How many language types are there?

Max Bane
bane@uchicago.edu

University of Chicago

January 7, 2011
85th Annual Meeting of the LSA, Pittsburgh

Statement of the problem

- Given a sample of languages
  - Ex: Chamorro, Malakmalak, Turkmen, . . .
- . . . each of which can be labeled with a type
  - Ex: penultimate stress, initial stress, final stress, . . .
- . . . how many types did we fail to see?

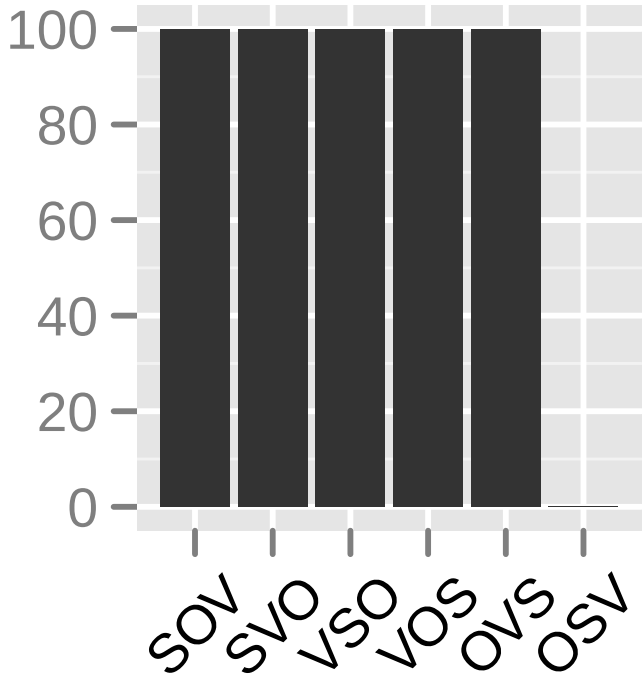
- Known in statistics as the Unseen Species Problem.

A hypothetical example

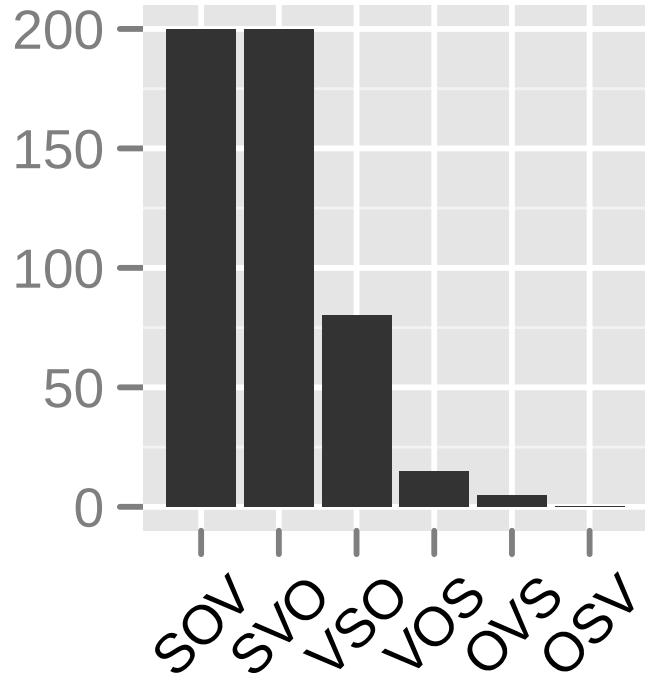
- Imagine a sample of 500 languages labeled for basic word order (subject, object, verb).
  - We know there are 6 logical possibilities.
  - SOV, SVO, VSO, VOS, OVS, OSV
- Consider two scenarios in which OSV is not attested in the sample. . .

