



Convergent Equilibria in Cross-Lingual Phoneme Distributions: Statistical & Simulation-Based Analysis

Most computational models of phoneme distributions overlook how *contrastive categories emerge diachronically*.

We identify a new cross-linguistic statistical pattern in the phonemic distributions of English and 166 Australasian languages.

We find a cross-lingual convergence in phoneme surprisal distributions – phoneme systems exhibit similar surprisal distributions despite typological diversity. **Do different statistical laws apply at different linguistic levels?**

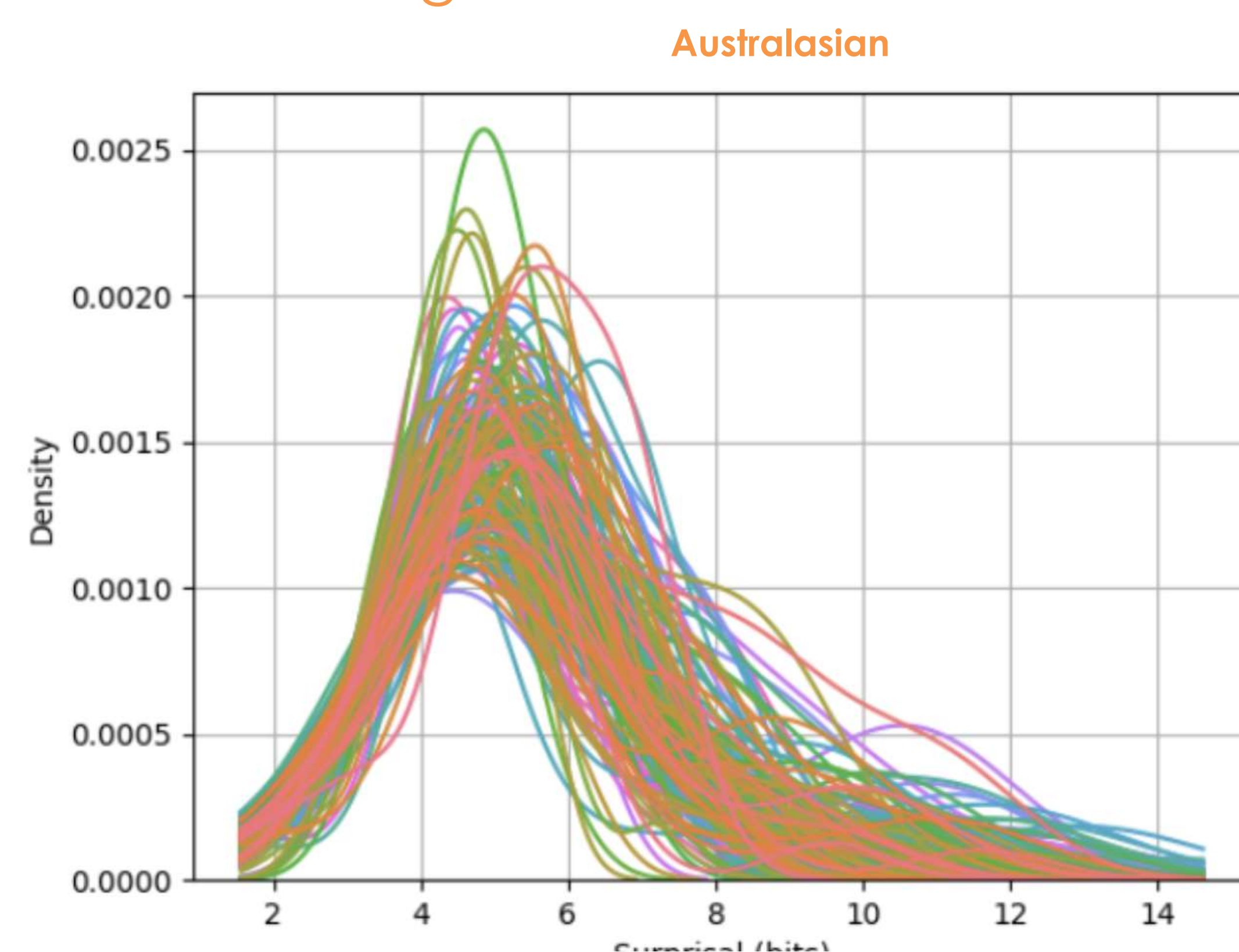
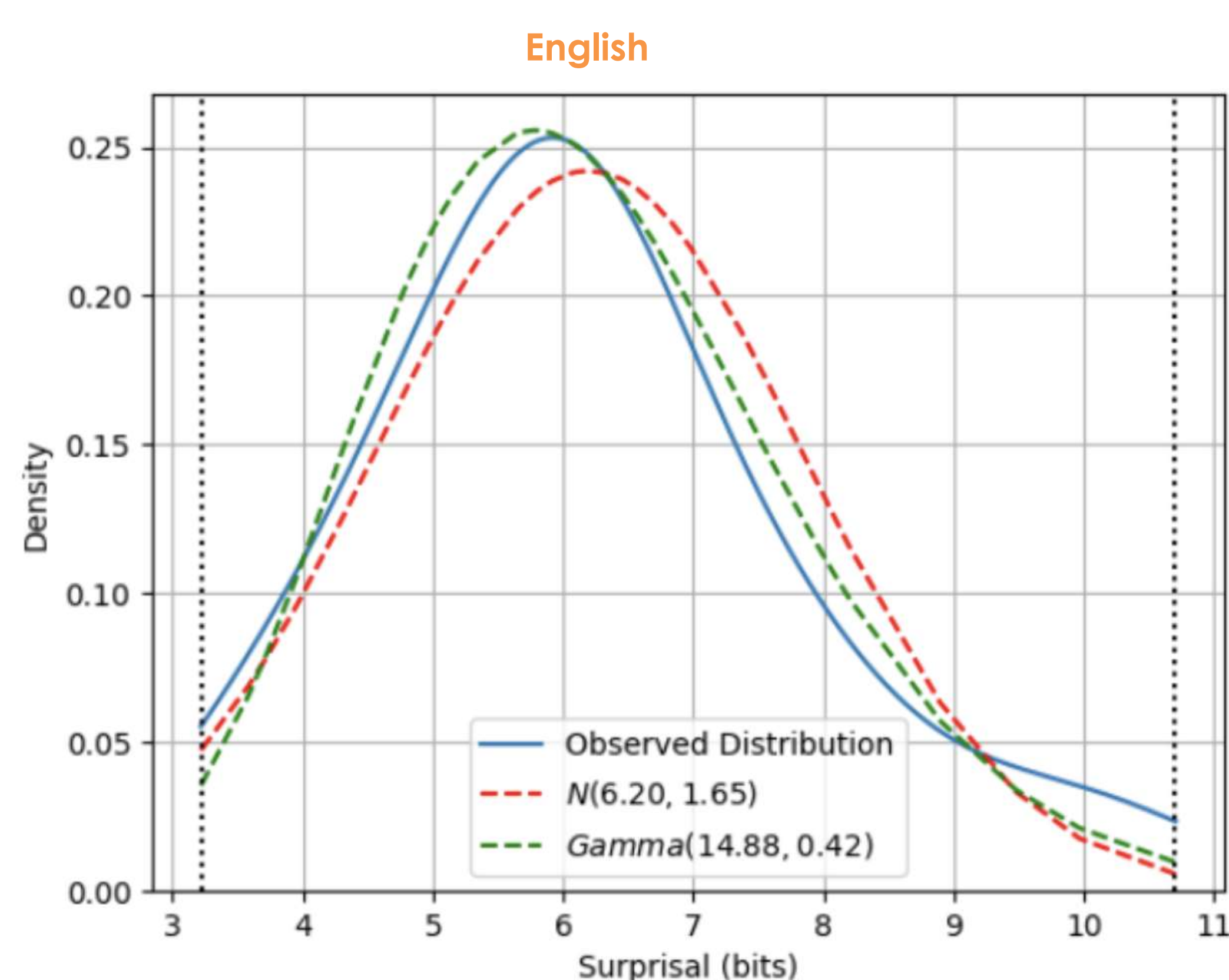


