Funded by
the European Union

European Research Council
Established by the European Commission

Zipfian Laws

Across Diverse Languages

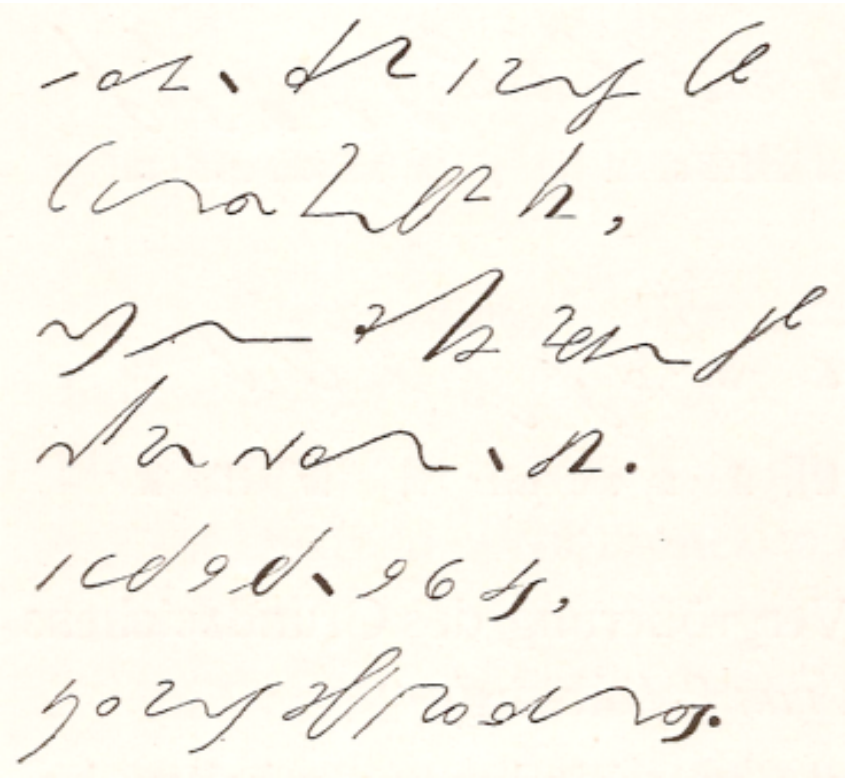
Christian Bentz

March 22, 2024

Department of General Linguistics, University of Tübingen

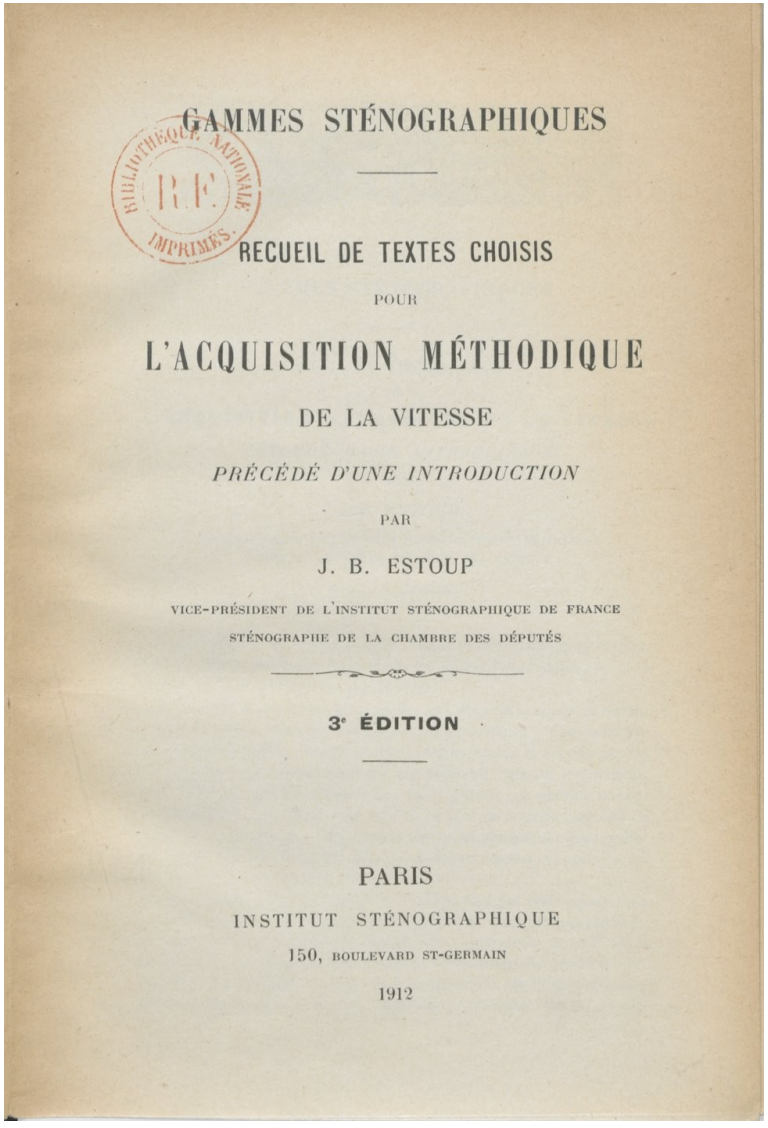
19th Century Information Technology

Stenography



Es reden und träumen die Menschen viel von besseren künftigen Tagen. Nach einem glücklichen goldenen Ziel sieht man sie rennen und jagen. Die Welt wird alt und wird wieder jung; doch der Mensch hofft immer Verbesserung.

Stenography



GAMMES STÉNOGRAPHIQUES

RECUEIL DE TEXTES CHOISIS

POUR

L'ACQUISITION MÉTHODIQUE

DE LA VITESSE

PRÉCÉDÉ D'UNE INTRODUCTION

PAR

J. B. ESTOUP

VICE-PRÉSIDENT DE L'INSTITUT STÉNOGRAPHIQUE DE FRANCE

STÉNOGRAPHE DE LA CHAMBRE DES DÉPUTÉS

3e ÉDITION

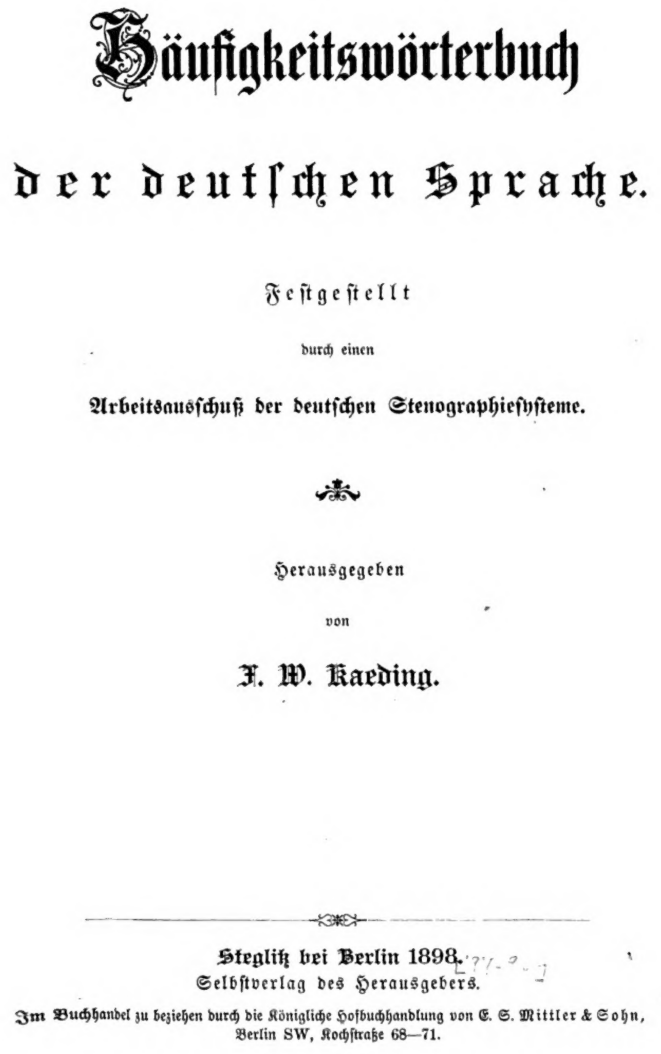
PARIS

INSTITUT STÉNOGRAPHIQUE

150, BOULEVARD ST-GERMAIN

1912

Estoup (1912). Gammes sténographiques.



Häufigkeitswörterbuch

der deutschen Sprache.

Festgestellt

durch einen

Arbeitsausschuß der deutschen Stenographiesysteme.

Herausgegeben

von

F. W. Kaeding.

Steglitz bei Berlin 1898.

Selbstverlag des Herausgebers.

