EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Language Evolution WiSe 2023/2024

Lecture 9: Quantitative Linguistics

21/11/2023, Christian Bentz

Overview

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Introduction

What is unique about human language?

"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 (1965). The psycho-biology of language, p. 187.

Competing Definitions of *Language*

- **Formal Language Theory**
- **Faculty of Language**
  - Recursion
  - Rich Language Faculty (Narrow Sense)
- **Minimalism**
  - Strong Minimalist Thesis
  - Minimalist Layers Hypothesis
- **Usage-Based Grammar**

Definition (Usage-Based)

"While all linguists are likely to agree that *grammar is the cognitive organization of language*, a **usage-based theorist** would make the more specific proposal that grammar is the cognitive organization of one's **experience with language**."

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

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.

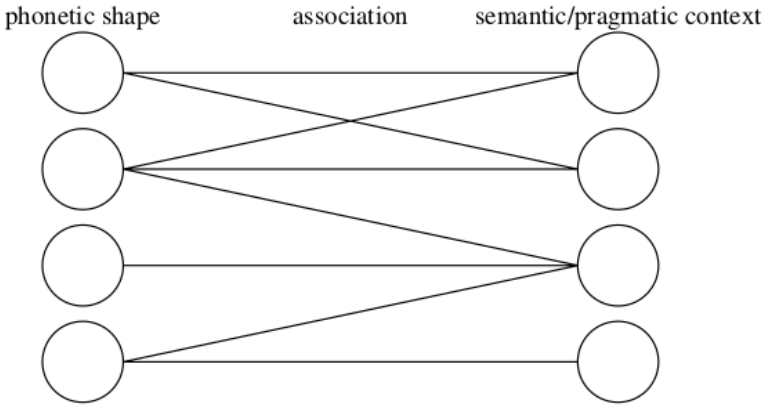


FIGURE 3. Variable associations of form and meaning in a linguistic sign.

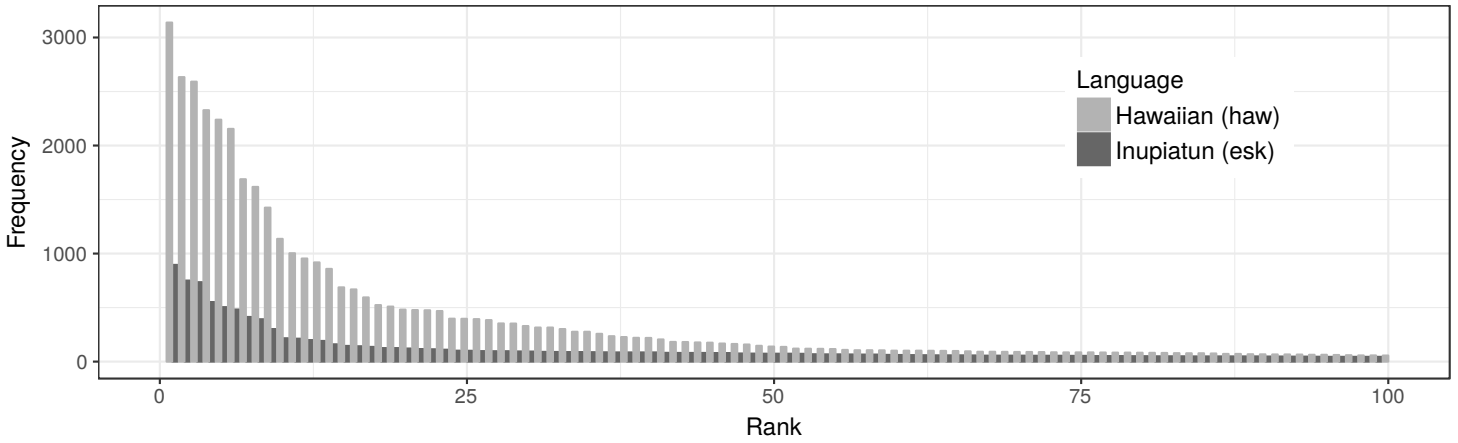
Bybee (2006). From usage to grammar: The mind's response to repetition.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Section 1: Word Frequency Distributions

Word Frequency Distributions



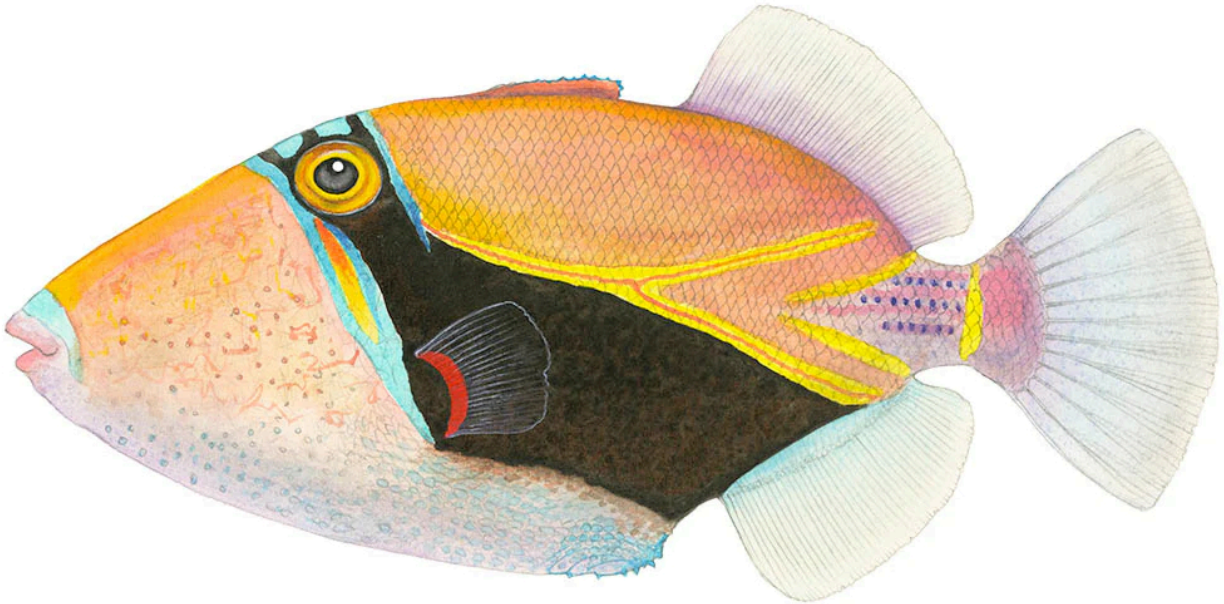
Hawaiian (haw)	Iñupiatun (esk)
40001001 O ke kuauhau na ka hanauna o Iesu Kristo , ka mamo a Davida , ka mamo a Aberahama.	40001001 Uvva ukua aglaŋich sivullianŋiñ Jesus Christ-ŋum , kinguvianŋupluni David-miñ Abraham-miñḷu .
40001002 Na Aberahama o Isaaka ; na Isaaka o Iakoba ; na Iakoba o Iuda a me kona poe hoahanau;	40001002 Abraham aapagigaa Isaac-ŋum , Isaac-li aapagigaa Jacob-ŋum , Jacob-li aapagigaa Judah-ŋum aniqataiñḷu .
[...]	[...]