A hypothetical example

Scenario 1

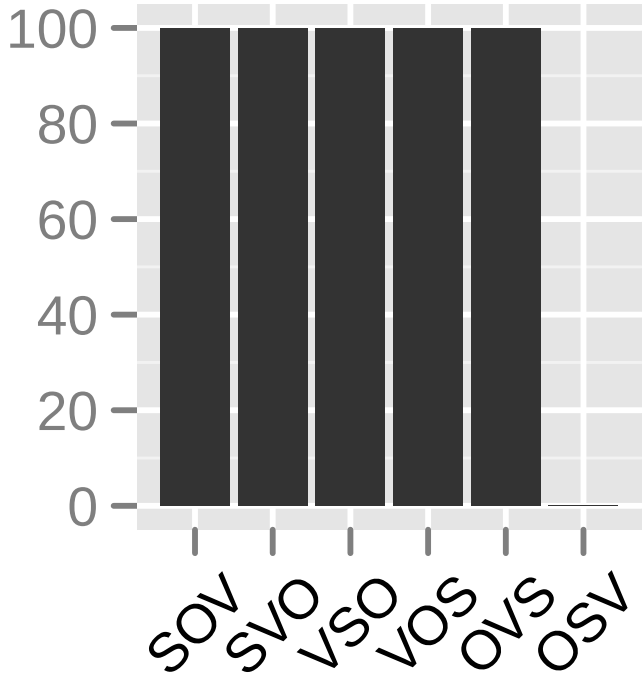


Scenario 2

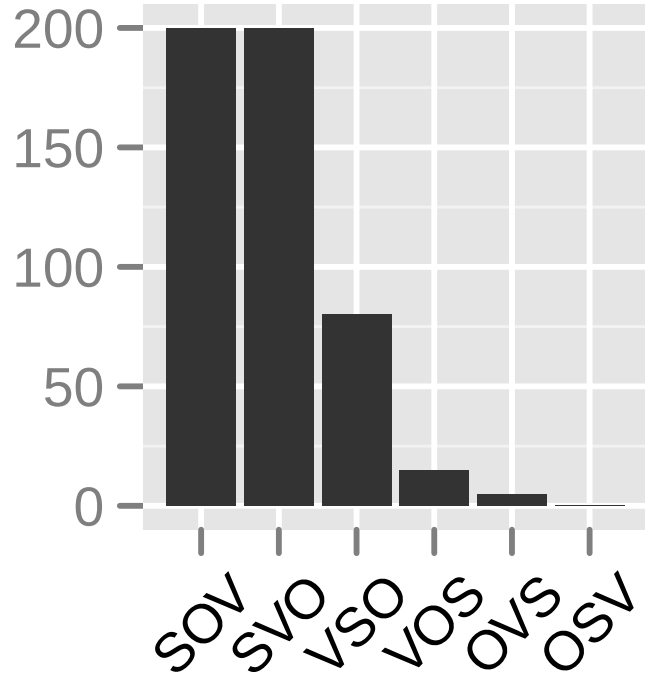


A hypothetical example

Scenario 1



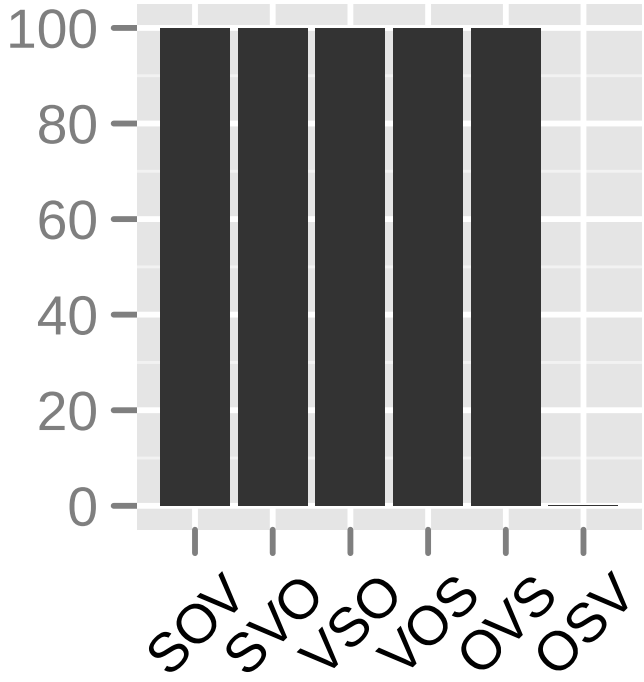
Scenario 2



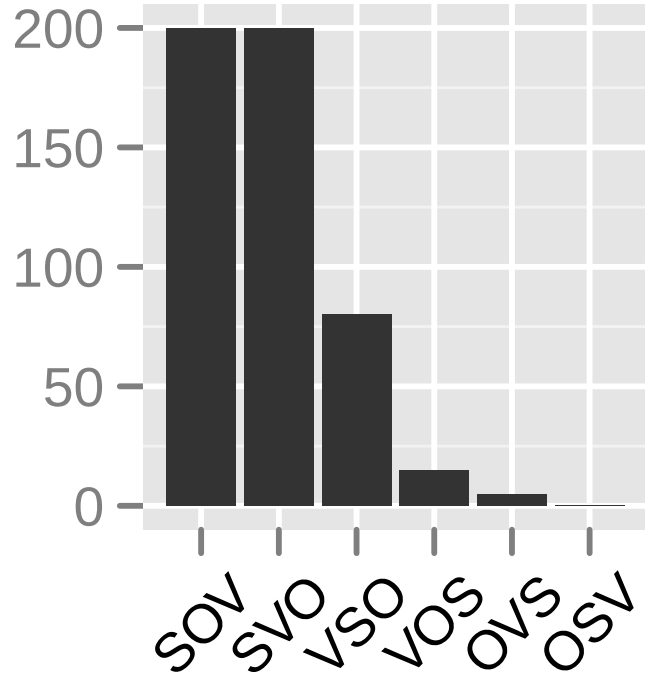
- Intuitively, the probability that we've missed OSV purely by chance seems greater in Scenario 2.

