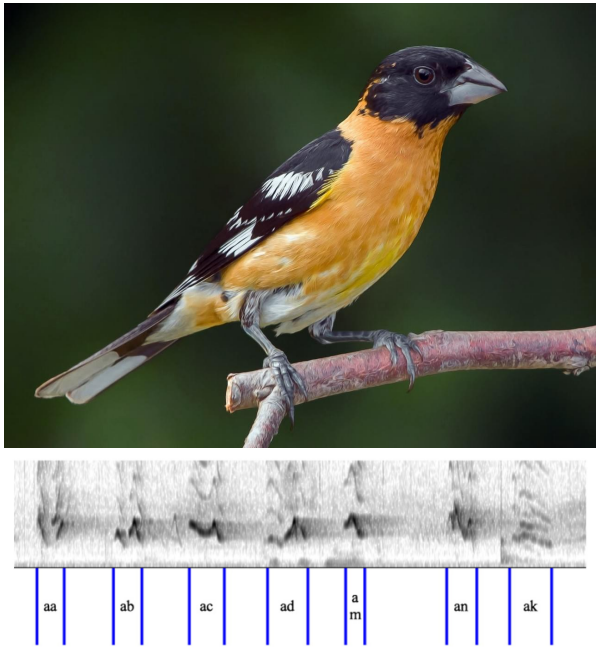


Fig. 1. Audio recording of black-headed grosbeak showing annotation with Praat.

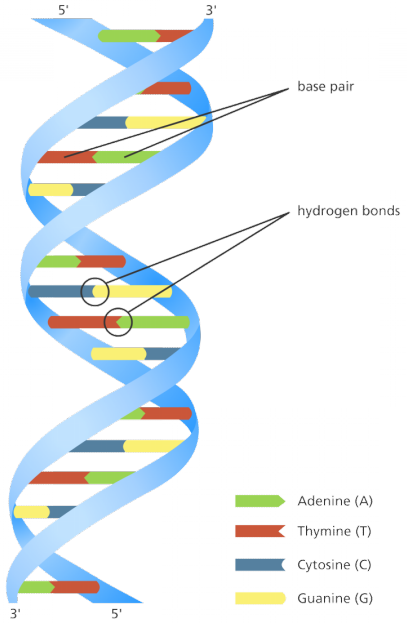
```
#type:          animal
#specification: Black-Headed Grosbeak (bhg)
#scriptcode:    Latn
#source:        http://taylor0.biology.ucla.edu/birdDBQuery/Files/Files_TextGrids/2013/Jun/GTNP13-69.TextGrid (last access: 13.07.2019),
phrases appended in bhg_Phrases.pdf
#encoding:      utf-8
#copyright:     NA
#comments:      Each "unit" corresponds to a sound produced by the bird. White spaces indicate intervals of different durations between the
units. These annotations resulted from a recording made in one piece from a specific bird. Comments that would sometimes follow the bird
recording were removed.

<line_1>        uj kd ro su sv sw sx gf jr dw kd tc jt ag ta tb uj kd no su sv sx sw sx gf jr dw kd tc jr ag ta td te uj kd no su sv sw sx gf
jr dw kd tc jt ag ta td te uj kd no su sv sx sw sw jr dw kd tc jt ag ta td te uj kd no su sv su sv sx sw sx gf jr dw kd tc jt ag ta td te uj
kd no no su sv sw sx gf jr dw kd tc jt ag ta td tc uj kd kd no su sv sx sw gf jr dw kd tc jt ag uj kd no su sv su sv sx sw sw jr dw kd tc jt
ag qe uj kd no no su sv sx sw sw jr ou ou ou ou ou uj kd no no su sv sx sw uk ou ou ou ou ou ou te uj kd no su sv sc sw sw uk ou ou ou ou ou
ou te uj kd no sx gf sw sx gf sw jr dw kd tc jt ag qe ta uj kd no no su sv sw sx gf jr dw kd tc jt ag ta td te uj kd no su sv sx gf sw jr dw
kd tc jt ag qe  uj kd uj kd no su sx sw sx gf sw jr dw kd tc jt ag ta td te
```