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

Differences in the Lexicon

"Centuries before there were marine biologists and scientific methods for classifying fish and other marine life, Pacific Islanders were passing on their accumulated knowledge about the behavior of each of hundreds of varieties of fish."

Nettle and Romaine (2000). Vanishing voices, p. 56.



humuhumunukunukuapua'a
"Triggerfish with a snout like a pig"

Differences in Morphological Typology

(1) Hawaiian (haw, PBC 41006018)

A ua olelo aku o loane ia ia [...]
Then PERF say to SUBJ Johan he.DAT [...]
"Then Johan said to him [...]"

(2) Turkish (tur, PBC 41006004)

Ýsa da on-lar-a [...] *de-di*
Jesus also 3P-PL-DAT [...] say-3SG.PERF
"Jesus also said to them [...]"

(3) Iñupiatun (esk, PBC 41006004)

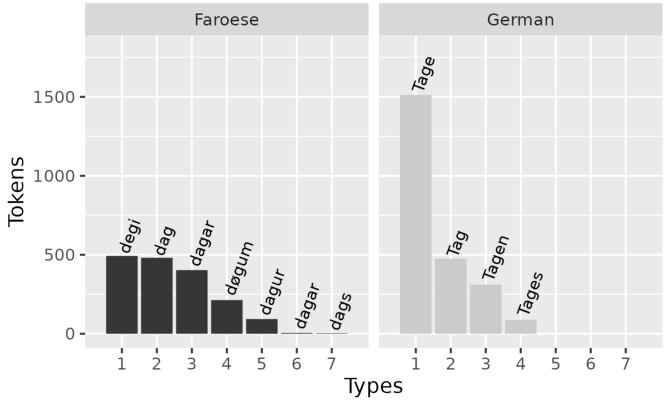
Aglaan Jesus-ngum itna-ġ-ni-ġai [...]
But Jesus-ERG this-say-report-3S.to.3PL
"But Jesus said to them (it is reported) [...]"

Bentz (2018). Adaptive languages: An information-theoretic account of language diversity.

Side Note: Distributional Typology

Typologists are moving away from **aggregated** and **paradigmatic** perspectives on morphology and rather towards a **distributional** view.

a) Distributional



b) Paradigmatic

Singular		Plural	
Nom.	*dagur*	Nom.	***dagar***
Gen.	*dags*	Gen.	*daga*
Dat.	*degi*	Dat.	*døgum*
Acc.	*dag*	Acc.	***dagar***

Singular		Plural	
Nom.	***Tag***	Nom.	***Tage***
Gen.	*Tages*	Gen.	***Tage***
Dat.	***Tage***	Dat.	*Tagen*
Acc.	***Tag***	Acc.	***Tage***

c) Aggregated (WALS)

Bentz and Verkerk (forthcoming). Sociolinguistic typology: complexification and simplification.

Wälchli (2012). Indirect measurement in morphological typology.

Differences in Writing Systems

ISO 15924	Script	Writ. Sys.	No.	Example*
Latn	Latin	Alphabet	38	And they remembered his words ,
Grek	Greek	Alphabet	1	Και ενεθυμηθησαν τους λογους αυτου .
Deva	Devanagari	Abugida	1	तब उस की बातें उन को स्मरण आईं ।
Geor	Georgian	Alphabet	1	და მოეგუსენნეს სიტყუანი მისნი .
Hang	Hangul	Alphabet	1	저희가 예수의 말씀을 기억하고
Mymr	Burmese	Abugida	1	မိန့်တော်မူခဲ့သောစကားများကိုပြန်သတိရ ၍ ၊ -
Cyrl	Cyrillic	Alphabet	1	И они вспомнили эти слова Его .

*Verse number 42024008 of the New Testament.

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

Differences in Writing Systems

Burmese (*mya*)

(4) [...] 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, and Samardžić (2023). Languages through the looking glass of BPE compression.

