Language Change as a (Random?) Walk in Entropy Space

Christian Bentz

November 19, 2019

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”

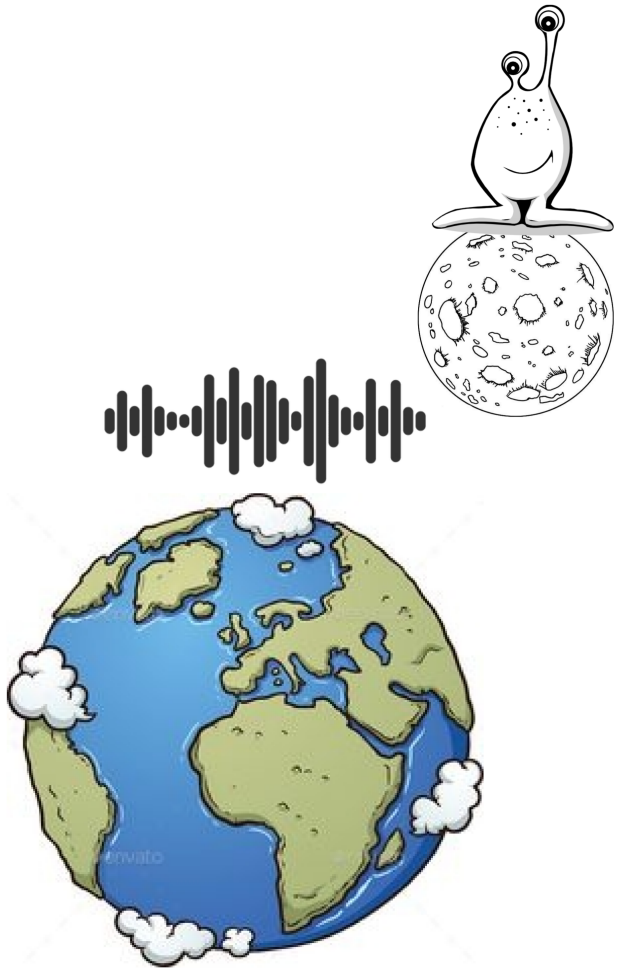
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Introduction

The Martian Linguist (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), p. 187.



Mapping out the Space of Human Languages

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.

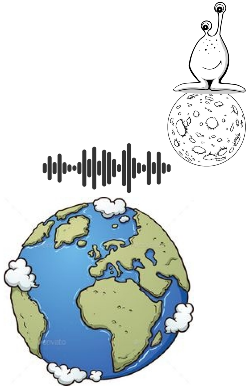
ހުރިހާ އިންސާނުންވެސް ދުނިޔެއަށް އުފަންވަނީ، މިނިވަންކަމުގައި، ހަމަހަމަ ޙައްޤުތަކަކާއެކު، ހަމަހަމަ ދަރަޖައެއްގައި ކަމުގައިވާނެއެވެ.

모든 인간은 태어날 때부터 자유로우며 그 존엄과 권리에 있어 동등하다.

يولد جميع الناس أحرارًا متساوين في الكرامة والحقوق.

የሰው፡ልጅ፡ሁሉ፡ሲወለድ፡ነጻና፡በክብርና፡በመብትም፡እኩልነት፡ያለው፡ነው።

人人生而自由,在尊严和权利上一律平等。



Methods

Methodological Choices

Modality: Written

Methodological Choices

Modality: Written

Alternatives: Spoken, Signed, Whistled

Methodological Choices

Modality: Written

Alternatives: Spoken, Signed, Whistled

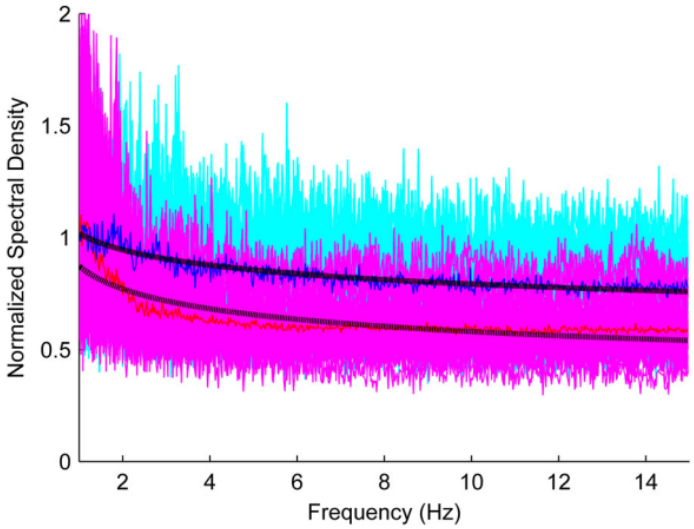


Figure 2. Example normalized spectral density for ASL (cyan/blue) and everyday motion (magenta/red). Cyan/magenta lines show raw data for Optical Flow between 0.20–0.25 px/sec, and the average for ASL (blue) and everyday motion (red) over all videos is also shown. Black lines show the respective fit according to Equation (1). Signing videos show greater fractal complexity.

Malaia, Borneman & Wilbur (2016). Assessment of information content in visual signal.

Methodological Choices

Modality: Written

Alternatives: Spoken, Signed, Whistled

Unit: Orthographic Word

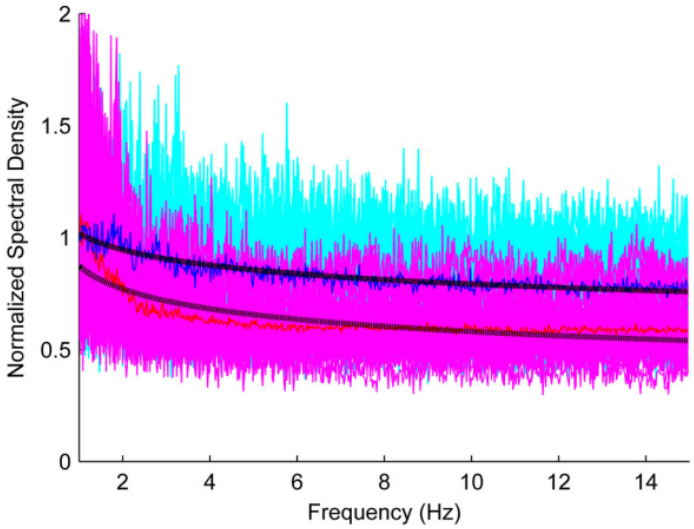


Figure 2. Example normalized spectral density for ASL (cyan/blue) and everyday motion (magenta/red). Cyan/magenta lines show raw data for Optical Flow between 0.20–0.25 px/sec, and the average for ASL (blue) and everyday motion (red) over all videos is also shown. Black lines show the respective fit according to Equation (1). Signing videos show greater fractal complexity.

Malaia, Borneman & Wilbur (2016). Assessment of information content in visual signal.

Methodological Choices

Modality: Written
Alternatives: Spoken, Signed, Whistled

Unit: Orthographic Word
Alternatives: Characters, Syllables, Morphemes, Phrases

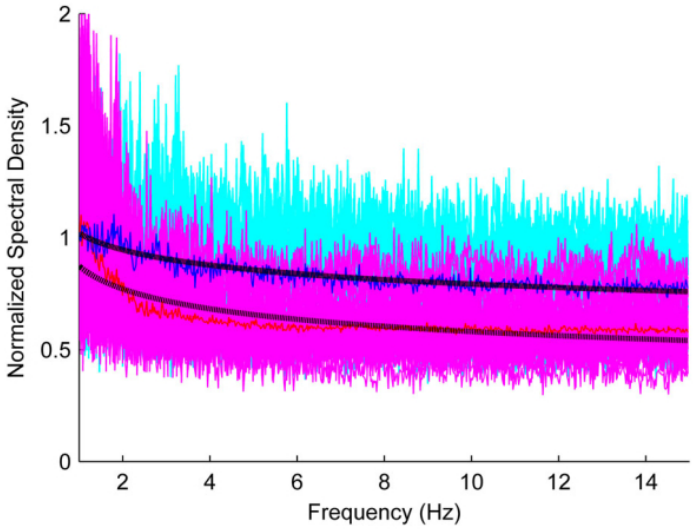


Figure 2. Example normalized spectral density for ASL (cyan/blue) and everyday motion (magenta/red). Cyan/magenta lines show raw data for Optical Flow between 0.20–0.25 px/sec, and the average for ASL (blue) and everyday motion (red) over all videos is also shown. Black lines show the respective fit according to Equation (1). Signing videos show greater fractal complexity.

Malaia, Borneman & Wilbur (2016). Assessment of information content in visual signal.