Natural

DNA

- 30 sequences




```
#type:          natural
#specification: DNA (dna)
#scriptcode:    Latn
#source:        https://www.ncbi.nlm.nih.gov/nuccore/BAOO01000001.1?report=fasta (last access: 04.07.2019), Input formats found on https://
blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp (last access: 04.07.2019)
#encoding:      utf-8
#copyright:     NA
#comments:      Raphanus Sativus (radish) DNA, "N" stands for any nucleic acid (see input formats)

<line_1>        GGTAGTTAGGGTCTGAAAAAGATTTTGCGTTTTGATAGTTAAATCGATGTAGAACTAAAGTCCCGGTGCA
<line_2>        AATGAGATTCATGCAGTTTGTAGAATAAGTGTGGATTTGGTAAAAATAATTTGCAGTTTTGATAAAAAAA
<line_3>        ATGTAAGACGATTTTGATTACAAAAATGATGAAAATATATATTTTAATATAATTAGAAATTAAATTTGTT
<line_4>        TTCTTAGAGATAATGTCTTAATATTTTTATTTTATATAATTTAAATATTAGAACTAATATGAAATAACAA
<line_5>        TTCTAATTATATATATTATTTTATATTATTAATTACTAAAATTATTTTGAAATAAAAATACAAATCAAAA
<line_6>        TGTAAATAAAGTGATAAATAAAAATAAAATTAATGATTTCATTTTCCAATAACATATTTATAACACAACT
<line_7>        GTTAAAATGCATCGTAACAATCATTTTATATATATATATATATATATATATATATATATTATTTTTGG
<line_8>        TAAAATGTCAAGATTACCATTTTCTGGTTTTTTCAATGTTGATACAACTTATTAGAAACAAACAACAACC
```

Methods

Terminology

- **Sequence**: All the characters in the entire file.
- **String**: A string of characters delimited by *white spaces*.
- **Character**: Individual UTF-8 character.
- **Unit**: Information encoding unit, i.e. *string* or *character*.

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.

ހުރިހާ ޔަޤީންކަމެއްގެ ކުރިން، ހުރިހާ ފަރާތްތަކުގެ ދަށުން، ހުރިހާ ދަތުރެއްގެ ކުރިން، ހުރިހާ ދަތުރެއްގެ ދަށުން، ހުރިހާ ދަތުރެއްގެ ދަށުން، ހުރިހާ ދަތުރެއްގެ ދަށުން.

모든 인간은 태어날 때부터 자유로우며 그 존엄과 권리에 있어 동등하다.

يولد جميع الناس أحرارًا متساوين في الكرامة والحقوق.

የሰው፡ልጅ፡ሁሉ፡ሲወለድ፡ነጻና፡በክብርና፡በመብት፡እኩል፡ነት፡ያለው፡ነው።

人人生而自由,在尊严和权利上一律平等。

Entropy Estimation: Unigrams

$$\widehat{H}^{ML}(X) = -\sum_{i=1}^{W} \widehat{p}^{ML}(x_i) \log_2 \widehat{p}^{ML}(x_i) \tag{1}$$

- ML: Maximum likelihood or “plug-in” estimator

Shannon, Claude E. (1948). A mathematical theory of communication.
Cover & Thomas (2006). Elements of information theory, p. 14.

Example: Characters as Units

$\mathrm{A}_1\ \mathrm{l}_2\ \mathrm{l}_3\ \mathrm{h}_4\ \mathrm{u}_5\ \mathrm{m}_6\ \mathrm{a}_7\ \mathrm{n}_8\ \mathrm{b}_9\ \mathrm{e}_{10}\ \mathrm{i}_{11}\ \mathrm{n}_{12}\ \mathrm{g}_{13}\ \mathrm{s}_{14}\ \mathrm{a}_{15}\ \mathrm{r}_{16}\ \mathrm{e}_{17}\ \mathrm{b}_{18}\ \mathrm{o}_{19}\ \mathrm{r}_{20}\ \mathrm{n}_{21}\ \mathrm{f}_{22}$
$\mathrm{r}_{23}\ \mathrm{e}_{24}\ \mathrm{e}_{25}\ \mathrm{a}_{26}\ \mathrm{n}_{27}\ \mathrm{d}_{28}\ \mathrm{e}_{29}\ \mathrm{q}_{30}\ \mathrm{u}_{31}\ \mathrm{a}_{32}\ \mathrm{l}_{33}\ \mathrm{i}_{34}\ \mathrm{n}_{35}\ \mathrm{d}_{36}\ \mathrm{i}_{37}\ \mathrm{g}_{38}\ \mathrm{n}_{39}\ \mathrm{i}_{40}\ \mathrm{t}_{41}\ \mathrm{y}_{42}$
$\mathrm{a}_{43}\ \mathrm{n}_{44}\ \mathrm{d}_{45}\ \mathrm{r}_{46}\ \mathrm{i}_{47}\ \mathrm{g}_{48}\ \mathrm{h}_{49}\ \mathrm{t}_{50}\ \mathrm{s}_{51}[\ldots]$