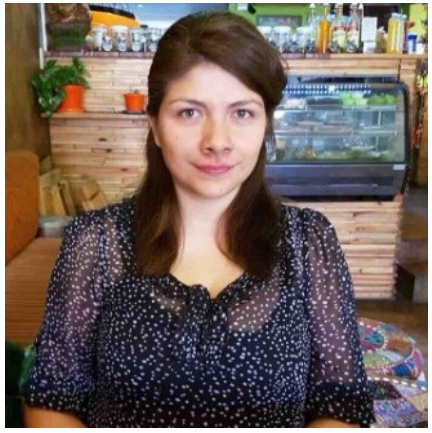
Im Buchhandel zu beziehen durch die Königliche Hofbuchhandlung von E. S. Mittler & Sohn, Berlin SW, Kochstraße 68—71.

Kaeding (1898). Häufigkeitswörterbuch der Deutschen Sprache.

Overview

Thank You

Tanja Samardžić



Ximena Gutierrez-Vasques

Olga Pelloni

Steven Moran

Ramon Ferrer-i-Cancho

Sonia Petrini

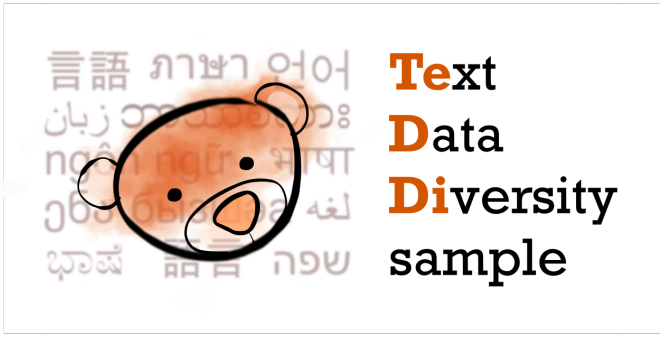
Julia Łukasiewicz-Pater

Tim Wientzek

Resources

UNICODE

CDLI Cuneiform Digital Library Initiative



opensubtitles.org

Introduction

Historical Background: Word Frequencies and Lengths

	Wörter	Prozent	Silben
1 silbig	5 426 326	49,76	5 426 326
2 =	3 156 448	28,94	6 312 896
3 =	1 410 494	12,93	4 231 482
4 =	646 971	5,93	2 587 884
5 =	187 738	1,72	938 690
6 =	54 436	0,50	326 616
7 =	16 993		118 951
8 =	5 038		40 304
9 =	1 225		11 025
10 =	461		4 610
11 =	59	0,22	649
12 =	35		420
13 =	8		104
14 =	2		28
15 =	1		15
	10 906 235*)		20 000 000

Kaeding (1898), p. 24.

Historical Background: Zipf's Original Formulations

Law of Abbreviation

"[...] the **magnitude of words** tends, on the whole, to stand in an **inverse** (not necessarily proportionate) relationship to the **number of occurrences**;"

Law of Word Frequencies (Zipf's Law)

"[...] the **number of different words** (i.e. variety) seems to be ever larger as the **frequency of occurrence** becomes ever smaller."[a]

Zipf (1935). The psycho-biology of language, p. 25.

[a] This is sometimes also referred to as number/frequency law, but it is equivalent to the rank/frequency law.

Methods

Tokenization

Language	Verse number 42024008 (New Testament)
English	And they remembered his words ,
Greek (Modern)	Και ενεθυμηθησαν τους λογους αυτου .
Hindi	तब उस की बातें उन को स्मरण आईं ।
Georgian	და მოეგონნეს სიტყუანი მისნი .
Korean	저희가 예수의 말씀을 기억하고
Burmese	မိန့်တော်မူခဲ့သောစကားများကိုပြန်သတိရ ၍ ၊ -
Russian	И они вспомнили эти слова Его .
Inuktitut	ᐊᖏᓂᓪ ᐃᖅᑭᓕᖅᐳᑦ ᐅᖃᐅᔨᖏᓐᓄᑦ ,
Kalaallisut	Taava oqaasii eqqaalerpaat .

Source: Parallel Bible Corpus (PBC)

Gutierrez-Vasques, Bentz, & Samardžić (2023). Languages through the looking glass of BPE compression.

Tokenization by non-alpha-numeric characters

Language	Verse number 42024008 (New Testament)
English	And they remembered his words ,
Greek (Modern)	Και ενεθυμηθησαν τους λογους αυτου .
Hindi	तब उस की बातें उन को स्मरण आईं ।
Georgian	და მოეგონნეს სიტყუანი მისნი .
Korean	저희가 예수의 말씀을 기억하고
Burmese	မိန့်တော်မူခဲ့သောစကားများကိုပြန်သတိရ ၍ –
Russian	И они вспомнили эти слова Его .
Inuktitut	ᐊᖏᓂᔪᓄ ᐃᖅᑭᓕᖅᐳᑦ ᐅᖃᐅᓯᖏᓐᓂᒃ ,
Kalaallisut	Taava oqaasii eqqaalerpaat .

Source: Parallel Bible Corpus (PBC)

Gutierrez-Vasques, Bentz, & Samardžić (2023). Languages through the looking glass of BPE compression.

Tokenization by language-specific algorithm

Burmese (mya)

(1) [...] မိန့်တော်မူခဲ့သောစကားများကိုပြန်သတိရ ၍ ၊ –
[...] မိန့် | တော် | မူ | ခဲ့ | သော | စ | ကား | များ | ကို | ပြန် | သ | တိ | ရ | ၍

[...] mín-tɔ-mu-gé θɔ sá.kà-mya-ko pyan θá.ti.rá ywé
[...] speak-HON-perform-DISPL REL word-PL-OBJ return remember and

“[...] and (they) remember the words that (he) spoke.”

Gutierrez-Vasques, Bentz, & Samardžić (2023). Languages through the looking glass of BPE compression.

Tokenization provided by resource

Mandarin Chinese (cmn)

(2) 學生通過實際運用來學習科學課程的內容。

學生	通過	實際	運用	來	學習	科學	課程
xuéshēng	tōngguò	shíjì	yùnyòng	lái	xuéxí	kēxué	kèchéng
student	through	practice	use.INF	PRT	learn.INF	science	course

的	內容	。
de	nèiróng	.
POSS	content	.

‘Students learn science content by applying it.’

Petrini et al. (in prep). The optimality of word lengths.

Frequencies (Probabilities)

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

Source: Universal Declaration of Human Rights (UDHR)

Frequencies (Probabilities)

Corpus (Word Tokens):

$$C = (t_1, t_2, \ldots, t_k)$$

Vocabulary (Word Types):

$$\mathcal{V} = \{w_1, w_2, \ldots, w_n\},\ n = |\mathcal{V}|$$

Raw Frequency:

$$f(w_i) = \sum_{j=1}^{k} [t_j = w_i]$$

Relative Frequency (ML):

$$\hat{p}(w_i) = \frac{f(w_i)}{\sum_{i=j}^{n} f(w_j)}$$

Surprisal in language model:

$$\hat{s}(w_i) = -\frac{1}{f(w_i)} \sum_{j=1}^{f(w_i)} \hat{p}(w_i|c_j)$$

Piantadosi et al. (2011). Word lengths are optimized for efficient communication.
Levshina (2022). Frequency, informativity and word length.
Pimentel et al. (2023). Revisiting the optimality of word lengths.

Magnitudes (Written Language)

Example	Unit	Length
English: limitation		
limit-ation	morpheme	2
li-mi-ta-tion	syllable	4
l-ɪ-m-ɪ-t-e-ɪ-ʃ-ə-n	phoneme	10
l-i-m-i-t-a-t-i-o-n	character	10
1-2-3-2-2-2-2-2-1-2	stroke	19
Chinese (simplified): 限度		
限-度	morpheme	2
限-度	syllable	2
ɕ-i-à-n-t-ù	phoneme	6
限-度	character	2
8-9	stroke	17
Chinese (Pinyin): xiàndù		
xiàn-dù	morpheme	2
xiàn-dù	syllable	2
ɕ-i-à-n-t-ù	phoneme	6
x-i-à-n-d-ù (x-i-a-`-n-d-u-`)	character	6 (8)
2-2-3-2-2-3 (2-2-2-1-2-2-2-1)	stroke	14 (14)

Magnitudes (Spoken Language)

Language	Word	No. char.	IPA	Duration
French	est	3	ɛ	0.14
	une	3	ynə	0.17
	sont	4	sõ	0.23
	comme	5	kɔmə	0.22
	informations	12	ɛ̃fɔrmasjõ	0.67
	Average	**5.4**		**0.29**
Spanish	es	2	es	0.19
	una	3	una	0.21
	son	3	son	0.25
	como	4	komo	0.27
	informaciones	13	informaθjones	0.78
	Average	**5**		**0.34**

Petrini, Casas-i-Muñoz, Cluet-i-Martinell, Wang, Bentz & Ferrer-i-Cancho (2023). Direct and indirect evidence of compression of word lengths.
Petrini et al. (in prep). The optimality of word lengths.