Methodological Choices

Modality: Written
Alternatives: Spoken, Signed, Whistled

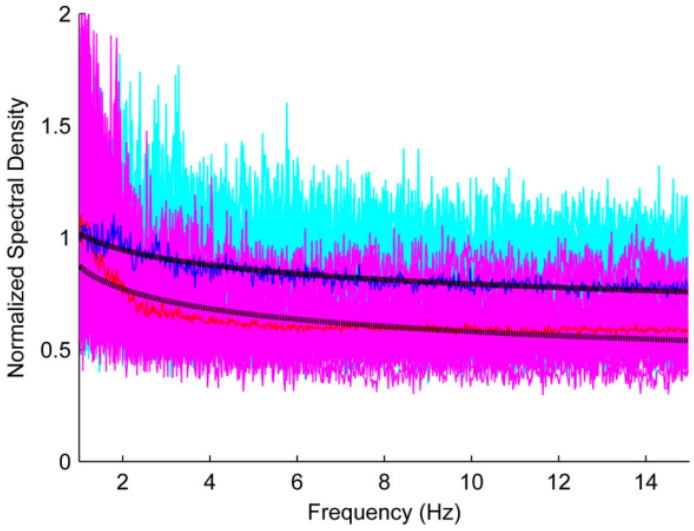
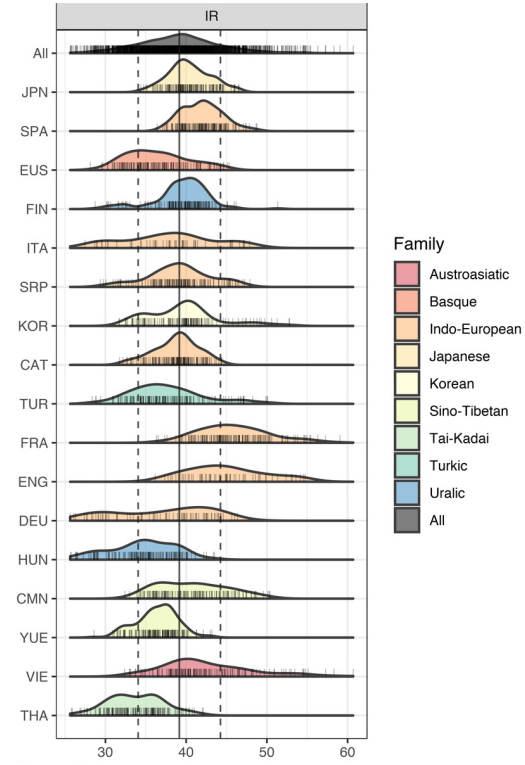


Figure 2. Example normalized spectral density for ASL (cyan/blue) and everyday motion (magenta/red). Cyan/magenta lines show raw data for Optical Flow between 0.20–0.25 px/sec, and the average for ASL (blue) and everyday motion (red) over all videos is also shown. Black lines show the respective fit according to Equation (1). Signing videos show greater fractal complexity.

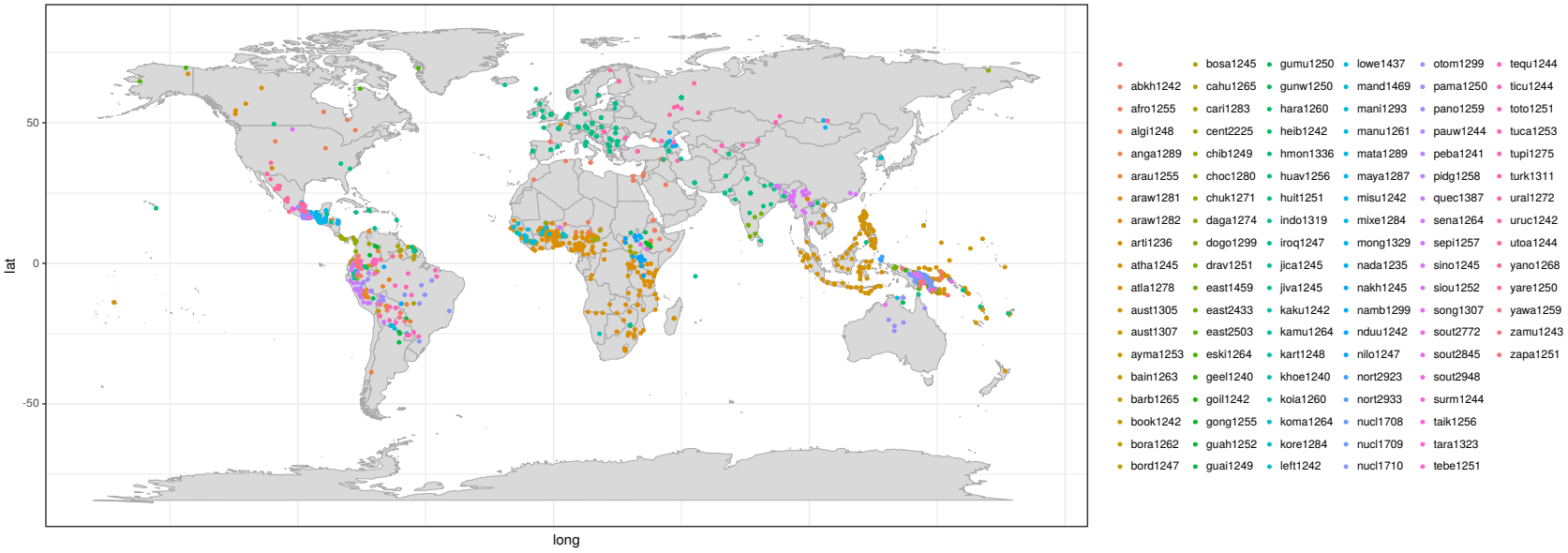
Malaia, Borneman & Wilbur (2016). Assessment of information content in visual signal.

Unit: Orthographic Word
Alternatives: Characters, Syllables, Morphemes, Phrases



Coupé, Oh, Dediu & Pellegrino. (2019). Different languages, similar encoding efficiency.

Data: Parallel Bible Corpus



- 1514 translations ($\leq$50K tokens)
- 1131 unique languages (ISO codes)
- 109 families (Glottolog)

Müller & Cysouw (2014). A massively parallel Bible corpus.
Hammarström, Harald & Forkel, Robert & Haspelmath, Martin. 2019. Glottolog 4.0.

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

in_{1} the_{2} beginning_{3} god_{4} created_{5} the_{6} heavens_{7} and_{8} the_{9} earth_{10} and_{11} the_{12} earth_{13} was_{14} waste_{15} and_{16} empty_{17} [...]

Example

$\text{in}_1\ \text{the}_2\ \text{beginning}_3\ \text{god}_4\ \text{created}_5\ \text{the}_6\ \text{heavens}_7\ \text{and}_8\ \text{the}_9\ \text{earth}_{10}$
$\text{and}_{11}\ \text{the}_{12}\ \text{earth}_{13}\ \text{was}_{14}\ \text{waste}_{15}\ \text{and}_{16}\ \text{empty}_{17}\ [...]$

$$\widehat{H}^{ML}(X) = -(\tfrac{4}{17}\log_2(\tfrac{4}{17}) + \tfrac{3}{17}\log_2(\tfrac{3}{17}) + \cdots + \tfrac{1}{17}\log_2(\tfrac{1}{17})) \sim 3.2$$

Example

$\text{in}_1\ \text{the}_2\ \text{beginning}_3\ \text{god}_4\ \text{created}_5\ \text{the}_6\ \text{heavens}_7\ \text{and}_8\ \text{the}_9\ \text{earth}_{10}\ \text{and}_{11}\ \text{the}_{12}\ \text{earth}_{13}\ \text{was}_{14}\ \text{waste}_{15}\ \text{and}_{16}\ \text{empty}_{17}\ [...]$

$$\widehat{H}^{ML}(X) = -(\tfrac{4}{17}\log_2(\tfrac{4}{17}) + \tfrac{3}{17}\log_2(\tfrac{3}{17}) + \cdots + \tfrac{1}{17}\log_2(\tfrac{1}{17})) \sim 3.2$$

Problem: natural language is **not an i.i.d process** ("bag-of-words" drawing with replacement) due to short and long range correlations, e.g. frequent n-grams in a text ("and the earth").

Entropy Estimation: Entropy Rate

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

- n: number of word tokens
- L_i: length (+1) of the longest contiguous substring 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).

