UNIVERSITY OF CAMBRIDGE

The Distribution of Phonemes across Languages: Chance, costs, and integration across linguistic tiers

Fermín Moscoso del Prado Martín[1,2] & Suchir Salhan[1,3]

[1]Department of Computer Science & Technology, University of Cambridge, UK
[2]Jesus College, University of Cambridge, UK
[3]Gonville & Caius College, University of Cambridge, UK

My Research Goals & General Method

The "Uninteresting" Aspects of Human Language

- In which aspects are human languages exactly as we should expect them to be?
- Information theory provides a powerful tool for this: the Principle of Maximum Entropy.

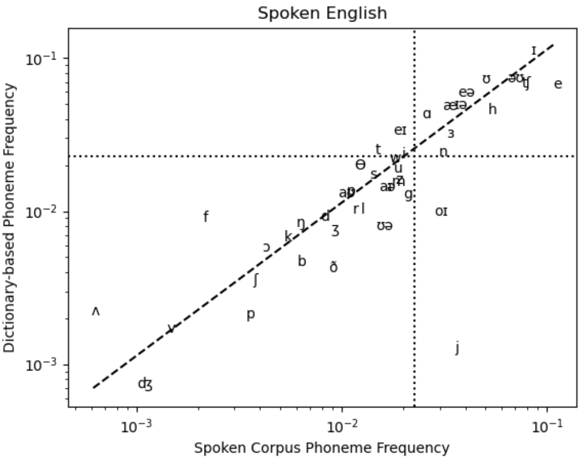
Making the Remaining Aspects less Interesting

- "Interesting" aspects indicate one is missing pieces of information: Constraints and costs
- Finding these missing bits brings us back to the "uninteresting" case

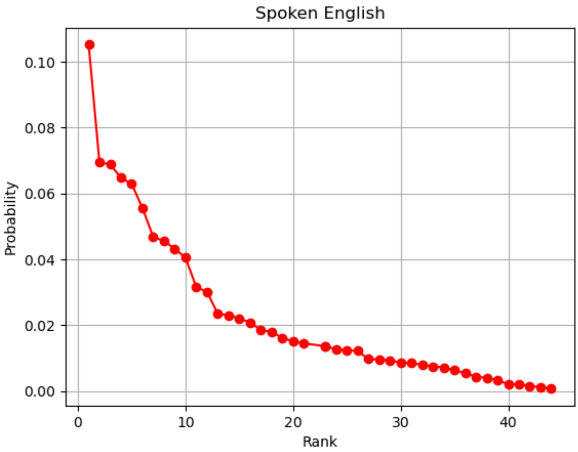
Areas of Interest

- Language processing and representation
- Dynamics of language at different timescales
  - Dialogue (seconds)
  - Acquisition and aging (years)
  - Language change (decades and beyond)
- What is/are the distribution(s) of linguistic structures across languages?
  - What is the distribution?
  - Why is it so?
  - What (if anything) do these distributions tell us about the nature and processing of human languages.

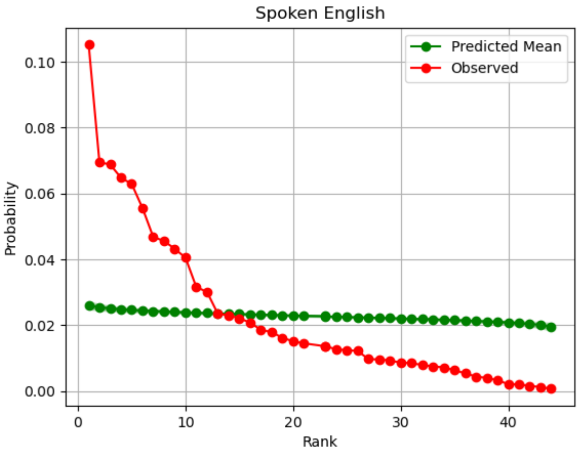
Phonemic Frequencies are Stable



Why are Phoneme Frequencies not plain Uniform?



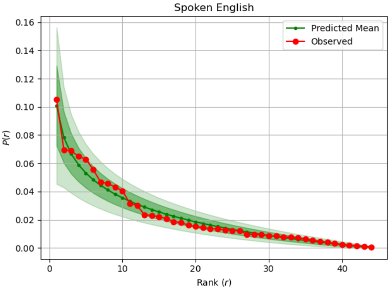
Why are Phoneme Frequencies not plain Uniform?



Questions

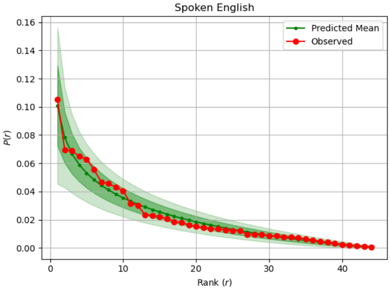
- What is the distribution of phoneme contrasts across languages?
- Is it a single distribution, or it depends on the language?
- Do such distributions reflect other aspects of language, beyond phonology?

This is what I would like



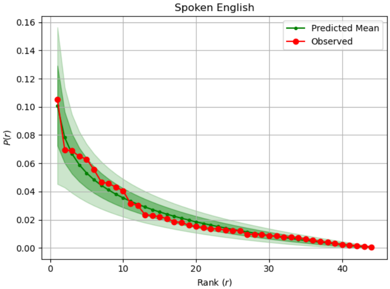
- I do not want to fit this curve

This is what I would like



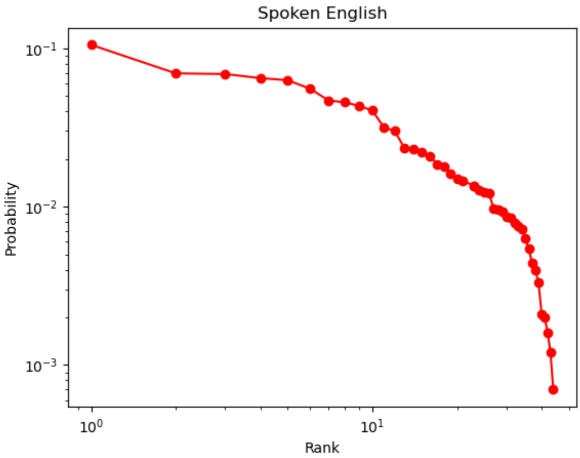
- I do not want to fit this curve
  "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk." (von Neumann, according to Fermi)

This is what I would like



- I do not want to fit this curve
  "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk." (von Neumann, according to Fermi)
- I want to compute this curve a priori

The Usual Tool: log-log Rank-Probability Plots



Proposed Distributions: We Want Power-laws!

Log-Series Model (Sigurd 1968)

$$p_{\mathrm{LS}}(r \mid \theta) = -\frac{\theta^{r}}{r \ln(1-\theta)}, \qquad 0 < \theta < 1$$

Yule-Simon Law (Martindale & Tambovtsev 2007)

$$p_{\mathrm{Y}}(r \mid \rho) = \rho \, \frac{\Gamma(r)\,\Gamma(\rho+1)}{\Gamma(r+\rho+1)}$$

‘Composite’ (Macklin-Cordes & Round 2020)

- Fit a power-law to the left tail (!?!) and something else for the right

How can V Phonemes Distributed?

