Modelling the Diachronic Emergence of Phoneme Frequency Distributions

Fermín Moscoso del Prado Martín
Department of Computer Science and Technology
University of Cambridge, UK
fm611@cst.cam.ac.uk

Suchir Salhan
Department of Computer Science and Technology
University of Cambridge, UK
sas245@cst.cam.ac.uk

Abstract

Phoneme frequency distributions exhibit robust statistical regularities across languages, including exponential-tailed rank-frequency patterns and a negative relationship between phonemic inventory size and the relative entropy of the distribution. The origin of these patterns remains largely unexplained. In this paper, we investigate whether they can arise as consequences of the historical processes that shape phonological systems. We introduce a stochastic model of phonological change and simulate the diachronic evolution of phoneme inventories. A naïve version of the model reproduces the general shape of phoneme rank-frequency distributions but fails to capture other empirical properties. Extending the model with two additional assumptions –an effect related to frequency and a stabilising tendency toward a preferred inventory size– yields simulations that match both the observed distributions and the negative relationship between inventory size and relative entropy. These results suggest that some statistical regularities of phonological systems may arise as a result of diachronic sound change instead of –or in addition to– explicit optimisation or compensatory mechanisms.

1 Introduction

The frequency distributions with which different phonemes occur in a language can provide insights into the nature, representation, and processing of human language. Nevertheless, despite its importance, only a handful of studies have investigated the nature of these distributions; most investigated these distributions from a macroscopic (in the sense of Mandelbrot, 1957) perspective, considering the shape of the rank-frequency plots (Sigurd, 1968; Good, 1969; Martindale et al., 1996; Martindale and Tambovtsev, 2007; Macklin-Cordes and Round, 2020; Moscoso del Prado Martín and Salhan, 2026).

Recently, it has been reported (Moscoso del Prado Martín and Salhan, 2026) that –at both macro- and microscopic levels of description– phoneme frequency distributions exhibit detectable effects consistent with the 'Compensation Hypothesis' (Hockett, 1955; Martinet, 1955). This hypothesis predicts that increased complexity in one domain of language is offset elsewhere in the language. Several studies had previously found evidence for this hypothesis involving balancing aspects of the phonological system with other aspects of language structure (Moran and Blasi, 2014; Pimentel et al., 2020, 2021). It is, however, remarkable that compensation is already detectable by examining the unigram phoneme frequency distributions alone.

An important related question has received little attention: How do phoneme frequency distributions arise in diachronic terms? In particular, one could also ask whether what appear as compensation effects might in fact be unexpected consequences of the historical processes that have shaped the phoneme inventories. This would open the possibility of compensation phenomena being epiphenomenal, rather than the result of any actual optimisation process. Inspired by the predictions of Evolutionary Phonology (Blevins, 2004), Ceolin and Sayeed (2019) and Ceolin (2020) introduce a stochastic split-and-merger model of sound change in which phoneme inventories evolve through repeated splitting and merging events. Their simulations show that statistical patterns commonly associated with phonological markedness can emerge from simple diachronic processes, without assuming markedness as a primitive property of phonological systems. In a similar vein, in this study, we introduce a model of phoneme change based on Hoenigswald's (1965) typology of phonological changes. We use this model to investigate whether two simple diachronic principles –frequency and a stabilising tendency in phoneme inventory size– are sufficient for generating both the observed phoneme rank-

Proceedings of the Society for Computation in Linguistics (SCiL) 2026, pages 138–146

frequency distributions and the negative relationship between phonemic inventory size and relative entropy.

In what follows, we first describe the empirical properties of phoneme frequency distributions that our model aims to explain (Section 2). We then introduce a stochastic model of phonological change (Section 3). We examine three incremental versions of the model of increasing complexity: a naïve baseline (Section 3.3), a version incorporating effects related to frequency (Section 3.4), and a model further introducing a central tendency in phonemic inventory size (Section 3.5). We show that the latter reproduces both the observed rank-frequency patterns of phoneme frequencies and the negative relationship between phonemic inventory size and relative entropy, suggesting that these macroscopic patterns may arise as natural consequences of diachronic phonological dynamics.

2 Properties of Phoneme Distributions

Exponential tails: Contrary to arguments for power-law distributions of phonemes (Sigurd, 1968; Martindale et al., 1996; Martindale and Tambovtsev, 2007; Ceolin and Sayeed, 2019; Ceolin, 2020), phoneme frequency distributions exhibit rank-frequency plots typical of exponential-tailed distributions (Good, 1969; Macklin-Cordes and Round, 2020; Moscoso del Prado Martín and Salhan, 2026): In the double-logarithmic plane, the right tails of these distributions fall abruptly, quickly deviating from the straight lines that are characteristic of power-law-tailed distributions (e.g., word frequency distributions; Condon, 1928). These patterns are illustrated in Figure 1a for the 166 languages of Macklin-Cordes and Round's (2020) dataset of Australian language varieties,[1] and for the 107 languages in the NorthEuraLex database (Dellert et al., 2019).[2]