Mathematical Formulation: The Law of Abbreviation

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

Word (w_i)	r_i	f_i	l_i
and	1	4	3
are	2	2	3
in	3	2	2
a	4	1	1
...	...	...	...
spirit	22	1	6
they	23	1	4
towards	24	1	7
with	25	1	4

Definitions

Pearson's r:

$$r_{pl} = \frac{\sum_{i=1}^{n}(p_i - \overline{p})(l_i - \overline{l})}{\sqrt{\sum_{i=1}^{n}(p_i - \overline{p})^2}\sqrt{\sum_{i=1}^{n}(l_i - \overline{l})^2}}.$$

Kendall's τ:

$$\tau(p_i, l_i) = \frac{n_c - n_d}{\binom{n}{2}}.$$

n_c: number of concordant pairs,
n_d: number of discordant pairs.

Petrini et al. (2023). Direct and indirect evidence of compression of word lengths.

Mathematical Formulation: The Law of Word Frequencies

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

Word (w_i)	r_i	f_i	l_i
and	1	4	3
are	2	2	3
in	3	2	2
a	4	1	1
...	...	...	...
spirit	22	1	6
they	23	1	4
towards	24	1	7
with	25	1	4

Definitions

Zipf's Law:

$$r \times f = C,$$

$$f = \frac{C}{r} = Cr^{-1},$$

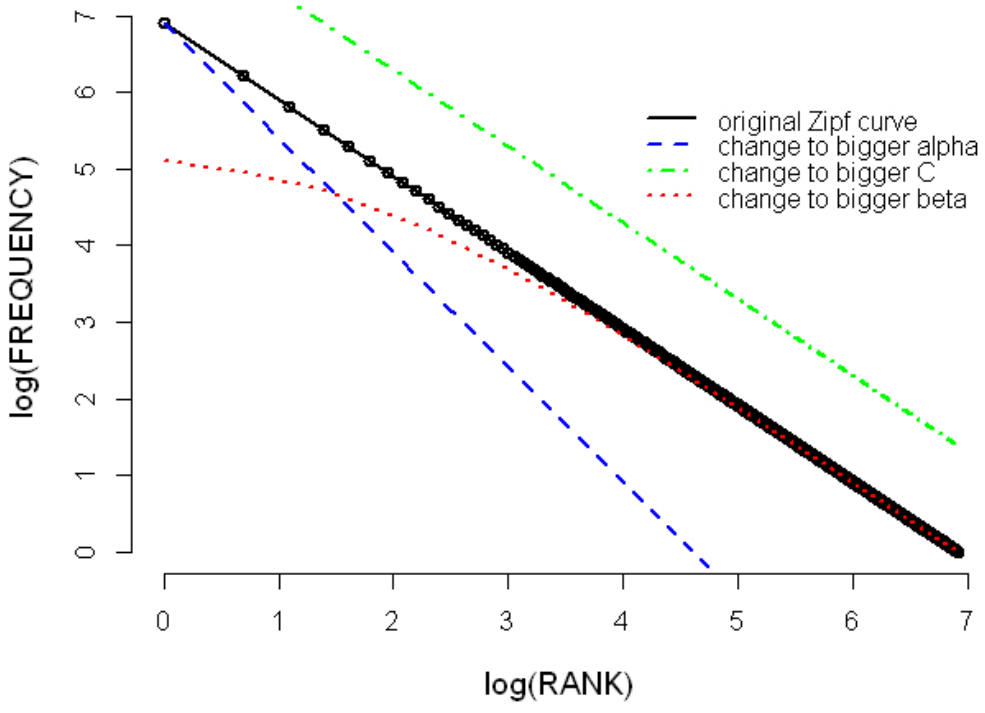
$$f \propto r^{-\alpha}.$$

Zipf-Mandelbrot Law:

$$f(r) \propto (r + \beta)^{-\alpha}.$$

Zipf (1949). The principle of least effort, p. 24.
Mandelbrot (1965). Information theory and psycholinguistics, p. 556.

Mathematical Formulation: The Law of Abbreviation



Bentz, Kiela, Hill, & Buttery (2014). Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts.

Definitions

Zipf's Law:

$$r \times f = C,$$

$$f = \frac{C}{r} = Cr^{-1},$$

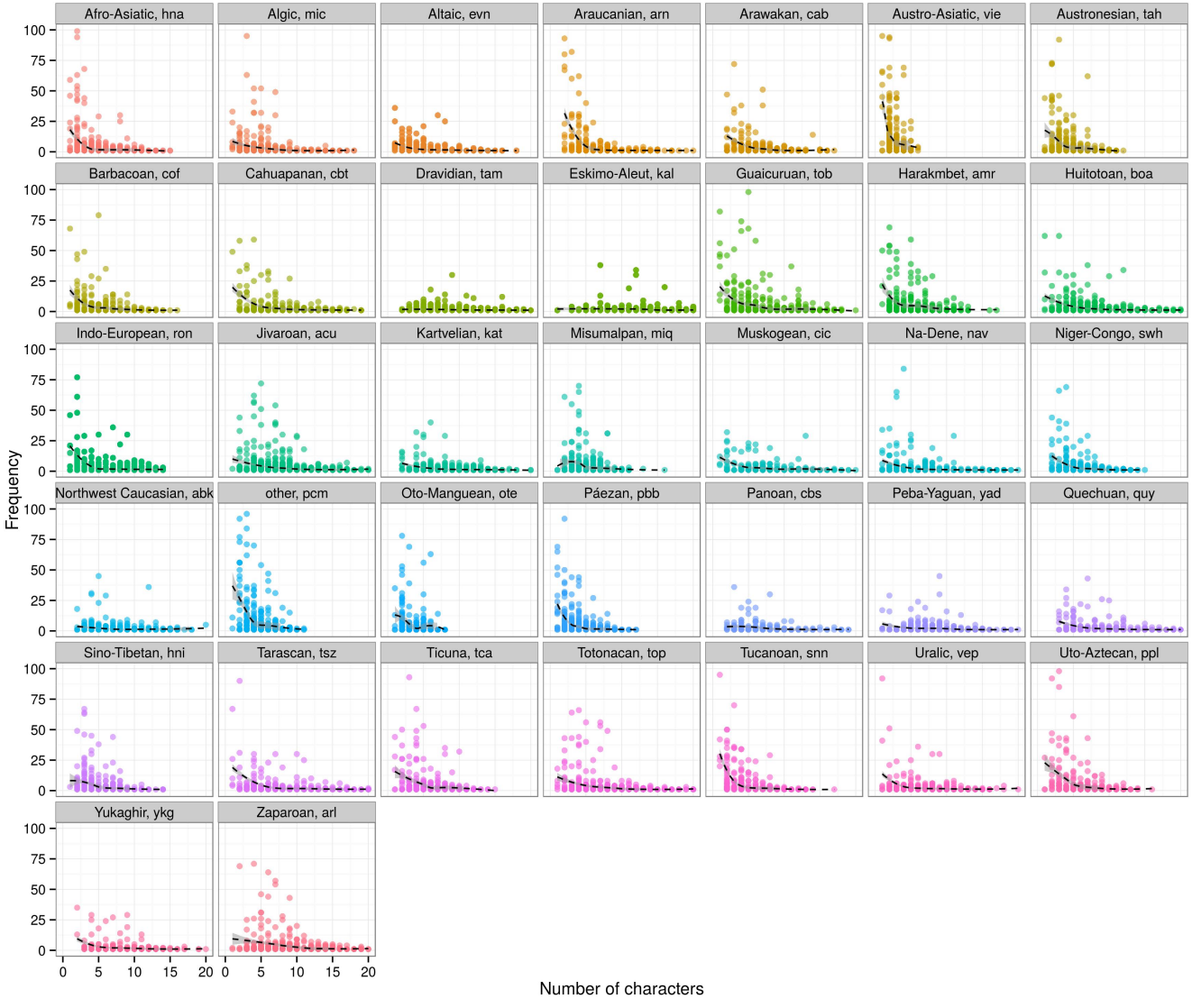
$$f \propto r^{-\alpha}.$$

Zipf-Mandelbrot Law:

$$f(r) \propto (r + \beta)^{-\alpha}.$$

Results

The Law of Abbreviation as a Language Universal



Bentz & Ferrer-i-Cancho (2016). Zipf's law of abbreviation as a language universal.