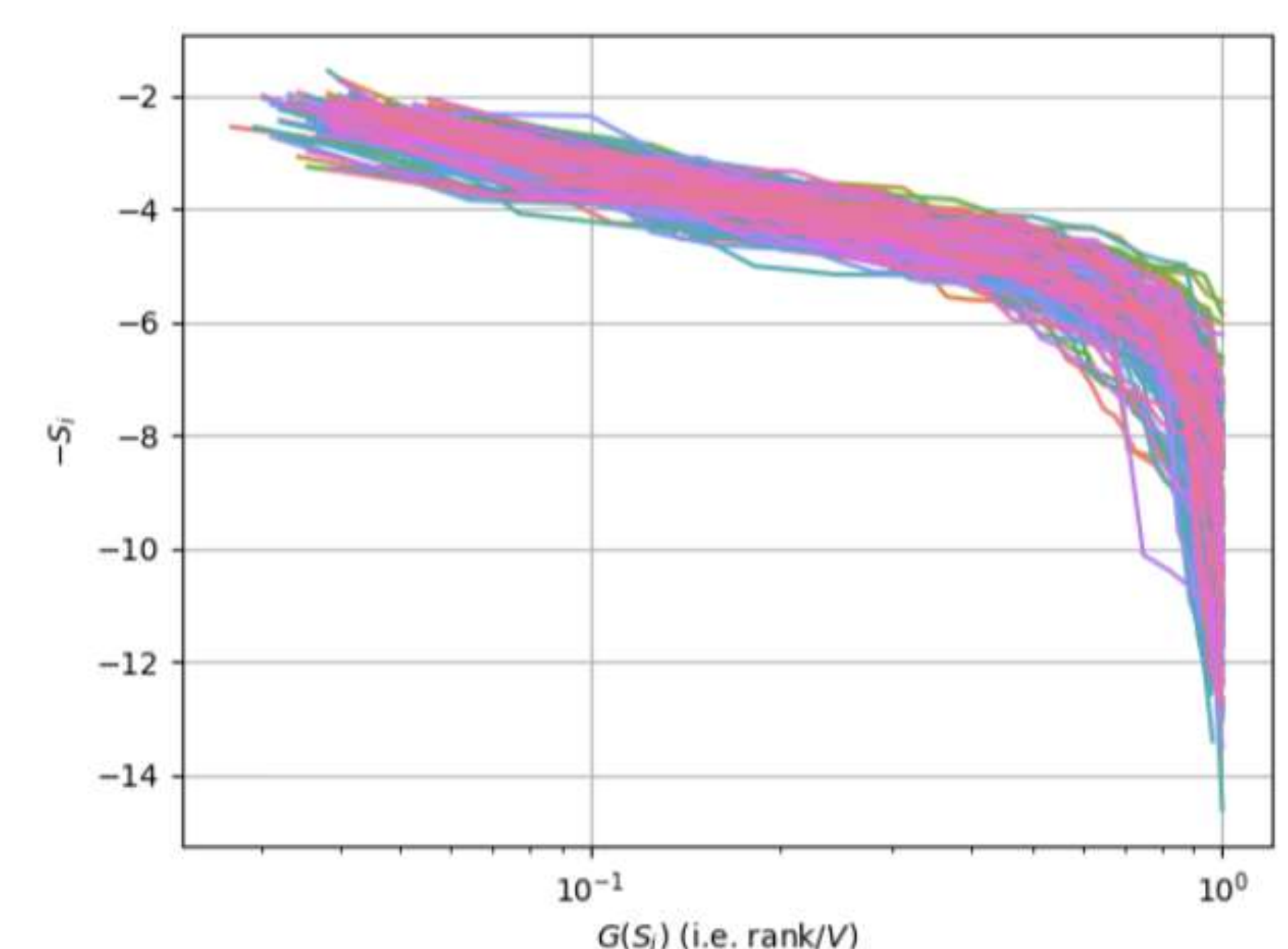
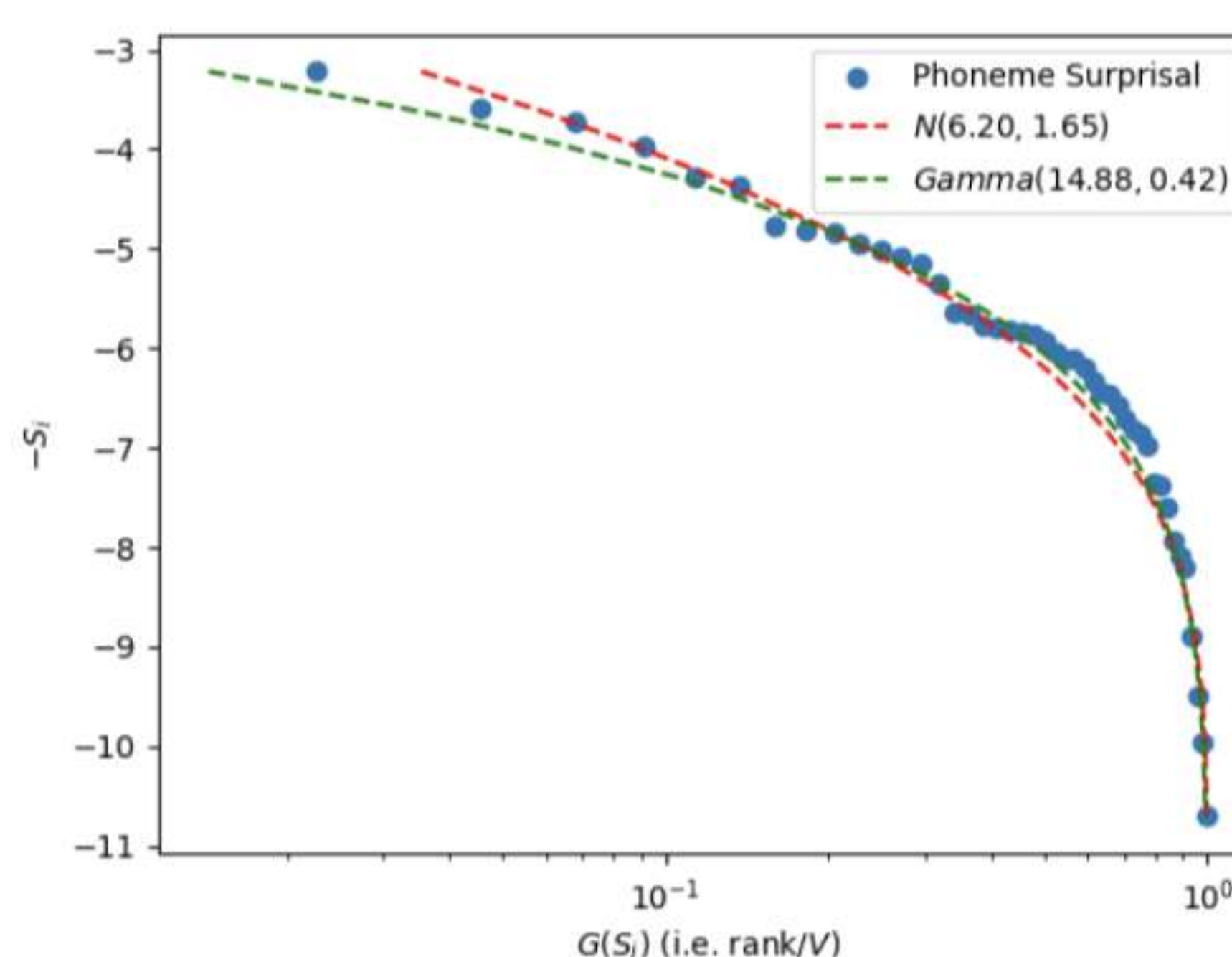
Figure: Estimated distributions of phoneme probabilities and their associated surprisal values (defined as negative log-probabilities)