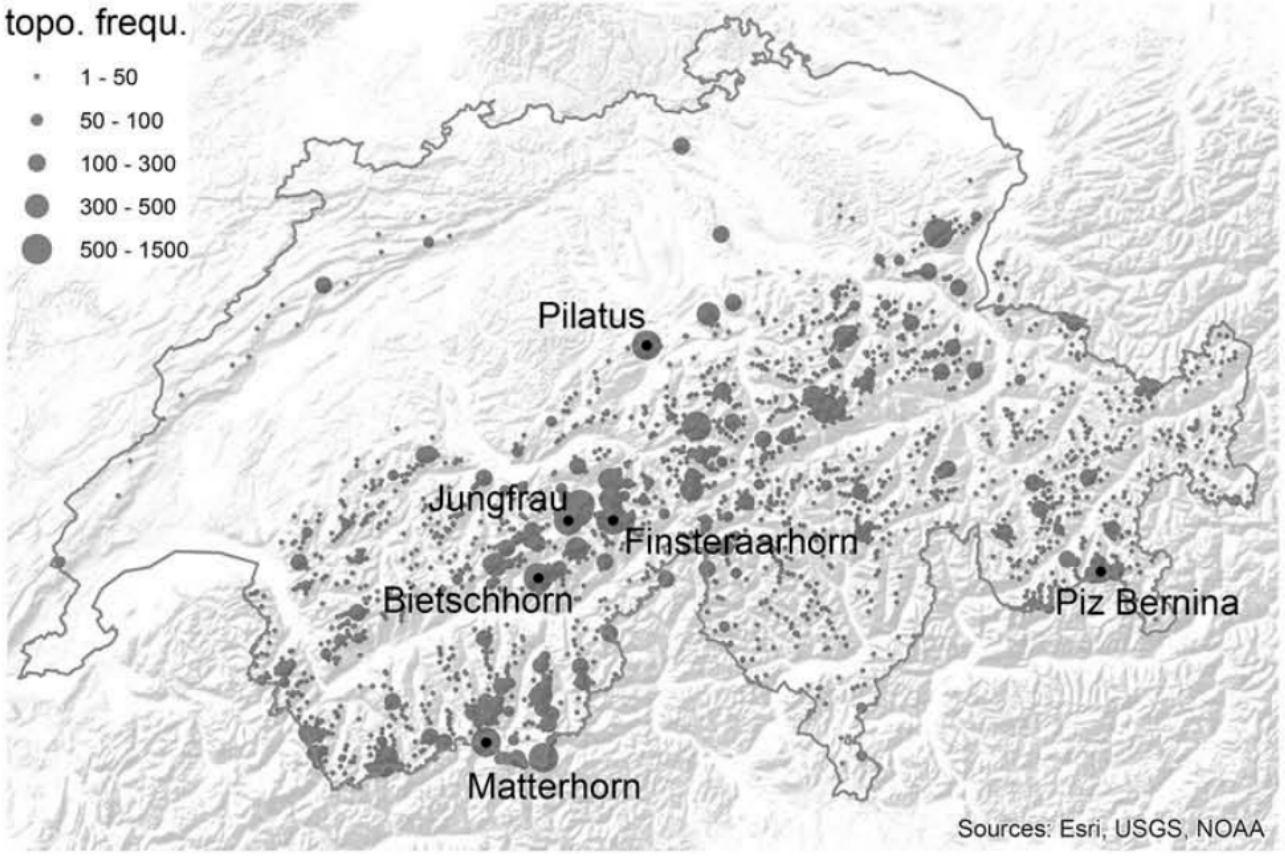
Differences in Translation

Korean (*kor*)

(5) 저희가 예수의 말씀을 기억하고
jeo-hui=ga yei-su=ui mal-sseum=eul gi-eog=ha-go
1P-PL=SUBJ Jesus=POSS speak.HON=OBJ remember=COM
Literal translation: “And we remember Jesus’ speech.”
English verse: “And they remembered his words.”

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

Differences in the “Real World”



Derungs & Samardžić (2017). Are prominent mountains frequently mentioned in text?

Differences in the “Real World”

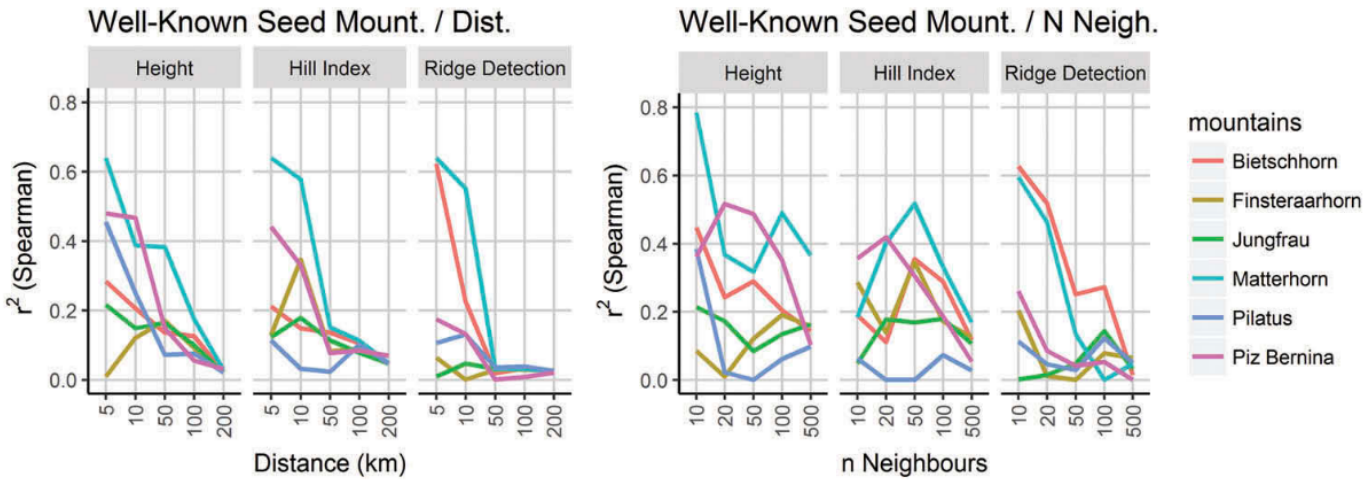


Figure 5. The relation toponym frequency: spatial measure tested for different spatial extents and a set of well-known seed mountains.