The Law of Abbreviation as a Language Universal

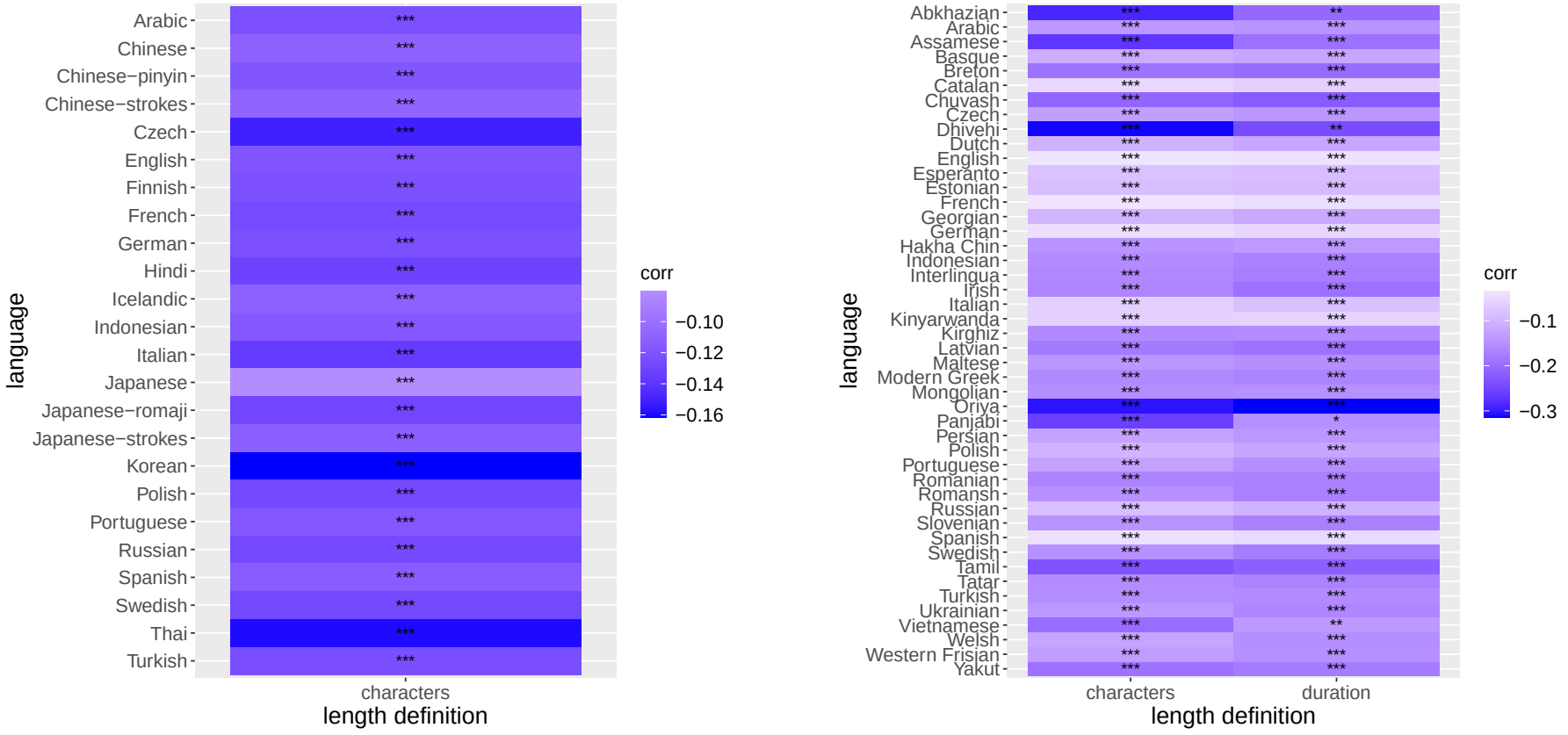
TABLE I

THE CONCORDANCE WITH ZIPF'S LAW OF ABBREVIATION ACROSS 986 LANGUAGES. FOR EACH DATASET, N IS THE NUMBER OF TEXTS OR LANGUAGES, N_α^- IS THE NUMBER OF NEGATIVE CORRELATIONS BETWEEN WORD FREQUENCY AND WORD LENGTH WITH P-VALUES NOT EXCEEDING α; N_α^+ IS THE CONVERSE OF N_α^- FOR POSITIVE CORRELATIONS.

	Texts		Languages	
	PBC	UDHR	PBC	UDHR
N	907	355	801	332
N_1^-	907	355	801	332
N_1^+	0	0	0	0
$N_{0.05}^-$	907	328	801	307
$N_{0.01}^-$	907	316	801	296
$N_{0.001}^-$	907	283	801	265
$N_{0.0001}^-$	907	245	801	230

Bentz & Ferrer-i-Cancho (2016). Zipf's law of abbreviation as a language universal.

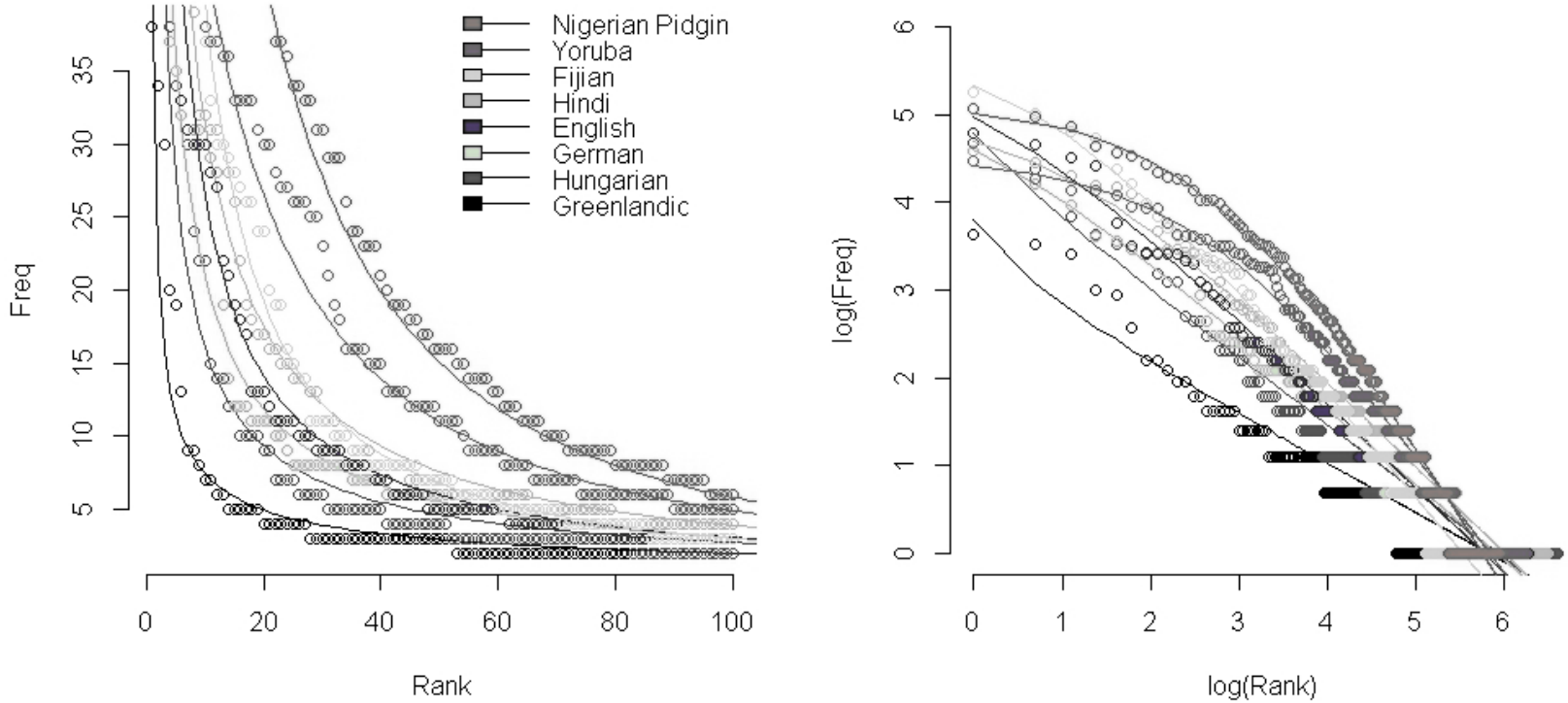
The Law of Abbreviation as a Language Universal



Left panel: Pearson correlations for PUD (Parallel Universal Depencencies).
Right panel: Pearson correlations for Common Voices (Spoken corpus).
$p < 0.01$ ***, $p < 0.05$ **, $p < 0.1$ *

Petrini et al. (2023). Direct and indirect evidence of compression of word lengths.