- Before modelling communicative efficiency, preferential attachment, Martian intervention, …

How can V Phonemes Distributed?

- Before modelling communicative efficiency, preferential attachment, Martian intervention, ...
- it is good to see how far one can go with mere chance

How can V Phonemes Distributed?

- Before modelling communicative efficiency, preferential attachment, Martian intervention, ...
- it is good to see how far one can go with mere chance

Fact 1: Probabilities must add up to the unit

$$p_1 + p_2 + \ldots + p_V = \sum_{i=1}^{V} = 1$$

How can V Phonemes Distributed?

Fact 1: Probabilities must add up to the unit

Consequences

How can V Phonemes Distributed?

Fact 1: Probabilities must add up to the unit

Consequences

- The probabilities of phonemes observed in a corpus of N of phonemes, must follow a V-dimensional multinomial distribution with parameters $p_1, p_2, \ldots, p_V$, and $n = N$.

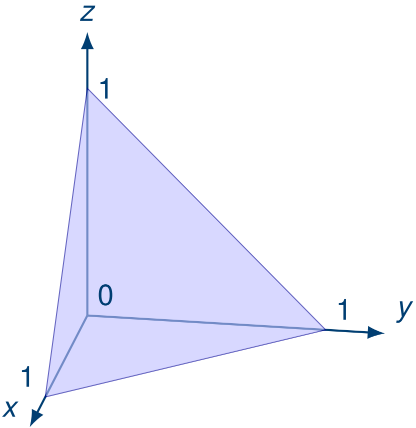
How can V Phonemes Distributed?

Fact 1: Probabilities must add up to the unit

Consequences

- The probabilities of phonemes observed in a corpus of N of phonemes, must follow a V-dimensional multinomial distribution with parameters $p_1, p_2, \ldots, p_V$, and $n = N$.
- We consider the p_i themsemselves as samples from a V-dimensional Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_N$.
  - Dirichlet is a distribution over distributions
  - All distributions on the $(V-1)$-simplex are possible samples from a V-dimensional Dirichlet distribution.

The (V-1)-Simplex



The 2-simplex (models 3 dimensional distributions)

How can V Phonemes Distributed? Start naïve

Fact 1: Probabilities must add up to the unit

How can V Phonemes Distributed? Start naïve

Fact 1: Probabilities must add up to the unit

Assumption: All phonemes are born equal

- We only know there are V distinct phonemic contrast

How can V Phonemes Distributed? Start naïve

Fact 1: Probabilities must add up to the unit

Assumption: All phonemes are born equal

- We only know there are V distinct phonemic contrast
- We have no additional information on the contrast themselves

How can V Phonemes Distributed? Start naïve

Fact 1: Probabilities must add up to the unit

Assumption: All phonemes are born equal

- We only know there are V distinct phonemic contrast
- We have no additional information on the contrast themselves
- Therefore, we cannot make any reasonable assumption that any phoneme i is more probable than phoneme j

How can V Phonemes Distributed? Start naïve

Fact 1: Probabilities must add up to the unit

Assumption: All phonemes are born equal

- We only know there are V distinct phonemic contrast
- We have no additional information on the contrast themselves
- Therefore, we cannot make any reasonable assumption that any phoneme i is more probable than phoneme j
- We must assume that the Dirichlet parameters $\alpha_1 = \alpha_2 = \ldots = \alpha_V = \alpha$

How can V Phonemes Distributed? Start naïve

Fact 1: Probabilities must add up to the unit

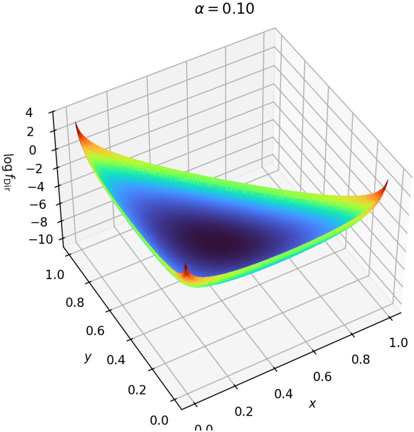
Assumption: All phonemes are born equal

- We only know there are V distinct phonemic contrast
- We have no additional information on the contrast themselves
- Therefore, we cannot make any reasonable assumption that any phoneme i is more probable than phoneme j
- We must assume that the Dirichlet parameters $\alpha_1 = \alpha_2 = \ldots = \alpha_V = \alpha$
- This is called a Symmetric Dirichlet Distribution with concentration parameter α.

Symmetric Dirichlet Distribution

- The single parameter α controls the likelihood of getting samples that are more or less centrally distributed within the simplex
  - $\alpha = 1$: all distributions within the simplex are equally probable (i.e., a uniform distribution over distributions)
  - $\alpha > 1$: more central (i.e., more uniform-like, higher entropy) distributions are preferred
  - $\alpha < 1$: more extreme (i.e., more skewed, lower entropy) distributions are preferred

Symmetric Dirichlet Distribution



Symmetric Dirichlet Distribution

- The single parameter α controls the likelihood of getting samples that are more or less centrally distributed within the simplex
  - $\alpha = 1$: all distributions within the simplex are equally probable (i.e., a uniform distribution over distributions)
  - $\alpha > 1$: more central (i.e., more uniform-like, higher entropy) distributions are preferred
  - $\alpha < 1$: more extreme (i.e., more skewed, lower entropy) distributions are preferred

Symmetric Dirichlet Distribution

- The single parameter α controls the likelihood of getting samples that are more or less centrally distributed within the simplex
  - $\alpha = 1$: all distributions within the simplex are equally probable (i.e., a uniform distribution over distributions)
  - $\alpha > 1$: more central (i.e., more uniform-like, higher entropy) distributions are preferred
  - $\alpha < 1$: more extreme (i.e., more skewed, lower entropy) distributions are preferred
- The marginals (i.e., the distribution of the individual p_i) are distributed according to a Beta$(\alpha, (V - 1)\alpha)$

Symmetric Dirichlet Distribution

- The single parameter α controls the likelihood of getting samples that are more or less centrally distributed within the simplex
  - $\alpha = 1$: all distributions within the simplex are equally probable (i.e., a uniform distribution over distributions)
  - $\alpha > 1$: more central (i.e., more uniform-like, higher entropy) distributions are preferred
  - $\alpha < 1$: more extreme (i.e., more skewed, lower entropy) distributions are preferred
- The marginals (i.e., the distribution of the individual p_i) are distributed according to a $\mathrm{Beta}(\alpha, (V-1)\alpha)$
- All p_i have the same distribution $\rightarrow$ an individual phoneme distribution (i.e., for a language) is a random sample of V values from a $\mathrm{Beta}(\alpha, (V-1)\alpha)$ distribution

Order Statistics (i.e., Rank values starting from below)

