Two Ising-Type Models of Language Evolution

Ludo Oliver

Supervised by Dr C. Houghton

Department of Engineering Mathematics

University of Bristol

Project Report submitted in support of the degree of Bachelor of Engineering

This work was carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol - http://www.bristol.ac.uk/acrc

April 2024

Abstract

The evolution and propagation of language is a complex process, occurring across vast stretches of time and driven by poorly understood mechanisms. Instead of attempting to accurately replicate this complexity, current models focus on using simple models that still capture some characteristics of these processes. This project explores two such models, a Preference Model proposed by Conor Houghton [1] and the Threshold Model novel to this project. Both are based upon the Ising model, used in statistical physics to generalise the behaviour of ferromagnetic materials. Whilst both models utilise vector representations of language, they differ in how they define the interactions of each speaker. The Preference Model utilises a communication preference between speakers of closely aligned languages, whilst the Threshold Model allows for interactions between all speakers of languages within a certain threshold. An exploration of the dynamics of both models was made, before an analysis of the language distribution produced by both models. It was shown each model reproduced some of the characteristic traits of the global language distribution, with the Preference Model ultimately providing an improved estimation of the real-world data. The success of both models offers compelling evidence of the potential for adapted Ising models within a linguistic context.

Contents

1 Introduction

1.1 Motivation

Evaluating hypotheses about the evolution of human language into its current form presents a huge challenge for the field of linguistics. The archaeological record is severely constrained, not only to those languages that postdate the development of writing, but also by the limits of our translations. With little physical evidence, inquiry into the earliest developments of language was historically dominated by unverifiable speculation. However, with the constant improvement seen in modern simulation methods, as well as an increase in computing resources, interdisciplinary research has increased. Models based on other physical and biological processes are now being adapted to linguistics with increasing success. These developments in both scope and scale have now also allowed for the creation of novel models to examine current hypotheses about language dynamics.

Improvements in modelling methods have also greatly benefited the field of statistical mechanics, allowing researchers to explore beyond the limits of analytical solutions. This is exemplified in the Ising model, a model of phase transitions created in the early 20th century. Although analytic solutions for both the 1 and later 2-dimensional cases were found, ≥ 3 dimensional instances were later proven NP-complete. Nevertheless, Monte-Carlo modelling methods have provided numerical solutions for many formulations of both Ising and Ising-like models. The complex emergent behaviour seen within the Ising model has been shown to mirror phenomena seen across both the natural and physical world, including human behaviour, language dynamics and cognition.

Building upon the initial work of Conor Houghton [1], this project intends to investigate two modified Ising-type models applicable to questions of language evolution. Both will be evaluated against the observed distributions of real languages, seeking to reflect some holistic characteristics of real language distribution. The aim of this project is to illustrate the potential in modifying the Ising model, with the hope of providing a foundation for further exploration of other linguistic hypotheses using related models.

1.2 Existing Work

A dominant paradigm in 20th and 21st century linguistics is built upon the work of Noam Chomsky, who argued for language as an innate faculty, present at a genetic level [2]. Under this naturist perspective, syntax and grammatical structure are seen an emergent property of the computational process present in human thought. Social interactions and environmental factors are considered as deciding only the semantics of a particular language. In his theory of universal grammar, Chomsky proposes a hierarchy of structural complexity within language, each increasing the power of language as a tool. Historically, research on the biological faculty of language often assumed a near Chomskian perspective, looking at evolutionary markers of the brain development necessary to move through this hierarchy, as shown in work such as Berwick et al.[3]. Competing perspectives instead argue for interactions between speakers as the driving force of both language acquisition and development within populations. These paradigms are often found at the core of agent-based modelling, an increasingly dominant tool in the exploration of language dynamics, as outlined in a review by Boissonneault et al. [4]. This review also highlights a clear shift towards population level models of language evolution, as opposed to the evolution of a speaker's language.

One of the key developments within the field of agent-based models was the work of Luc Steels and his introduction of the naming game [5]. Steels begins his paper with an explicit rejection of Chomskian theory and instead declares his belief in an interaction paradigm of language evolution. Steels then introduces a game-like model, in which the ability to communicate between speakers is a trait selected for over time through simple inter-agent interactions. Many implementations and modifications of this model have seen great success, with emergent properties such as grammar [6] and the ability to introduce the evolution of new language [7]. These results show a broad applicability of this method and its potential within implementing novel or unexplored theories about the evolution of language. Of particular relevance to this project is the work of Dorota Lipowska and Adam Lipowski [8], and their implementation of a communication preference within a lattice formulation of the naming game model. Although not fully explored, their results show strong similarities between the phase transitions within this modified naming game and the Ising model. This connection between these two models further establishes the relevance of applying the Ising model to questions of language evolution.

Created in the 1920s, the Ising model was initially proposed in order to model the properties of ferromagnetic materials and analyse their behaviour under different temperature regimes [9]. Below a certain temperature, ferromagnets can spontaneously align their constituent electrons, a property which is lost at higher temperatures. This shift in behaviour, based on a small change in one of a system's parameters, constitutes a phase transition, a property of many complex systems. Indeed, whilst most classical examples of phase transitions focus on physical phenomena, a review by Heffern et al. [10] illustrates how prevalent this behaviour is within both a biological and ecological context. Research into these systems often focuses on behaviour at this boundary, known as the system's critical region. One such paper, by Marinazzo et al. [11], outlines a particular instance of an Ising like model of neural computation. In this paper,

information transfer across the model is maximised at just below the system's critical point. The complex emergent behaviour of even simple systems in this region is a key reason for the Ising model's success at replicating seemingly unconnected phenomena.