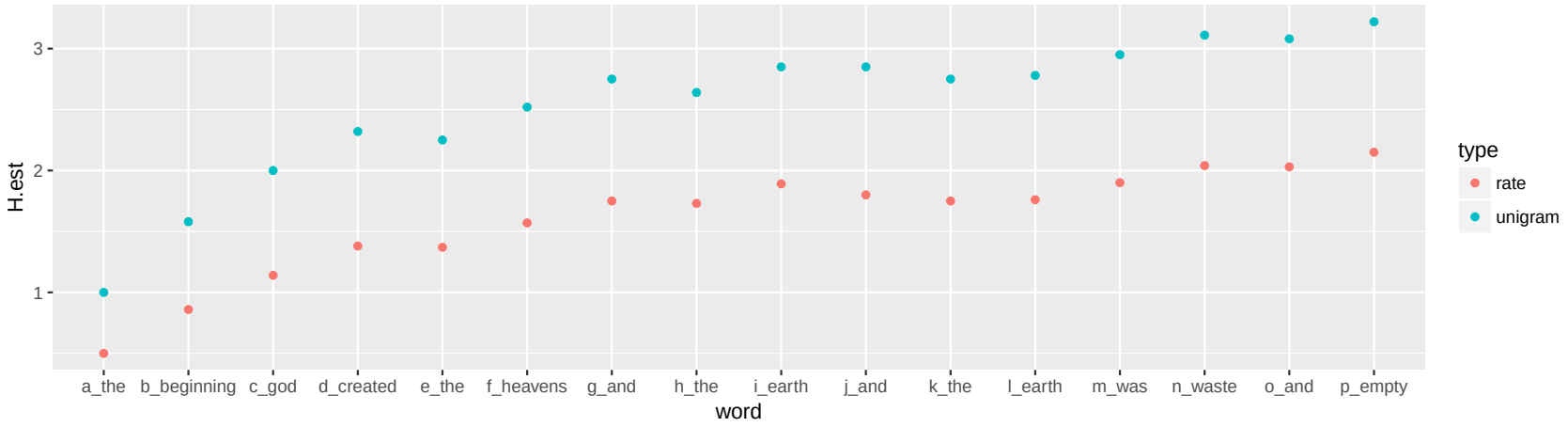
Example

$\text{in}_1\ \text{the}_2\ \text{beginning}_3\ \text{god}_4\ \text{created}_5\ \text{the}_6\ \text{heavens}_7\ \text{and}_8\ \text{the}_9\ \text{earth}_{10}$
$\text{and}_{11}\ \underline{\text{the}}_{12}\ \underline{\text{earth}}_{13}\ \text{was}_{14}\ \text{waste}_{15}\ \text{and}_{16}\ \text{empty}_{17}\ [\ldots]$

$$L_{11} = 3(+1) = 4$$

$$\frac{\log_2(11)}{4} \sim \frac{3.46}{4} \sim 0.87$$

Example: Unigram and Rate Comparison

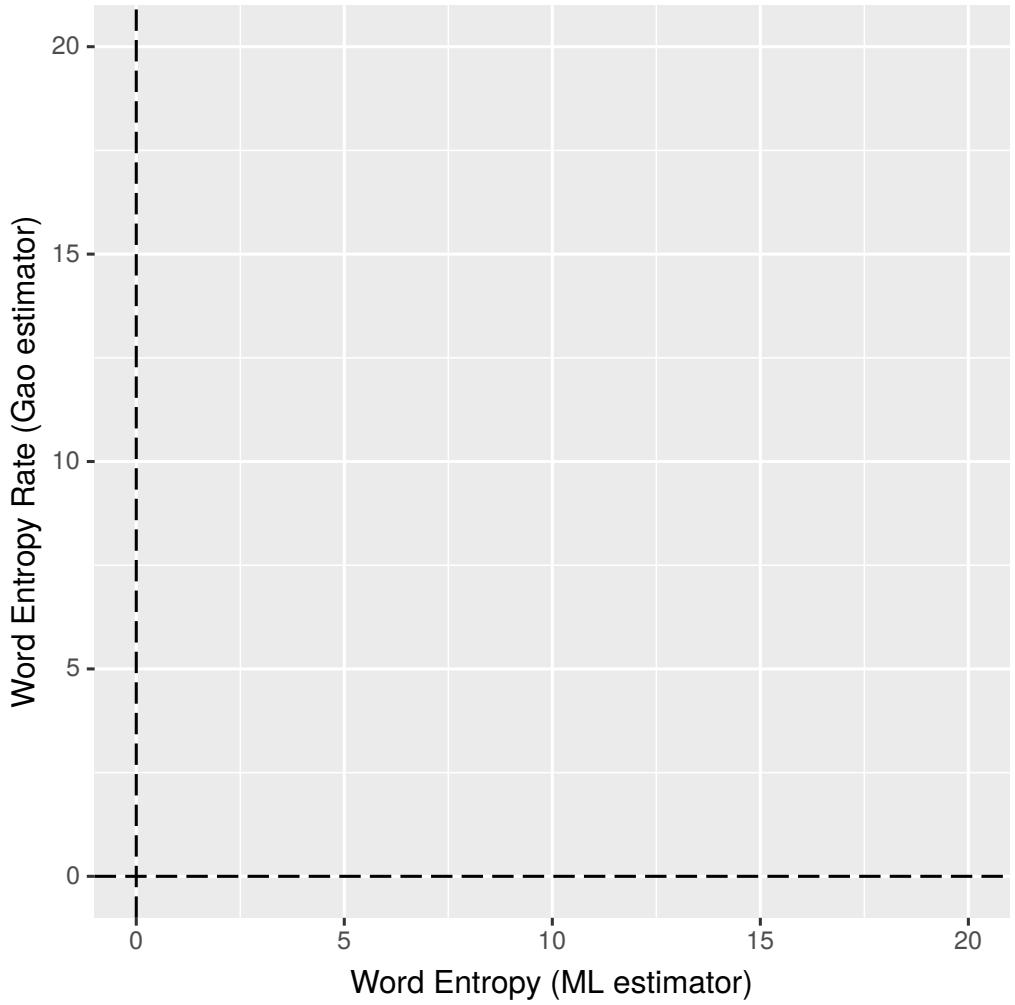


R package *Hrate* (https://github.com/dimalik/Hrate)

Bentz, Alikaniotis, Cysouw, & Ferrer-i-Cancho (2017). The entropy of words - learnability and expressivity across more than 1000 languages.

Results

Delimiting the Space of Possible Languages



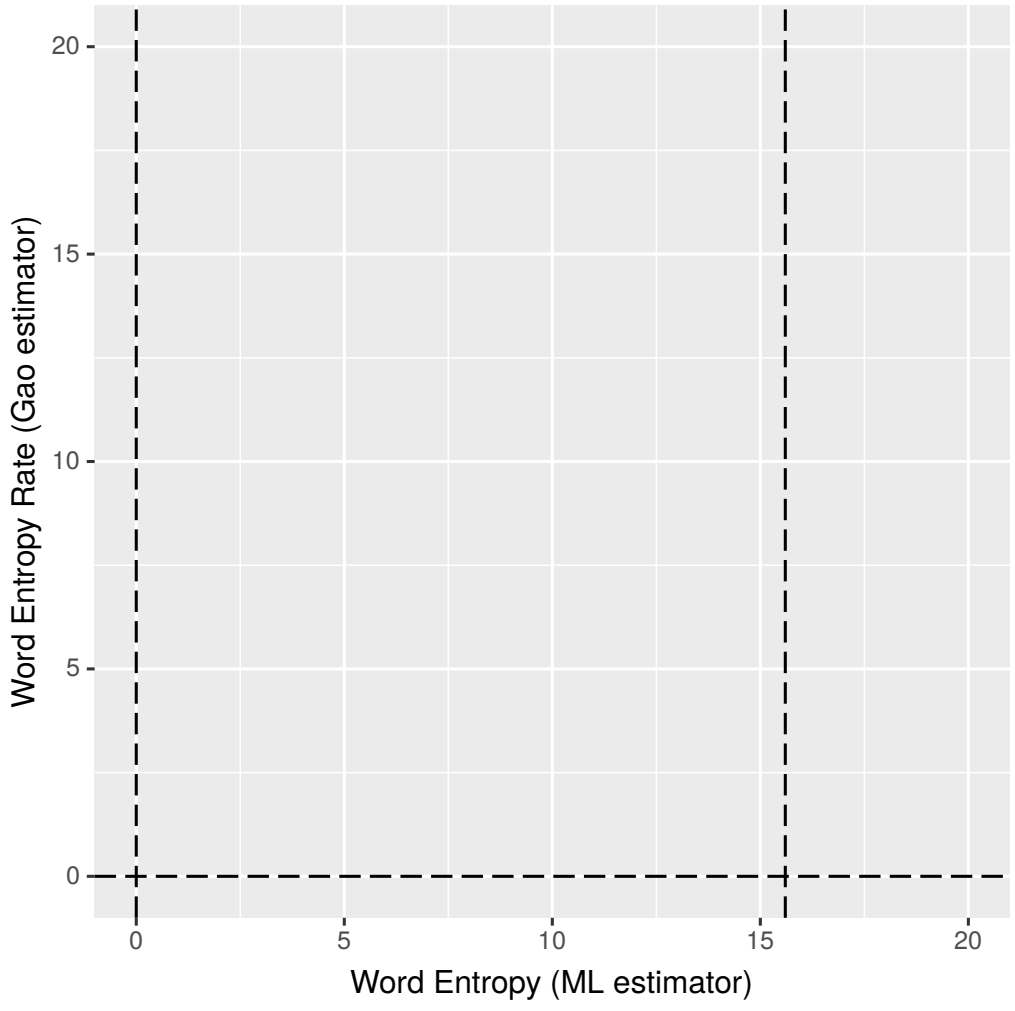
Word Unigram Entropy:

$H(X) \geq 0$

Word Entropy Rate:

$h(\mathcal{X}) \geq 0$

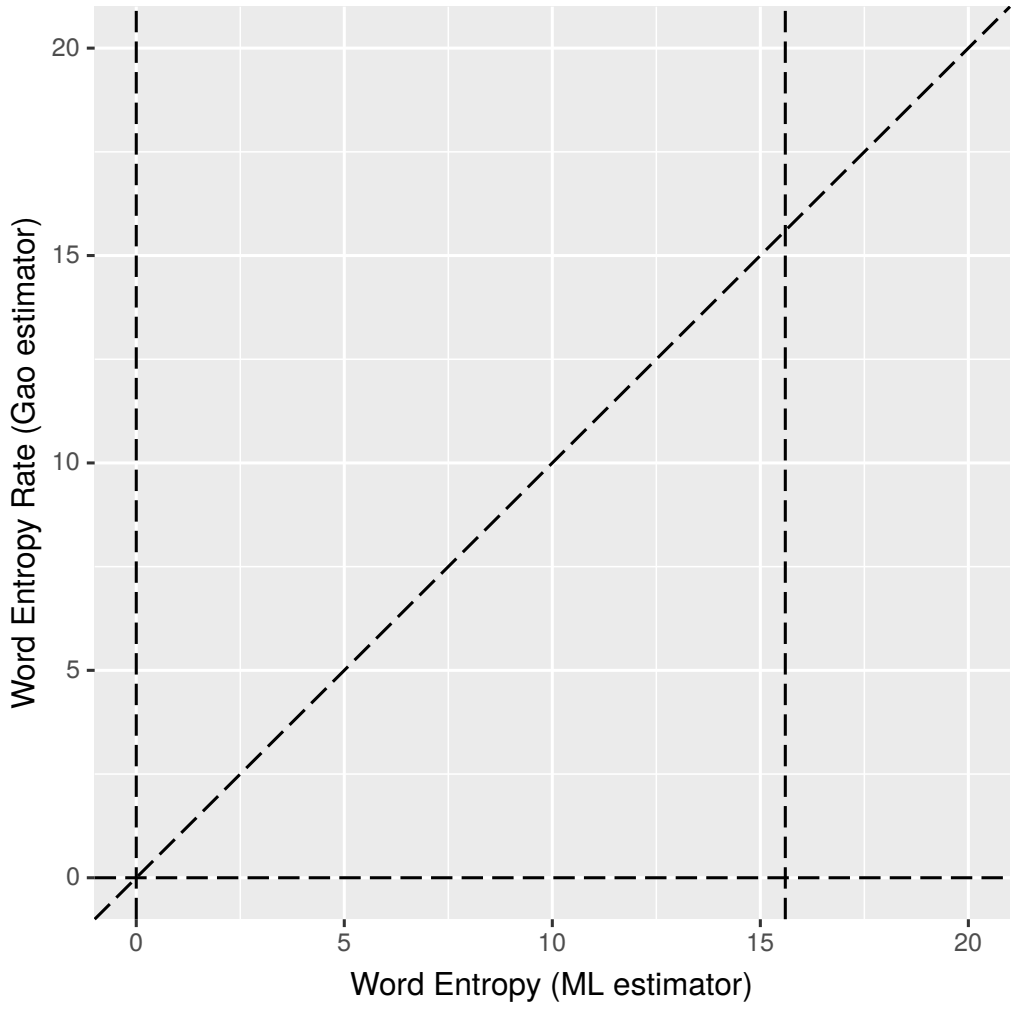
Delimiting the Space of Possible Languages



Maximum Word Unigram Entropy at 50K tokens:

$H^{max}(X) =$
$-\sum_{i=1}^{V} p(\frac{1}{5\times10^4}) \log_2 p(\frac{1}{5\times10^4}) =$
$\log_2(5 \times 10^4) \sim \mathbf{15.6}$

Delimiting the Space of Possible Languages



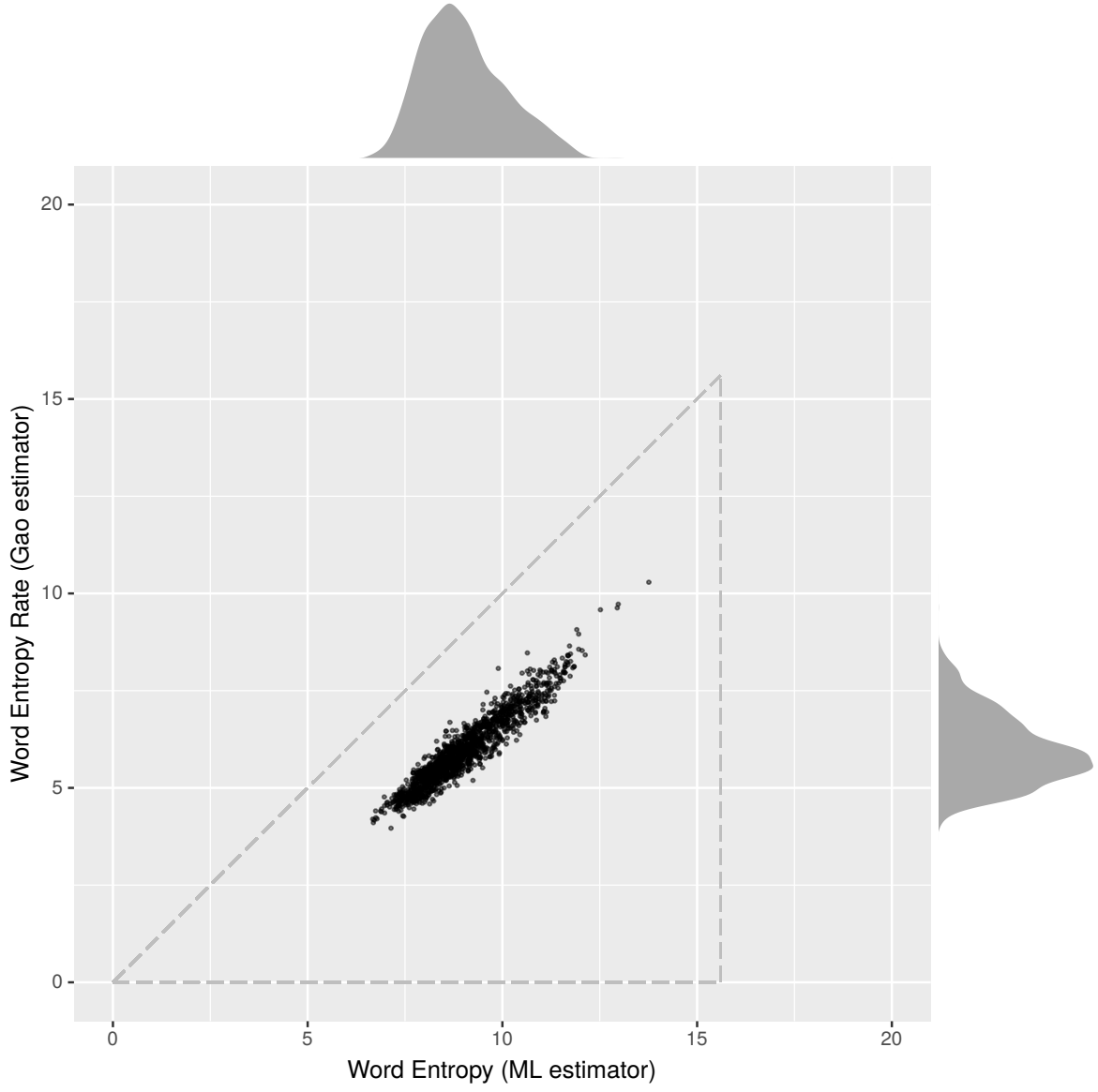
Lemma 1:
$h(\mathcal{X}) = H(X)$ for i.i.d variables

Lemma 2:
$h(\mathcal{X}) =$
$\lim_{n\to\infty} H(X_n|X_1, X_2, ..., X_{n-1})$.