A hypothetical example

Scenario 1



Scenario 2



- Intuitively, the probability that we've missed OSV purely by chance seems greater in Scenario 2.
- Whereas in Scenario 1 it's more likely to be missing because it doesn't exist.

Open classes

- In many cases we face an open class of possible types (Greenberg 1978).
  - No logical bound on the number of possible types (or it is very large).
- Ex: stress systems, syllable inventories, segment inventories.
- We can't expect a sample of languages to capture all existing types.
- Don't necessarily know which missing types to look for.
- Are types missing because they don't exist, or because they are rare?

The question

- Given the observed number of types in a sample, and their frequencies, how many total types are likely to be in the sampled population?
- Unseen species (types) estimation methods (Gandolfi and Sastri 2004, Bunge and Fitzpatrick 1993)

The relevance for linguists

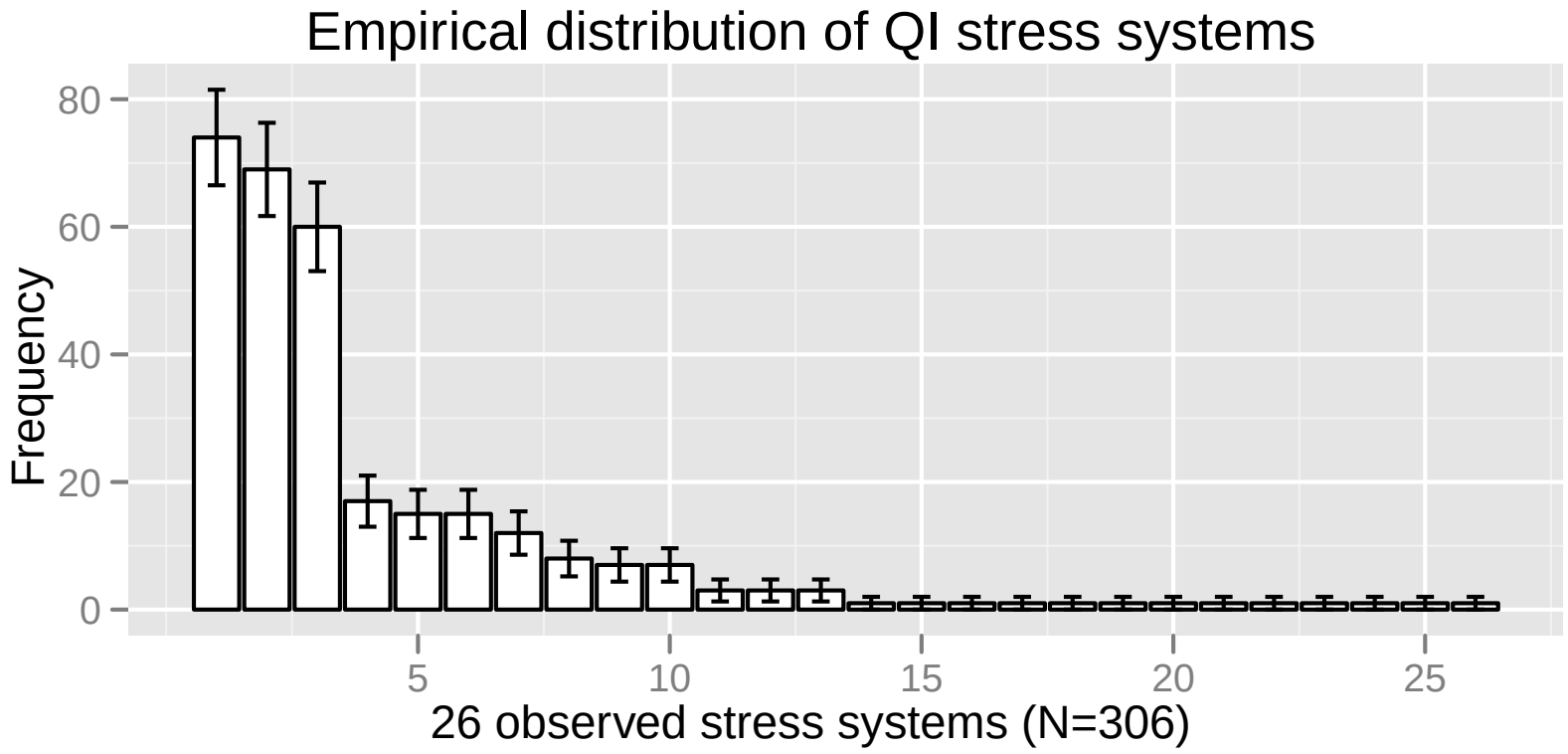
- Lexical richness
  - Estimating the size of a lexicon given a sample of observed word types and their frequencies (Baayen 2001).
- Typology
  - Estimating the total number of language types in the world given the observed types and their frequencies.
  - Assessing how representative a typological sample is.
  - Implications for theories that make predictions about the number of possible types.