Although based upon a separate paradigm to the one adopted within this project, the work of Stauffer et al. [12] draws clear parallels in model design. This model uses a 2D grid, occupied by agents assigned with a vector language, with two possible states for each element of this vector. The agent-based interactions of this model draw from a differential equation model of language competition created by Oliveira et al. [13] in which focus is placed heavily on inter-language competition. Oliveira's model also contains a phase transition between regimes of linguistic homogeneity and complete disorder, with their models best approximating the real-world language distribution within the system's critical region. This also reinforces the potential offered by the emergent behaviour of such systems at the point of phase transition. Interactions within the system take place through a mechanism of 'colonisation', in which agents spread themselves into unoccupied regions of the lattice. This paper also implements random mutation to the language vector and a fitness metric, allowing for the simulation of evolutionary processes by favouring higher fitness populations.

Another lattice based linguistics model is seen in the work of Siva et al. [14], in which a model of the propagation of grammar is proposed. This paper also adapts an Ising-type model, with a focus on structural relationships between the nodes instead of the system's global properties. However, it is purely a model of grammar, and as such is of little direct relevance to the models proposed by this paper.

Stauffer's paper evaluates their results according to their ability to replicate the language distributions observed by Gomes et al. [15]. This paper outlines multiple distribution laws which closely match the global distribution of spoken languages and their respective number of speakers. However, despite the Stauffer paper's success in matching these distributions, it sits opposed to the current paradigm of most models of linguistic dynamics. The lack of agent-agent interactions ignores the use of language as a method of communication, whilst the static nature of language in colonised nodes fails to reflect observed shifts in language that occur even within isolated communities. Whilst the binary vectors used to represent language introduce similar linguistic diversity, no consideration is made for implementing relatedness between languages, instead these vectors simply serve as the identifying labels for each population.

This discretisation of languages, prevalent in many comparable models, sits at odds with the current understanding of language as a continuum, as illustrated by the existence of dialects. Dialects exist as distinct linguistic objects, each somewhat arbitrarily defined not as a separate language but as a subset of another. This distinction is made not only by linguistic differences, but also by shifts in identity, as documented in work such as that of R. Greenberg [16]. Following the dissolution of Yugoslavia in the 1980s, Greenberg notes the individual dialects once perceived as members of a common Yugoslavian language splintered into what are now regarded as four distinct languages, each with separate national identifies.

Despite the difficultly in distinguishing languages and dialects, a paper by S. Wichmann [17] attempts to quantify this distinction numerically. This boundary is calculated using the Levenshtein Edit Distance (LVD), a measure of how many substitutions, deletions and insertions are required to map from one string to another initially proposed in 1966 [18]. Using generally agreed upon classifications of observed languages and dialects, Wichmann finds this distinction is typically made at a mean LVD of 0.51, a threshold which will be used throughout this paper. Other research has tried to apply the LVD measure to create a threshold from which two languages become unintelligible, such as in the work of Gooskens et al. [19] and Tang et al. [20]. Both these papers find a statistically significant correlation between LVD and intelligibility, however neither establish a clear boundary. Both studies are also constrained to small dialect families of less than 15 languages, as opposed to the global scope of the Wichmann paper.

The boundary used by Wichmann to distinguish between languages and dialects based this distinction, like much of contemporary linguistics, was found using the classifications made within the Ethnologue database [21]. First published in 1951, the project has since expanded and now catalogues several thousand languages and their properties. Of particular interest to this project is the population data given for each language, used within papers such as Stauffer et al. [12] to evaluate the language distributions produced within their own models. Language populations fit an approximately log-normal distribution, a property also seen in other population dynamics, at both a cellular [22] and ecological [23] level.

The log-normality of language may also explain the scaling laws described by Gomes et al. [15], as discussed in work such as that of Mitzenmacher [24]. This paper demonstrates how the sum of several log-normal distributions can closely replicate a power law distribution, with both analytic and empirical evidence. However, this connection has only been tentatively established, as such the global log-normal distribution of language will be the main empirical comparison for this project, leaving power law relationships for future investigation.

2 Method

2.1 The Classical Ising Model

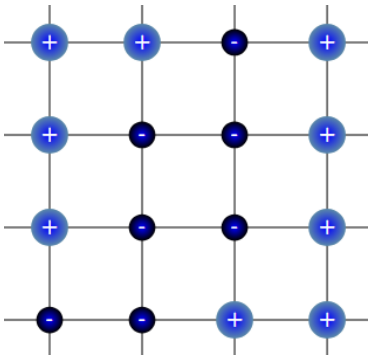

Figure 1: An illustration of a 4×4 subsection of a populated lattice, containing both up and down spin sites. Nodes that share an edge are considered nearest neighbours

The Ising model was first proposed as a simple model of ferromagnetic materials [9], substances that can undergo spontaneous magnetisation. This behaviour occurs when, below a critical temperature, the spins of a material's constituent electrons instantaneously align, even without the presence of an external magnetic field. This sudden self alignment results in the magnetisation of these materials and is key to many industrial processes. However, above a threshold temperature, ferromagnetic materials will no longer exhibit spontaneous magnetisation. This shift in behaviour, emerging from a small change in the system's parameters, is a classic example of a phase transition. Such changes in dynamics are a complex process, with novel behaviour emerging when the system occupies the critical region between these two temperature regimes