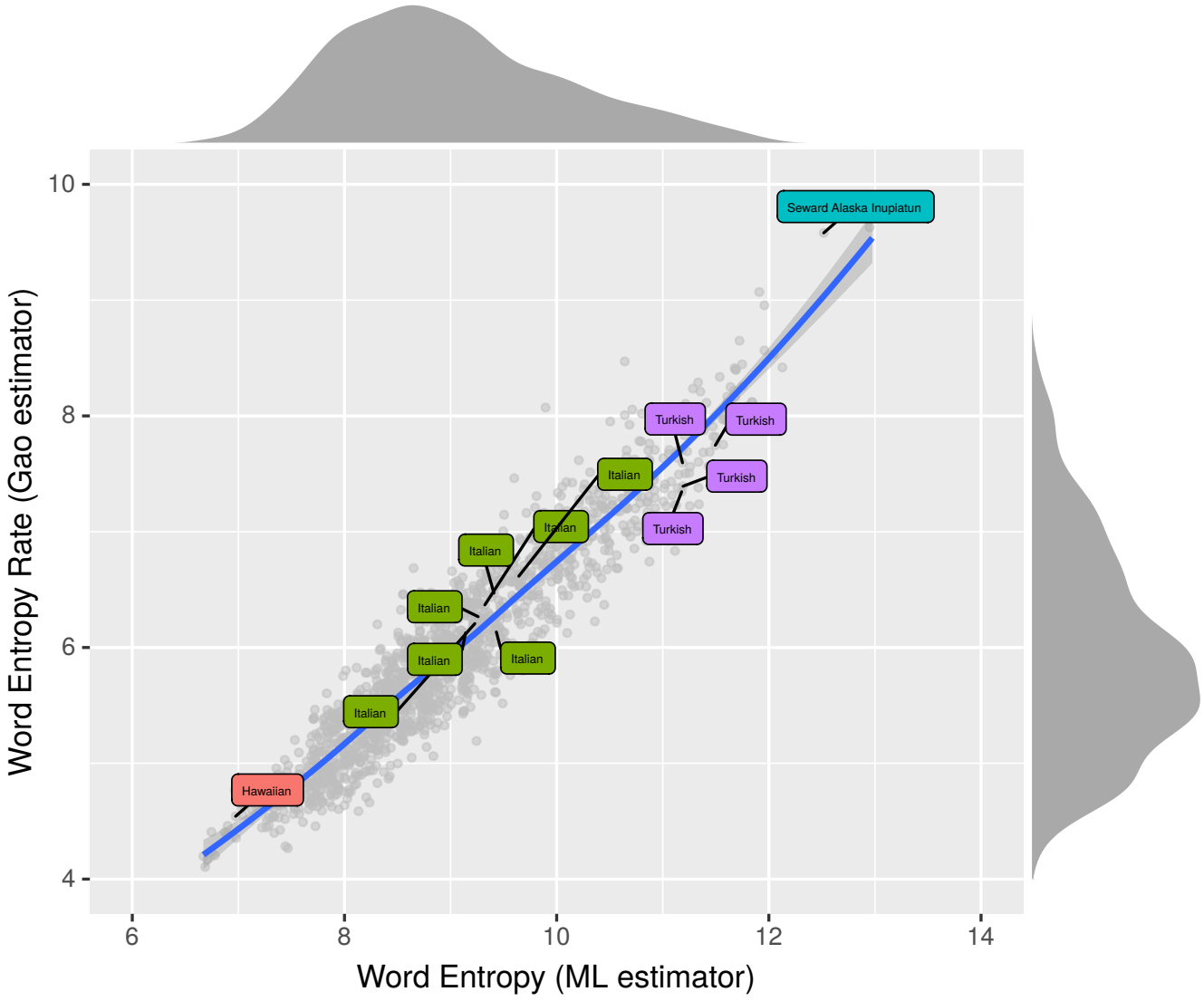
Hence:
$h(\mathcal{X}) \leq H(X)$

Dębowski (forth.) Information theory meets power laws.

The 1131 Language Sample



Zooming Into the Range



Differences in Morphological Encoding (Among Other Factors)

(1) Hawaiian (haw, PBC 41006018)

A ua olelo aku o Ioane 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 [...]"

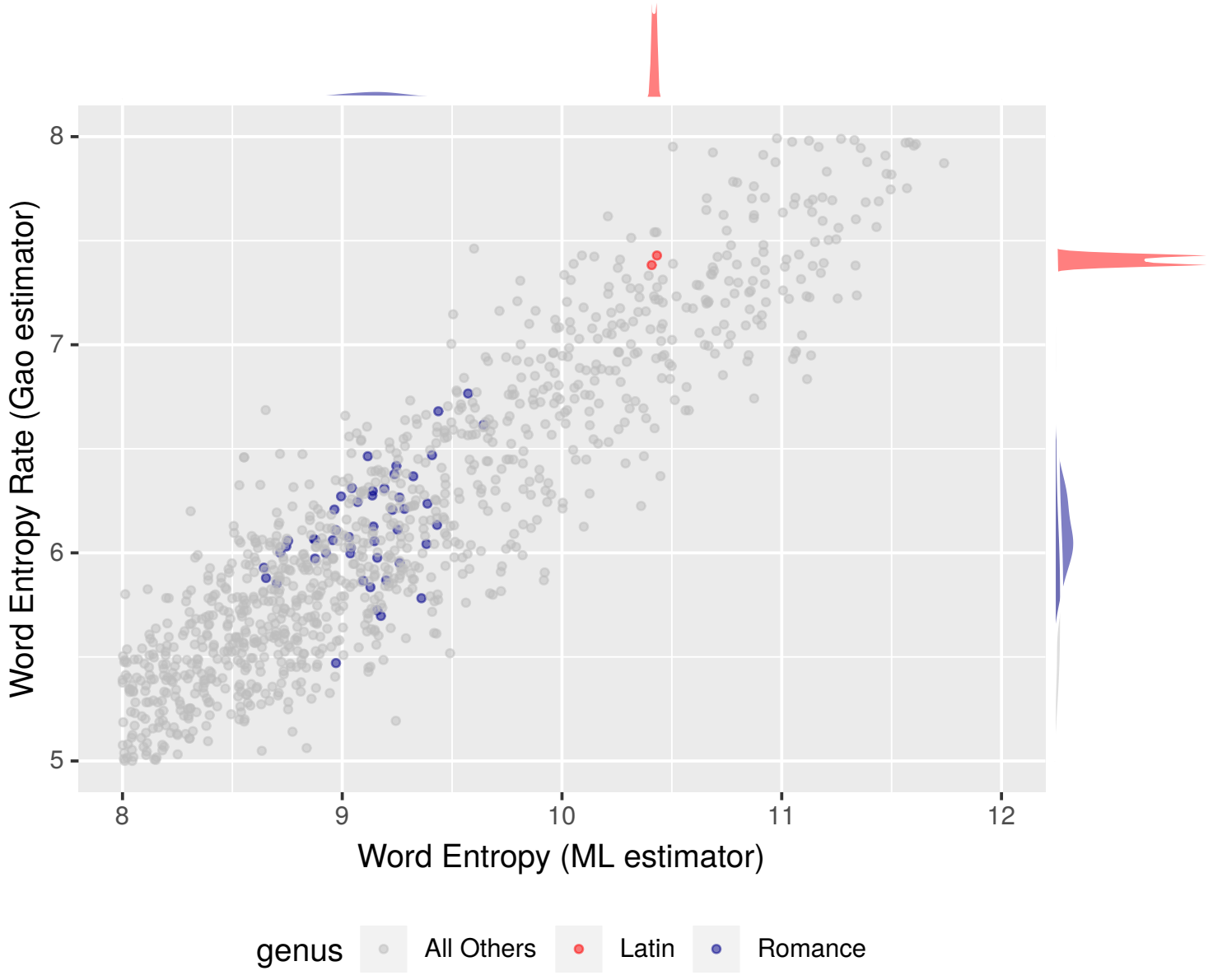
(3) Iñupiatun (esk, PBC 41006004)

Aglaan Jesus-ŋum 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.