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

$$\widehat{H}^{ML}(X) = -\left(\tfrac{5}{51}\log_2\left(\tfrac{5}{51}\right) + \tfrac{1}{51}\log_2\left(\tfrac{1}{51}\right) + \ldots\right) \sim 3.97$$

Example: Strings as Units

$\text{All}_1\ \text{human}_2\ \text{beings}_3\ \text{are}_4\ \text{born}_5\ \text{free}_6\ \text{and}_7\ \text{equal}_8\ \text{in}_9\ \text{dignity}_{10}\ \text{and}_{11}\ \text{rights}_{12}\ [\ldots]$

unit	freq
All	1
and	2
are	1
beings	1
born	1
dignity	1
...	...

$$\widehat{H}^{ML}(X) = -(\tfrac{1}{12}\log_2(\tfrac{1}{12}) + \tfrac{2}{12}\log_2(\tfrac{2}{12}) + \ldots) \sim 3.41$$

Entropy Rate: Theory

The entropy rate as "per symbol entropy of n random variables".

$$h(\mathcal{X}) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) \tag{2}$$

Entropy Rate: Theory

The entropy rate as "per symbol entropy of n random variables".

$$h(\mathcal{X}) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) \tag{2}$$

Alternative definition as "the conditional entropy of the last random variable given the past":

$$h(\mathcal{X}) = \lim_{n \to \infty} H(X_n | X_1, X_2, ..., X_{n-1}). \tag{3}$$

Assumptions:

- $\mathcal{X}$ is a *stationary* stochastic process
- n: number of word tokens

Cover & Thomas (2006) Elements of information theory, p. 74-75.
Dębowski (2020) Information theory meets power laws.

Entropy Rate: Estimator

$$\hat{h}(X) = \frac{1}{n} \sum_{i=2}^{n} \frac{\log_2 i}{L_i}, \tag{4}$$

- n: number of unit tokens
- L_i: length (+1) of the longest contiguous subsequence starting at position i which is also present in $i = 2$ to $i - 1$

Gao, Kontoyiannis & Bienenstock (2008). Estimating the entropy of binary time series, equation (6).

Example

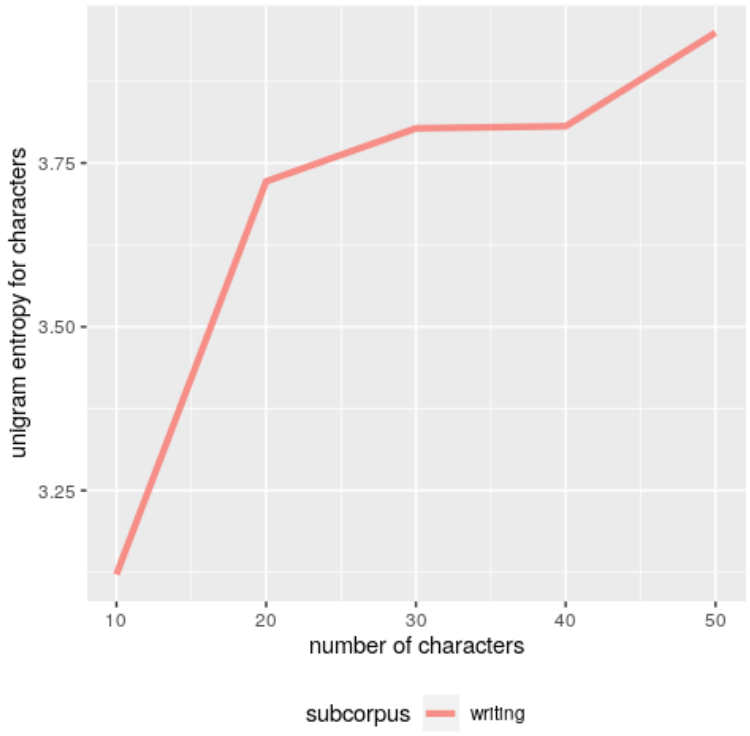
$A_1\ l_2\ l_3\ h_4\ u_5\ m_6\ a_7\ n_8\ b_9\ e_{10}\ i_{11}\ n_{12}\ g_{13}\ s_{14}\ a_{15}\ r_{16}\ e_{17}\ b_{18}\ o_{19}\ r_{20}\ n_{21}\ f_{22}$
$r_{23}\ e_{24}\ e_{25}\ a_{26}\ n_{27}\ d_{28}\ e_{29}\ q_{30}\ u_{31}\ a_{32}\ l_{33}\ i_{34}\ n_{35}\ d_{36}\ i_{37}\ g_{38}\ n_{39}\ i_{40}\ t_{41}\ y_{42}$
$a_{43}\ n_{44}\ d_{45}\ r_{46}\ i_{47}\ g_{48}\ h_{49}\ t_{50}\ s_{51}[...]$