In order to replicate the observed behaviour, the Ising model first generalises these materials as lattices - connected grids, in which each node is assigned a spin state of $s_i = \pm 1$. Each node may be interpreted as an electron fixed within a metal's structure, although this analogy is not intended to be completely rigorous. An example of a lattice is shown in Figure 1 - here it should be noted only the 4 directly connected nodes are considered neighbours within the classical Ising model. The energy of the system is then defined bas the sum of the local interactions of each individual node i and its neighbours, as well as the external magnetic field at that location H_i. For a system of N nodes, with $\langle i, j \rangle$ denoting the set of nearest neighbours and J_{ij} representing the interaction strength of nodes i and j, [25] defines the system's Hamiltonian as

$$H = \sum_{\langle i,j \rangle} J_{ij} \mathbf{s}_i \mathbf{s}_j + \sum_{i=1}^{N} H_i \mathbf{s}_i. \tag{1}$$

Within this system, pairs of adjacent nodes with opposite spins are energetically unfavourable, as are nodes that are in opposition to their local magnetic field. Over time, the system tends towards a more energetically favourable state, with nodes aligning themselves towards their local environment. In opposition to this trend towards uniformity, thermodynamic effects cause fluc-

tuations in spins within the system, moving it away from a possible energetic minimum. Above the system's critical temperature, T_c, the entropic contribution of these fluctuations surpass any magnetic contributions, and the system no longer self aligns. The systems' temperature is a relative value and is not intended to rigorously represent temperature, although once again it remains broadly analogous to our intuitive understanding of heat. It should also be noted that, below this critical temperature, the system's equilibria are comprised of only a single spin state, a significant motivation for the modifications presented in section 2.3. The Ising model's broad generalisation of a phase transition into a model with minimal complexity has made it a popular example of such behaviour. As a result, its phase behaviour is well documented in current literature. Analytic solutions exist for the 1 and 2-dimensional instances and, although not analytically solvable [26], higher complexity versions can still be numerically evaluated using Monte-Carlo methods. A solution in this context is a description of the systems dynamics and critical behaviour.

2.2 The Metropolis Hastings Algorithm

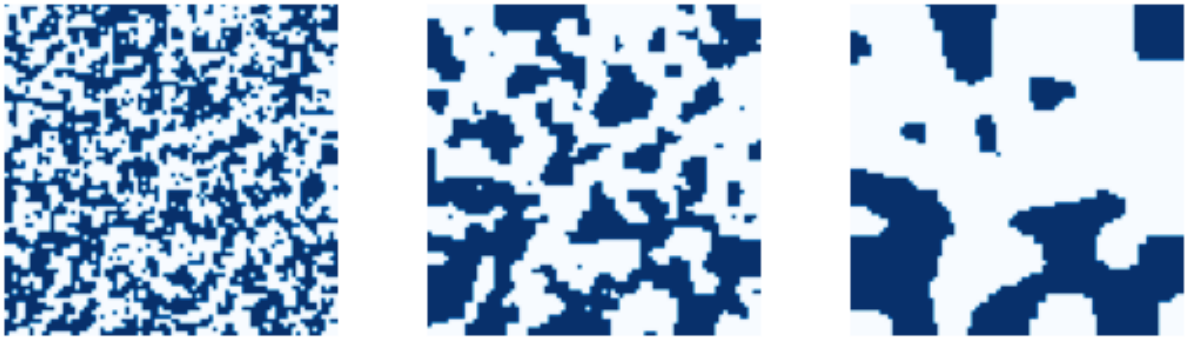

Figure 2: A visualisation of spin states within the classical Ising model, across 3 steps of the Metropolis Hastings method, carried out on a 60x60 lattice and at T=0.1. The equilibrium state is not included, as the lattice eventually converges to a monochrome, homogeneous state. Spins $+1, -1$ are coloured blue and white respectively

Devised as a method of Monte-Carlo simulation, the Metropolis-Hastings (MH) algorithm serves as a powerful tool for approximating samples of complex distributions [27], and although it may be generalised to many problems it is predominately used within statistical mechanics. Contextualised to the Ising model, the algorithm begins with the initialisation of a lattice, where each node is randomly assigned a spin value of ± 1. For each time-step the spin of a random node is flipped and the resulting difference in the system's energy, ΔE, is evaluated. If energetically favourable, i.e. $\Delta E < 0$, this change is accepted and the algorithm progresses to a new, randomly selected node. Thermodynamic effects are then introduced by accepting state changes in the $\Delta E > 0$ case, at a probability $P = e^{-\frac{\Delta E}{T}}$. Since the probability of these thermodynamic effects increases with temperature, the phase transition at T_c can be understood as the point at which the frequency of the random fluctuations exceeds the rate of self-alignment. This process of randomly modifying spins is continued over the given time period, or until the system's energy converges to its equilibrium value.

Despite the algorithm's reference to time-steps, it should be noted that it does not seek to replicate the temporal behaviour of an Ising type system. Instead, MH is used to sample possible lattice configurations at a given temperature. As such, no analysis of either model's behaviour within the MH algorithm has been presented within this project. The implementation of this algorithm for the proposed modified Ising models is further detailed in 2.4.

2.3 Modifications to the Classical Ising Model

The current paradigm for modelling of both the evolution and propagation of language focuses its efforts on producing or adapting simple models, questioning how well they can reflect language dynamics on a more holistic scale. Such models often include stochastic effects, such as adding random the mutations to language, as seen in phenomena such as lexical drift. These models also often focus on the local interactions of individual speakers as the driving force for the emergence of a common language. Such models also often focus on producing their behaviour through minimal complexity, with their dynamics emerging from seemingly simple rules. The Ising model presents a strong candidate for such modelling, well motivated through an anthropomorphism of the model.