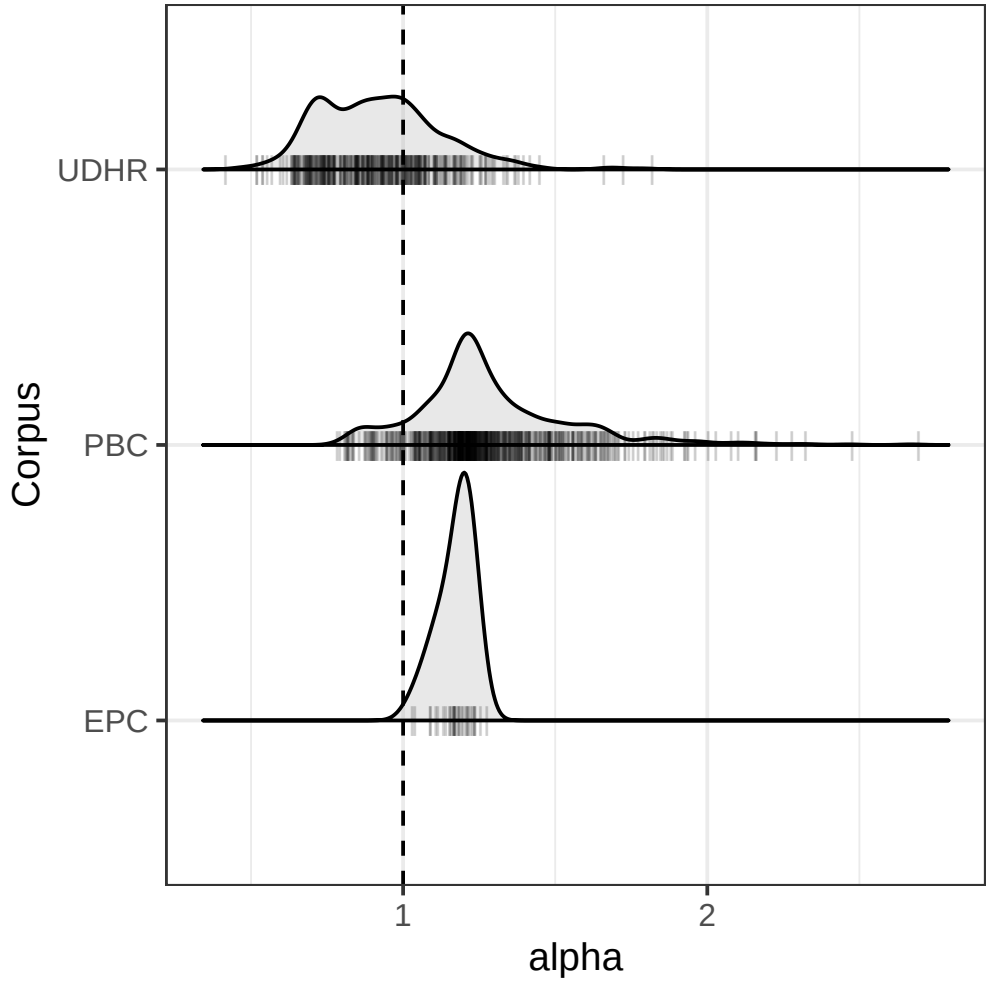
The Law of Word Frequencies as a Language Universal (?)



Source: Universal Declaration of Human Rights (UDHR).

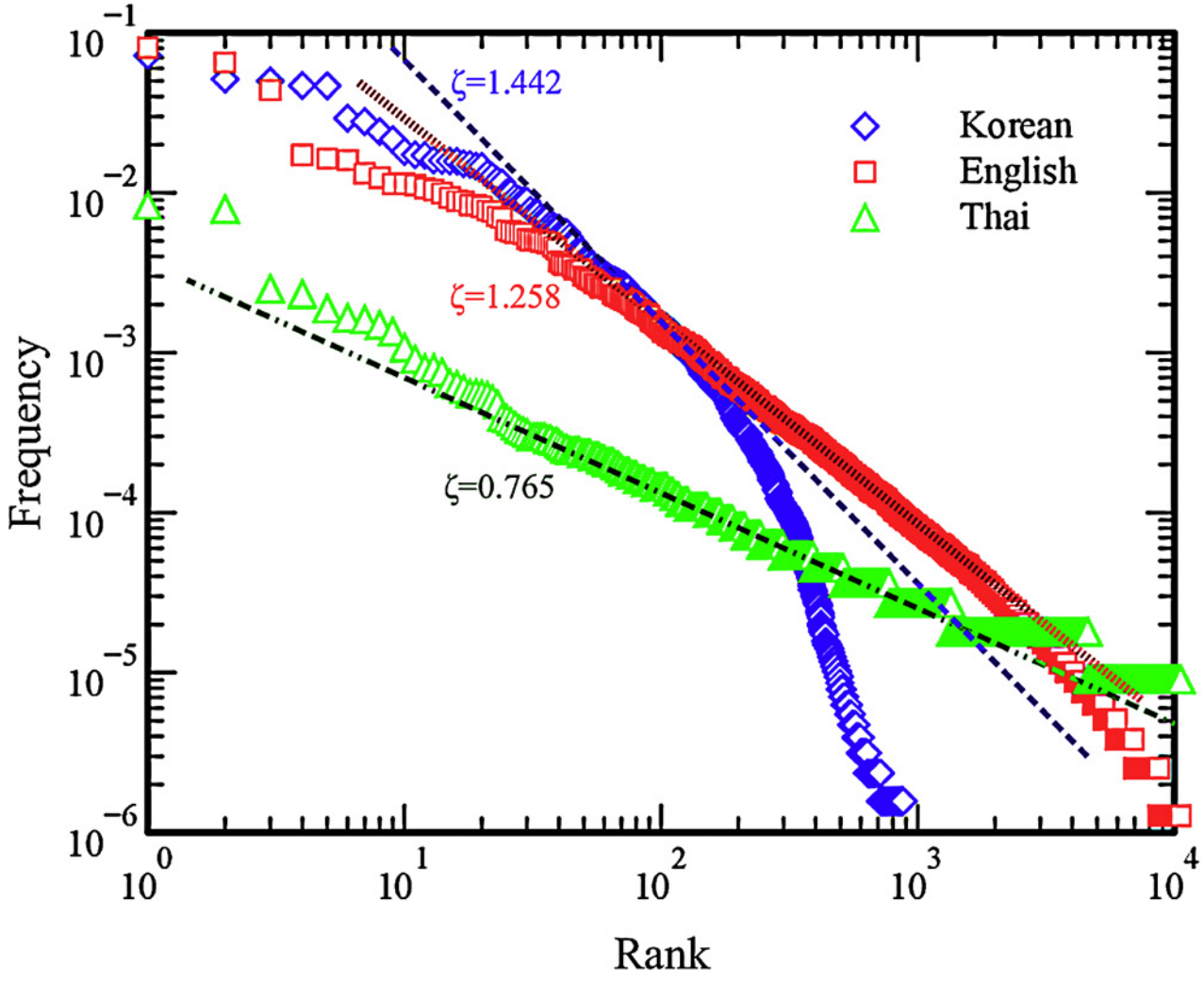
Bentz & Kiela (2014). Zipf law across languages of the world.

The Law of Word Frequencies as a Language Universal (?)



Reanalysis of: Bentz, Verkerk, Kiela, Hill, & Buttery (2015). Adaptive communication: Languages with more non-native speakers tend to have fewer word forms.

The Law of Word Frequencies as a Language Universal (?)



Mehri & Jamaati (2017). Variation of Zipf's exponent in one hundred live languages: A study of the Holy Bible translations.

Problem: Statistical Tests

Book	Rank: $f(r)$		Frequency: $P(f)$		Linear: $\log f(\log r)$	
	$\hat{\alpha}_Z$	p-value	$\hat{\alpha}_Z$	p-value	$\hat{\alpha}_Z$	R^2
Alice's Adventures in Wonderland (L. Carroll)	1.22	$< 10^{-4}$	1.46	$< 10^{-4}$	1.21	0.97
The Voyage Of The Beagle (C. Darwin)	1.20	$< 10^{-4}$	1.59	$< 10^{-4}$	1.29	0.97
The Jungle (U. Sinclair)	1.21	$< 10^{-4}$	1.45	$< 10^{-4}$	1.22	0.98
Life On The Mississippi (M. Twain)	1.20	$< 10^{-4}$	1.38	$< 10^{-4}$	1.16	0.98
Moby Dick; or The Whale (H. Melville)	1.19	$< 10^{-4}$	1.38	$< 10^{-4}$	1.15	0.98
Pride and Prejudice (J. Austen)	1.21	$< 10^{-4}$	1.66	$< 10^{-4}$	1.35	0.98
Don Quixote (M. Cervantes)	1.21	$< 10^{-4}$	1.70	$< 10^{-4}$	1.38	0.98
The Adventures of Tom Sawyer (M. Twain)	1.21	$< 10^{-4}$	1.29	$< 10^{-4}$	1.12	0.98
Ulysses (J. Joyce)	1.18	$< 10^{-4}$	1.15	$< 10^{-4}$	1.03	0.97
War and Peace (L. Tolstoy)	1.20	$< 10^{-4}$	1.84	$< 10^{-4}$	1.44	0.97
English Wikipedia	1.17	$< 10^{-4}$	1.60	$< 10^{-4}$	1.58	0.99

Note: The null hypothesis for the statistical test is that the distribution follows the Zipfian proposal, i.e. $f(r) \propto r^{-1}$.

Altmann & Gerlach (2016). Statistical laws in linguistics.

Discussion

Random Typing...

Miller (1957). Some effects of intermittent silence.