This talk

- Two running examples
  - Quantity insensitive (QI) stress.
  - Syllable structure.
- One method of unseen species estimation (Zhang and Stern 2009).
- Applied to QI stress and syllable structure
- Implications for typologists and theoretical linguists
  - Overgeneration and “under-overgeneration”.

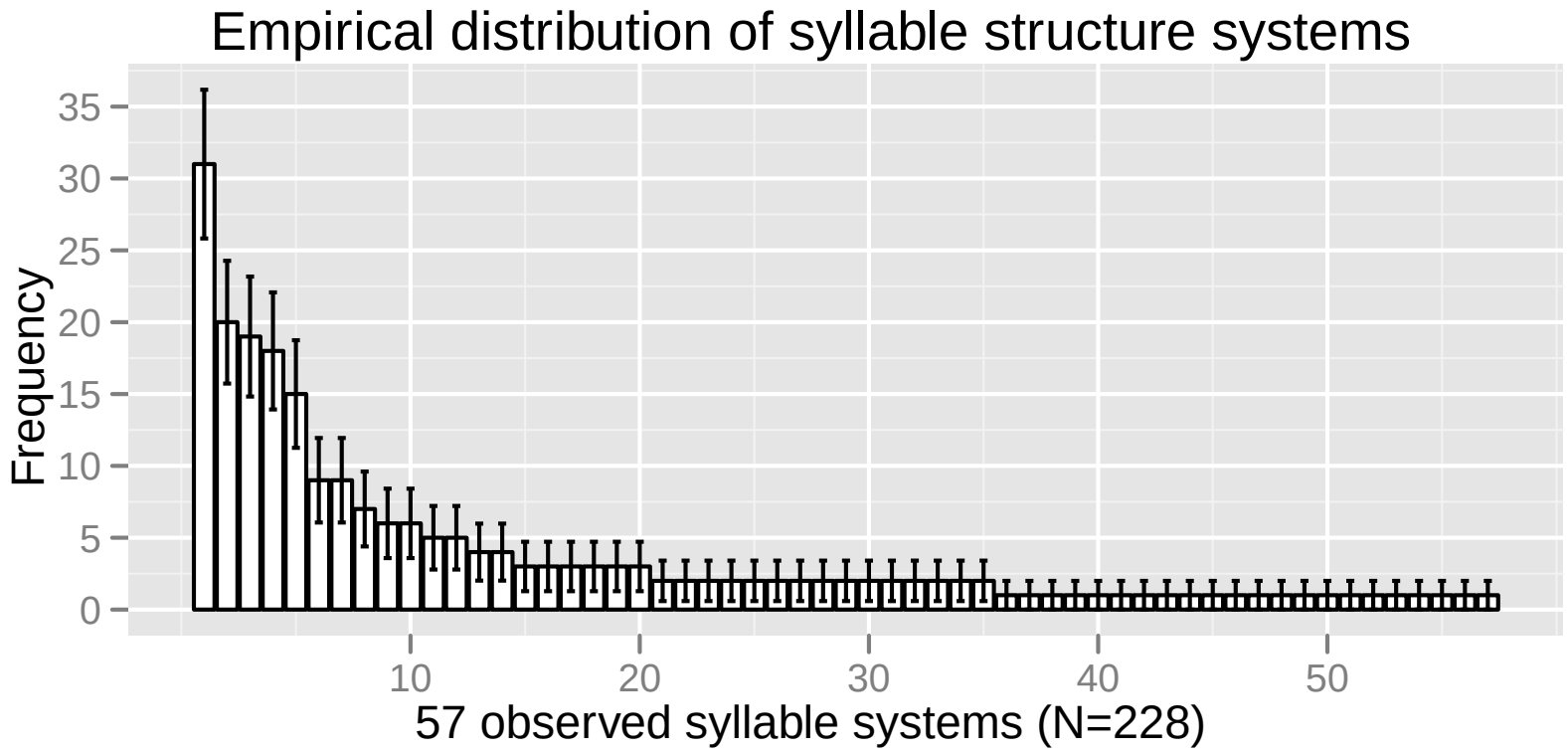
A typological sample of QI stress systems

- From the U Delaware Stress Pattern Database (Heinz et al. 2010).
- 306 languages, 26 observed types.
- Primary and secondary stress placement.
- Fixed, dual, binary, ternary systems.
- Example types:
  - Fixed final primary (74 languages).
    - $\sigma\acute{\sigma}$, $\sigma\sigma\acute{\sigma}$, $\sigma\sigma\sigma\acute{\sigma}$, etc.
  - Odd from left, leftmost primary (15 languages).
    - $\acute{\sigma}\sigma$, $\acute{\sigma}\sigma\grave{\sigma}$, $\acute{\sigma}\sigma\grave{\sigma}\sigma$, $\acute{\sigma}\sigma\grave{\sigma}\sigma\grave{\sigma}$, etc.
  - Initial primary if $\leq$ 3 syllables, else even from right with leftmost primary (1 language; Malakmalak).
    - $\acute{\sigma}$, $\acute{\sigma}\sigma$, $\acute{\sigma}\sigma\sigma$, $\acute{\sigma}\sigma\grave{\sigma}\sigma$, $\sigma\acute{\sigma}\sigma\grave{\sigma}\sigma$, $\acute{\sigma}\sigma\grave{\sigma}\sigma\grave{\sigma}\sigma$, etc.