By considering each node as the speaker of an individual spin language and interpreting the interactions of individual nodes as conversations, many properties of the Ising model seem well suited for linguistics based models. The system's temperature parameter introduces the necessary stochastic effects, and a framework for simulating these conversations exists within the MH-algorithm. However, under this anthropomorphisation, the classical Ising model could only represent two disjoint languages, with no room for bilingual speakers. This model would also inevitably converge to either linguistic homogeneity below its critical temperature, and complete disorder above it. Even ignoring the sociological implications of a model in which linguistic homogeneity is inevitable, this clearly fails to capture the $\approx 10^4$ separate languages seen today, as well as the sometimes continuous nature of real language boundaries. This diversity is therefore introduced through a simple modification, assigning each node not with a single spin state $s_i = \pm 1$ but instead a n length spin vector $\mathbf{s_i} = (s_{i1}, s_{i2}, \ldots, s_{in})$ - allowing for 2^n spin languages. However, this modification still presents the same previously seen convergence towards uniform settled states. As the individual layers are not directly linked, this would only create n uncoupled Ising models, instead of a cohesive system. Two methods of coupling these models are examined within this paper - the Preference and Threshold models - both with the goal of allowing heterogeneous language distribution at an equilibrium state.

The Threshold Model

In this project I propose a novel model, inspired by the Preference Model previously introduced in [1], which I will refer to this model as the Threshold Model (TM). The motivation for this model is a conjecture that a language speaker will not select a single preferred conversational

partner, rather they will converse with, and hence learn from, any other speaker whose language is sufficiently close to their own. During each time-step of the MH algorithm, the selected node will evaluate the Hamming distance to all adjacent nodes. Only adjacent nodes within a distance θ are used to evaluate the nearest neighbour interactions of each MH time step. In the case that no neighbouring node is within the threshold distance this interaction instead utilises all four adjacent nodes. . Normalised Hamming distance was implemented as a simple metric to quantify this closeness, with a threshold value of 0.51 being selected based on the work of Wichmann [17]. Hamming distance is a measure of how many substitutions are required to edit one string into another, without the deletion or insertion operations characteristic of the LVD measure used by Wichmann. Given the reduced number of possible operations, the Hamming distance serves as an upper bound of the LVD between two strings of the same length and so will be used in place of LVD for this project. Normalising for the length of the language vectors n, the Hamming distance between two nodes $\mathbf{s}_i, \mathbf{s}_j$, may be calculated as

$$h(\mathbf{s}_i, \mathbf{s}_j) = \frac{1}{n} \sum_{x=1}^{n} \mid s_{ix} - s_{jx} \mid . \tag{2}$$

Formalised further, the Threshold model redefines a node's set of nearest neighbours $\langle i, j \rangle$ as

$$\langle i, j \rangle = \{ j \in J \mid h(i, j) < \theta \}, \tag{3}$$

for the four adjacent nodes J and with the Hamming distance metric defined in Equation 2. This relation allows for separate clusters of disjoint languages without influencing each other, avoiding the trend towards uniformity seen in the classical model.

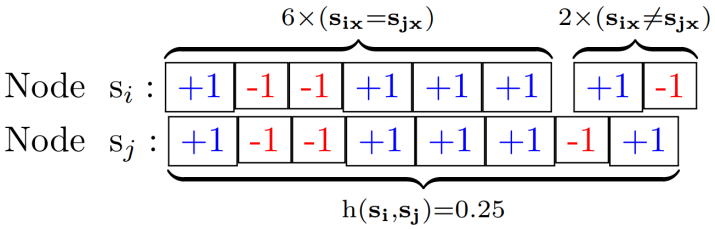

Figure 3: An example of two nodes with n=8 language vectors, at a distance 0.25

The Preference Model

Proposed by Conor Houghton in [1], the Preference Model (PM) redefines a node's nearest neighbours as the single neighbouring node with a language vector closest to itself under the hamming distance measure. Anthropomorphising once more, this models individual speakers as only interacting with those they share the most with linguistically, allowing distinct clusters to emerge as a result. Once again, this relation avoids the system reaching complete homogeneity as, by choosing to ignore each other, speakers of dissimilar languages can still share an edge without interacting.

Formalising this relationship the preferred nearest neighbour of the PM, $\langle i, j \rangle$, is redefined as

$$\langle i, j \rangle = j \mid \max_{j \in J} h(i, j), \tag{4}$$

once again for the set of adjacent nodes J and Hamming distance defined in Equation 2. If multiple nodes are equidistant from i, the neighbour j is chosen randomly.

2.4 Implementation and Simulation

Both models implemented using no external magnetic field, a constant connection strength between each node and upon a $L \times L$ square lattice. Nodes along the edge of the lattice were considered as having a reduced neighbourhood, rather than the periodic boundary conditions sometimes implemented in Ising-type models. The energy was calculated as the sum of the of interactions between each node i and its 4 adjacent nodes J, formalised as

$$E = -\frac{1}{nL^2} \sum_{i,J} (\mathbf{s}_i \cdot \mathbf{s}_j) \ . \tag{5}$$

The energy value is also normalised for n - the length of the language vector, as well as the lattice size L. This simplified form of Equation 1 also allows for a more intuitive understanding of the system's energetic behaviour. The systems' energy is at its lowest when the dot product between each of the two vectors is at its maximum for each element of the sum. This corresponds to a lattice populated completely with a single spin vector and a normalised energy of exactly -1.