Correlation between phonemic inventory size and relative entropy: In a recent study, Moscoso del Prado Martín and Salhan (2026) report that –across the world's languages– there is a negative correlation between a language's *Phonemic Inventory Size* (PIS), and the relative entropy (Shannon, 1948) of its phoneme distribution. The entropy of the phoneme distribution is an indicator of the per-phoneme informational content of a language (Cherry et al., 1953). Generally speaking, increasing the PIS increases the potential value of the distribution's entropy. However, the reduction in relative entropy attenuates the effect of increasing the PIS on the inventory's entropy, as would be predicted by the Compensation Hypothesis (Hockett, 1955; Martinet, 1955). This is illustrated in Figure 1b. The figure plots the negative correlation between PIS and relative entropy across the languages in the Australian languages and NorthEuraLex datasets mentioned above. Note that this relation is rather robust, it is present on the Australian language dataset –which was one of the datasets used by Moscoso del Prado Martín and Salhan–, however, the relation also extends to the NorthEuraLex languages –not used in the previous study. In fact, as shown in the figure, the precise prediction made by Moscoso del Prado Martín and Salhan (dashed line in the figure) extrapolates into the NorthEuraLex observations remarkably well.

3 Modelling Phonological Change

3.1 Phonological changes

Phonological changes (also known as phonemic changes) are diachronic changes within a language's system of contrastive sounds. Such changes go beyond mere shifts in pronunciation, which are referred to as *phonetic changes*. They affect the oppositions among phonemes, and may result in changes to a language's phoneme inventory: Contrasts may arise, disappear, or be reorganised, resulting in a restructuring of the language's sound system.

Phonological changes can be of different types. Using Hoenigswald's (1965) systematisation, we can distinguish between three main types: *Primary splits* (also known as conditioned mergers) occur when some instances of a phoneme A become an already existing phoneme B in particular contexts. In these cases, the number of phonemes remains the same, but their distribution changes. *Secondary splits* (also known as phonemic splits) occur when some instances of A become a new phoneme B. Changes of this type result in new contrast being created. In turn this increases the size of phonemic inventories. Finally, *unconditioned mergers* (also known simply as mergers) occur when all instances of phonemes A and B collapse into a single phoneme A. In these cases, a contrast is eliminated and the number of phonemes decreases. Note that,

[1] Downloaded from https://zenodo.org/records/4104116 on May 1, 2025.

[2] v0.9, downloaded from the LexiBank Project at https://github.com/lexibank/northeuralex on March 3, 2026.

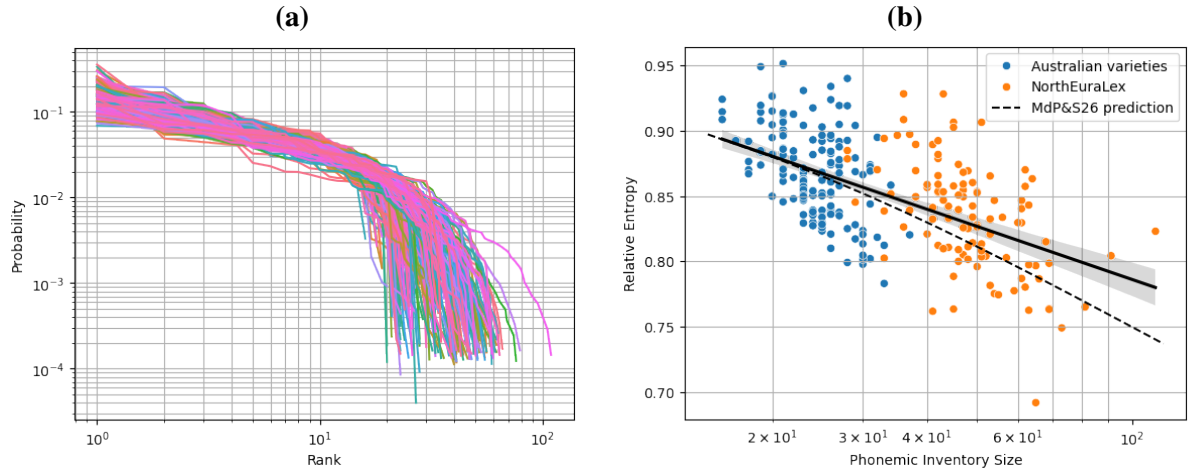

Figure 1: **(a)** Rank-frequency plots (note the log-log scale) for the phoneme frequency distributions across the languages in Macklin-Cordes and Round's (2020) dataset of Australian language varieties, and those in NorthEuraLex. Each line plots one language. **(b)** Relationship between PIS and relative entropy for the languages in Macklin-Cordes and Round's (2020) dataset of Australian language varieties, and those in NorthEuraLex. Each point plots one language. The solid line plot a log-linear regression, and the shading plots its 95% C.I.. The dashed line plots the relation as predicted by Moscoso del Prado Martín and Salhan (2026).

within this classification, there is no specific type for phoneme losses; these are just considered unconditional mergers with a zero phoneme.