- In a sample of size V, sort the observations in non-decreasing order:

$$p_{(1)} \le p_{(2)} \le \cdots \le p_{(V)}.$$

  The notation $p_{(k)}$ (parentheses) distinguishes the ordered values from the raw observations $p_1, \ldots, p_V$.
- $\mathbf{p_{(1)}}$: first order statistic (the sample minimum).
- $\mathbf{p_{(V)}}$: V-th order statistic (the sample maximum).
- $\mathbf{p_{(k)}}$ for $1 \le k \le V$: the k-th smallest value in the sample (often interpreted as an empirical quantile).
- For a Beta distribution the order statistics involve difficult integrals, but they are easy to compute numerically

Symmetric Dirichlet Distribution: The Role of Entropy

- A sample of a V-dimensional symmetric Dirichlet distribution are probabilities $p_1, \ldots, p_V$

Symmetric Dirichlet Distribution: The Role of Entropy

- A sample of a V-dimensional symmetric Dirichlet distribution are probabilities $p_1, \ldots, p_V$
- A sample generated from this sampled distribution (i.e., the observed phonemes in a corpus) is expected to have an entropy:

$$H = \psi(\alpha V + 1) - \psi(\alpha + 1)$$

where ψ is the digamma function

Symmetric Dirichlet Distribution: The Role of Entropy

- A sample of a V-dimensional symmetric Dirichlet distribution are probabilities $p_1, \ldots, p_V$
- A sample generated from this sampled distribution (i.e., the observed phonemes in a corpus) is expected to have an entropy:

$$H = \psi(\alpha V + 1) - \psi(\alpha + 1)$$

  where ψ is the digamma function
- Estimation algorith for α. Given a sample of phonemes (ie., a corpus):
  1. Estimate H and V (possibly, with bias correction; for H, Chao et al., 2013; for V, Chao & Lee, 1992)
  2. Solve the equation numerically

Symmetric Dirichlet Distribution: The Role of Entropy

- A sample of a V-dimensional symmetric Dirichlet distribution are probabilities $p_1, \ldots, p_V$
- A sample generated from this sampled distribution (i.e., the observed phonemes in a corpus) is expected to have an entropy:

$$H = \psi(\alpha V + 1) - \psi(\alpha + 1)$$

  where ψ is the digamma function
- Estimation algorith for α. Given a sample of phonemes (ie., a corpus):
  1. Estimate H and V (possibly, with bias correction; for H, Chao et al., 2013; for V, Chao & Lee, 1992)
  2. Solve the equation numerically
- With α, we can compute the mean and variance of the predicted order statistics

Datasets I will Use

Universal Declaration of Human Rights

- 53 languages, transcribed using XPF (Cohen Priva et al., 2021)
- Typologically, genetically, and geographically diverse sample
- Imprecisions and missing phonemes (but bias corrections help)

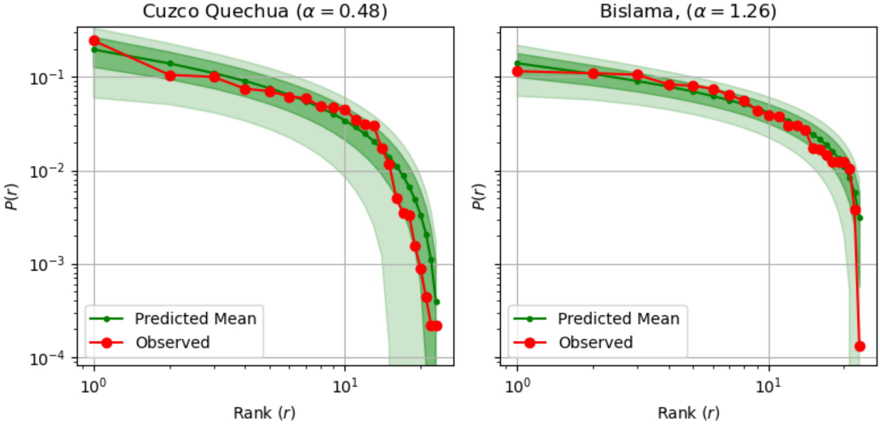
Australian Languages (Macklin-Cordes & Round, 2020)

- 166 Australian Language varieties
- Typologically, genetically, and geographically limited
- Accurate (each inventory curated by an expert)

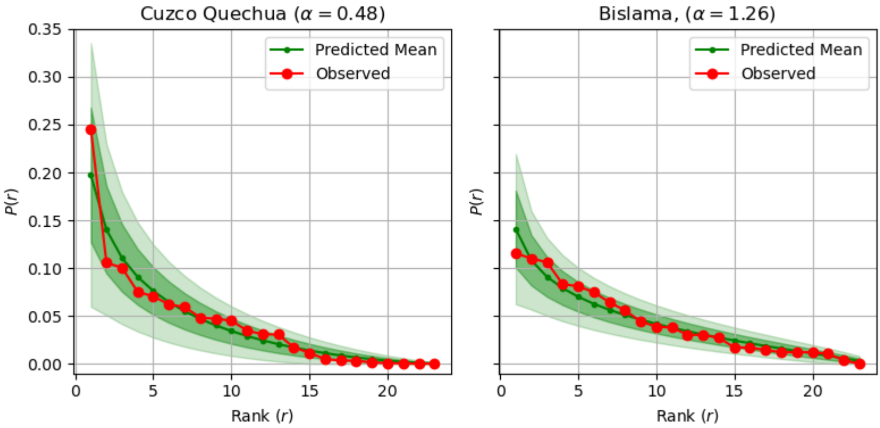
PHOIBLE (Moran & McCloy, 2019)

- removed duplicates, keeping most likely in case of disagreement
- 2,681 inventories, but no frequency distributions

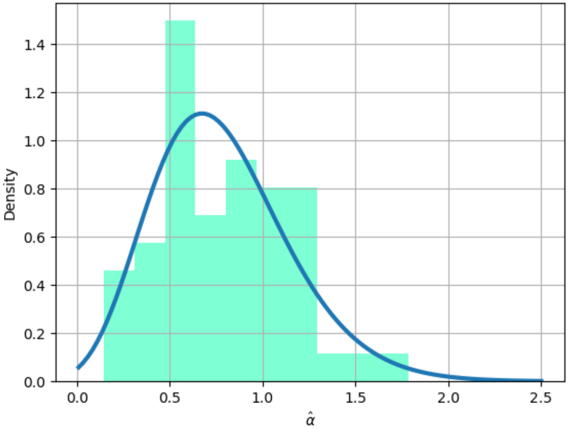
It FITS all 53 languages rather well



It FITS all 53 languages rather well



Do we really even need to fit it?

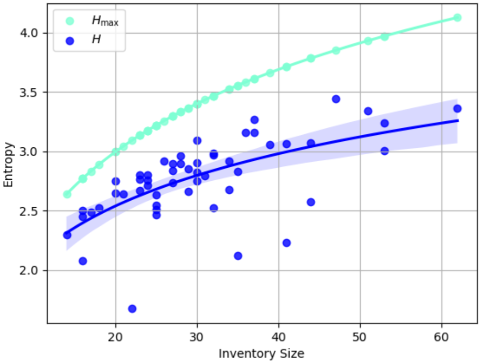


Do we really even need to fit it?