As per the standard MH algorithm, each node of the lattice was initialised with random spin vectors, using a uniform distribution of ± 1 elements. For each randomly selected node $\mathbf{s}_i$,the indices of its nearest neighbours $\langle i, j \rangle$ were calculated, as per either Equation 4 or Equation 3 for the PM and TM respectively. A single element x of the spin vector for this node was then randomly flipped. The localised change in energy ($\Delta \hat{E}$) caused by this flip could then be calculated. This value is not the overall change in system energy, as per the standard MH implementation, and instead it only considers the change within a node's preferred neighbourhood. Formalising this relationship gives $\Delta \hat{E}$

$$\Delta \hat{E} = \frac{2}{n} s_{ix} \sum_{j \in \langle i,j \rangle} s_{jx} \ . \tag{6}$$

In the $\Delta \hat{E} < 0$ case this change in spin is instantly accepted. Conversely, changes with positive $\Delta \hat{E}$ are randomly accepted with probability $P(accept) = e^{-\frac{\Delta \hat{E}}{T}}$.

To run each simulation, every individual speaker was represented as a class instance, with each step of the MH algorithm being defined as a method on that class. Additionally, these classes stored the language vector of each speaker and the indices of its adjacent nodes. Simulations were conducted in large batches using the BluePebble1 high performance computing facilities. Each simulation was initialised as part of a job array- a large collection of individual scripts.

Parameters such as the grid size and temperature values could then be automatically passed in as arguments. This allowed for large numbers of simulations (≥ 100) to be run in parallel. Each parameter and model combination investigated was simulated 10 times during initial investigations, before moving to 15 runs once a basic analysis of the dynamics was completed. This was vital to investigate the systems' behaviour in any degree of detail, as each individual simulation could take several hours to converge into a steady state. Results from each run were outputted with the parameter values encoded, and then imported into locally run visualisation programs, as the current BluePebble configuration does not allow for the use of visualisation software. The visualisations of the Probability Density Functions (PDFs) and time derivatives utilised interpolation functions to allow for a clearer representation of the shape of both, however these functions were not used for the statistical analysis of any of the models

Unless explicitly stated, each simulation was conducted on a 2 dimensional 300×300 lattice, with length $n = 8$ spin vectors and across 2×10^9 time steps. Justification for the $L = 300$ grid size is provided in section 3.1, whilst the language vector length of eight served as a balance between language diversity and practical limitations. The use of 2×10^9 was decided upon as an upper bound for convergence following a brief examination of the relationship between time to converge, language vector length and grid size. A moving average of energy was evaluated every $\approx 6 \times 10^7$ time steps - if the mean of the last three energy values was within 1% of the current energy value the simulation was concluded, as it was assumed to have converged.

Later implementations of the algorithm also introduced some optimisations. Since the relationship between nodes is symmetric, the efficiency of the energy calculation could be significantly improved by only evaluating $s_i \cdot s_j$ once for each pair of i, j, moving across the lattice in a series of diagonal lines. This halved the time taken to evaluate the energy, one of the slowest operations in the procedure.

3 Results

3.1 Phase behaviour

As the defining feature of the Ising model is the presence of a phase transition, initial efforts focused on establishing an understanding of each system's phase behaviour. This included identifying if such a phase transition was even present and, if so, locating the point of this transition. However, due to the discontinuous boundary conditions implemented within both model, there was also a clear need to first understand the effects of lattice length L upon the system's behaviour. The settled state energies of each model were evaluated across five separate grid lengths in the interval $[60, 300]$ with temperatures varying between $(0, 1]$, as illustrated in figure 4. In both models, the energy temperature relationship appeared to converge towards some asymptotic curve with an increase in grid size, probably due to the reduced contribution of boundary effects from nodes on the lattice. For smaller grid sizes both models also reached their minimum energy, not at $T = 0$ as expected. This behaviour disappeared for larger values of L within the

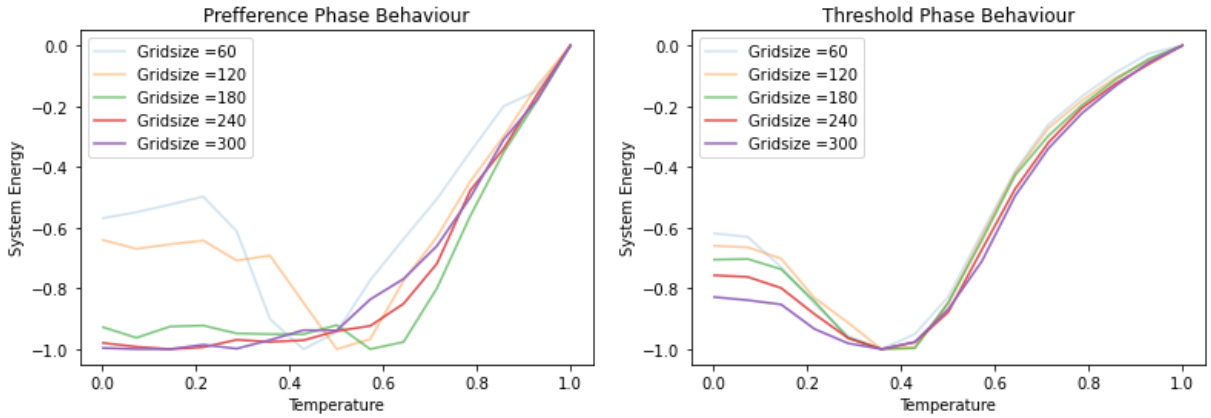

Figure 4: The normalised settled energy and temperature relationships of the preference and threshold Ising models, evaluated across 5 separate grid-sizes in the interval $L = [60, 300]$. Each curve was calculated for $0 < T \leq 1$.