Empirical distribution of QI stress systems
Frequency
80
60
40
20
0
5
10
15
20
25
26 observed stress systems (N=306)

A typological sample of syllable structure systems

- From the Syllable Typology Database (van der Hulst and Goedemans 2009)
- Syllable inventory reduced to CV-template.
  - All consonants collapsed to **C**, vowels to **V**.
  - Length distinct from sequence: **V:** vs **VV**
  - Optionality and disjunction encoded.
- 228 languages, 57 observed types.
- Example types:
  - (C)V (20 languages)
  - CV(C) (19 languages)
  - (C)V(:)(C) (7 languages)
  - C(C)V(:|C)(C) (1 language; Poqomchi')


Empirical distribution of syllable structure systems
Frequency
35
30
25
20
15
10
5
0
10
20
30
40
50
57 observed syllable systems (N=228)

Many rare types

- Note that both distributions are very skewed.
- A few common types, and a long tail of many rare types.
- Seems to be a common situation in human language (cf. word distributions, segment inventories, etc.)
- Therefore:
  - We don't expect all types in the population to be present in the samples.
  - Many will be missed because most types are rare.

Many rare types

- Note that both distributions are very skewed.
- A few common types, and a long tail of many rare types.
- Seems to be a common situation in human language (cf. word distributions, segment inventories, etc.)
- Therefore:
  - We don't expect all types in the population to be present in the samples.
  - Many will be missed because most types are rare.

- But the distribution of observed types can tell us something about the plausible number of types in the population!

Approaches to the unseen species problem

- Good-Turing estimate (Good 1953)
  - Estimate of the probability mass of unseen types:

$$p_{\text{unseen}} = n_1/N$$

  - n_1 = number of types observed only once ("hapax legomena")
  - N = size of the sample
- QI stress:

$$p_{\text{unseen}} = 13/306 = 4.2\%$$

- Syllable structure:

$$p_{\text{unseen}} = 22/228 = 9.6\%$$

Approaches to the unseen species problem

- But Good-Turing only gives an estimate of unseen probability mass.
  - The probability that we would encounter a new type if we added one more language to our sample.
- Not the number of types.
- Need more sophisticated techniques.
  - Commonly used in ecology (counting animal species), genetics (counting alleles).

Zhang and Stern's (2009) model: overview

- A recent approach by Zhang and Stern (2009) is particularly suited to linguistic typology.
  - Shown to work well when many types are rare.
- Dirichlet-mulitnomial probability model for type frequencies in sample.
  - Assume population is large relative to sample.
  - Assume each language type (including unobserved types) has some fixed probability of being sampled; they all sum to 1.

Zhang and Stern's (2009) model: data distribution

- S: number of types in the population,
- S_o: number of types observed in the sample.
- $\vec{\theta} = (\theta_1, \ldots, \theta_S)$, the true probability of each type.
- $\vec{y}$: the S-length vector of counts of each type in the sample (unobserved types have count 0); then

$$\vec{y}|S, \vec{\theta} \sim \text{Multinomial}(N, \vec{\theta})$$

Zhang and Stern's (2009) model: data distribution

- S: number of types in the population,
- S_o: number of types observed in the sample.
- $\vec{\theta} = (\theta_1, \ldots, \theta_S)$, the true probability of each type.
- $\vec{y}$: the S-length vector of counts of each type in the sample (unobserved types have count 0); then

$$\vec{y}|S, \vec{\theta} \sim \text{Multinomial}(N, \vec{\theta})$$

- But we don't know S, so we observe $\vec{y}$ without the 0-counts; call it $\vec{y_o}$ (length S_o).
- We don't know how the elements of $\vec{\theta}$ correspond to the observed species!

Zhang and Stern's (2009) model: data distribution

$$\vec{y_o}|S,\vec{\theta} \sim \sum_{\text{all possible } \vec{\theta_o}} \text{Multinomial}(N,\vec{\theta_o})$$

- Multinomial summed over all possible correspondences between $\vec{\theta}$ and observed species.
- This is the data likelihood.
- Goal: obtain posterior inferences about S
  - Assuming some priors on S and $\vec{\theta}$.

Zhang and Stern's (2009) model: priors

- Prior distribution on S: Geometric with parameter f.

$$P(S) = f(1-f)^S$$

  - Can choose f by saying we have some confidence that $S \leq S_{\text{max}}$, some maximum value.
  - E.g., 99.9% confidence that $S \leq 10000 \Rightarrow f = 0.0007$.