3.2 The general model

Consider language at a particular time point τ in its history. It uses V_τ distinct contrastive phonemes $\pi_1, \pi_2, \ldots, \pi_{V_\tau}$. The probabilities with which each phoneme occurs in the language at time τ are given by the vector $\mathbf{p}_\tau = \{\mathrm{p}_\tau(\pi_1), \mathrm{p}_\tau(\pi_2), \ldots, \mathrm{p}_\tau(\pi_{V_\tau})\}$, where $\sum_{i=1}^{V_\tau} \mathrm{p}_\tau(\pi_i) = 1$. Time is treated as a discrete sequence of phonological changes c_τ. Each change is of one of three types: primary split (p), secondary split (s), or merger (m). Formally, c_τ is sampled at each time step from the alphabet $\Sigma = \{p, s, m\}$. Time intervals are defined so that each contains exactly one change event.

We construct stochastic models in which, at every time point τ, the change type c_τ is sampled randomly from Σ, with probabilities $P_\tau(p)$, $P_\tau(s)$, and $P_\tau(m)$, so that $P_\tau(p) + P_\tau(s) + P_\tau(m) = 1$. Independently, a phoneme $\pi_i \in \{\pi_1, \ldots, \pi_{V_\tau}\}$ is sampled (according to some predefined scheme, details below) as the target of change. In the case of primary splits and mergers, a second phoneme π_j is also sampled, representing the phoneme toward which π_i shifts (primary split) or with which it merges (merger). In addition, a proportion parameter $\alpha_\tau \in (0, 1]$ is sampled to determine how much probability mass is transferred between phonemes. We assume α_τ is uniformly distributed, reflecting the fact that phonological changes can affect contrasts to varying degrees.

The three types of changes are defined as transformations on the probability vector:

Primary split: a proportion α_τ of $\mathrm{p}(\pi_i)$ is reassigned to an existing phoneme π_j:

$$\begin{aligned} \mathrm{p}_{\tau+1}(\pi_i) &= (1 - \alpha_\tau)\mathrm{p}_\tau(\pi_i), \\ \mathrm{p}_{\tau+1}(\pi_j) &= \mathrm{p}_\tau(\pi_j) + \alpha_\tau\, \mathrm{p}_\tau(\pi_i). \end{aligned} \tag{1}$$

For $0 < \alpha_\tau < 1$, the inventory size remains $V_{\tau+1} = V_\tau$; if $\alpha_\tau = 1$, π_i disappears and $V_{\tau+1} = V_\tau - 1$.

Secondary split: a new phoneme $\pi_{V_\tau+1}$ is created by splitting π_i:

$$\begin{aligned} \mathrm{p}_{\tau+1}(\pi_i) &= (1 - \alpha_\tau)\mathrm{p}_\tau(\pi_i), \\ \mathrm{p}_{\tau+1}(\pi_{V_\tau+1}) &= \alpha_\tau\, \mathrm{p}_\tau(\pi_i). \end{aligned} \tag{2}$$

For $0 < \alpha_\tau < 1$, $V_{\tau+1} = V_\tau + 1$; if $\alpha_\tau = 1$, π_i disappears and $V_{\tau+1} = V_\tau$

Unconditioned merger: two phonemes π_i and π_j collapse completely:

$$\mathrm{p}_{\tau+1}(\pi_j) = \mathrm{p}_\tau(\pi_j) + \mathrm{p}_\tau(\pi_i), \tag{3}$$

and π_i is removed, therefore $V_{\tau+1} = V_\tau - 1$.

Over time, this process generates stochastic trajectories both in inventory size V_τ and in the internal distributional structure of the phonological

system (the probability vector). As we show below, the method by which the change probabilities $P_\tau(p)$, $P_\tau(s)$, and $P_\tau(m)$, the choice of phonemes, and the α_τ are chosen at each time point can substantially affect the behaviour of these models. In summary, this model treats phonological change as a stochastic process redistributing probability mass across phonemes while allowing contrasts to be created or eliminated.

3.3 Simulation 1: a naïve model

As a baseline, we simulated the basic model in Section 3.2, using its simplest instantiation. We considered constant equal probabilities $P_\tau(p) = P_\tau(s) = P_\tau(m) = 1/3$ for all timepoints τ. At each time step, the phonemes involved in the change (π_i and π_j in the notation above) were randomly sampled from the current phoneme inventory with uniform probabilities (i.e., any phoneme was equally likely to be chosen), and the proportion of phonemes involved in the change (α_τ), when necessary, was sampled from a uniform [0,1] distribution. We simulated the evolution of 400 distinct languages, each for 1,000 time steps. Initially, all languages were set to uniform distributions of 34 phonemes (this number was taken as the approximate mean PIS for the world's languages according to PHOIBLE 2.0; Moran and McCloy, 2019).