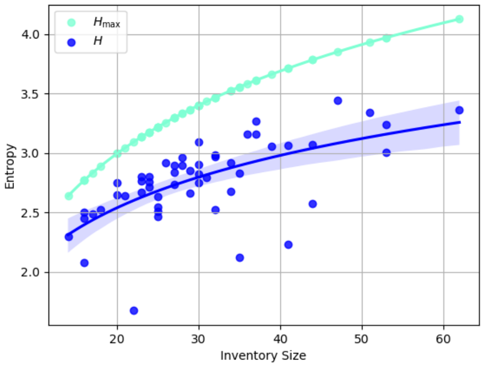
Prior	Dirichlet parameter	Interpretation
Laplace	$\alpha = 1$	Principle of indifference All probability distributions are equally likely Used for letters by Gusein-Zade (1988)
Jeffreys	$\alpha = .5$	Principle of consistency Any parametrisation should lead to the same choice
Observed	$\langle\hat{\alpha}\rangle = .78 \pm .05$	Very close to both priors

- Entropy and inventory size are inter-correlated across languages

- Entropy and inventory size are inter-correlated across languages

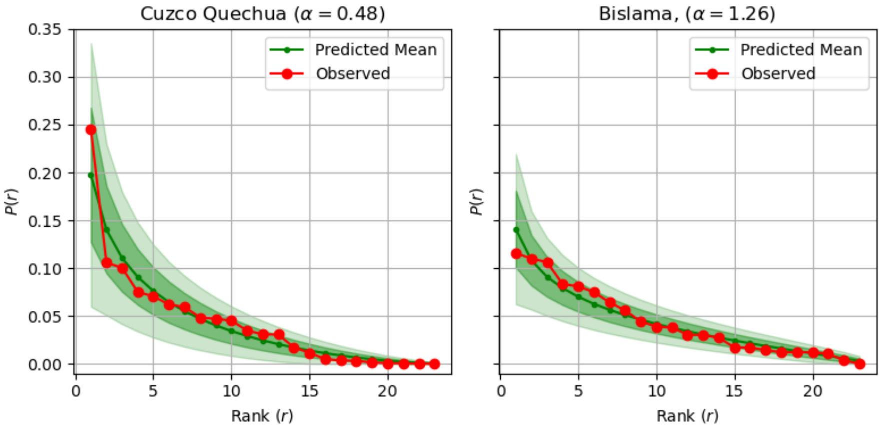


- Entropy and inventory size are inter-correlated across languages

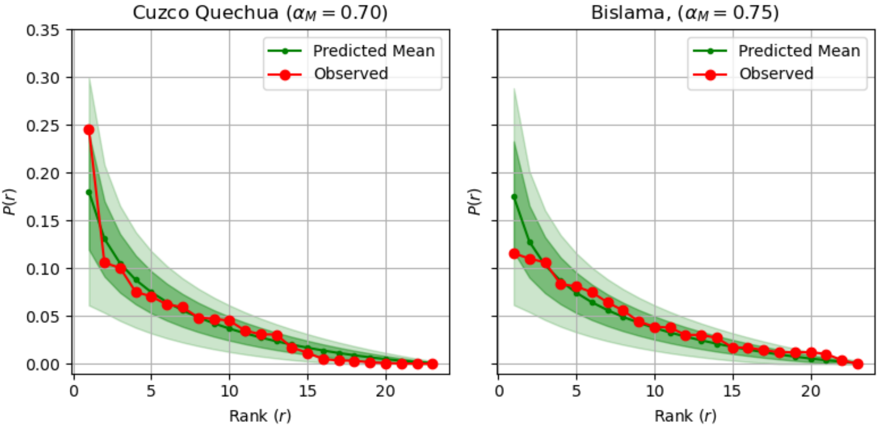


- One can predict H from V with a log-linear regression ($H \approx .64 \log V + .64$)
- Just two parameters, common for all languages (no more fitting)

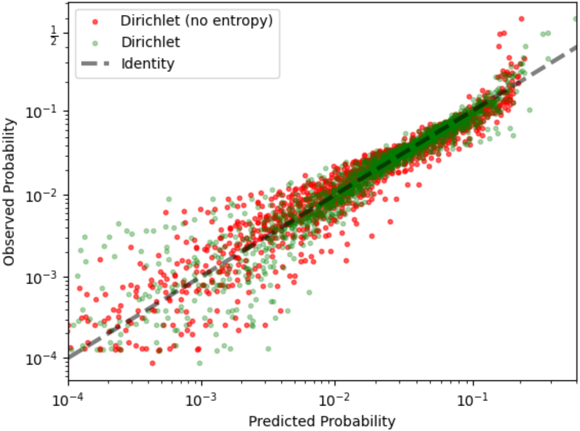
It FITS all 53 languages rather well



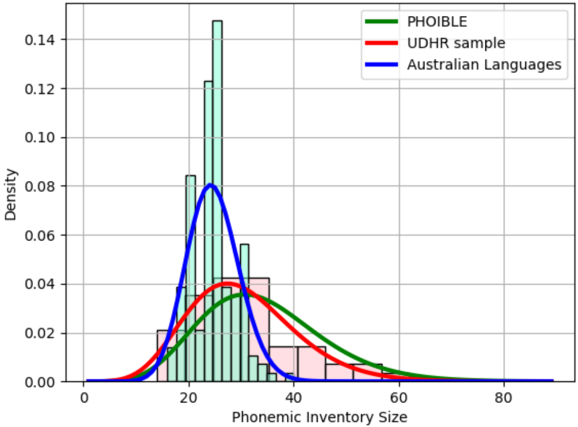
It PREDICTS all 53 languages rather well