… is not a valid baseline

Cognitive Plausibility: Humans do *not* produce characters, syllables, words (etc.) by drawing randomly from an alphabet with uniform or non-uniform probabilities (neither do chimpanzees).

Piantadosi (2014). Zipf's word frequency law in natural language.

... is not a valid baseline

Learning: There is an increasing number of studies which illustrate the link of Zipfian distributions to human learning and cultural evolution.

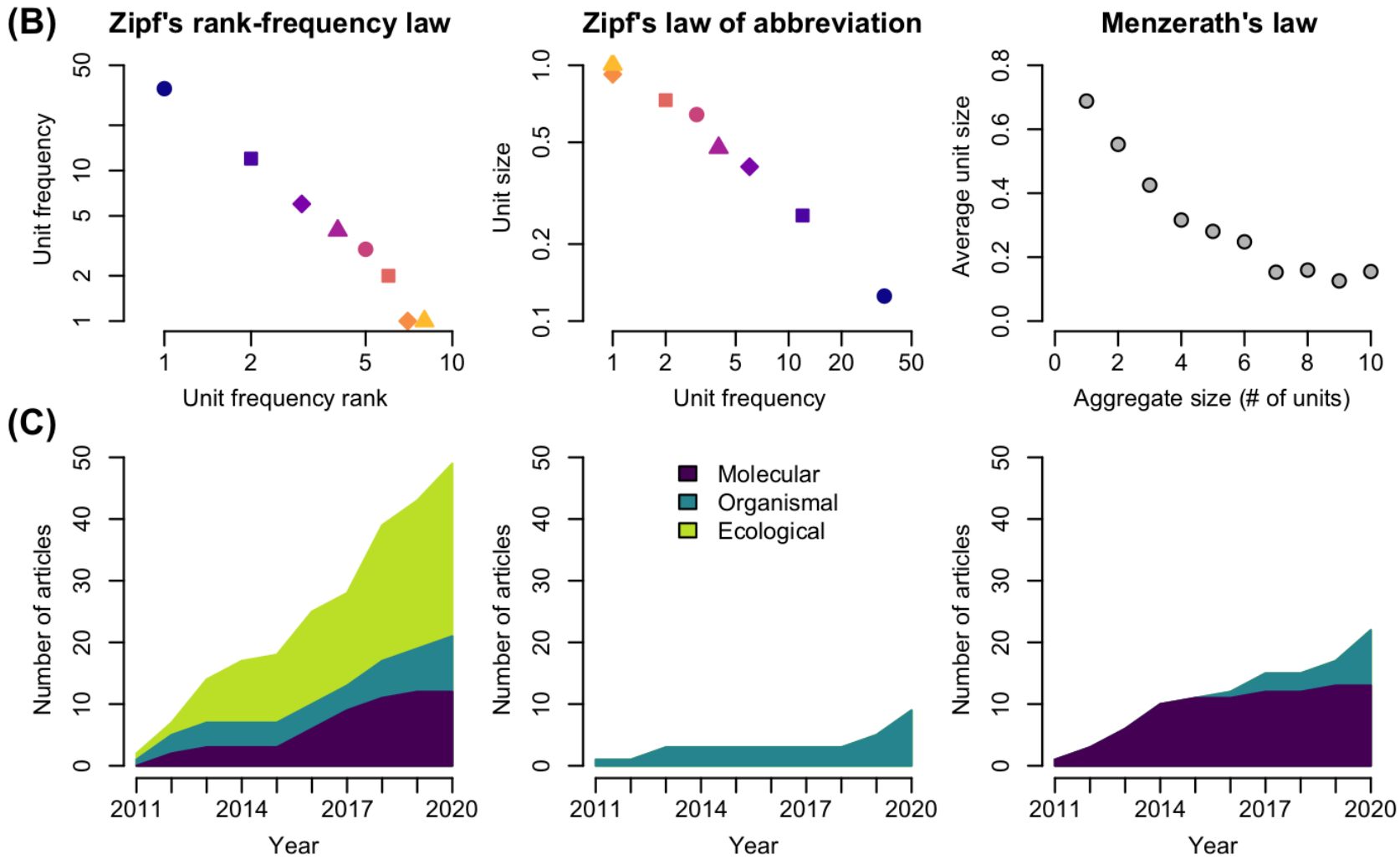
Kanwal et al. (2017). Zipf's law of abbreviation and the principle of least effort.
Lavi-Rotbain & Arnon (2022). The learnability of Zipfian distributions in language.
Lavi-Rotbain & Arnon (2023). Zipfian distributions in child-directed speech.
Arnon & Kirby (2024). Cultural evolution creates the statistical structure of language.

… is not a valid baseline

Optimality: Counter intuition, randomly generated strings are actually information-theoretically *optimal* for encoding a set of meanings.

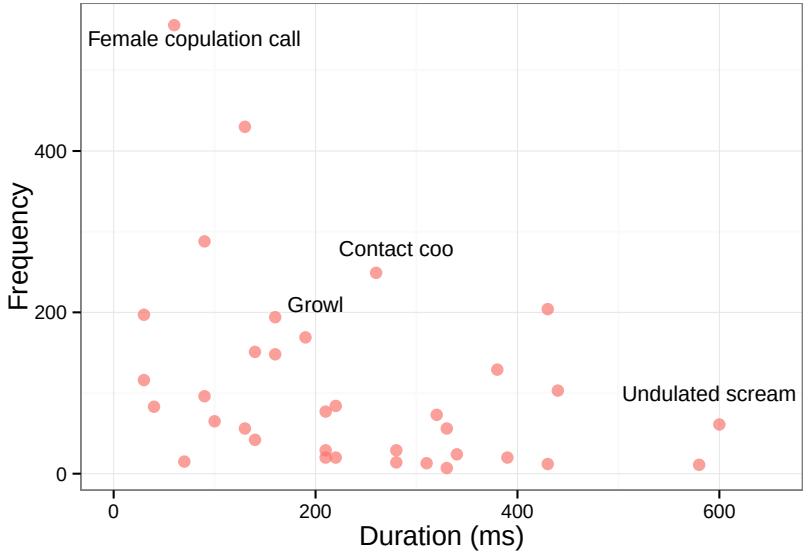
Ferrer-i-Cancho, Bentz, & Seguin (2022). Optimal coding and the origins of Zipfian laws.

Zipfian Laws in Animal Communication



Semple, Ferrer-i-Cancho, & Gustison (2022). Linguistic laws in biology.

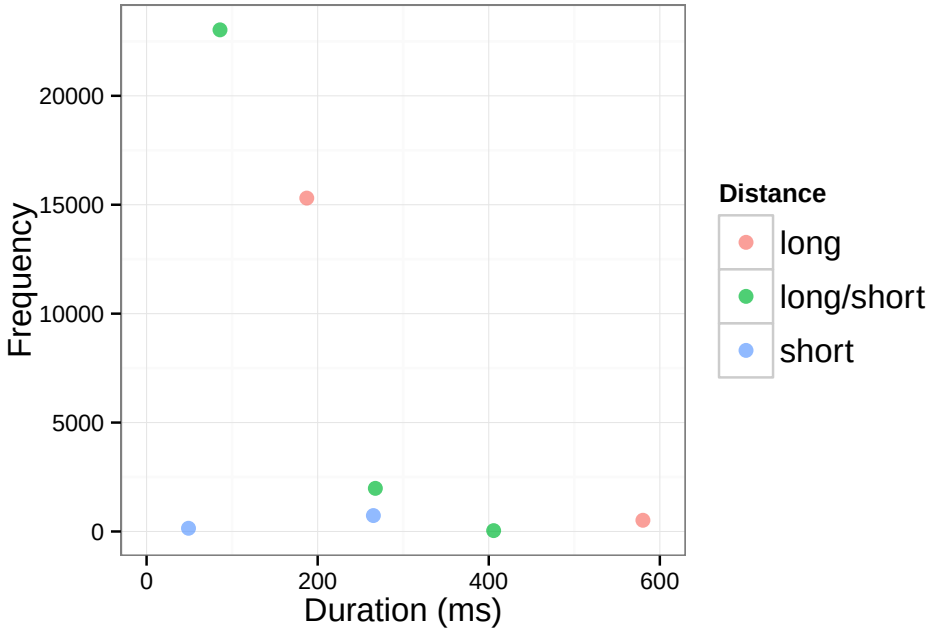
Law of Abbreviation: Formosan Macaques



- Repertoire Size: 35
- r = -0.42
- p = 0.011

Semple, Hsu, & Agoramoorthy (2010). Efficiency of coding in macaque vocal communication.