The frequency of occurrence of so-called toponyms (in this case names of famous mountains) in texts is significantly correlated with measures of spatial salience (e.g. height), especially if a text is written in a location close-by.

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 *Word Frequency Distributions*?

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Philosophy
General Linguistics

Section 2: Simple Measures

Type-Token Ratio (TTR)

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

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

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
...	...

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

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

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

Repetition Rate (R)

$$R = \frac{r}{\sum_{i=1}^{V} f_i - 1},$$

- r: number of **adjacent repetitions**,
- V: number of types,
- f_i: token frequency of a given type.

Note: r is the number of *actual repetitions*, and the term $\sum_{i=1}^{V} f_i - 1$ in the denominator is the number of *possible repetitions* given the token frequencies.

Example

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

char.types	freq
a	5 - 1 = 4
A	1 - 1 = 0
b	2 - 1 = 1
d	3 - 1 = 2
e	5 - 1 = 4
f	1 - 1 = 0
g	3 - 1 = 2
...	...

word.types	freq
All	1-1 = 0
human	1-1 = 0
beings	1-1 = 0
are	1-1 = 0
born	1-1 = 0
free	1-1 = 0
and	2 -1 = 1
...	...

$$R = \frac{2}{32} = 0.0625$$

$$R = \frac{0}{1} = 0$$

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

Exercise

Take the strings of characters below and calculate the *TTR* and *R* for characters (white spaces are not counted). What is different/similar for the English character sequence compared to the others? What is the problem with the repetition measure *R*?

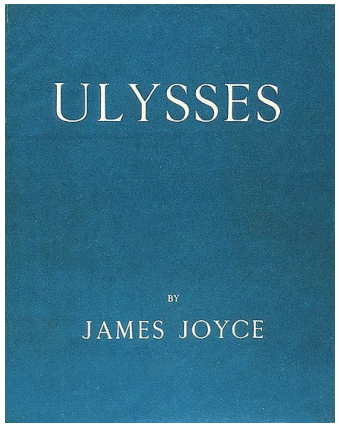
All human beings
uj kd ro su sv sw sx
GGTAGTTAGGGTCT
N01 N01 N01 ZATU6
SWCCSSSSSSSSS

Section 3: Quantitative Laws

Quantitative Linguistics

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

George Kingsley Zipf



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

"[...] we have found a clearcut correlation between the number of different words in the *Ulysses* and the frequency of their usage [...]":

$$r \times f = C \qquad (1)$$

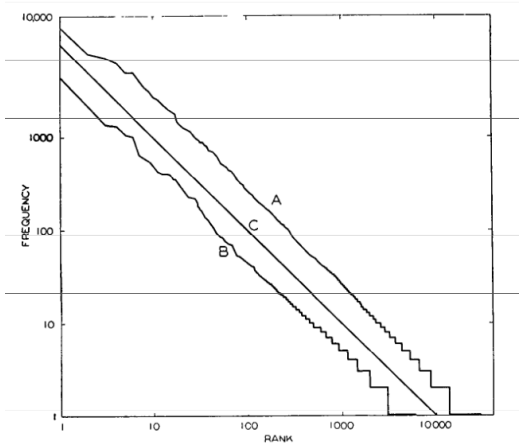


Fig. 2–1. The rank-frequency distribution of words. (A) The James Joyce data; (B) the Eldridge data; (C) ideal curve with slope of negative unity.

Zipf, G. K. (1949). Human behavior and the principle of least effort, p. 24.

Zipf's Law (of Word Frequencies)

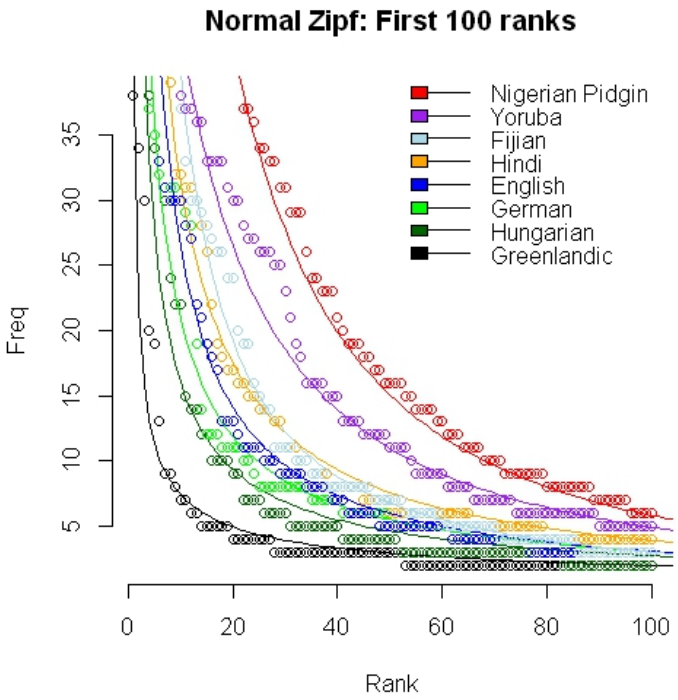
Word	Rank	Freq	Char
the	1	12539	3
and	2	9964	3
of	3	7459	2
to	4	7317	2
in	5	3985	2
you	6	3747	3
for	7	3014	3
is	8	2957	2
he	9	2925	2
a	10	2862	1
…	…	…	…
work–then	2742	1	10
world-rulers	2743	1	12
worm	2744	1	4
wormwood	2745	1	8
wounding	2746	1	8
writer	2747	1	6
writers	2748	1	7
zarephath	2749	1	9
zenas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