Zhang and Stern's (2009) model: priors

- Prior distribution on S: Geometric with parameter f.

$$P(S) = f(1-f)^S$$

  - Can choose f by saying we have some confidence that $S \leq S_{\text{max}}$, some maximum value.
  - E.g., 99.9% confidence that $S \leq 10000 \Rightarrow f = 0.0007$.
- Prior distribution on $\vec{\theta}$: Dirichlet with parameter $\alpha \cdot \vec{1}_S$.

$$\vec{\theta} | S, \alpha \sim \text{Dirichlet}(\alpha \cdot \vec{1}_S)$$

  - α controls how skewed the distribution of types is. $\alpha \to 0 \Rightarrow$ more skewed.
  - Noninformative, improper hyperprior on α (but proper posterior).

The bottom line

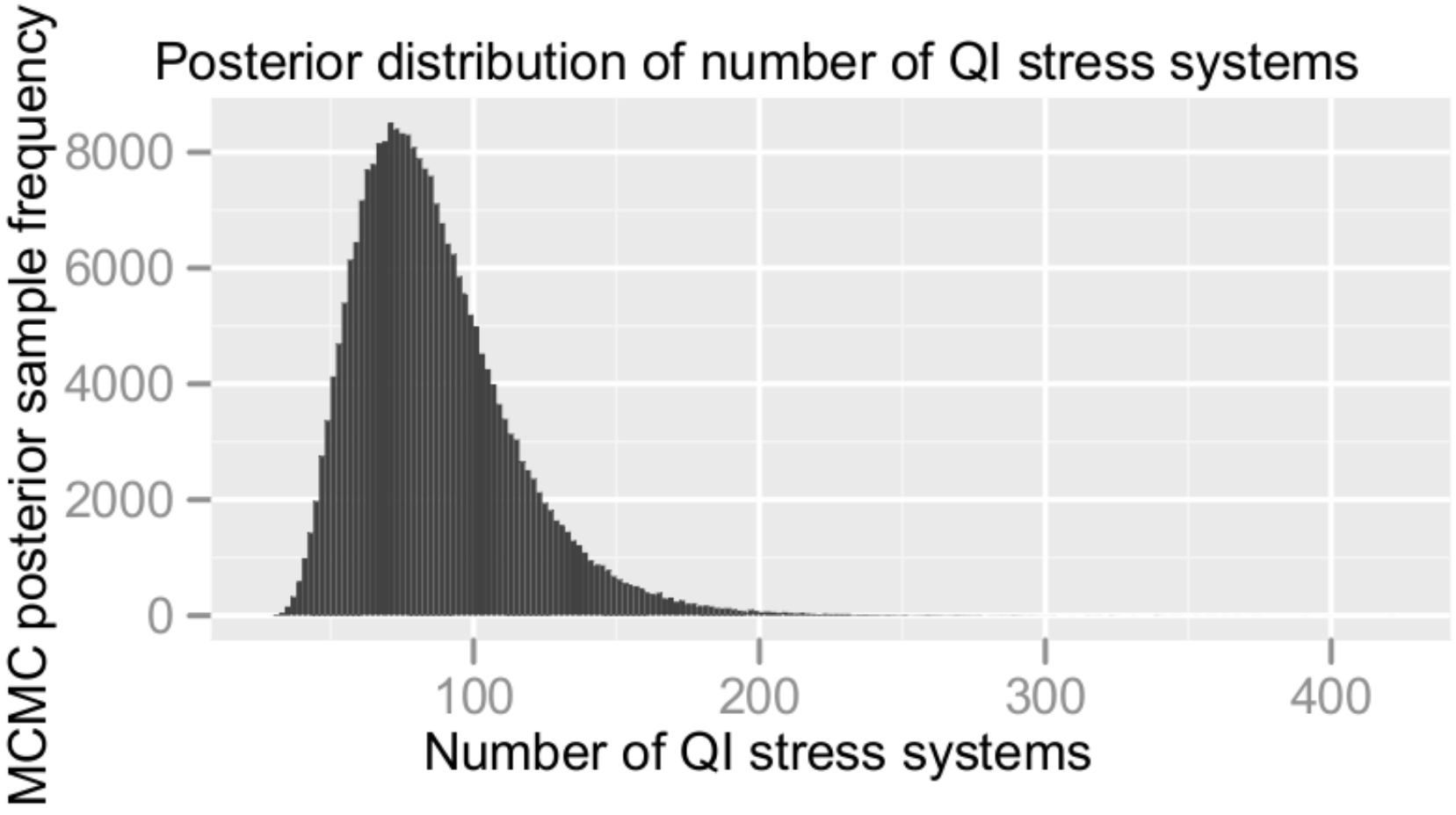
- A model of how observed types should be distributed given S, total number of population types, and α how skewed the population is.
- Flip it around: posterior inferences about S and α, given observed type distribution.
- Caveat: using no information about the types themselves; types are just labels

Posterior sampling

- Posterior inferences obtained using Markov chain Monte Carlo.
  - Gibbs sampling around S and α, with Metropolis-Hastings updates for each.
- 5 chains.
- 1 million samples each, 500k burn-in.
- 1/10 thinning.
- Non-convergence diagnostics (Gelman et al. 2003).
- 250k final posterior samples (50k per chain).