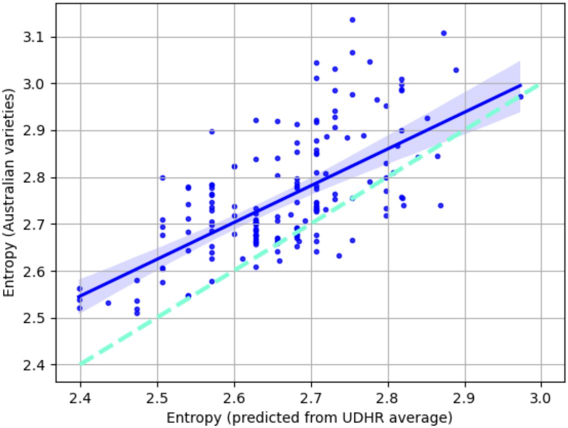
We can predict phoneme frequencies



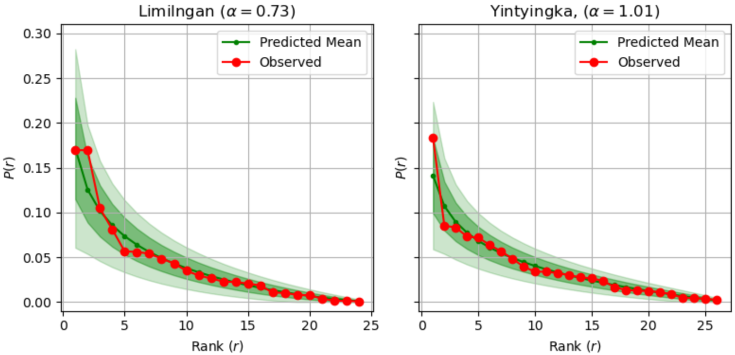
Also for the Australian languages



Also for the Australian languages

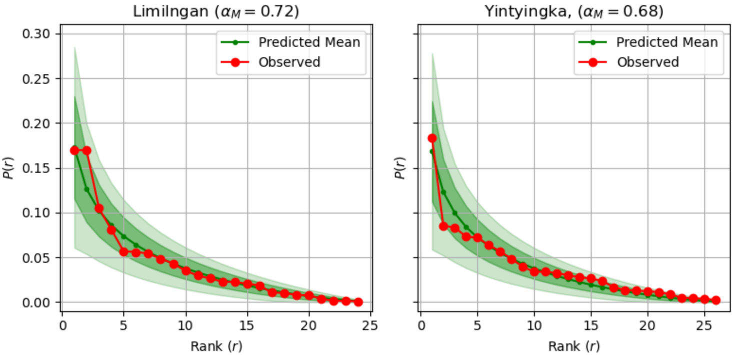


Also for the Australian languages



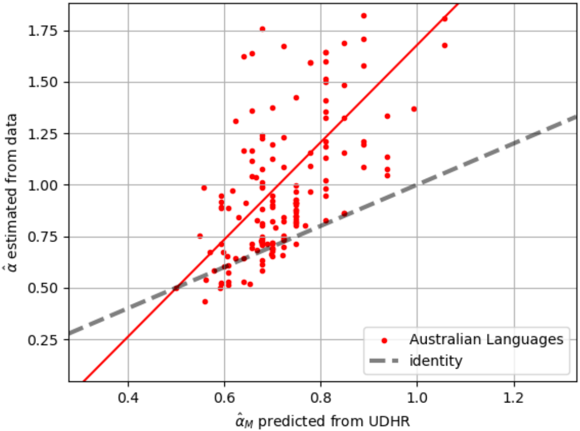
Fitted using the distributions

Also for the Australian languages

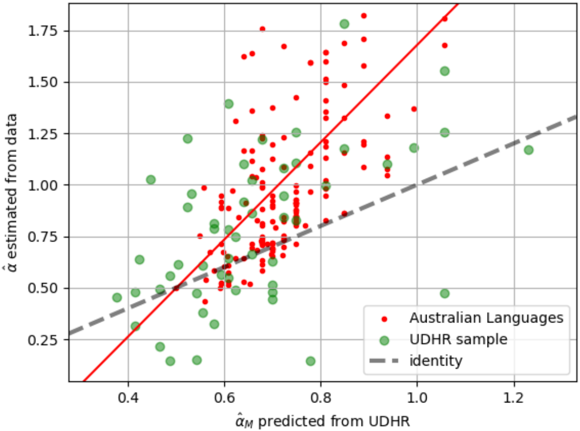


Computed directly using the regression parameters from the UDHR

Also for the Australian languages



Also for the Australian languages



Interim summary (I)

- The distribution of phonemes in the World's languages is roughly what would be expected by chance

Interim summary (I)

- The distribution of phonemes in the World's languages is roughly what would be expected by chance
- $H \approx .64 + .64 \log V \quad \rightarrow \quad H/H_{\max} \approx .64 + .64/\log V$
  - The entropy of phoneme distributions lies between 64% and 90% of the maximum entropy (i.e., with $V = 11$)
  - With a slight upper adjustment (larger inventories have lower entropies in relative terms)
  - More formal version of observations by Ladefoged & Maddieson (1996), Pierrehumbert (2001), and Moran & Blasi (2014).

Interim summary (I)

- The distribution of phonemes in the World's languages is roughly what would be expected by chance
- $H \approx .64 + .64 \log V \quad \rightarrow \quad H/H_{\max} \approx .64 + .64/\log V$
  - The entropy of phoneme distributions lies between 64% and 90% of the maximum entropy (i.e., with $V = 11$)
  - With a slight upper adjustment (larger inventories have lower entropies in relative terms)
  - More formal version of observations by Ladefoged & Maddieson (1996), Pierrehumbert (2001), and Moran & Blasi (2014).
- These two pieces of information, are sufficient for computing the probability distribution of a phonemic inventory a priori, with just two assumptions:
  1. Probabilities sum to one
  2. All phonemes are a priori equal


BARBRADY
ALRIGHT PEOPLE
Move along. There's nothing to see here.

The Principle of Maximum Entropy

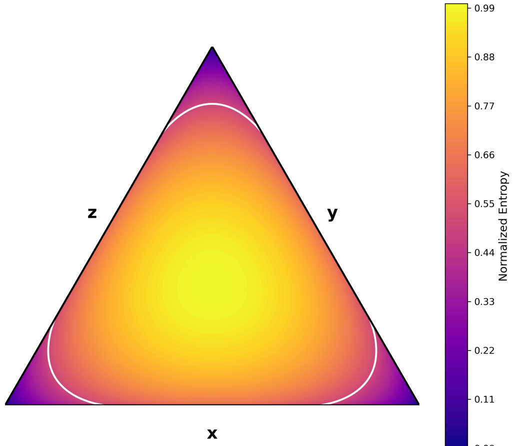
Maximum Entropy

- Among the possible distributions of $p_1, \ldots, p_V$
- The one that maximizes the entropy:

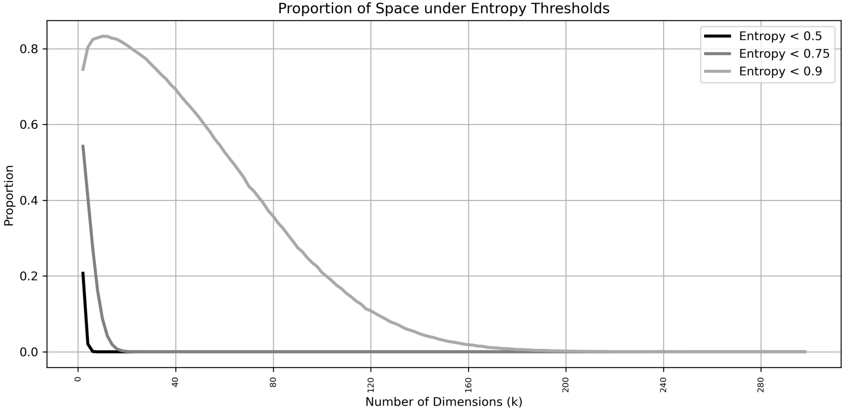
$$H = -p_1 \log p_1 - p_2 \log p_2 - \ldots - p_V \log p_V$$

is the most probable

Low-Entropy Distributions are Extremely Unlikely



Low-Entropy Distributions are Extremely Unlikely



The Principle of Maximum Entropy

Maximum Entropy

- Among the possible distributions of $p_1, \ldots, p_V$
- The one that maximizes the entropy is the most probable

$$\begin{aligned} \arg\max H[p_1, \ldots, p_V] &= -p_1 \log p_1 - \ldots - p_V \log p_V \\ &\text{s.t. } p_1 + \ldots + p_V = 1 \end{aligned}$$