Historic Change: Latin and Modern Romance



Simple Example: Word for “Brother” in the Bible

Classical Latin

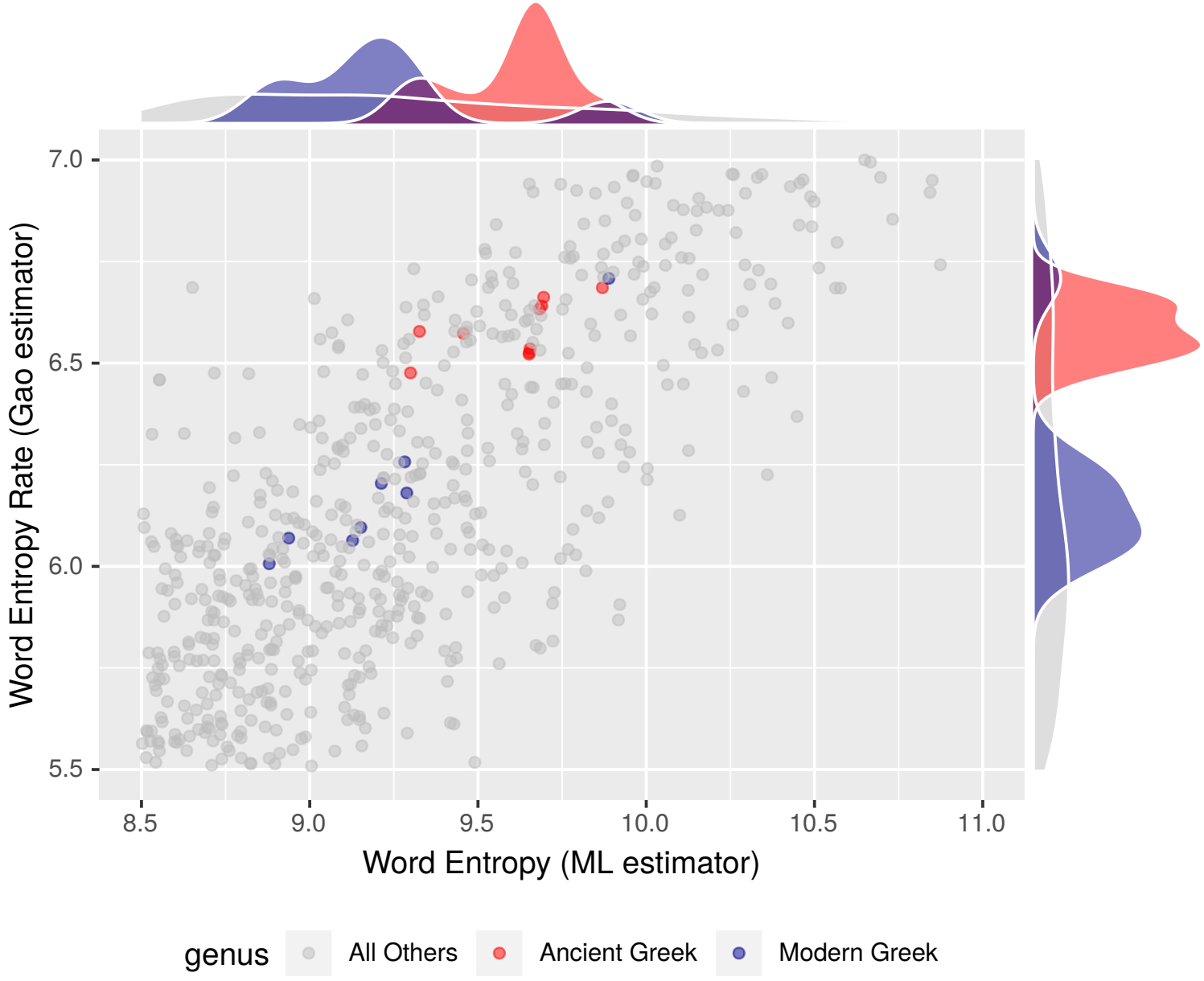
01004008 Dixitque Cain ad Abel **fratrem** suum [...]
01004009 Ubi est Abel **frater** tuus?
01004011 [...] suscepit sanguinem **fratris** tui de manu tua!

Italian

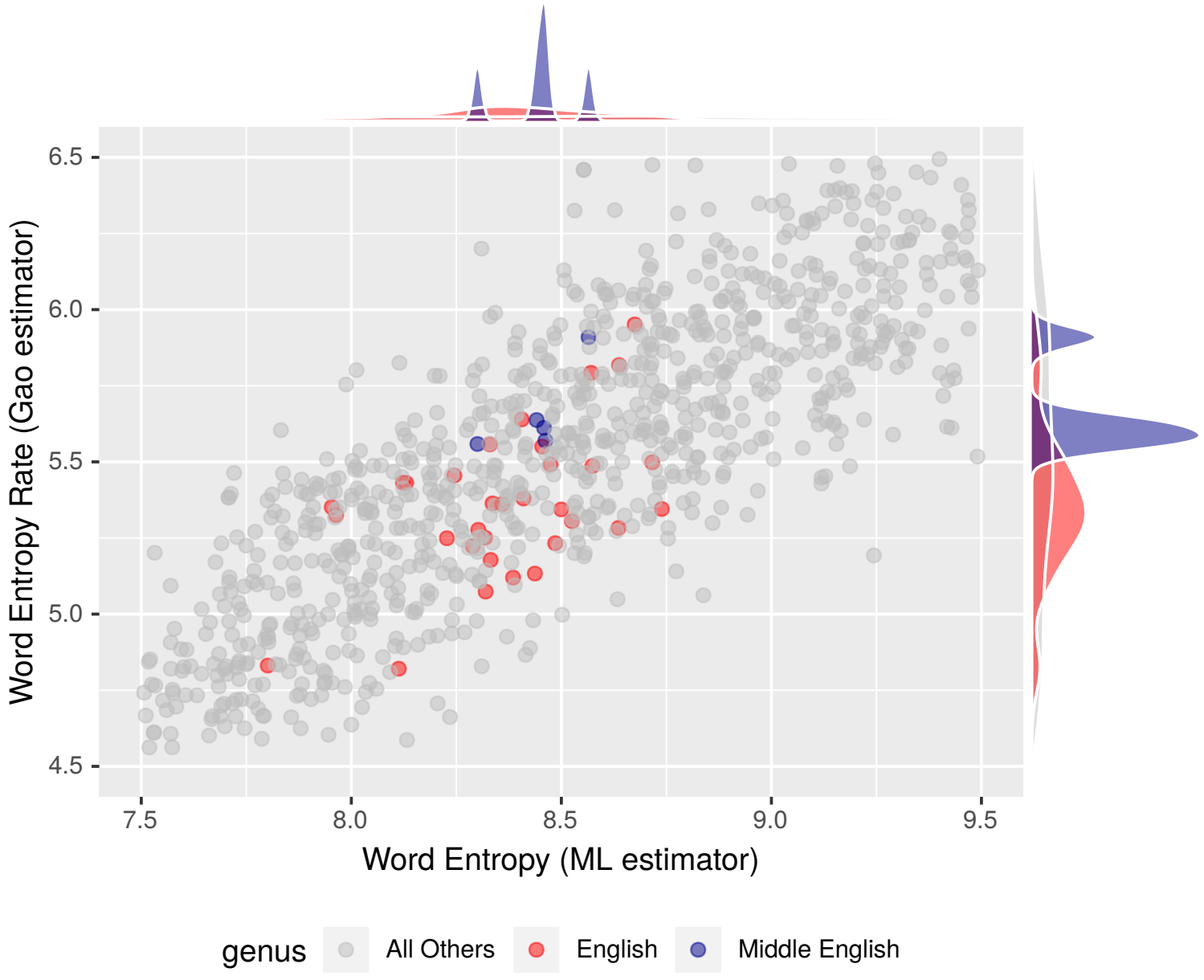
01004008 Caino disse al **fratello** Abele [...]
01004009 Dov’è Abele , tuo **fratello**?
01004011 [...] ha bevuto il sangue di tuo **fratello!**

Bentz & Berdicevskis (2016). Learning pressures reduce morphological complexity: linking corpus, computational and experimental evidence.