PM, but remained present throughout the TM. However, in the TM this energetic minimum appears consistent throughout increasing grid sizes, although the difference between the energy at this location and the energy at the expected minima of $T = 0$ is reduced significantly for each increase in the value of L, clear evidence this is due to the finite boundaries used.

Based on these results for the TM, balanced with the practical limitations imposed by the $\mathcal{O}(L^2)$ time and space complexity of each simulation, a grid size of $L = 300$ was decided upon. This was deemed an appropriate balance between these boundary effects and the practical demands of running each individual simulation. As shown by Ostmeyer et al. [28], similar conclusions have also been drawn for the classical Ising model - a lattice length of 300 is shown to converge to an energy value of within $\leq 5\%$ of a $L = 900$ lattice. This $L = 300$ value also produces stable results within the preference model, and unless explicitly stated otherwise will be used as the default grid size throughout the rest of the paper.

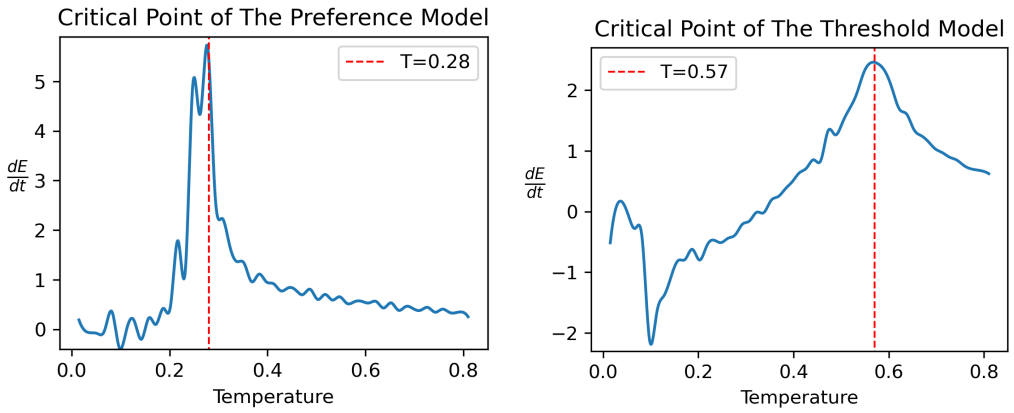

Figure 5: The rate of change of energy with respect to temperature, evaluated using finite difference methods. Clear peaks in this value are shown for $T = 0.28$ in the PM and $T = 0.57$ in the TM The curve shown was fitted using SciPy's interpolate function to aid in visualisation. This was evaluated for $L = 300, 0 < T \leq 0.825$ and with 2×10^9 time-steps

Having established a value for grid size for the rest of the project investigation into the dynamics of both systems began. The settled energy of each system was evaluated at 55 steps, 0.015 apart, across the interval $T = (0, 0.825]$. Here, the derivative of energy with respect to temperature was evaluated using finite difference methods, as shown in Figure 5. Each model has a clear maximum at values of $T = 0.28$ and $T = 0.57$ for the PM and TM respectively.

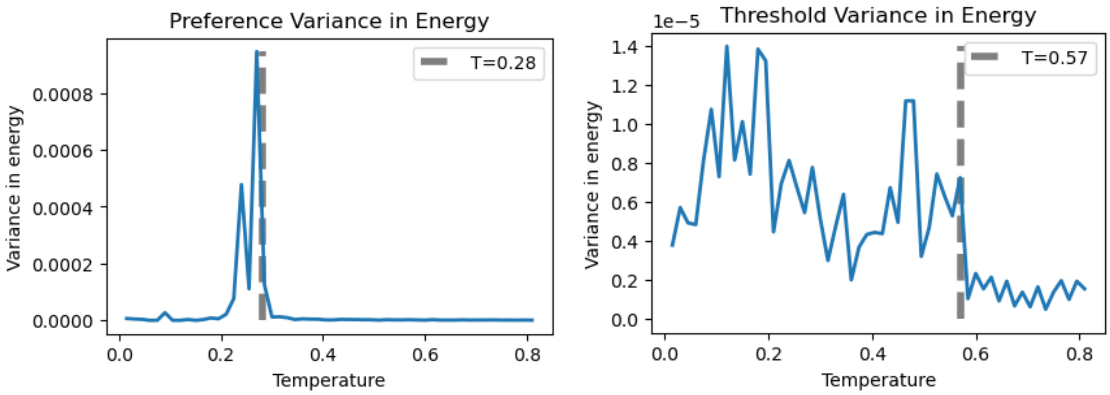

Figure 6: The variance in energy across 15 runs of each model. In both models, sudden decrease in variance could be seen at $T = 0.28$ for the PM and $T = 0.57$ for the TM. This was evaluated for $L = 300, 0 < T \leq 0.825$ and with 2×10^9 time-steps.

This location for the critical point was also evidenced by other changes in dynamics at these locations. Above the identified critical values, both models showed a clear decrease in variance, as illustrated in Figure 6.