As shown in Figure 2a, even this naïve simulation already generates rank-frequency patterns resembling those observed in empirical phoneme distributions (Figure 1a). This is so even when all distributions started out as uniforms. However, examining the simulated distributions in more detail reveals crucial differences with real phoneme frequency distributions. Figure 2b plots the correlation between (log) PIS and the relative entropy of the distribution, which is significant (Pearson's $r = .47$, $p < .01$). However, this goes in exactly the opposite direction than it did in real phoneme distributions: Whereas simulated data show a positive correlation, in actual distributions it is negative. This difference indicates that, in order to capture the properties of phoneme frequencies, our diachronic models need to consider further details.

3.4 Simulation 2: considering frequency

Simulation 1 was too naïve in assuming that all phonemes are equally likely to take part in phonological change. This is in fact known not to be the case. The *Functional Load Hypothesis* –tracing back to Gilliéron (1918)– proposes that phonological contrasts that distinguish many words (i.e., have high functional load) are less likely to be lost through sound change, whereas contrasts that contribute little to lexical differentiation are more prone to merger (for details see, Hockett, 1955, 1967). This hypothesis has received empirical support (Wedel et al., 2013a,b). Our models are too coarse-grained to explicitly consider a direct measure of functional load (see Surendran and Niyogi, 2003). We can however, indirectly take it to account to some degree: Wedel et al. (2013b) showed that –in general– a phoneme's functional load is positively correlated with its frequency of occurrence.

In order to take phoneme frequency into consideration, we ran new simulations. As before, we considered constant equal probabilities $P_\tau(p) = P_\tau(s) = P_\tau(m) = 1/3$ for all timepoints τ, and the proportion of phonemes involved in the change (α_τ), when necessary, was sampled from a uniform [0,1] distribution. Again, we simulated the evolution of 400 distinct languages, each for 1,000 time steps. Initially, all languages were set to uniform distributions of 34 phonemes.

This time, however, at each time step, the phonemes involved in the change were randomly sampled from the current phoneme inventory with different probabilities. The phonemes that would be split or merged into another one (i.e., the phonemes whose probabilities are decreased in Equations 1, 2, and 3) were sampled with probabilities directly proportional to their surprisals (i.e., less frequent phonemes were sampled more often). On the other hand, the phonemes whose probabilities would increase (in Equations 1 and 3) were sampled uniformly (i.e., all phonemes are equally likely to be chosen). In this way, we have increased the probability that the less frequent phonemes are lost in historical change.

The resulting distributions retain the characteristic shape of phoneme rank-frequency curves observed in real languages, albeit with a substantially increased variance; the introduction of the frequency bias causes a rich-get-richer effect, resulting in many extremely skewed distributions with close to zero relative entropies (see Figure 3a). The modification did not fix the incorrect sign of the PIS–relative entropy correlation (Pearson's $r = .68$, $p < .01$; see Figure 3b).

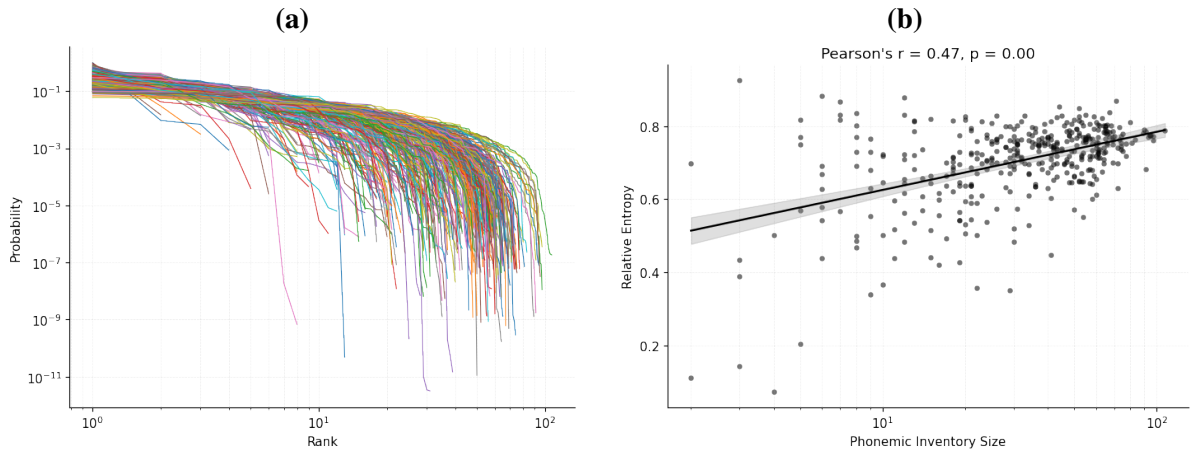

Figure 2: **(a)** Rank-frequency plots (note the log-log scale) for the final phoneme frequency distributions in Simulation 1. Each line plots one language. **(b)** Relationship between PIS and relative entropy for the final distributions in Simulation 1. Each point plots one simulated language. The solid line plots a log-linear regression, and the shading plots its 95% C.I..