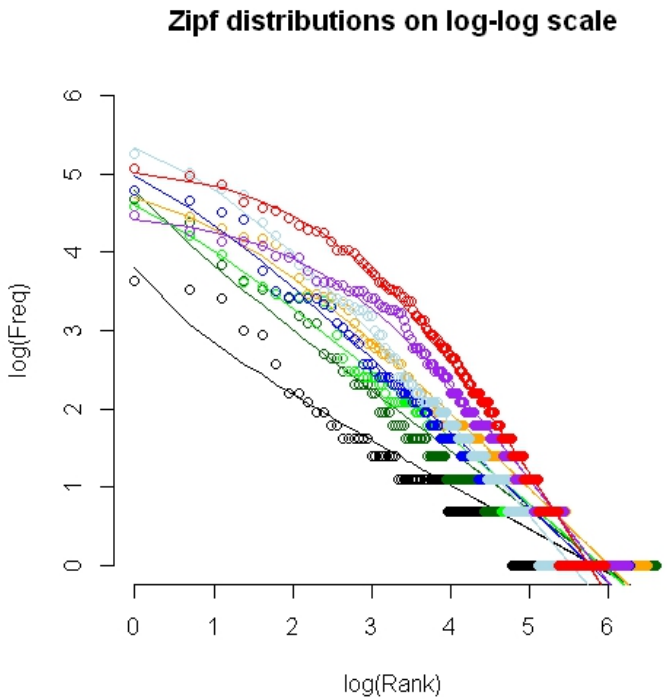
Another (more common) formulation of the law:

$$f(w) \propto \frac{1}{r^{\alpha}} \tag{2}$$

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.

Zipf's Law across Languages





Note: This illustrates another version of the law, the *Zipf-Mandelbrot Law*, which has an extra parameter (β) accounting for the deviations from linearity in high frequency items.

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

Zipf's Law of Abbreviation

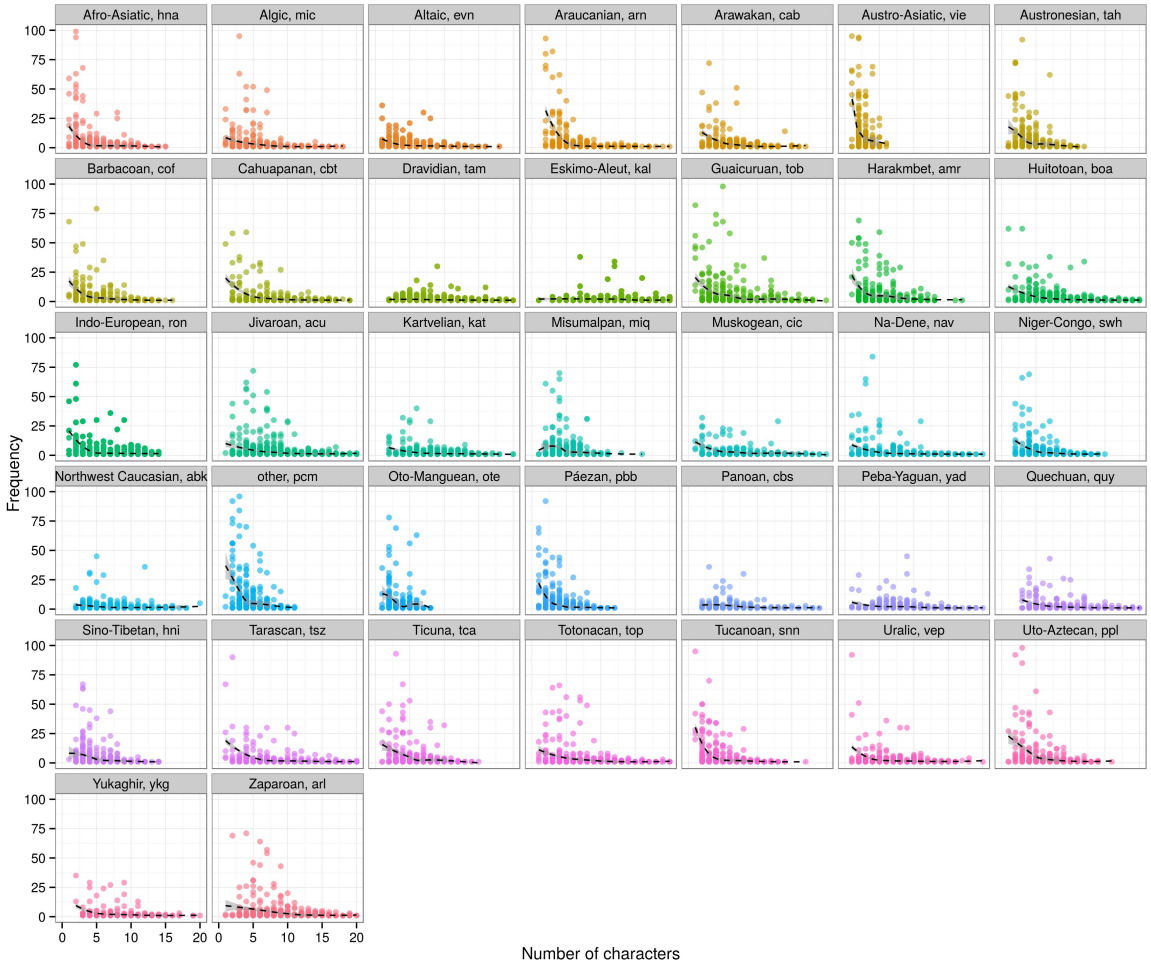
Word	Rank	Freq	Char
the	1	12539	3
and	2	9964	3
of	3	7459	2
to	4	7317	2
in	5	3985	2
you	6	3747	3
for	7	3014	3
is	8	2957	2
he	9	2925	2
a	10	2862	1
…	…	…	…
work–then	2742	1	10
world-rulers	2743	1	12
worm	2744	1	4
wormwood	2745	1	8
wounding	2746	1	8
writer	2747	1	6
writers	2748	1	7
zarephath	2749	1	9
zenas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

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

Zipf, George K. (1965). The psycho-biology of language: An introduction to dynamic philology, p. 25.