The Principle of Maximum Entropy

Maximum Entropy

- Among the possible distributions of $p_1, \dots, p_V$
- The one that maximizes the entropy is the most probable

$$\begin{aligned} \arg\max H[p_1, \dots, p_V] &= -p_1 \log p_1 - \dots - p_V \log p_V \\ &\text{s.t. } p_1 + \dots + p_V = 1 \end{aligned}$$

- We should expect $p_1 = p_2 = \dots = p_V = 1/V$.

The Principle of Maximum Entropy

Maximum Entropy

- Among the possible distributions of $p_1, \ldots, p_V$
- The one that maximizes the entropy is the most probable

$$\arg\max H[p_1, \ldots, p_V] \quad = -p_1 \log p_1 - \ldots - p_V \log p_V$$
$$\text{s.t. } p_1 + \ldots + p_V = 1$$

- We should expect $p_1 = p_2 = \ldots = p_V = 1/V$.
- As long as there aren't any additional constraints

The Principle of Maximum Entropy

Maximum Entropy

- Among the possible distributions of $p_1, \ldots, p_V$
- The one that maximizes the entropy is the most probable

$$\arg\max H[p_1, \ldots, p_V] = -p_1 \log p_1 - \ldots - p_V \log p_V$$
$$\text{s.t. } p_1 + \ldots + p_V = 1$$

- We should expect $p_1 = p_2 = \ldots = p_V = 1/V$.
- As long as there aren't any additional constraints
- Additional constraints can only decrease maximum entropy

The Principle of Maximum Entropy

Maximum Entropy subject to Constraints

- One can introduce costs for each alternative

$$c_1, c_2, \ldots, c_V \geq 0$$

- This restricts the possible distributions to those that have a given average cost (or, functionally equivalent, to those that optimise the average entropy to cost ratio)

$$\begin{aligned} \arg\max H[p_1, \ldots, p_V] \quad &= -p_1 \log p_1 - \ldots - p_V \log p_V \\ &\text{s.t.} \begin{cases} p_1 + \ldots + p_V = 1 \\ p_1 c_1 + \ldots + p_V c_V = \mathbb{E}[C] \end{cases} \end{aligned}$$

Solution to Maximum Entropy subject to k Constraints

The Principle of Maximum Entropy

Solution to Maximum Entropy subject to k Constraints

- Gibbs-Boltzman distribution

$$p_i = \frac{1}{Z[\boldsymbol{\lambda}]} \mathrm{e}^{\sum_{j=1}^{k} \lambda_j c_{i,j}}$$

where $\boldsymbol{\lambda} = \lambda_1, \ldots, \lambda_k$ are Lagrange multipliers

The Principle of Maximum Entropy

Solution to Maximum Entropy subject to k Constraints

- Gibbs-Boltzman distribution

$$p_i = \frac{1}{Z[\boldsymbol{\lambda}]} e^{\sum_{j=1}^{k} \lambda_j c_{i,j}}$$

where $\boldsymbol{\lambda} = \lambda_1, \ldots, \lambda_k$ are Lagrange multipliers

- matches a log-linear regression model without interactions:

$$\log p_i = \lambda_0 + \sum_{j=1}^{k} \lambda_j c_{i,j}$$

where $\lambda_0 = -\log Z[\boldsymbol{\lambda}]$

Possible Costs of Phonemes

"Physical" Costs (articulatory/perceptual)

Possible Costs of Phonemes

"Physical" Costs (articulatory/perceptual)

- Some phonemes are more difficult to perceive and/or articulate

Possible Costs of Phonemes

"Physical" Costs (articulatory/perceptual)

- Some phonemes are more difficult to perceive and/or articulate
- These costs should be relatively language independent

Possible Costs of Phonemes

"Physical" Costs (articulatory/perceptual)

- Some phonemes are more difficult to perceive and/or articulate
- These costs should be relatively language independent
- Consider the distinction (Gotelli & Chao, 2013)

Abundance : Frequency of a phoneme in a given language. e.g., 3.2% of the phonemes in spoken English are instances of /m/ (the 10th most frequent)

Incidence : Frequency of a phoneme across languages. e.g., 96.8% of the world's languages contain the contrast /m/ (the most frequent)

Possible Costs of Phonemes

"Physical" Costs (articulatory/perceptual)

- Some phonemes are more difficult to perceive and/or articulate
- These costs should be relatively language independent
- Consider the distinction (Gotelli & Chao, 2013)

  Abundance: Frequency of a phoneme in a given language. e.g., 3.2% of the phonemes in spoken English are instances of /m/ (the 10$^{\text{th}}$ most frequent)

  Incidence: Frequency of a phoneme across languages. e.g., 96.8% of the world's languages contain the contrast /m/ (the most frequent)
- Hypothesis: Incidence and abundance frequencies are correlated

Possible Costs of Phonemes

"Physical" Costs (articulatory/perceptual)

Phonotactic costs

- Human languages are redundant (Shannon, 1951)

Possible Costs of Phonemes

"Physical" Costs (articulatory/perceptual)

Phonotactic costs

- Human languages are redundant (Shannon, 1951)
- The more a phoneme is predictable from its context, the more it's likely to be elided (Cohen Priva, 2015)
  - Diachronically this would leave traces in the distribution.
  - Hypothesis: more predictable phonemes should be less frequent

Possible Costs of Phonemes

"Physical" Costs (articulatory/perceptual)

Phonotactic costs

- Human languages are redundant (Shannon, 1951)
- The more a phoneme is predictable from its context, the more it's likely to be elided (Cohen Priva, 2015)
  - Diachronically this would leave traces in the distribution.
  - Hypothesis: more predictable phonemes should be less frequent
- Phonotactic surprisal of phonemes (van Son & Pols, 2003)

$$I(/p/) = \left\langle -\log \frac{\text{Frequency}(<\text{word onset}> + /p/)}{\text{Frequency}(<\text{word onset}> + /k/)} \right\rangle,$$

Possible Costs of Phonemes

“Physical” Costs (articulatory/perceptual)

Phonotactic costs

Lexical costs

Possible Costs of Phonemes

“Physical” Costs (articulatory/perceptual)

Phonotactic costs

Lexical costs

- Phonemic contrasts serve to distinguish words

Possible Costs of Phonemes

"Physical" Costs (articulatory/perceptual)

Phonotactic costs

Lexical costs

- Phonemic contrasts serve to distinguish words
- Hypothesis Phonemes need to resolve word indentities ($H_{\text{phonemes}} \propto H_{\text{words}}$)

Possible Costs of Phonemes

“Physical” Costs (articulatory/perceptual)

Phonotactic costs

Lexical costs

- Phonemic contrasts serve to distinguish words
- Hypothesis Phonemes need to resolve word indentities ($H_{\text{phonemes}} \propto H_{\text{words}}$)
- Lexical surprisal of phonemes

$$I_L(/p/) = \left\langle -\log \frac{\text{Frequency}(\text{<word onset>} + /p/)}{\text{Frequency}(\text{<word onset>})} \right\rangle$$

Possible Costs of Phonemes

"Physical" Costs (articulatory/perceptual)

- Hypothesis: Incidence and abundance frequencies are correlated.