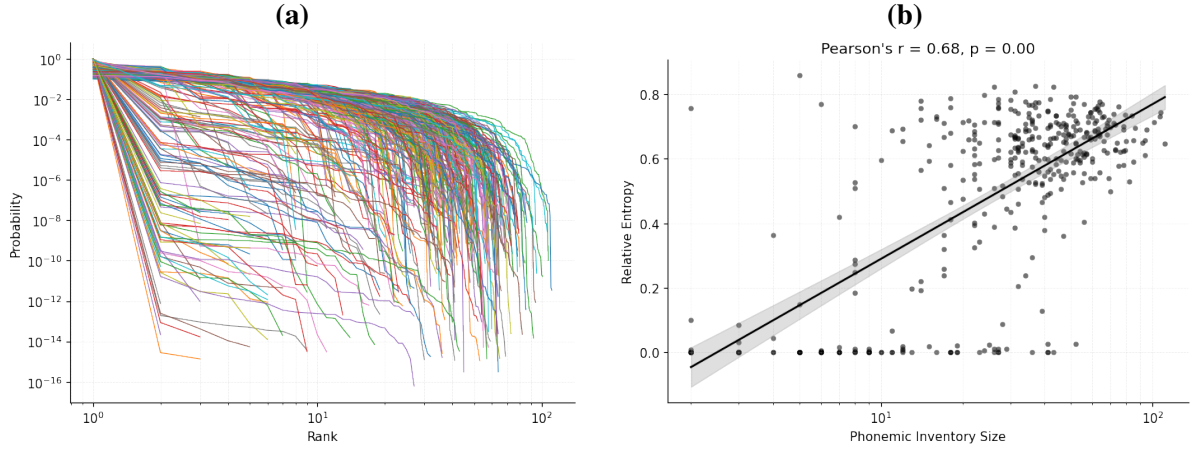

Figure 3: **(a)** Rank-frequency plots (note the log-log scale) for the final phoneme frequency distributions in Simulation 2. Each line plots one language. **(b)** Relationship between PIS and relative entropy for the final distributions in Simulation 2. Each point plots one simulated language. The solid line plots a log-linear regression, and the shading plots its 95% C.I..

3.5 Simulation 3: adding a central tendency

Simulation 2 still presents two problems with respect to actual phoneme frequency distributions. First, it does not show the correct relationship between PIS and relative entropy. By itself, this might not be so much of a problem. It could very well be that this correlation truly arises as a consequence of actual compensation. Indeed, Moscoso del Prado Martín and Salhan (2026) find that such correlation could be explained microscopically by compensation at the level of phonotactics, by which languages with larger inventories tend to have phonemes that appear in more predictable contexts. As our models do not include any microscopic details on phonotactics, it could be the case that we are not able to model the emergence of compensation.

Second, and more critically, in Simulations 1 and 2, PIS values evolve as random walks on the positive integers. At each step, the step size probabilities depend on the probabilities of the three change types. In the idealised case in which $0 < \alpha_\tau < 1$ almost surely,

$$\begin{aligned} V_{\tau+1} - V_\tau &= \Delta V_\tau, \\ P(\Delta V_\tau = 0) &= P_\tau(p), \\ P(\Delta V_\tau = +1) &= P_\tau(s), \\ P(\Delta V_\tau = -1) &= P_\tau(m). \end{aligned} \tag{4}$$

If α_τ can take the boundary value 1, primary splits and secondary splits may also reduce or preserve

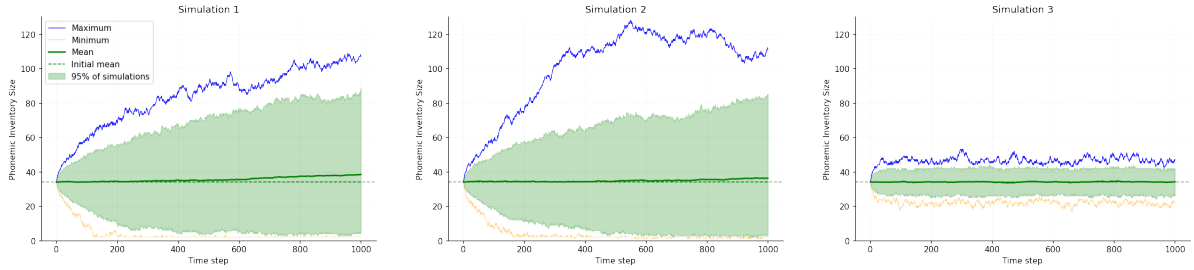

Figure 4: Convergence of the PIS in the simulations. The blue and orange lines respectively plot the maximum and minimum values at each time point across the 400 simulated languages. The solid green lines plot the mean PIS at each time point, the dashed green lines are the predicted mean μ, and the shaded areas contain 95% of the runs in each of the simulations.