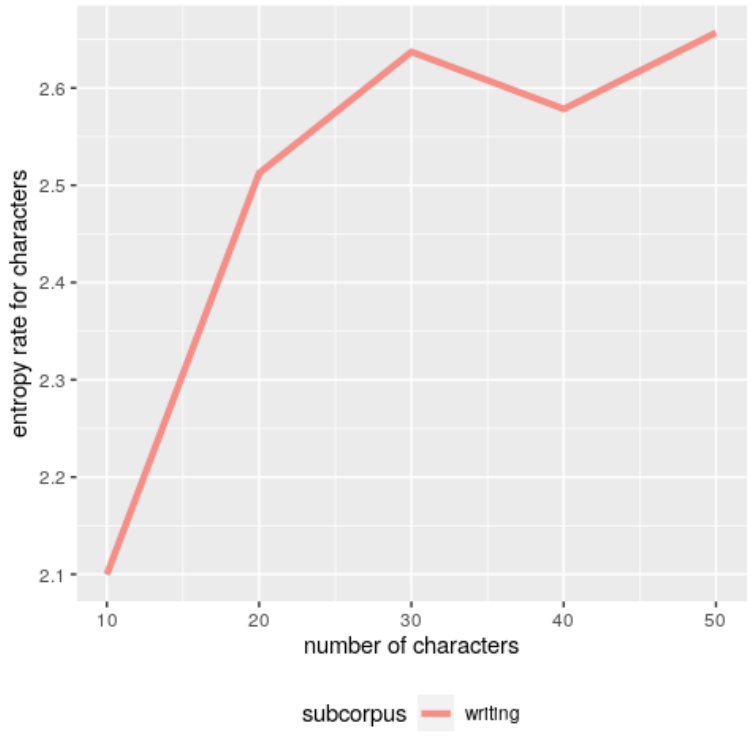
$$L_{43} = 3(+1) = 4$$

$$\frac{\log_2(43)}{4} \sim \frac{5.43}{4} \sim 1.36$$

R package *Hrate* (https://github.com/dimalik/Hrate)
Bentz, Alikaniotis, Cysouw, & Ferrer-i-Cancho (2017). The entropy of words - learnability and expressivity across more than 1000 languages.

Example Plots (UDHR First Sentence in English)