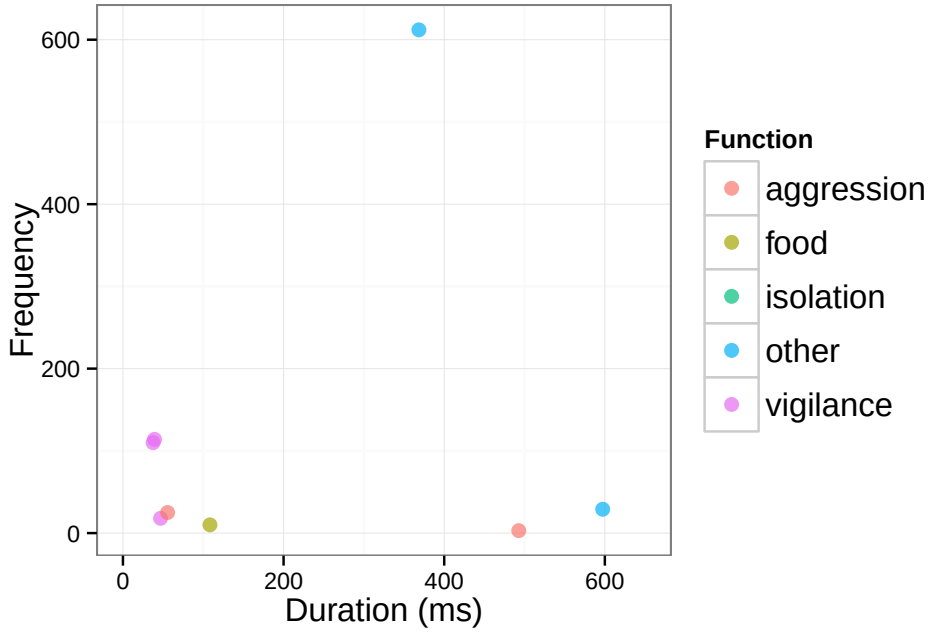
Law of Abbreviation: Golden-Backed Uakaris



- Repertoire Size: 7
- r_s = -0.36
- p = 0.43

Bezzera et al. (2011). Brevity is not always a virtue in primate communication.

Law of Abbreviation: Common Marmosets



- Repertoire Size: 12
- $r_s = 0.056$
- p = 0.86

Bezzera et al. (2011). Brevity is not always a virtue in primate communication.

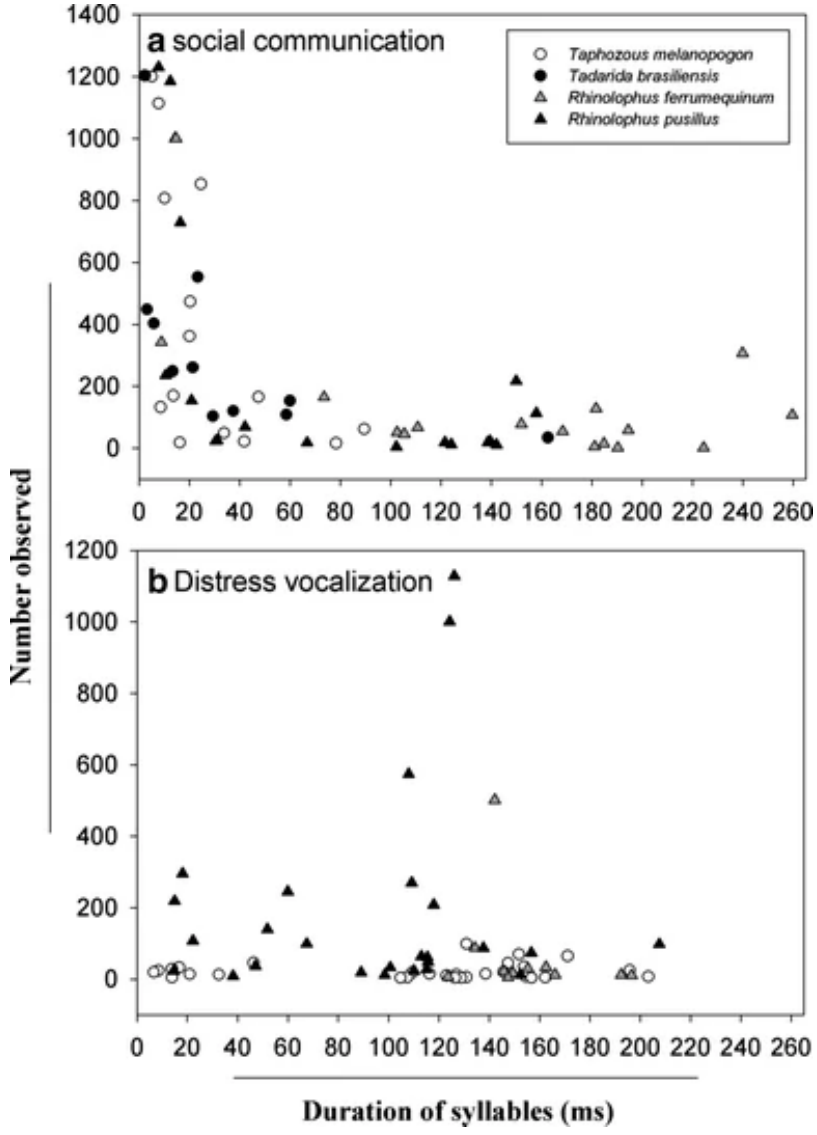
Law of Abbreviation: Four Species of Bats

Social Communication

- Repertoire Size: 14, 10, 17, 17
- r_p = -0.55, -0.63, -0.66, -0.55
- p = 0.04, 0.05, 0.004, 0.02

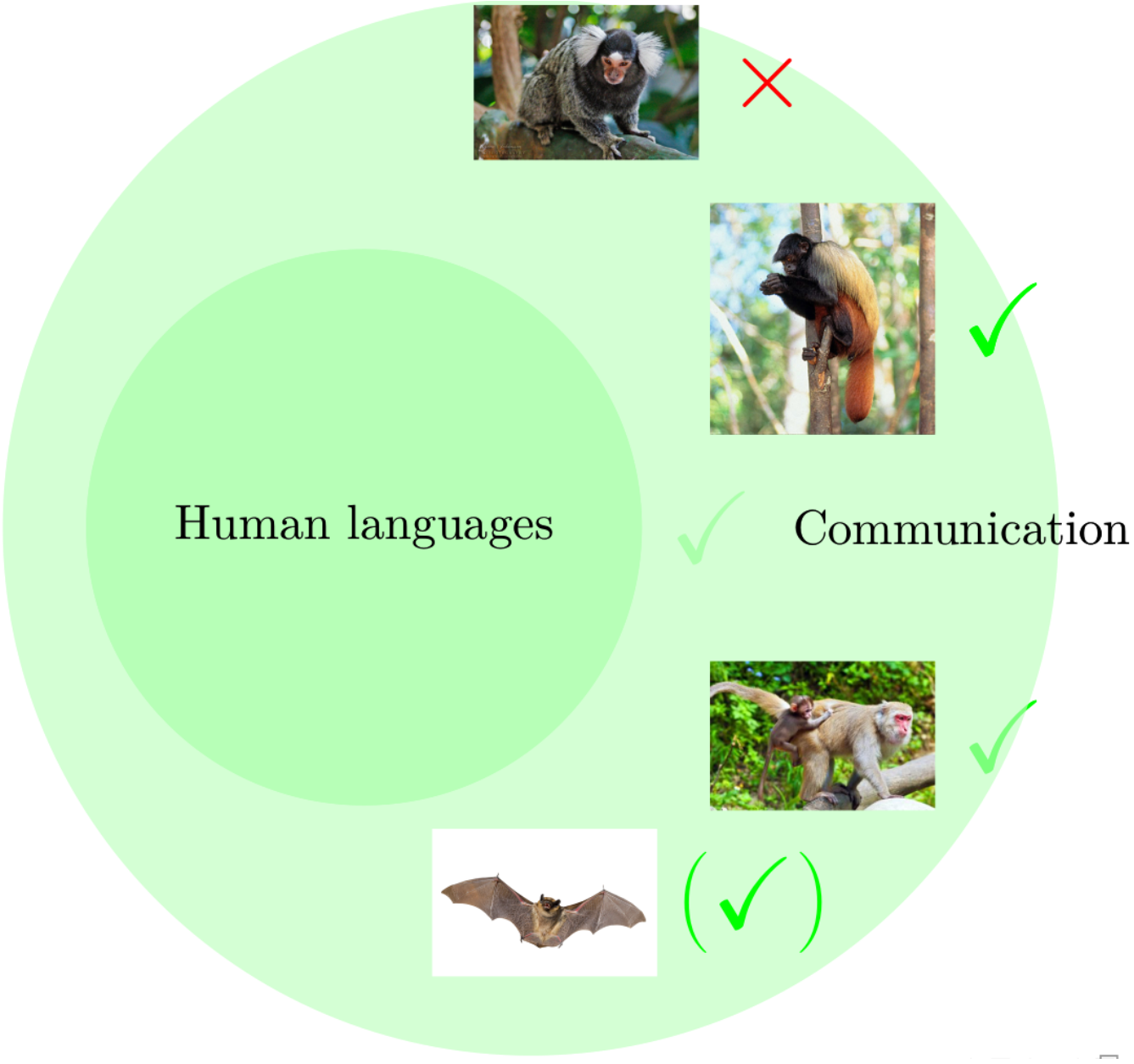
Distress Vocalization

- Repertoire Size: 29, 11, 26
- r_p = 0.06, -0.25, 0.12
- p = 0.75, 0.46, 0.56



Luo et al. (2013). Brevity is prevalent in bat short-range communication.