V_τ through elimination of the source phoneme. Moreover, mergers are only defined when V_τ exceeds a minimal value, which induces a reflecting constraint at small inventories. Whether this random walk has a stationary mean or it has a drift depends on the specific values of the different parameters. However, as long as the probabilities $P_\tau(p)$, $P_\tau(s)$, and $P_\tau(m)$ remain constant in time, the *variance* of V_τ is bound to increase with time. As a consequence, in this situation, as time goes on, there is no limit to the size that V_τ can take, and the minimum goes all the way down to two phonemes.[3] The resulting random walks for Simulations 1 and 2 are summarised in Figure 4. Notice that in both simulations, even after 1,000 time steps the ranges of PIS values keep increasing without any upper bounds.

This property is not desirable for a model of phonological change. First, there is no evidence that the range of values in the PIS is increasing with time in the world's languages. Previous work has noted that the number of phonemic contrasts across languages is concentrated within a relatively narrow range. Extremely small or large inventories are rather rare (Maddieson, 1984; Anderson et al., 2023). For instance, according to Phoible v2.0 (Moran and McCloy, 2019), the minimum value of the PIS is that of some variants of Rotokas, which may have as few as eleven contrasts (if vowel length is not considered), and the maximum is capped at the 160 contrasts attributed to East Taa. Were the range of PIS left to increase freely in an unconstrained random walk, we would expect more extreme values to be observed among thousands of languages. Also as a result, as can be appreciated comparing Figure 1a with Figures 2a and 3a, the variability in the lower rank phonemes is substantially larger in the simulations than in the real language data.

The only way of avoiding such growing variance of the V_τ is to make the P_τ values dependent on the value of V_τ itself. We assume there is some optimal value of V_τ, which we denote by μ. The further V_τ is from μ, the less likely it should be to get even further. From Equation 4 we see that $P_\tau(s)$ should decrease in value when $V_\tau > \mu$, and in turn, whenever $V_\tau < \mu$, it is $P_\tau(m)$ that should be lowered. We implemented this adaptive strategy using exponential functions, implementing a smooth bias toward μ while preserving positive probabilities,

$$\begin{aligned} P_\tau(p) &= \frac{1}{k(\tau)} \\ P_\tau(s) &= \frac{1}{k(\tau)}\, \mathrm{e}^{(\mu - V_\tau)/\mu} \\ P_\tau(m) &= \frac{1}{k(\tau)}\, \mathrm{e}^{(V_\tau - \mu)/\mu} \\ k(\tau) &= 1 + \mathrm{e}^{(\mu - V_\tau)/\mu} + \mathrm{e}^{(V_\tau - \mu)/\mu} \end{aligned} \tag{5}$$

We modified Simulation 2 to include the adaptive scheme above, with $\mu = 34$ (the mean PIS across the world's languages). As before, we simulated 400 languages for 1,000 time steps. Figure 4 shows that the simulations have now converged in terms of maximum, minimum and range of values of the PIS.[4]

Figure 5a plots the rank-frequency distributions of the simulated inventories. Notice that the variability on the lower ranks has considerably decreased in relation to Simulations 1 and 2 (compare with Figures 2a and 3a), the distributions are more

[3] This was set as a hard limit in the simulations, as it wouldn't make any sense to have languages with a single contrastive phoneme.

[4] Note that, as seen is the results of Simulations 1 and 2, convergence in terms of the shape of the rank-frequency plot is achieved in all cases, whether the PIS has converged or not.

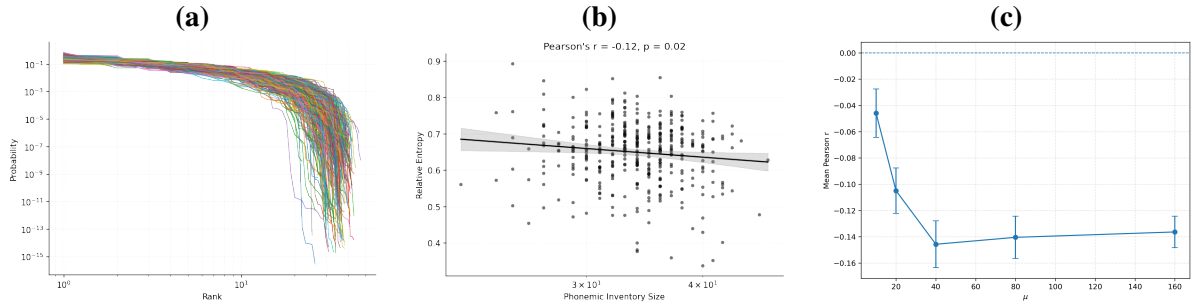

Figure 5: **(a)** Rank-frequency plots (note the log-log scale) for the final phoneme frequency distributions in Simulation 3. Each line plots one language. **(b)** Relationship between PIS and relative entropy for the final distributions in Simulation 3. Each point plots one simulated language. The solid line plots a log-linear regression, and the shading plots its 95% C.I.. **(c)** Sensitivity analysis as for parameter μ: mean and standard error of the Pearson correlation coefficient between PIS and relative entropy for different values of μ across ten simulations for each of the μ values.