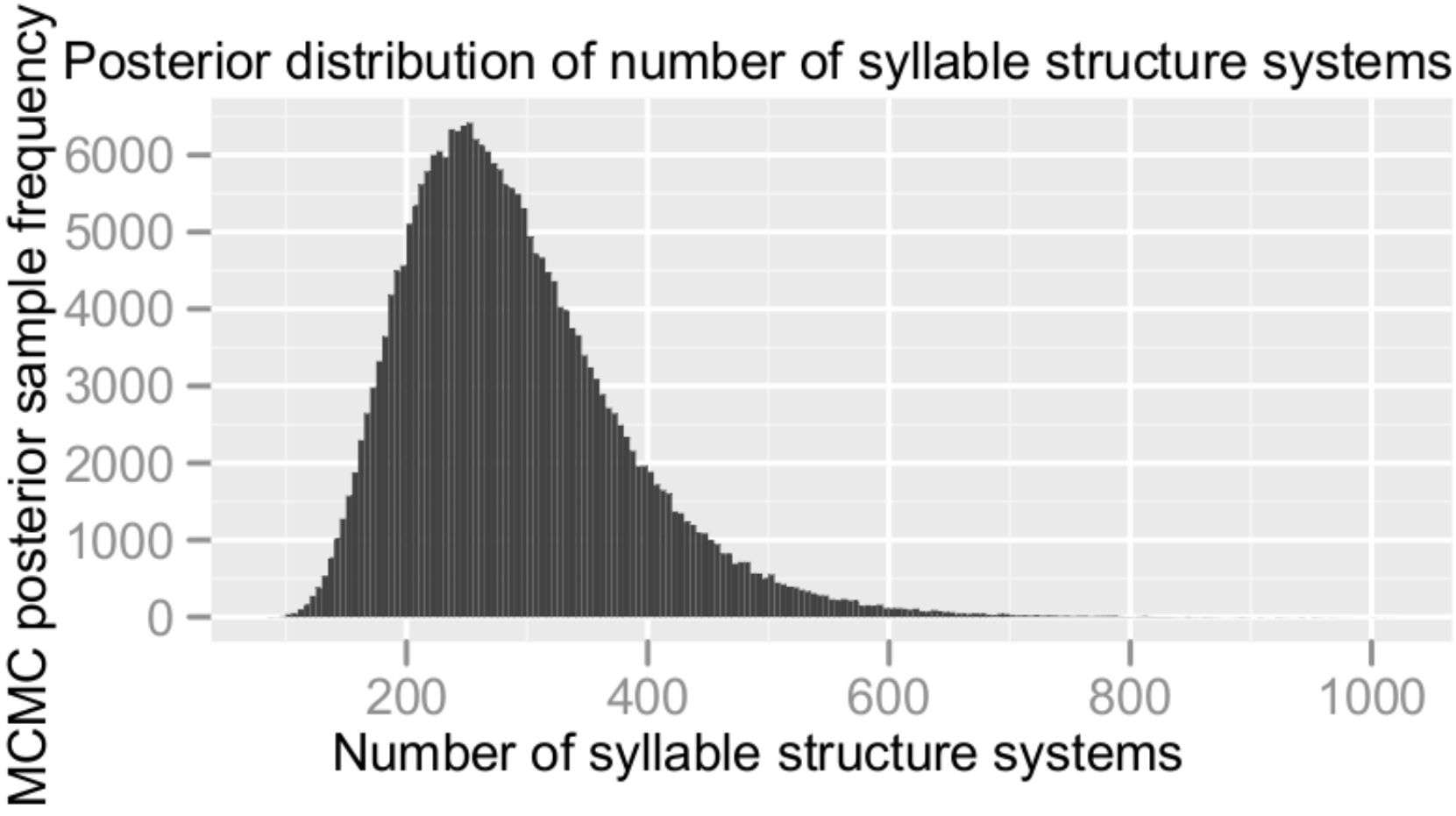
Posterior inference for QI stress

95% central interval: [47, 155] (cf. 26 observed)



Posterior inference for Syllable structure

95% central interval: [156, 507] (cf. 57 observed)



Implications: for the typology builder

- Distribution of observed types suggests samples are far from representing all types.
  - More so for syllable inventories than stress.
- Can help guide database construction.

Implications: for the theoretical linguist

- Many grammatical models make typological predictions.
- Can shed some light on the problem of overgeneration.
- Example: Gordon's (2002) OT model of QI stress generates factorial typology of 152 possible stress systems.
  - Captures attested systems, but "overgenerates" by $\sim$130 systems.
  - Is this reasonable?

- Yes, plausible. Falls within 95% interval [47, 155].