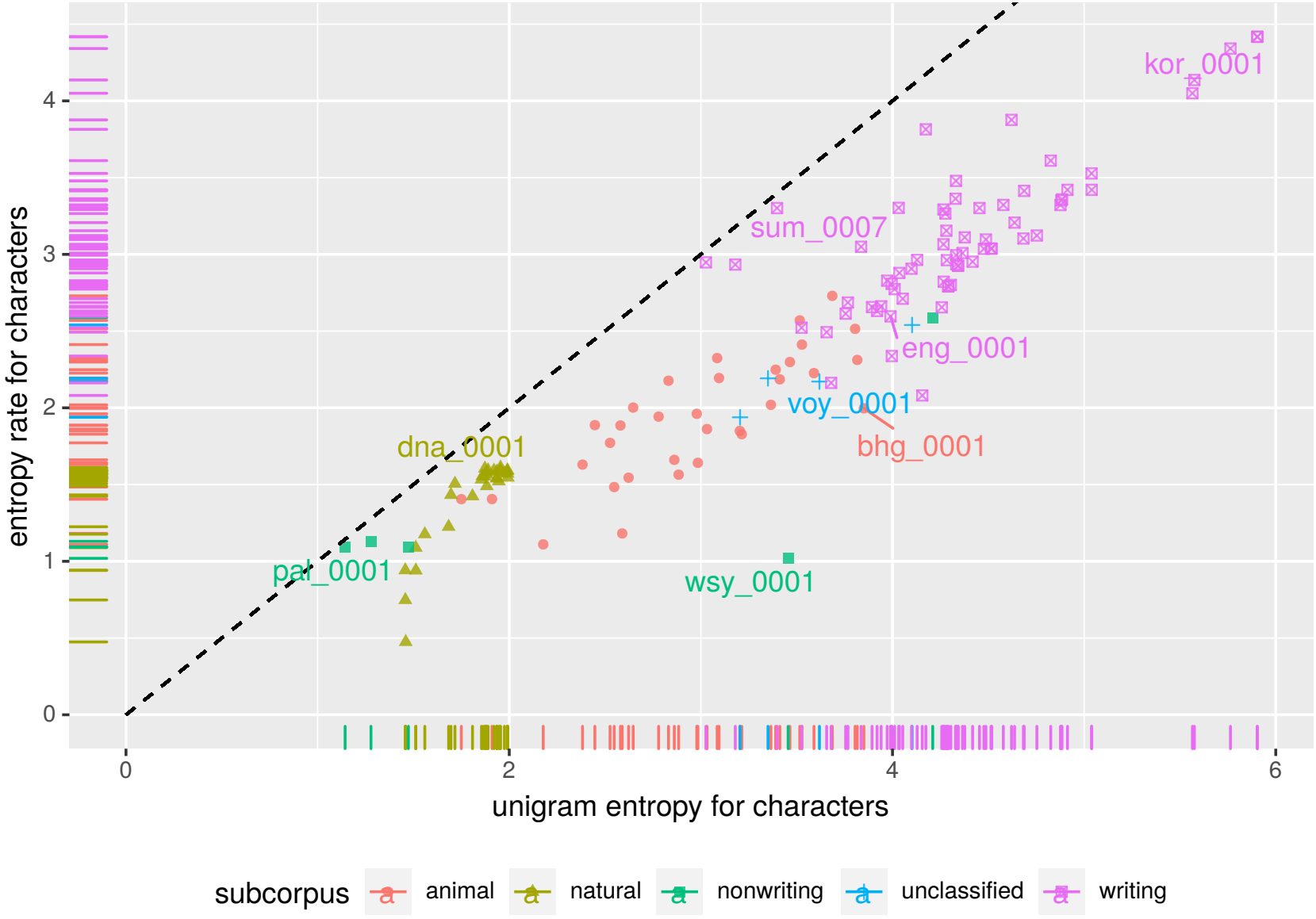




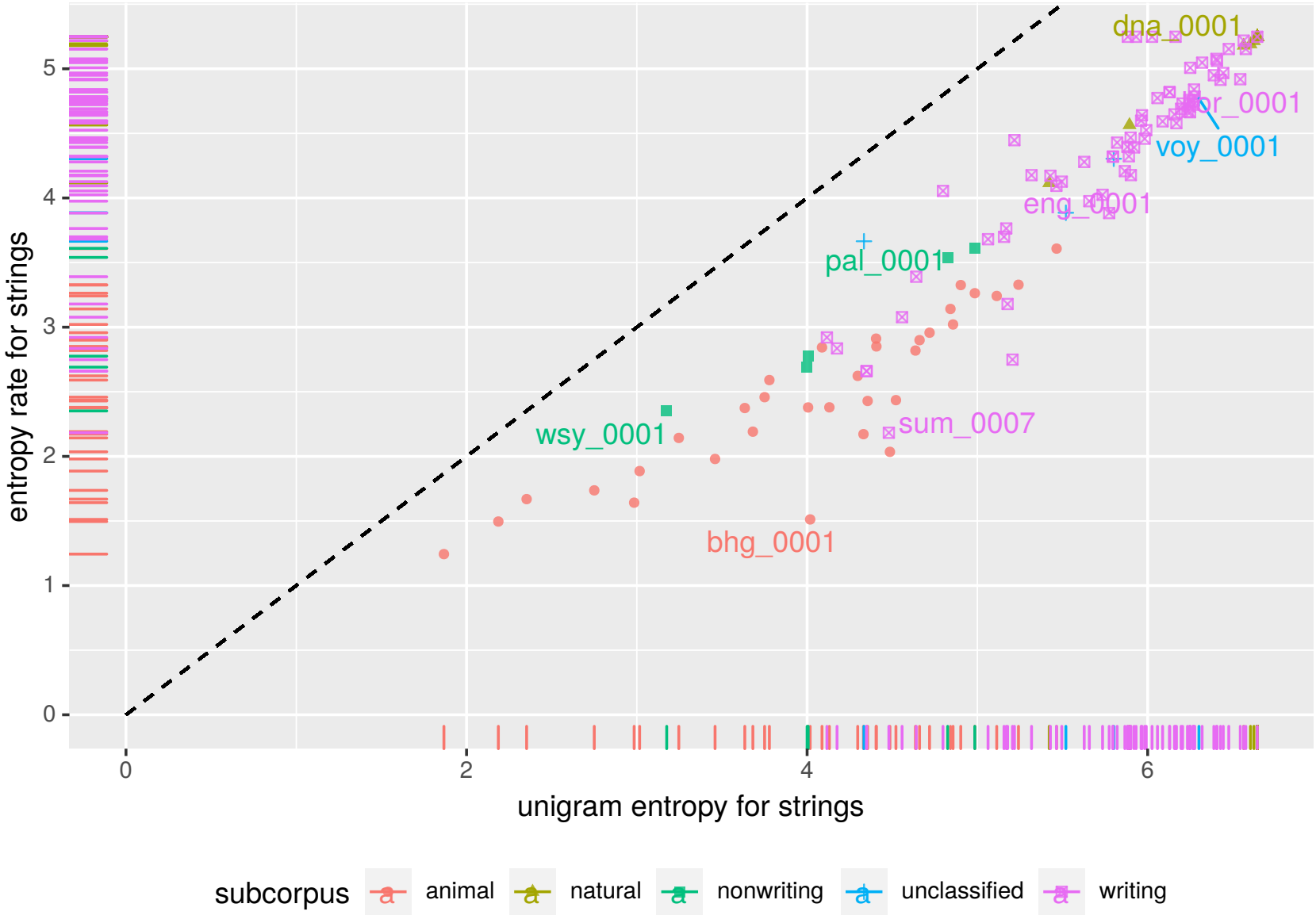
Results

Stabilization Analyses (100 Units)

Character Entropies (100 Units)