similar to those observed in real languages (compare with Figure 1a). Most crucially, as Figure 5b plots, a significant negative correlation between PIS and relative entropy has now emerged (Pearson's $r = -.12$, $p = .02$). In short, both of the problems in the previous simulations have been solved by introducing the central tendency. Admittedly, this is a very slight correlation, substantially weaker than that obtained for real languages. Nevertheless, it is quite robust: Repeating the simulation 100 times resulted in a negative r value every single time ($\bar{r} = -.15 \pm .005$), and these correlations reached significance in 78 out of the 100 runs. In addition, as shown by the sensitivity analysis summarised in Figure 5c, this negative correlation seems to arise irrespective of the value of the central tendency parameter μ.

One could fit the different simulation parameters to make the resulting values match human languages more closely, but this is not necessary for the general patterns to arise. Remarkably, the negative relationship between PIS and relative entropy can arise without any actual compensation mechanisms being involved. In our simulations, it is just an unexpected side effect of the interaction between phonological changes, and a central tendency towards a specific number of contrasts. In this sense, what appears as compensation could in fact be epiphenomenal.

4 Discussion

The simulations presented here show that several macroscopic properties of phoneme frequency distributions can arise from relatively simple diachronic mechanisms. In particular, stochastic phonological change acting on phoneme inventories naturally generates rank-frequency distributions with exponential tails similar to those observed in real languages. Such distributions arise naturally from the historical processes that continually reshape phonological systems. Our results should be interpreted as providing a generative explanation for the statistical patterns. The goal of the models is not to capture all mechanisms involved in phonological change. Rather, we test whether simple diachronic processes are sufficient to generate the observed macroscopic regularities of phoneme frequency distributions.

Naïve versions of the model fail to capture important empirical patterns. In particular, they predict a positive relationship between phonemic inventory size (PIS) and relative entropy –contrary to the findings of Moscoso del Prado Martín and Salhan (2026)– and they produce unbounded variation in inventory sizes over time. Introducing a central tendency in the evolution of PIS substantially improves the models. This constrains the random walk behaviour of inventory size, producing stationary dynamics that are more consistent with the observed range of phoneme inventories across languages.

Not including a central attractor –as in Simulations 1 and 2, or in the models of Ceolin and Sayeed (2019) and Ceolin (2020)– results in phonemic inventory sizes that are too broadly distributed. Instead, our assumption of a central tendency in phonemic inventory size is consistent with previous research on phonological systems. Typological surveys have long noted that the number of phonemic contrasts across languages is concentrated within

a relatively narrow range, with extremely small or extremely large inventories being rare (Maddieson, 1984; Anderson et al., 2023). At a theoretical level, models such as Adaptive Dispersion Theory propose that phonological systems evolve under competing pressures for perceptual distinctiveness and articulatory economy, which tend to produce stable configurations of contrasts (Liljencrants and Lindblom, 1972; Lindblom, 1986; Lindblom and Maddieson, 1988). From this perspective, a stabilising tendency in the evolution of phonemic inventories can be interpreted as a coarse-grained reflection of underlying pressures shaping phonological systems. Our model does not attempt to represent these mechanisms explicitly, but instead captures their aggregate effect through a simple probabilistic bias toward a preferred inventory size.

5 Conclusion

Taken together, these results suggest that some statistical properties of phonological systems may emerge from the interaction of diachronic pressures –in line with the results of Ceolin and Sayeed (2019) and Ceolin (2020). In particular, the negative relationship between PIS and relative entropy was previously interpreted as evidence for compensatory organisation within the phonological system (Moscoso del Prado Martín and Salhan, 2026). However, it could also arise –perhaps in part– as a by-product of stochastic sound change operating under a stabilising tendency toward a preferred inventory size. In this possibility, rather than balancing complexity across organisational tiers, compensation might arise in diachronic time from balancing pressures between perception and articulation. This does not rule out the possibility that genuine compensatory mechanisms between tiers could exist. However, it highlights the importance of considering whether apparent optimisation effects may instead reflect epiphenomenal consequences of the dynamics of sound change, much in the spirit of Evolutionary Phonology (Blevins, 2004, 2019).

References

Cormac Anderson, Tiago Tresoldi, Simon J. Greenhill, Robert Forkel, Russell D. Gray, and Johann-Mattis List. 2023. Variation in phoneme inventories: Quantifying the problem and improving comparability. *Journal of Language Evolution*, 8(2):149–168.

Juliette Blevins. 2004. *Evolutionary phonology: The emergence of sound patterns*. Cambridge University Press, Cambridge, UK.

Juliette Blevins. 2019. Evolutionary phonology as human behavior. In Nancy Stern, Ricardo Otheguy, Wallis Reid, and Jaseleen Sackler, editors, *Columbia School Linguistics in the 21st Century*, pages 281–300. John Benjamins Publishing Company.