In other words: **more frequent words** tend to be **shorter**.

Zipf's Law of Abbreviation across Languages



Zipf's Law of Abbreviation across Languages

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.

Menzerath-Altmann Law

Paul Menzerath (right) with Paulo Lazerda (left) in the Bonn phonetics laboratory.

bə'ʦɐ̯ɪ:çnʊŋ: „'platdia¸lɛktɪʃə ¸ʔǫo:sʃpra·xə“ 'za·gən ¸vələn. ɪn 'dǫy:ʧlan:t ¸trɪft ʔe·bən 'nɪçt ʦu·, vas hɛr Palmer fy:r 'ja·pan 'fɛstʃtɛltə, un:t vas hɛr Graham mu'ta·ti:s mu'tandi:s (zo·!) fy:r 'ʔalə ʃpra·xən 'ʔanʦu·ne:mən ʃɐ̯ɪnt': 'nɛ·mlɪç, das 'ʔɐ̯ɪ:nfax di· 'ʔǫo:sʃpra·xə 'je·dəs ɪndi'vi:duums dər gə'bɪldətən 'klasə al:s 'mustər gɛl:tən dar:f. 'das bə¸ʦvɐ̯ɪflə ɪç ʃo·n fy:r gro·sbri'tʼanıən, əs ʃtɪmt 'zɪçər ¸nɪçt fy:r 'fraŋ:krɐ̯ɪ:ç, fy:r ɪ'tʼa:li̯ən 'ʔǫo:x nɪçtʼ, ʔun:t ¸gan:ʦ un:t 'ga:r nɪçt fy:r 'dǫy:ʧlan:tʼ. 'mustərhaftəs 'dǫy:ʧ vɪrt ɪn 'kʼɐ̯ɪnər 'dǫy:ʧən 'ʃtat o·dər 'lan:tʃaft gə'ʃprɔxən; 'ʔǫo:x dər gə'bɪldətə bɛr'li:nər, 'mʏn:çənər, 'vi·nər, 'fraŋ:kfʊrtər, 'kʼœlnər ¸ʃprɪçt ʔal:zo· nɪçt ʔo·nə 'vɐ̯ɪ:tərəs 'mustərhaftʼ. ʔǫo:f 'kʼɐ̯ɪnən

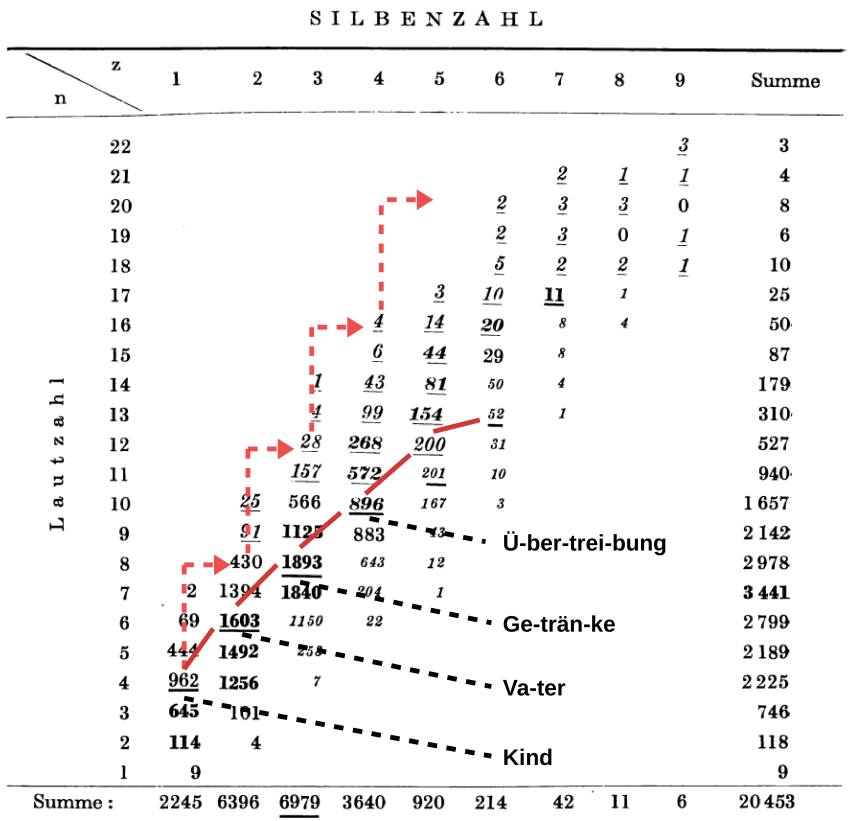
‘For Germany it is not true what Mr. Palmer noted for Japan and what Mr. Graham seems to assume mutatis mutandis for all languages, i.e. that the pronunciation of any educated individual may serve as a model. I doubt this for Great Britain, it is certainly not true for France, neither for Italy and least of all for Germany. Exemplary German is not spoken in any German city or region; neither the educated man from Berlin nor from Munich, Vienna, Frankfurt or Cologne can be considered to pronounce in an exemplary way’

Braun & Möbius (2022).