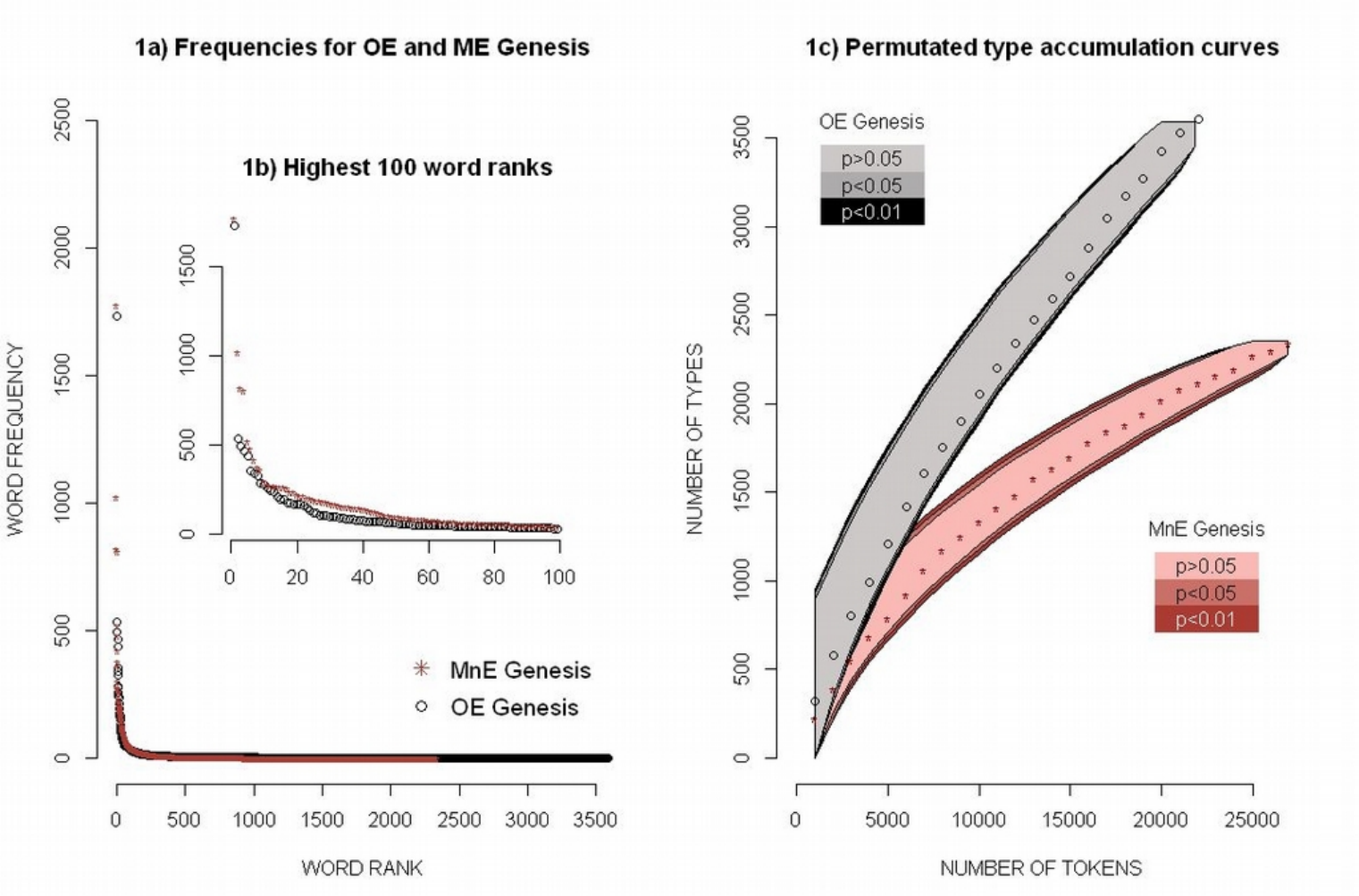
Historic Change: Ancient and Modern Greek



Historic Change: English



Historic Change: English



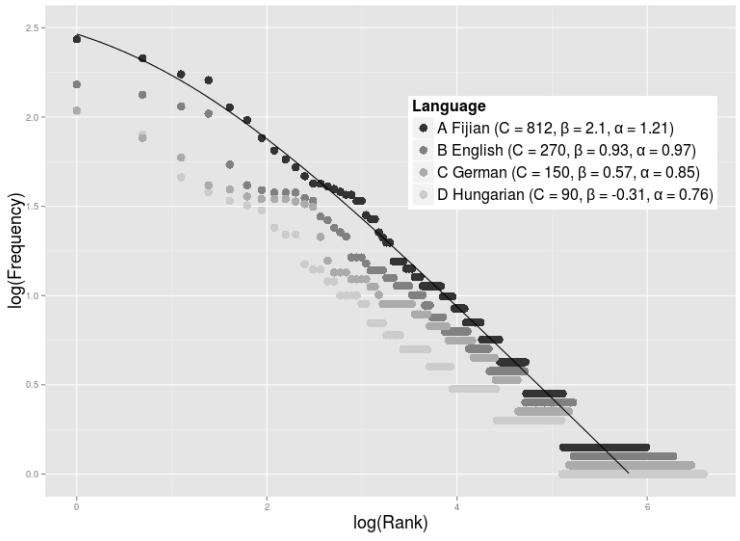
Bentz, Kiela, Hill, & Buttery (2014). Zipf's law and the grammar of languages.

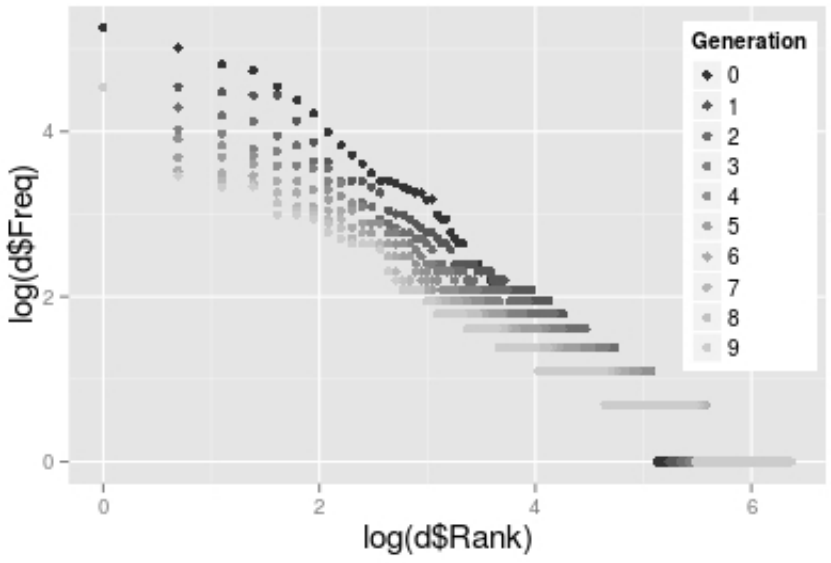
Discussion

Why do languages move towards low word entropy?

Author(s) & Year	Sociolinguistic Variable(s)	Language Structure
Sinnemäki (2009)	population size	argument marking.
Szmrecsanyi & Kortmann (2009)	L2 vs. L1 varieties	analyticity
Lupyan & Dale (2010)	population size	morphological compl.
Trudgill (2011)	various	morphological compl.
Bentz & Winter (2013)	L2 percentage	case compl.
Nichols (2013)	Altitude	morphological opacity
Bentz, Kiela, Hill & Buttery (2015)	L2 percentage	lexical diversity
Atkinson, Smith, & Kirby (2018)	L1 accommodation	morphological compl.
Sinnemäki & De Garbo (2018)	L1 and L2 sizes	gender, verbal morph.
Jon-And & Aguilar (2019)	L1 and L2 sizes	verbal morph.
Koplenig (2019)	L1 population size	morphological compl.
Raviv, Meyer, & Lev-Ari (2019)	population size	language structure
McWhorter (2019)	L2 influence	morphological compl.
Meinhardt, Malouf, & Ackerman (forth.)	neutral drift	morphological compl.

How do languages gain high word entropy?

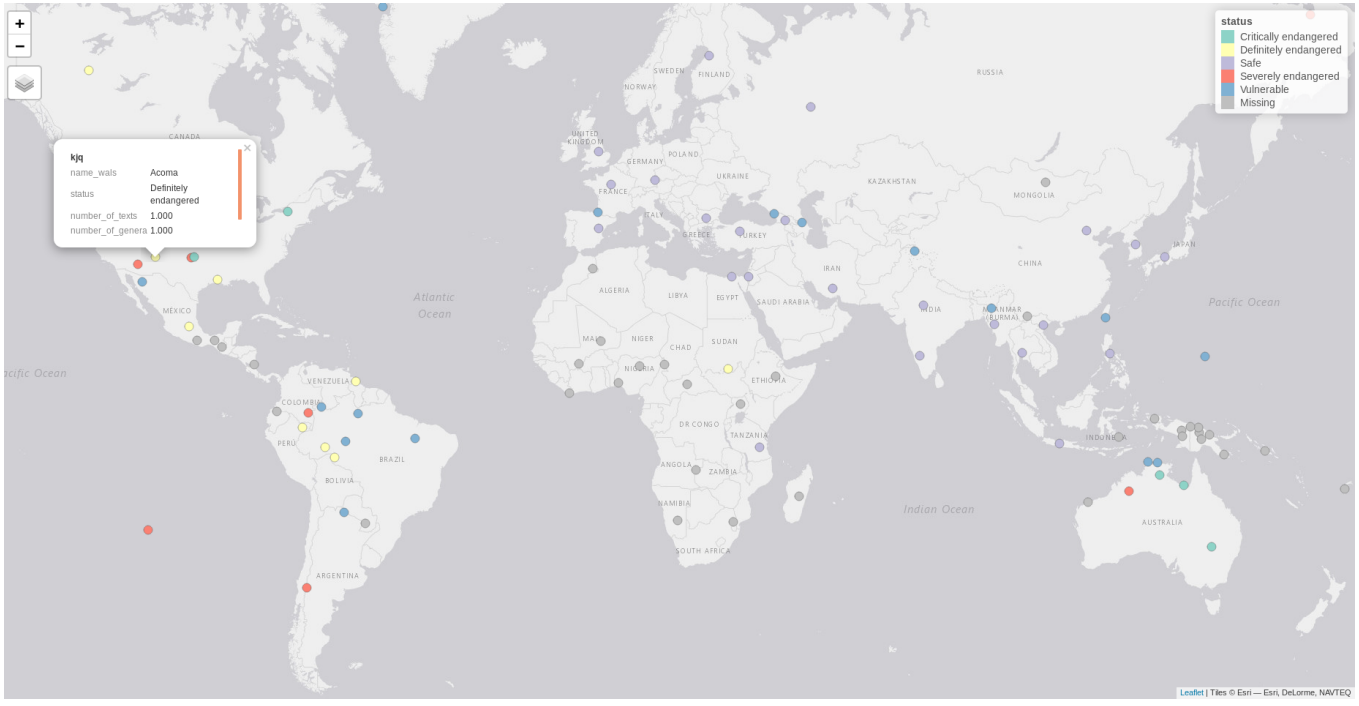




Bentz & Buttery (2014). Towards a computational model of grammaticalization and lexical diversity.