The Law of Abbreviation in Animal Communication



Statistical Fingerprint?

Can we robustly classify (short) character strings into **writing** and **non-writing**?

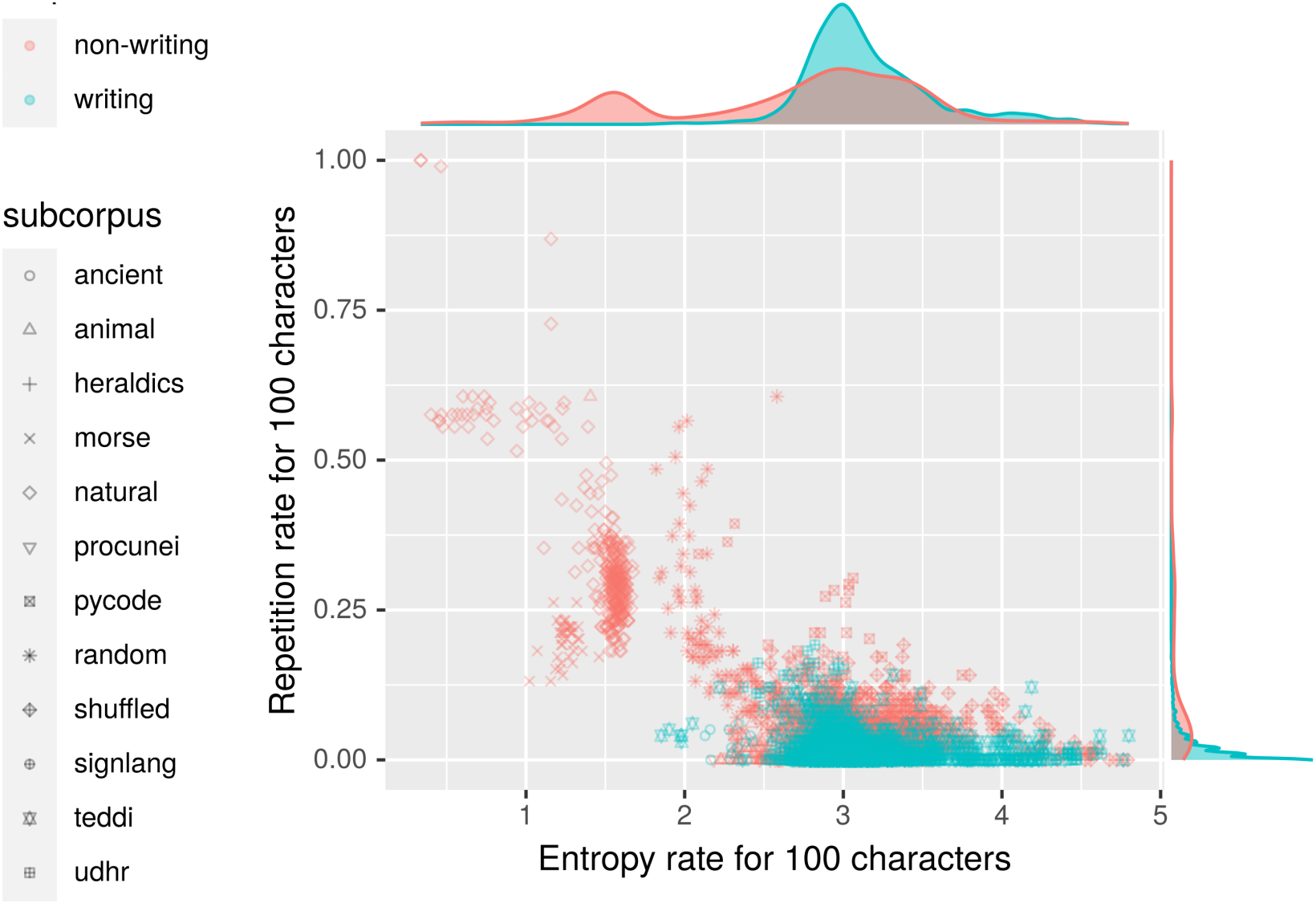
šum-ma a-wi-lum ✓	uj kd ro su sv sw ×
序言 鉴于对人类家庭所✓	.– .... . -. .. -. ×
Preamble Whereas ✓	AAAAGGTAGTTA ×
ПРЕАМБУЛА Принимая ✓	N19 N19 SZE∼a LU2 ×
AALLAQQAASIUTA ✓	pass p = Person() ×
전 문 모든 인류 구✓	hihhe bh fif cd ×
前文 人類社会のすべて✓	swr a j e eitimii ×
Isandulelo Ngokunjalo ✓	SWCCSSSSSSSSS ×

Bentz (2023). The Zipfian Challenge: Learning the statistical fingerprint of languages.
Sproat (2023). Symbols: An evolutionary history from the Stone Age to the future.
Sproat (2014). A statistical comparison of written language and nonlinguistic symbol systems.

Statistical Features

- Type-Token-Ratio (TTR)
- Unigram Entropy (huni)
- Entropy Rate (hrate)
- Repetition Rate (R)

Estimations (100 characters)



Classification Methods

- **K-Nearest-Neighbours** (KNN)
  - k: number of neigbours
- **Logistic Regression** (LR)
- **Support Vector Machines** (SVM)
  - kernel: linear, radial, sigmoid, polynomial
- **Multi-Layer Perceptron** (MLP)
  - hidden depth: no. of hidden layers
  - hidden size: overall no. of hidden units
  - error function: CE, SSE
  - activation function: sigmoid, logistic, tanh, relu
  - etc.

Results

Classifier	Chars.	Hyperparam.	Acc.	F1
Baseline	10	k = 1	0.69	0.63
KNN	10	k = 6	0.71	**0.73**
	100	k = 5	0.92	0.92
	1000	k = 7	**0.98**	0.95
LogRegr.	10	–	0.72	0.72
	100	–	0.79	0.77
	1000	–	0.93	0.84
SVM	10	kernel: linear	0.72	0.70
	100	kernel: radial	0.88	0.89
	1000	kernel: radial	0.92	0.82
MLP	10	hidden: 5, 4; tanh; SSE; rprop+	**0.73**	**0.73**
	100	hidden: 4, 4; tanh; SSE; rprop+	**0.93**	**0.93**
	1000	hidden: 4, 5, 2; tanh; SSE; rprop+	**0.98**	**0.96**

Bentz (2023). The Zipfian Challenge: Learning the statistical fingerprint of natural languages.

Conclusions

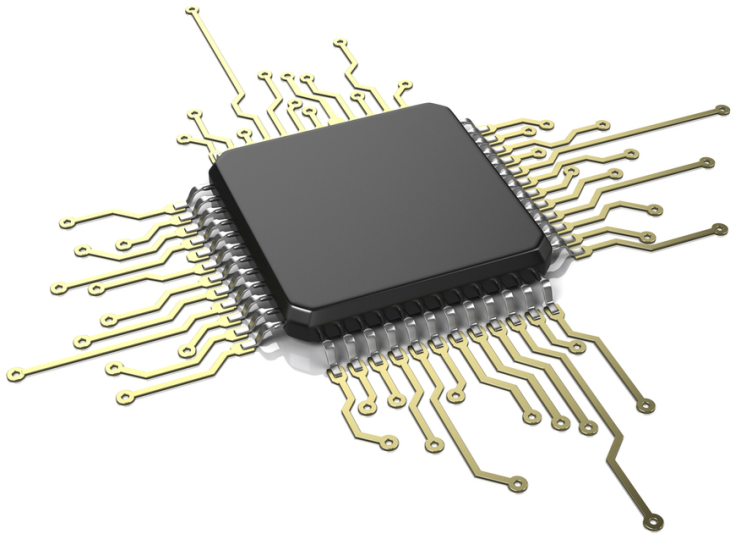
Conclusion: Universality in Human Languages

- Zipf's law of abbreviation ✓
- Zipf's law of word frequencies (✓)
- However: *Universals* are not necessarily **statistical fingerprints**

Conclusion: Explanations of Zipfian Laws

- Random typing ×
- Learning and cultural evolution ✓
- Context of communication ✓

21st Century Information Technology



Es reden und träumen die Menschen viel von besseren künftigen Tagen. Nach einem glücklichen goldenen Ziel sieht man sie rennen und jagen. Die Welt wird alt und wird wieder jung; doch der Mensch hofft immer Verbesserung.

01010100 01101000 01100001
01101110 01101011 00100000
01111001 01101111 01110101

Thank You