Menzerath-Altmann Law

"The greater the whole the smaller its parts". For example, as the **number of syllables** in words of a language increases, the *relative* number of **phonemes per syllable** decreases.

Note: Menzerath counted the number of words in a German lexicon with a given number of phonemes ("Lautzahl") and a given number of syllables ("Silbenzahl"). We might expect that the number of phonemes increases linearly with the number of syllables (dashed lines), but this is not the case. Instead, the *relative* (i.e. average) number of phonemes per syllable decreases.



Vértes (1955).

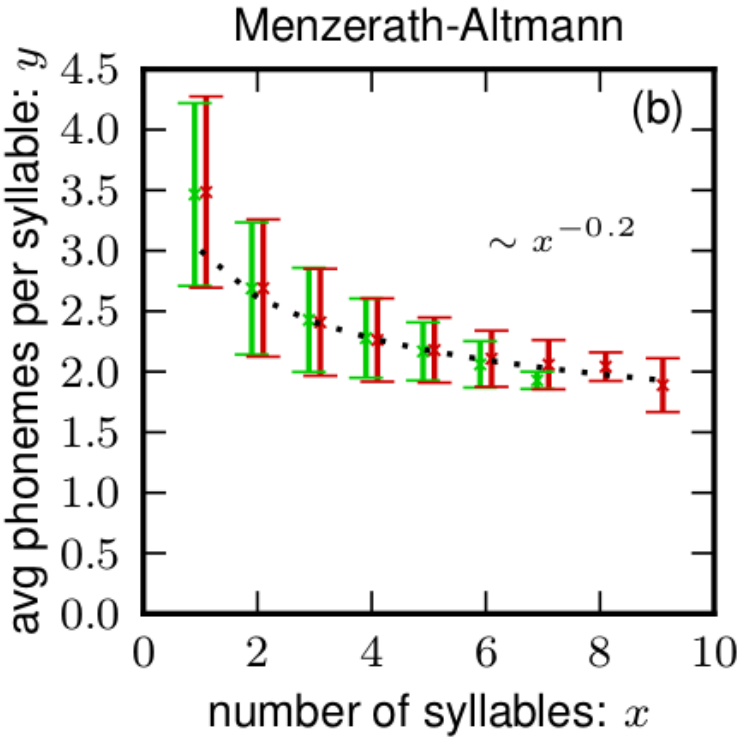
Menzerath-Altmann Law

The modern version of the law is formulated as:

$$y = \alpha x^{\beta} e^{-\gamma x}, \qquad (3)$$

with y representing the size of the parts (e.g. in number of phonemes), x representing the length of the whole (e.g. number of syllables, and α, β, and γ are parameters (different from the ones in other laws of course).

Gerlach & Altmann (2015), p. 3.



green: Moby Dick in English,
red: Wikipedia in English.

Heaps' Law (Herdan's Law)

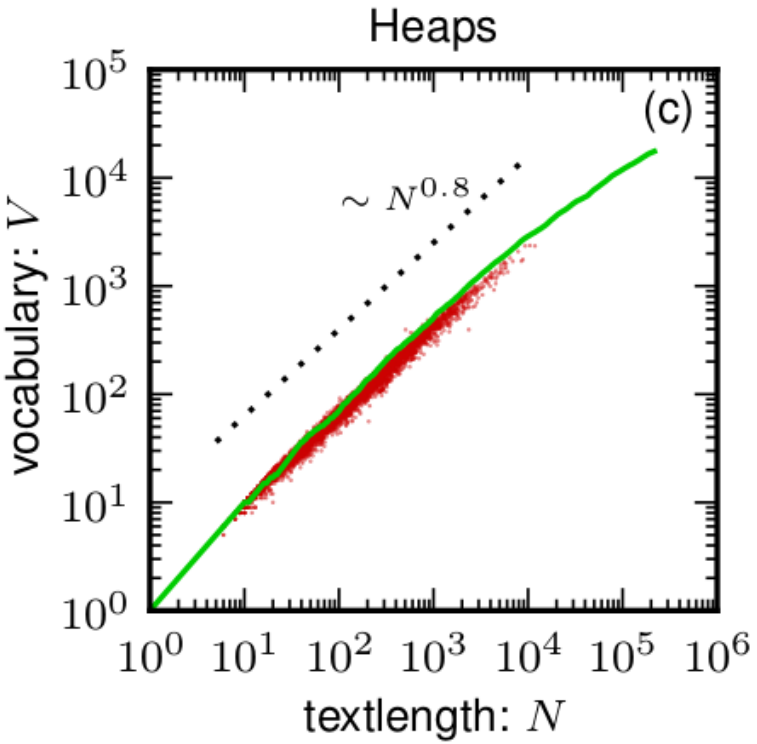
The number of *word types* in the vocabulary (V) grows with the number of *word tokens* in a text (N) according to

$$V \sim N^{\alpha}, \qquad (4)$$

with $0 < \alpha < 1$.

Gerlach & Altmann (2015), p. 3-4.

Note: This does not hold for characters in writing, as there will be a finite number of characters in any writing system. Once this number is reached, V cannot grow further.



green: Moby Dick in English,
red: Wikipedia in English.

Important Implication

If Heaps' law holds generally, it would mean that languages are **infinitely productive** at the word level. However big a corpus, we would never see all words of a given language, let alone all languages.

Current Research Question

Are “linguistic” laws also represented in other species?