Phonotactic costs

- Hypothesis: Phonemes appearing in more predictable contexts should be less frequent

Lexical costs

- Hypothesis $H_{\text{phonemes}} \propto H_{\text{words}}$

Testing the Hypotheses

- Log-linear Mixed Effects Regression model

  Indep. var. : Abundance of a phoneme in a language (log)

  Dependent vars.
  - Average phonotactic surprisal on UDHR
  - (log) Incidence frequency from PHOIBLE
  - Average lexical surprisal on UDHR

  Random effect: Language variety (and possible random slopes)

Testing the Hypotheses

- Log-linear Mixed Effects Regression model

  Indep. var. : Abundance of a phoneme in a language ($\log$)

  Dependent vars.
  - Average phonotactic surprisal on UDHR
  - (log) Incidence frequency from PHOIBLE
  - Average lexical surprisal on UDHR

  Random effect: Language variety (and possible random slopes)
- Result (without interactions, and a random slope of phonotactic surprisal)

	β	t	p		
Phonotactic Surprisal	0.23	4.11	0.00	$\rightarrow$	Phonotactic cost ✓
$\log P_{\text{incidence}}$	0.76	22.84	0.00	$\rightarrow$	Physical cost ✓
Lexical Surprisal	-0.80	-20.78	0.00	$\rightarrow$	Lexical cost ✓

(Note: $\log V$ was residualised out of both other predictors to avoid collinearity, and all were standardised to $N[0, 1]$ for effect magnitude comparability.)

From the costs, we guess the phoneme probabilities

MaxEnt as a very extreme regression

From the costs, we guess the phoneme probabilities

MaxEnt as a very extreme regression

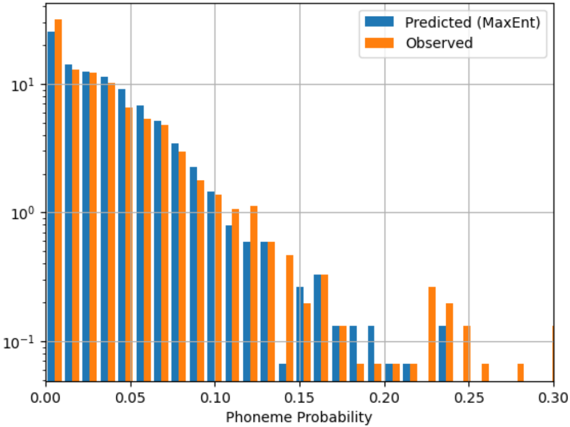
- As in regression, we are given the values of k independent variables (the costs $c_{i,j}$)
- We also have their average values ($C_j = \mathbb{E}_i[c_{i,j}]$)

From the costs, we guess the phoneme probabilities

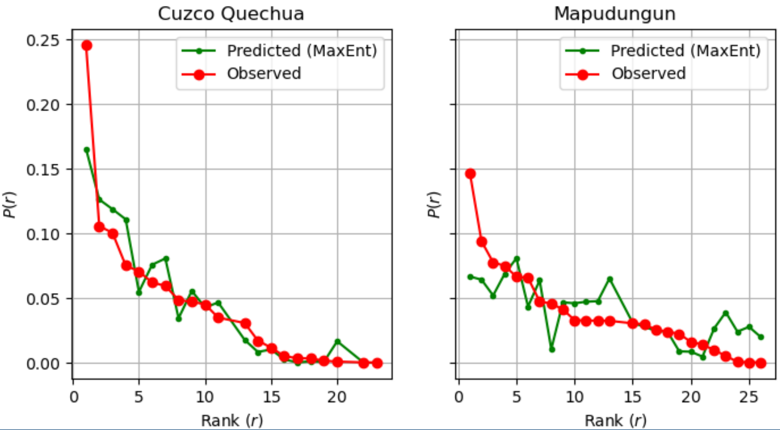
MaxEnt as a very extreme regression

- As in regression, we are given the values of k independent variables (the costs $c_{i,j}$)
- We also have their average values ($C_j = \mathbb{E}_i[c_{i,j}]$)
- Our task is to guess
  - the regression coefficients (λ_j) and
  - the actual probabilities (p_i)

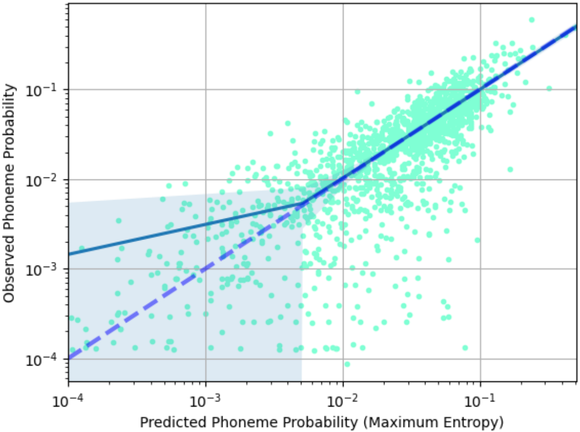
From the costs, we guess the phoneme probabilities



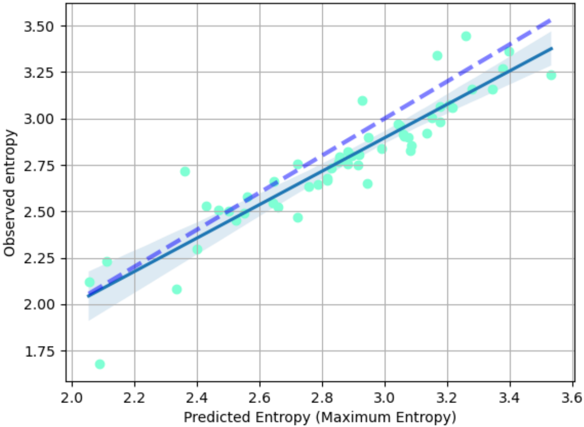
From the costs, we guess the phoneme probabilities



From the costs, we guess the phoneme probabilities



We are missing some constraints



"Costs"

“Costs”

- “Cost” is a term used in the maximum entropy literature. It refers to the individual values of constraints.

“Costs”

- “Cost” is a term used in the maximum entropy literature. It refers to the individual values of constraints.
- It does not necessarily correspond to the cognitive concept of cost.