Surprisal distributions are quasi-normal in shape, tightly clustered around a central mean with exponential thin tails.

This suggests that **phoneme inventories occupy a constrained statistical equilibrium**.

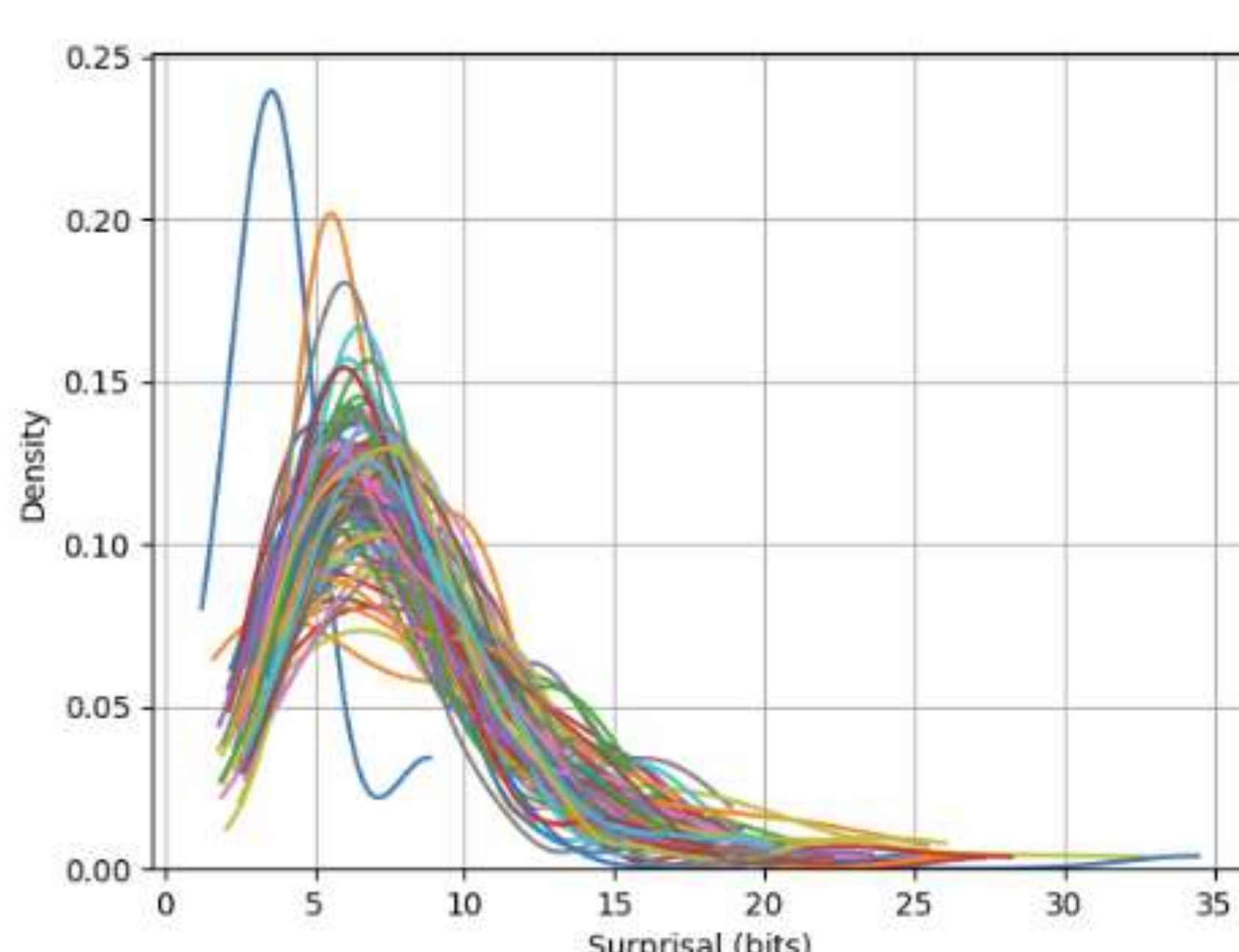
Rank-Frequency (Zipf-like plots) show that surprisal decreases with rank and **low-frequency phonemes decay exponentially**, not power law.

Thin-tailed phoneme distributions differ sharply from long-tailed (Zipf-Mandelbrot) distributions of morphemes/words.