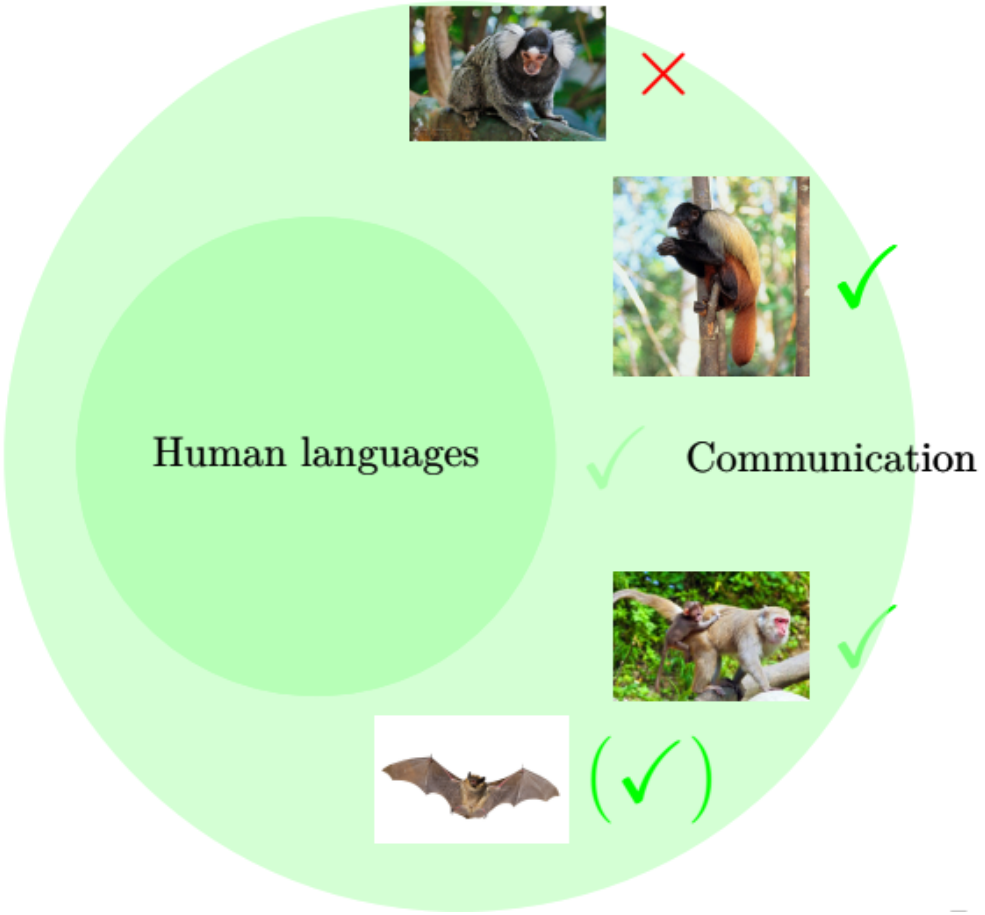



photo by Abdullah Mohiuddin
CC BY-NC-SA Hajime Yamauchi
GGCCTGACACCTTCGCACC

Summary

Summary

- **Word frequency distributions** are affected by a host of factors: *lexical typology*, *morphological typology*, *real world salience*, *translation*, *writing system*, etc.
- Differences in these distributions can be measured by simple measures such as **type-token-ratios** (TTR).
- More complicated models include **Zipf's law of word frequencies**.
- There are further **quantitative laws** found in natural languages: *Zipf's law of abbreviation*, *Menzerath-Altmann law*, *Heap's law*.
- It is a recurrent research question to what extent these are **specific to human language**.


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

References

References

Bentz, C. (2018). *Adaptive languages: An information-theoretic account of linguistic diversity*. Berlin, Boston: De Gruyter Mouton.

Bentz, C. and Ferrer-i-Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In: Bentz, C., Jäger, G. and Yanovich, I. (eds.) *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. University of Tübingen, online publication system.

Bentz, C., and Kiela, D. (2014). Zipf's law across languages of the world: towards a quantitative measure of lexical diversity. In *Evolution of Language: Proceedings of the 10th International Conference (EVOLANG10)* (pp. 385-386).

Bentz, C. and Verkerk, A. (forthcoming). Sociolinguistic typology: complexification and simplification.

Braun, A. and Möbius, B. (2022). Armando de Lacerda and his contemporaries: Paul Menzerath. In *Proceedings of Fifth International Workshop on the History of Speech Communication Research (HSCR)*.

Bybee, J. (2006). From usage to grammar: the mind's response to repetition. *Language*.

Derungs, C., and Samardžić, T. (2018). Are prominent mountains frequently mentioned in text? Exploring the spatial expressiveness of text frequency. *International Journal of Geographical Information Science*, 32(5), 856-873.

Gerlach, M. and Altmann, E. (2015). Statistical laws in linguistics. *arXiv*.

Gutierrez-Vasques, X., Bentz, C., and Samardžić, T. (2023). Languages Through the Looking Glass of BPE Compression. *Computational Linguistics*, pp. 1-59.

Nettle, D. and Romaine, S. (2000). *Vanishing voices: The extinction of the world's languages*. Oxford: Oxford University Press.

Mayer, T. and Cysouw, M. (2014). Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sproat, R. (2014). A statistical comparison of written language and nonlinguistic symbol systems. *Language*, 457-481.

Vértes, E. (1955). Review: Die Architektonik des deutschen Wortschatzes. In *Acta Linguistica Academiae Scientiarum Hungaricae*, Vol. 5, No. 3/4, pp. 415-433.

Wälchli, B. (2012). Indirect measurement in morphological typology. In A. Ender, A. Leemann and B. Wälchli (Ed.), *Methods in Contemporary Linguistics* (pp. 69-92). Berlin, Boston: De Gruyter Mouton.

Zipf, G. K. (1949). *Human behavior and the principle of least effort. An introduction to human ecology.* Cambridge: Addison-Wesley Press.

Zipf, G. K. (1965). *The psycho-biology of language. An introduction to dynamic philology*. Cambridge: MIT Press.

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