Future Research

The 100LC



SNF project “Non-randomness in morphological productivity”

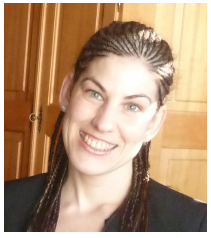
Tanja Samardžić

Olga Sozinova

Meta-Analyses of Morphosyntactic Complexity Measures

Participants

- Dominique Brunato & Giulia Venturi
- Ximena Gutierrez-Vasques
- Yoon Mi Oh
- Taraka Rama & Çağrı Çöltekin
- Kaius Sinnemäki & Vilijami Haakana
- Arturs Semenuks
- Olga Sozinova, Tanja Samardžić & Christian Bentz

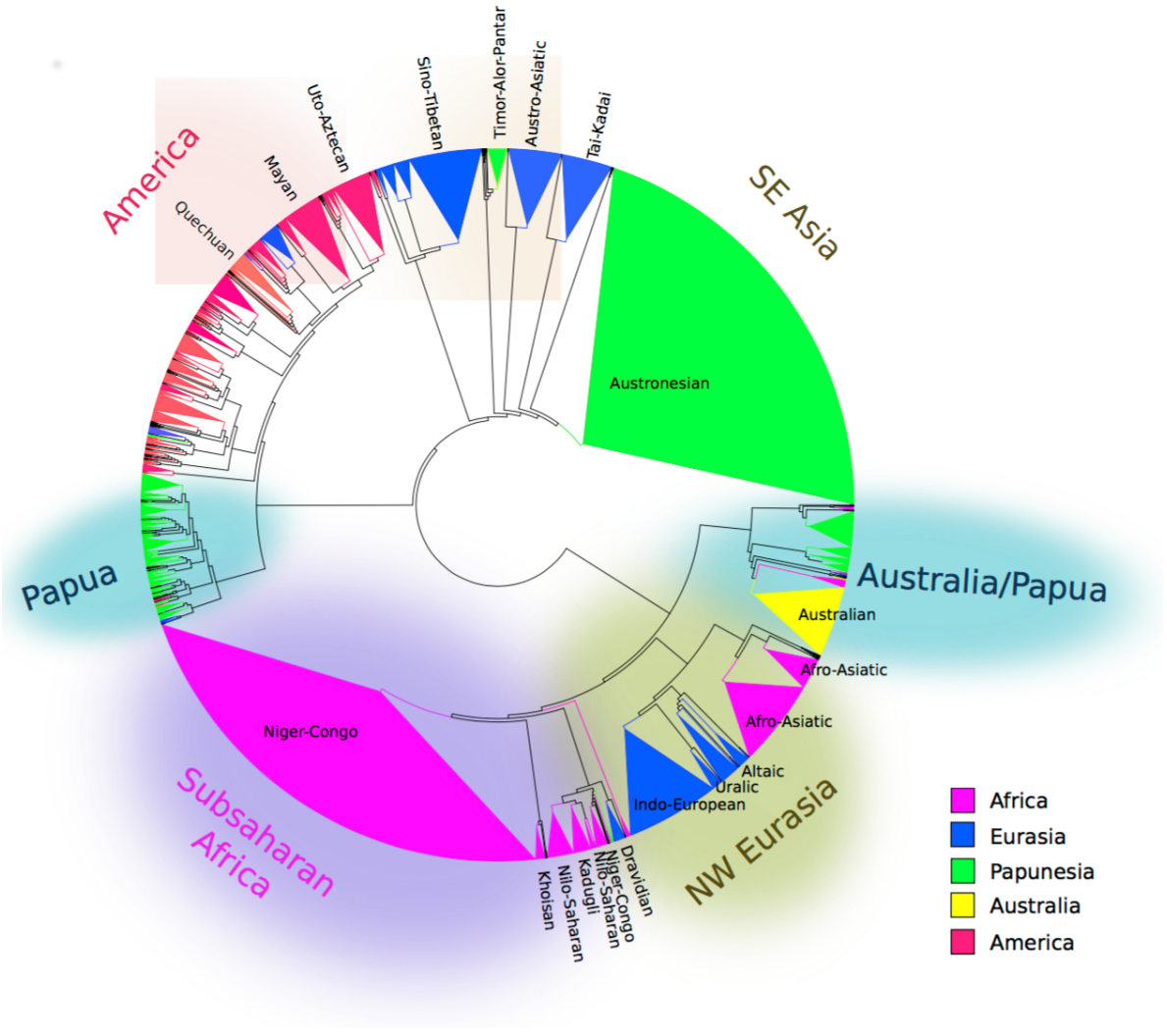


Katharina Ehret

Alice Blumenthal-Dramé

Aleksandrs Berdicevskis

Phylogenetic Analyses

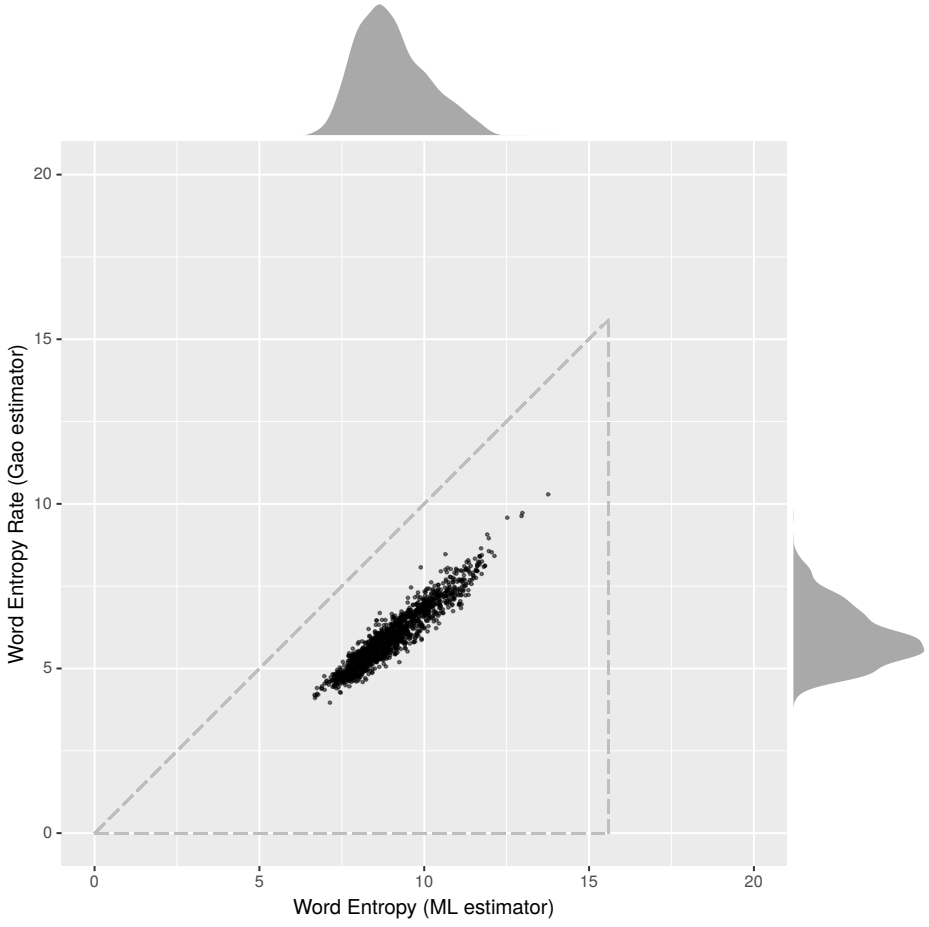


Gerhard Jäger

ERC Advanced Grant: *CrossLingference - Cross-Linguistic statistical inference using hierarchical Bayesian models*

Conclusion

Universality and Diversity



- Why are human languages constrained to relatively narrow entropy bands?
- What drives entropy differences within these bands?

Thank You