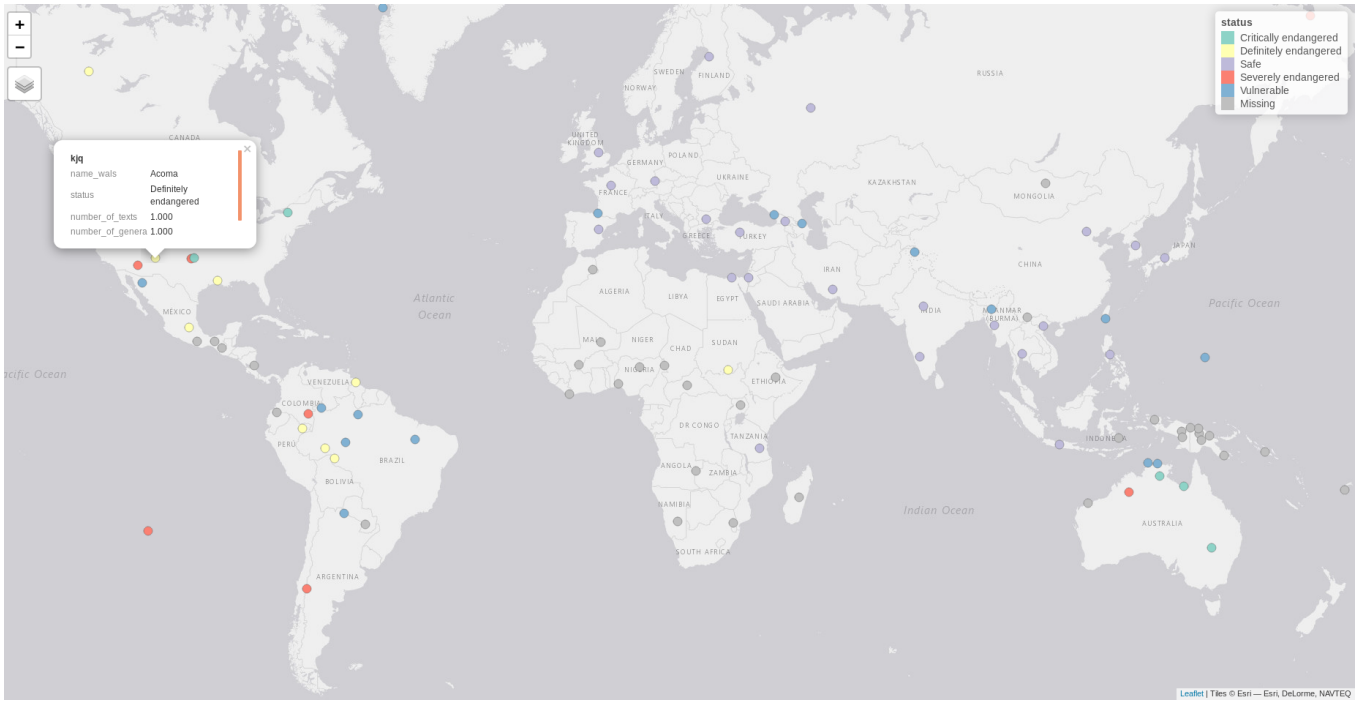
String Entropies (100 Units)

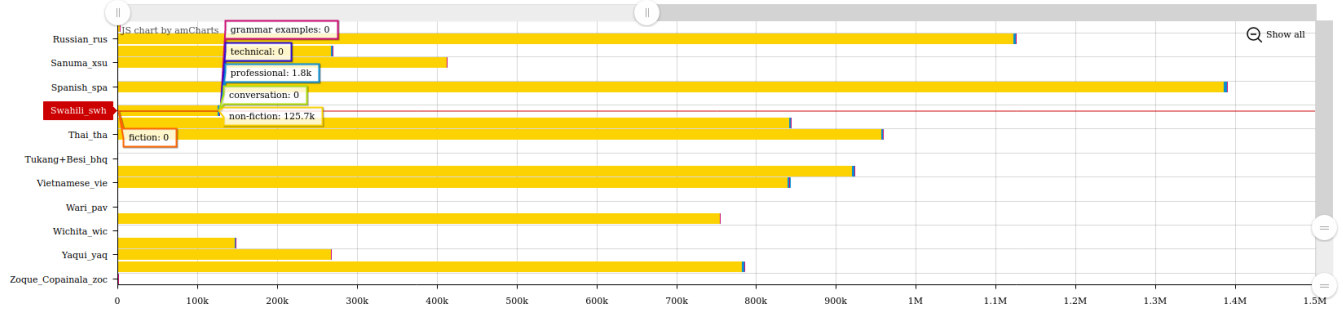


Future Directions

- Extending the database
- Use further measures (repetition rate, mutual information, Zipf parameters etc.)
- Create and test classification system