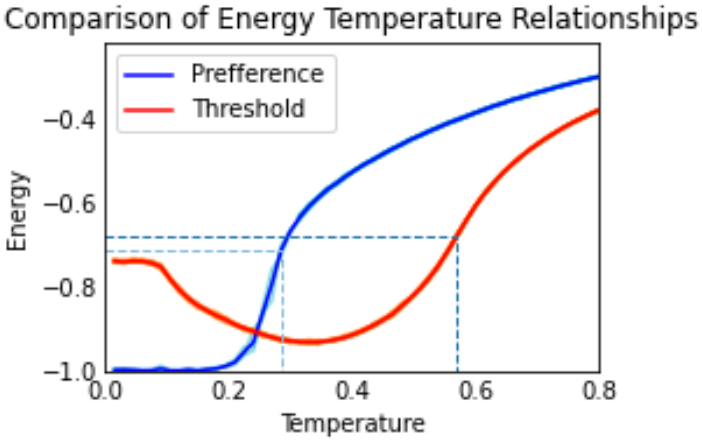

Figure 7: The energy temperature relationships of both models, with the critical point of each highlighted. Each curve is the mean of 15 runs, with each of the 15 shadowed around them. This was evaluated for $L = 300, 0 < T \leq 0.825$ and with 2×10^9 time-steps

3.2 Language Distribution

As established in previous work such as [12], one main way of evaluating the success of population-level models of language is comparing the relationship of n_s, the number of languages, with N their number of speakers. These curves are often referred to throughout the literature simply

as the n_s curve of each model.

To begin estimating these distributions, population counts from each of the 15 batches of simulations were aggregated together, producing 3840 sample languages with a total of 1.35×10^6 speakers at each temperature value. This aggregation is consistent with the standard MH procedure, as the lattices produced by this algorithm are intended only as samples of the specific model configuration. Using this data, a discrete estimate of the n_s Probability Density Function (PDF) was created using the histogram of population counts for each language, with the area rescaled to 1 in order to ensure it was a true PDF. Next, both log-normal and normal PDFs were fitted using maximum likelihood estimation, in which the parameters describing each distribution were optimised to be close as possible to the curve. As outlined in [29], log-normal distributions are defined by the following probability density function,

$$p(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}}. \tag{7}$$

In this distribution, it is the logarithm of the random variable x that follows a normal distribution of mean μ and variance σ^2. For a set of n samples $(x_1, \ldots, x_n)$, these parameters are best estimated as $\mu = \frac{1}{n}\sum_{i=1}^{n} ln(x_i)$ and $\sigma = \frac{1}{n}\sum_{i=1}^{n}(ln(x_i) - u)^2$.

Two goodness of fit tests were then performed under two separate hypotheses; $h_0 : N$ was normally distributed and $h_a : log(N)$ was normally distributed. Each test was implemented using the *scipy.stats.goodness_of_fit* module and are adapted from Stute et al. [30]. In this test, parameter estimations for the whole distribution are calculated before several subsets of the original data are taken as samples. The JSD divergence measures later outlined in Equation 10 are then used to calculate how well these estimated distributions describe the data they were fit to. Next, the parameters that best fit each sample are calculated, along with the corresponding JSD divergence. The p-value is then calculated as the proportion of samples with divergence higher than the initial distribution.

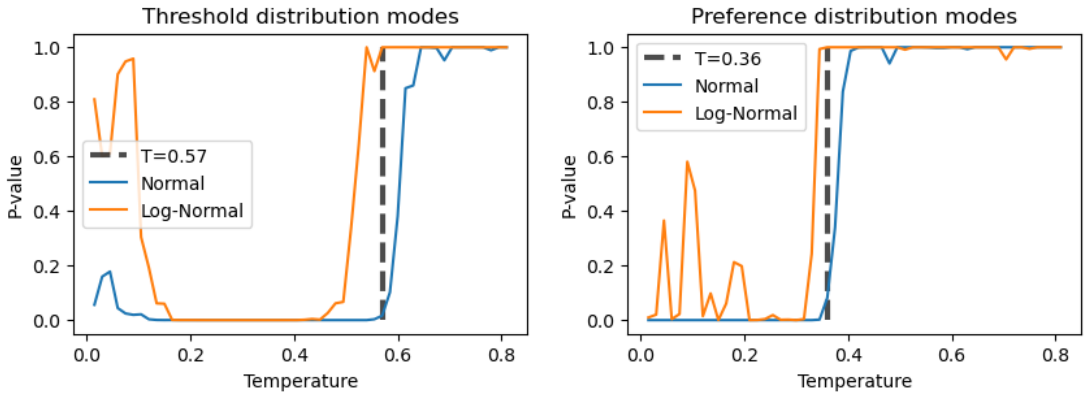

Figure 8: The p-values of the two separate hypotheses tests across the Temperature range $T = (0, 0.825]$. Values of $T = 0.57$ and $T = 0.36$ were chosen as the points where both models showed log-normal characteristics, with enough evidence to still reject normality. This was evaluated for $L = 300$ and with 2×10^9 time steps