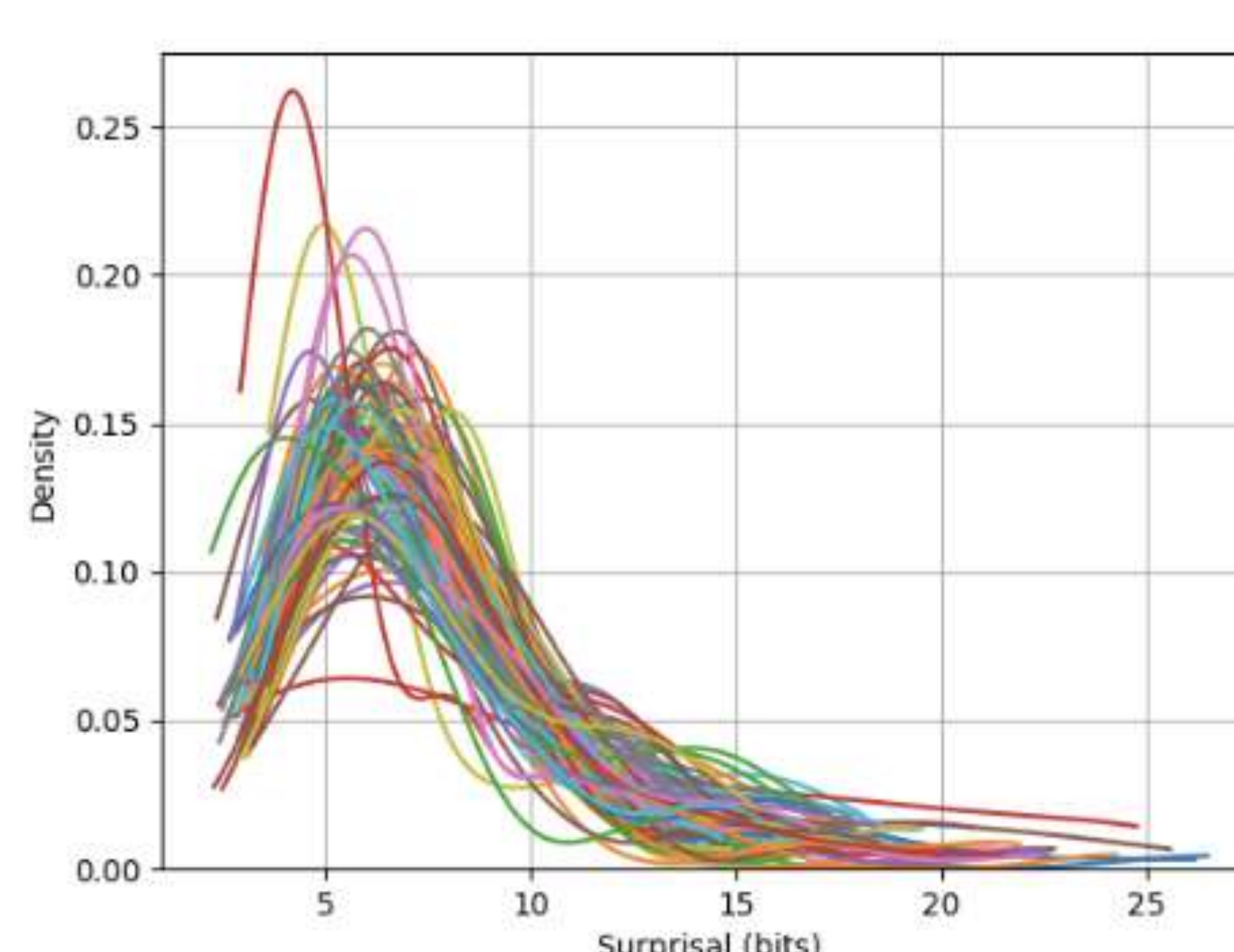


We can straightforwardly simulate this from a naïve computational model!

(a) Simulation 1 (Naive)



(b) Simulation 2 (Robustness Factors)



Our computational model simulates phoneme inventory evolution with iterative application of three possible diachronic operations:

primary split, redistributing mass between two phonemes

secondary split, duplicating a phoneme

Phoneme Merger

Either with (a) equal probability or (b) robustness weighting