Andrea Ceolin. 2020. *Functionalism, Lexical Contrast and Sound Change*. Ph.D. thesis, University of Pennsylvania.

Andrea Ceolin and Ollie Sayeed. 2019. Modeling markedness with a split-and-merger model of sound change. In Nina Tahmasebi, Lars Borin, Adam Jatowt, and Yang Xu, editors, *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 67–70. Association for Computational Linguistics, Florence, Italy.

E. Colin Cherry, Morris Halle, and Roman Jakobson. 1953. Toward the logical description of languages in their phonemic aspect. *Language*, 29(1):34–46.

Edward Uhler Condon. 1928. Statistics of vocabulary. *Science*, 67(1733):300.

Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Hizniye Isabella Boga, Zalina Baysarova, Roland Mühlenbernd, Johannes Wahle, and Gerhard Jäger. 2019. NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. *Language Resources and Evaluation*, 54(2):273–301.

Jules Gilliéron. 1918. *Généalogie des Mots qui Désignent l'Abeille d'après l'Atlas Linguistique de la France [The Genealogy of the Words for 'Bee' According to the Linguistic Atlas of France]*, volume 225 of *Bibliothèque de l'École des Hautes Études: Sciences Historiques et Philologiques*. Librairie Ancienne Honoré Champion, Paris.

Irving John Good. 1969. Statistics of language. In Andrew R. Meetham and Richard A. Hudson, editors, *Encyclopaedia of Linguistics, Information and Control*, pages 567–581. Pergamon Press, Oxford, England.

Charles F. Hockett. 1955. *A Manual of Phonology*. Waverly Press, Baltimore, MD.

Charles F. Hockett. 1967. The quantification of functional load. *Word*, 23(1–3):320–339.

Henry M. Hoenigswald. 1965. *Language Change and Linguistic Reconstruction*. University of Chicago Press, Chicago, IL.

Johan Liljencrants and Björn Lindblom. 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 48(4):839–862.

Björn Lindblom. 1986. Phonetic universals in vowel systems. In John J. Ohala and Jeri J. Jaeger, editors, *Experimental Phonology*, pages 13–44. Academic Press, Orlando, FL.

Björn Lindblom and Ian Maddieson. 1988. Phonetic universals in consonant systems. In Charles N. Li and Larry M. Hyman, editors, *Language, Speech and Mind*, pages 62–78. Routledge, London.

Jayden L. Macklin-Cordes and Erich R. Round. 2020. Re-evaluating phoneme frequencies. *Frontiers in Psychology*, 11:570895.

Ian Maddieson. 1984. *Patterns of Sounds*. Cambridge University Press, Cambridge, England.

Benoît B. Mandelbrot. 1957. Linguistique statistique macroscopique [Macroscopic statistical linguistics]. In Léo Apostel, Benoît B. Mandelbrot, and Albert Morf, editors, *Logique, Langage, et Théorie de l'Information*. Presses Universitaires de France, Paris, France.

Colin Martindale, Sabir M. Gusein-Zade, Dean McKenzie, and Mark Yu. Borodovsky. 1996. Comparison of equations describing the ranked frequency distributions of graphemes and phonemes. *Journal of Quantitative Linguistics*, 3:106–112.

Colin Martindale and Yuri Tambovtsev. 2007. Phoneme frequencies follow a Yule distribution. *SKASE Journal of Theoretical Linguistics*, 4:1–11.

André Martinet. 1955. *Économie des Changements Phonétiques [Economy of Phonetic Changes]*. A. Francke, Bern, Switzerland.

Steven Moran and Damián Blasi. 2014. Cross-linguistic comparison of complexity measures in phonological systems. In Frederick J. Newmeyer and Laurel B. Preston, editors, *Measuring Grammatical Complexity*, pages 217–240. Oxford University Press, Oxford, England.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena, Germany. Accessed on 04/02/2026.

Fermín Moscoso del Prado Martín and Suchir Salhan. 2026. The distribution of phoneme frequencies across the world's languages: Macroscopic and microscopic information-theoretic models. *Preprint*, arXiv:2603.02860.

Tiago Pimentel, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell. 2021. A surprisal–duration trade-off across and within the world's languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 949–962, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tiago Pimentel, Brian Roark, and Ryan Cotterell. 2020. Phonotactic complexity and its trade-offs. *Transactions of the Association for Computational Linguistics*, 8:1–18.

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Bengt Sigurd. 1968. Rank-frequency distributions for phonemes. *Phonetica*, 18:1–15.

Dinoj Surendran and Partha Niyogi. 2003. Measuring the functional load of phonological contrasts. Technical Report TR-2003-12, University of Chicago, Department of Computer Science.

Andrew Wedel, Scott A. Jackson, and Abby Kaplan. 2013a. Functional load and the lexicon: Evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. *Language and Speech*, 56(3):395–417.

Andrew Wedel, Abby Kaplan, and Scott A. Jackson. 2013b. High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, 128(2):179–186.