"Costs"

- "Cost" is a term used in the maximum entropy literature. It refers to the individual values of constraints.
- It does not necessarily correspond to the cognitive concept of cost.
- This is not evidence for cost optimization or efficiency in the cognitive sense
- Rather, it is just a way of indicating that some magnitudes must have a finite mean
  - Whether this is the result of evolutionary optimisation is a different question.

Interim Summary (II)

- Maximum Entropy can reconstruct the distribution of phonemes

Interim Summary (II)

- Maximum Entropy can reconstruct the distribution of phonemes
- The distribution of phonemes, in isolation reflects aspects of
  - Perceptual/articulatory factors
  - The phonotactic structure of a language
  - The lexical richness of a language

Interim Summary (II)

- Maximum Entropy can reconstruct the distribution of phonemes
- The distribution of phonemes, in isolation reflects aspects of
  - Perceptual/articulatory factors
  - The phonotactic structure of a language
  - The lexical richness of a language
- The different representational tiers are interrrelated

Interim Summary (II)

- Maximum Entropy can reconstruct the distribution of phonemes
- The distribution of phonemes, in isolation reflects aspects of
  - Perceptual/articulatory factors
  - The phonotactic structure of a language
  - The lexical richness of a language
- The different representational tiers are interrrelated
- However, a plain chance analysis seems to achieve the best reconstruction
  - However, the MaxEnt approach is addressing a tougher problem, not just guessing the probabilities for each rank, but also which specific phoneme occupies each rank position.
  - Entropy is the single crucial aspect of the phoneme distribution

$$H = \psi(V\alpha + 1) + \psi(\alpha + 1)$$

$$H = \psi(V\alpha + 1) + \psi(\alpha + 1)$$

Back to Symmetric Dirichlet: a Costs Model

$$H = \psi(V\alpha + 1) + \psi(\alpha + 1)$$

Physical costs

$$H = \psi\big(V\alpha + 1\big) + \psi\big(\alpha + 1\big)$$

Physical costs

Phonotactic costs

Back to Symmetric Dirichlet: a Costs Model

$$H = \psi(V\alpha + 1) + \psi(\alpha + 1)$$

Physical costs

Phonotactic costs

Lexical costs

Summary

- We are able to compute the distribution of phonemes (not just fit it)

Summary

- We are able to compute the distribution of phonemes (not just fit it)
  - From two basic assumptions: Normalisation and symmetry,

Summary

- We are able to compute the distribution of phonemes (not just fit it)
  - From two basic assumptions: Normalisation and symmetry,
  - or from considerations on the properties of individual phonemes

Summary

- We are able to compute the distribution of phonemes (not just fit it)
  - From two basic assumptions: Normalisation and symmetry,
  - or from considerations on the properties of individual phonemes
- Constraining the probability space works by fixing the entropy of the distribution to values lower than its maximum.

A Sneak Peak

A Sneak Peak

- This work is part of a larger project, which I call The –right– end of Zipf's Laws

A Sneak Peak

- This work is part of a larger project, which I call The –right– end of Zipf's Laws
- In further work, I mathematically demonstrate under which conditions can power-law and non-power-law distribution of linguistic structures arise.

A Sneak Peak

- This work is part of a larger project, which I call The –right– end of Zipf's Laws
- In further work, I mathematically demonstrate under which conditions can power-law and non-power-law distribution of linguistic structures arise.
- One of the consequences is that what appear as power-laws in language, are most likely illusions

A Sneak Peak

- This work is part of a larger project, which I call The –right– end of Zipf's Laws
- In further work, I mathematically demonstrate under which conditions can power-law and non-power-law distribution of linguistic structures arise.
- One of the consequences is that what appear as power-laws in language, are most likely illusions
- This also enables developing a single unified distribution for linguistic structures at any tier.

Thank you!

(image by Bob Tubbs, Public domain, via Wikimedia Commons)

References

- Chao, A., & Lee, S.-M. (1992). Estimating the Number of Classes via Sample Coverage. Journal of the American Statistical Association 87, 210–217
- Chao, A. et al. (2013). Entropy and the species accumulation curve: A novel entropy estimator via discovery rates of new species. Methods in Ecology and Evolution 4.
- Cohen Priva, U. (2015) Informativity affects consonant duration and deletion rates. Laboratory Phonology 6, 243–278.
- Cohen Priva, U. et al. (2021). The Cross-linguistic Phonological Frequencies (XPF) Corpus manual
- Gotelli, N. & Chao, A.. (2013). Measuring and Estimating Species Richness, Species Diversity, and Biotic Similarity from Sampling Data. Encyclopedia of Biodiversity 5 (pp. 195-211)
- Gusein-Zade, S. M. (1988). О распределении букв русского языка по частоте встречаемости [On the distribution of Russian language letters by frequency]. Проблемы передачи информации 24(4), 102–107.
- Ladefoged, P., & Maddieson, I. (1996) The sounds of the world's languages. Oxford: Blackwell Publishers.
- Moran, S. & McCloy, D. (2019) PHOIBLE 2.0. Jena: Max Planck Institute for the Science of Human History.
- Moran, S., & Blasi, D. E. (2014). Cross-linguistic comparison of complexity measures in phonological systems. In Proc. 10th International Seminar on Speech Production (ISSP 2014), Cologne.

References

- Macklin-Cordes, J. L., & Round, E. R. (2020). Re-evaluating phoneme frequencies. Frontiers in Psychology 11, 570895
- Martindale, C., & Tambovtsev, Y. (2007). Phoneme frequencies follow a Yule distribution. SKASE Journal of Theoretical Linguistics 4(2), 1–11.
- Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. L. Bybee & P. Hopper (Eds.), Frequency and the emergence of linguistic structure (pp. 137–157). Amsterdam: John Benjamins.
- Shannon, C. E. (1951). Prediction and Entropy of Printed English. Bell System Technical Journal 30(1), 50–64.
- Sigurd, B. (1968). Rank-frequency distributions for phonemes. Phonetica 18(1), 1–15.
- van Son, R. J. J. H. & Pols, L. C. W. (2003). How efficient is speech? Proceedings of the Institute of Phonetic Sciences 25 pp. 171–184.

Order Statistics

- General pdf of the k^{th} order statistic

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!\,(n-k)!}\left[F(x)\right]^{k-1}\left[1-F(x)\right]^{n-k} f(x), \qquad 0 < x < 1.$$

- Expected value (solved numerically)

$$\mathbb{E}[X_{(k)}] = \int_0^1 x\,\frac{n!}{(k-1)!\,(n-k)!}\,[F(x)]^{k-1}[1-F(x)]^{n-k} f(x)\,dx.$$

- Variance & standard error (solved numerically)

$$\mathrm{Var}[X_{(k)}] = \int_0^1 x^2\,\frac{n!}{(k-1)!\,(n-k)!}\,[F(x)]^{k-1}[1-F(x)]^{n-k} f(x)\,dx - \left(\mathbb{E}[X_{(k)}]\right)^2,$$