Extending the Database: The 100 Language Corpus

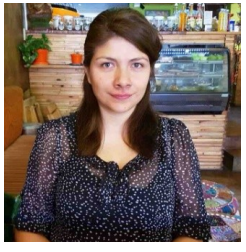




https://www.spur.uzh.ch/en/departments/research/textgroup/MorphDiv.html


Tanja Samardžić


Olga Sozinova


Ximena Gutierrez-Vasques

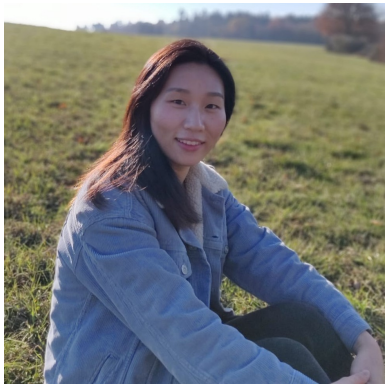

Steven Moran

Extending the Database: Palaeolithic Signs

Current

Ewa Dutkiewicz

Chris Bentz



Saetbyul Lee

Gabriele Russo

Former

Stephanie Samson

David Matzig

SignBase

SignBase

object number: gkl0025
object name: 58-264 Adorant
site name: Geissenklösterle
techno-complex: Aurignacien
layer: II
date BP min-max: 30625 ± 796 - 35700 ± 700
dating method: C14
excavation year: 1979
material: ivory
carrier: statuette
length_mm: 38
width_mm: 14
depth_mm: 5
preservation: almost complete
short description:
general_literature: Hahn J. 1977; Hahn J. 1986; Dutkiewicz E. 2017: Taf. 8a-8b

picture

picture_reference: Dutkiewicz E. 2017: Taf. 8a-8b

sign sequence + number: dot*13*_dot*10=_dot*12=_dot*13= | notch*13= | notch*13= | notch* 7= | notch *6= | radnotch*4= | obnotch*4=_obline*2= | obnotch*3=br

line	☐	cross	☐	vulva	☐
obline	☑	rhombus	☐	anthropomorph	☐
radline	☐	rectangle	☐		
dashline	☐	hashtag	☐	zoomorph	☐
circumline	☐	grid	☐	paw	☐
concenline	☐	hatching	☐	hoofprint	☐
notch	☑	v	☐		
obnotch	☑	zigzag	☐		
radnotch	☑	zigzagrow	☐		
dashnotch	☐	star	☐		
circumnotch	☐	maccaroni	☐		
crossnotch	☐				
dot	☑	pin	☐		
cupule	☐	pinleft	☐		
circle	☐	pinright	☐		
spiral	☐				
circumspiral	☐				

Dutkiewicz, et al. (2020) SignBase, a collection of geometric signs on mobile objects in the Paleolithic.

www.signbase.org

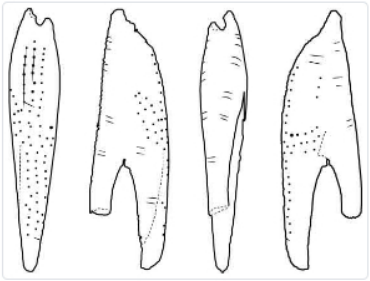
SIGN BASE

Home | Objects | References | Description | Contributors | Download | Imprint | Data Protection | Contact

Welcome to SignBase!

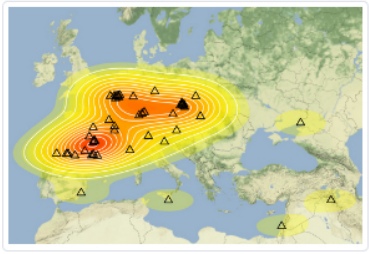
SignBase is an open access database for geometric motifs on mobile objects in Prehistory. Its focus lies on finds of the Eurasian Paleolithic and African Middle Stone Age. In these time periods, geometric motifs – also referred to as signs, patterns, or marks – are abundant in parietal art as well as on mobile objects. The term "geometric" denotes simple non-figurative forms such as dots, lines, and crosses, as well as more complex patterns. This includes frequent semi-abstract depictions such as vulvae, but excludes figurative depictions of animals, humans, etc. Decorated mobile objects are mostly made of osseous material, like ivory, bone or antler, while also featuring other organic and inorganic materials.

SIGN BASE



The relevant artifacts come from stratified archaeological contexts and are assigned to the particular techno-complexes. The objects are the core elements of the database, carrying a unique identifier. With this identifier comes information about geographic and archaeological provenience, the type of object and material, size and preservation, literature references, as well as a picture if available. The geometric motifs on each object are described in detail using a specifically developed encoding. The database aims to enable quantitative comparative studies on the development of graphical expressions before the emergence of writing systems.

SignBase gives latitudinal and longitudinal information for each object, and hence allows analyses of geographical distribution and geographic clustering of signs. Densities of sites with geometric signs, or densities of particular sign types give an overview of the abundance and distribution of signs. In the long run, this might be linked to population turn-overs, and respective cultural entities of the Paleolithic.



Thank You