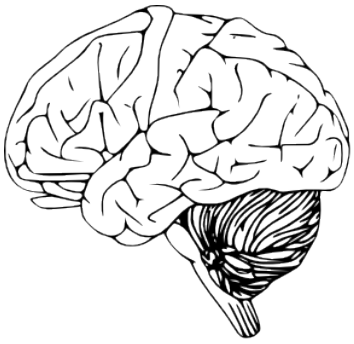
What would be implausible? Too little overgeneration

- A model that generates only 35 possible stress systems.
  - "overgenerates" in the classical sense (more than 26 attested systems).
  - But doesn't overgenerate enough.
  - 95% posterior belief that there are between [47, 155] systems in the world.
- Under-overgeneration

Too much overgeneration

- A model that generates 300 possible stress systems.
  - Possibility of too much overgeneration.
  - 95% posterior belief that there are between [47, 155] systems in the world.
- But not necessarily bad
  - Depends how much discrepancy we're comfortable with between:
    - number of "possible" language types
    - number of existing language types

Language types: possible vs existing vs observed

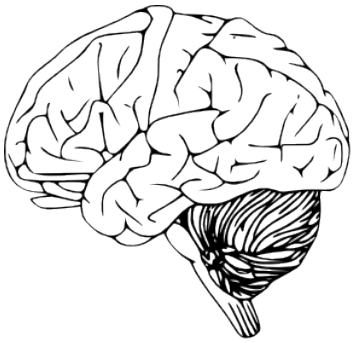


Possible language types

Existing language types

Observed language types

Language types: possible vs existing vs observed



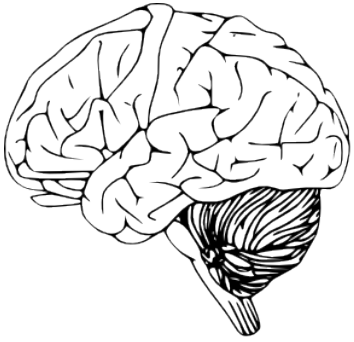
Possible language types

Existing language types

unseen species estimation

Observed language types

Language types: possible vs existing vs observed



Possible language types

theory of instantiation?

Existing language types

unseen species estimation

Observed language types

More work

- How much do we trust sample frequencies?
  - Genetic/areal relatedness, etc.
- Incorporating linguistic structure into estimate.

The end

Thank you!

Data

- Typological data
  - U Delaware Stress Typology Database
    `http://phonology.cogsci.udel.edu/dbs/stress/`
  - Syllable Typology Database
    Available in the Typological Database System at
    `http://languagelink.let.uu.nl/tds/index.html`
  - The specific sub-datasets used here available at
    `http://clml.uchicago.edu/~max`

Appendix: Top attested types

Stress system	languages
fixed final	74
fixed initial	69
fixed penultimate	60

Syllable template	languages
(C)V(C)	31
(C)V	20
CV(C)	19
(C)V(V)(C)	18
(C)(C)V(C)	15

Appendix: Number of types from Good-Turing

- If all types were *a priori* equiprobable, Good-Turing estimate of unseen probability mass could give an estimate of total number of types in population:

$$S = S_o/(1 - n_1/N)$$

  (S_o = number of types observed in sample)
- But we expect type probabilities to vary, necessitating more sophisticated methods.

Appendix: Precise form of data distribution

$$P(\vec{y_o}|\vec{\theta}, S) = \frac{N!}{y_1! \cdots y_{S_o}!} \frac{1}{\prod_{x=1}^{N} n_x!} \sum_{\{i_1,\ldots,i_{S_o}\} \in W(S_o)} \theta_{i_1}^{y_1} \cdots \theta_{i_{S_o}}^{y_{S_o}}$$

where $W(S_O)$ generates all size-S_O subsets of the population type labels $\{1, \ldots, S\}$.

References I

- Baayen, R. Harald. 2001. *Word frequency distributions.* Dordrecht: Kluwer Academic Publishers.
- Bunge, J. and M. Fitzpatrick. 1993. Estimating the number of species: a review. *J. Amer. Stats. Assn.* 88(421):364–373.
- Gandolfi, Alberto and C.C.A. Sastri. 2004. Nonparametric estimations about species not observed in a random sample. *Milan Journal of Mathematics* 72:81–105.
- Gordon, Matthew. 2002. A factorial typology of quantity-insensitive stress. *Natural Language and Linguistic Theory* 20(3):491–552.

References II

- Greenberg, Joseph. 1978. Typology and cross-linguistic universals. In Joseph Greenberg, Charles Ferguson, and Edith Moravcsik, eds., *Universals of human language,* volume 1, 33–60. Stanford University Press.
- Heinz, Jeffrey Nicholas. 2007. Inductive learning of phonotactic patterns. Ph.D. thesis, UCLA.
- Zhang, Hongmei and Hal Stern. 2009. Sample size calculation for finding unseen species. *Bayesian analysis* 4(4):763–792.