The results obtained for these tests on both models are shown in Figure 8. Behaviour within the

lower temperature regions of both models was not further investigated, as these regions showed significant instability in their dynamics. These tests showed a clear transition between types of distribution for both models. For the TM this happened for $0.55 \leq T \leq 0.6$. In this region the p-value of < 0.001 was sufficient evidence to reject normality but not enough to reject a log-normal distribution. The proximity to the T_c value identified suggests this is due to critical behaviour of the system. A similar transition is also shown within the PM, although this does not seem to occur at the identified critical temperature of the model. Instead, for $T = 0.28$, both hypotheses can be rejected with confidence. The new location for potential log-normality was instead found to be $T = 0.36$. Further analysis of the dynamics of the PM therefore focused on this location instead, although the model's behaviour at $T = 0.28$ was still explored. The estimated distributions of both these models can be seen in Figure 9. Whilst visually similar, at these temperatures both models had low enough p-values to reject normality. Although the p-values of approximately 1 do not confirm the model's log-normality, they do suggest the distributions can be well approximated by some log-normal curve, much like human language. Below the temperature regions identified for both models there was sufficient evidence to reject either hypothesis. Above these proposed locations, tests for both normality and log-normality reported p-values of ≈ 1, and as such were not explored further.

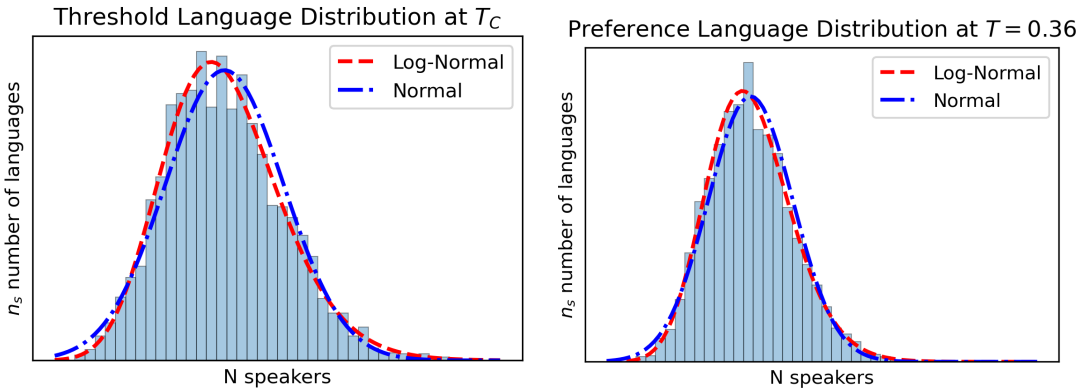

Figure 9: Curve fitting normal and log-normal distributions to the TM at $T = 0.57$ and PM at T= 0.28 respectively. Despite the visual similarity, goodness of fit tests gave a p-value of ≤ 0.001 for either variable being normally distributed. Also shown is the area = 1 normalised histogram of each distribution, using 40 discrete bins.

3.3 Comparison to Empirical Data

As the principal language dataset for linguistic analysis, the Ethnologue dataset [21] offers detailed statistics on the $\approx 7 \times 10^4$ globally spoken languages and the $\approx 8 \times 10^9$ people who speak them. Given the difference in scale between this dataset and the results of each model, adjustments were made to allow for a direct comparison. Normalising the logarithm of N to be within the range $[0, 1]$ for each dataset brought the results of each model into the same range, without modifying the shape of each PDF. Having been brought into the same axes, results could then be directly compared as shown in Figure 10. Three separate configurations were chosen for analysis: the PM at $T = 0.36$ ($PM\,|_{0.36}$), the PM at $T = 0.28$ ($PM\,|_{0.28}$) and the

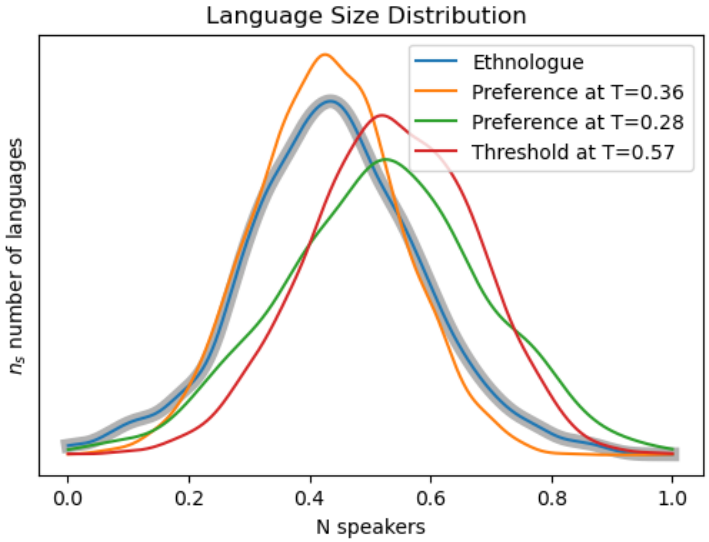

Figure 10: A comparison of the Ethnologue language dataset to the results obtained for the PM at $T = 0.36$, the PM at $T = 0.28$ and the TM at 0.57. Each curve shows the PDF of a different model configuration, comparing of the logarithm of N, the number of speakers, with n_s the number of languages. Each PDF was normalised between 0 and 1 to allow for direct visual comparison.

TM at 0.57 ($TM\ |_{0.57}$).

Although both the TM and two PM configurations show an approximately normal distribution of $log(N)$, the correlation between the PM and the real-world data is much higher, as shown in Table 1. The $PM\ |_{0.36}$ and Ethnologue PDFs share both a modal and mean (μ) value of 0.43 ± 0.01. This is significantly lower than the modes observed for $PM\ |_{0.28}$ or the $TM\ |_0$ .57 at 0.52 and 0.53 respectively. However, the $PM\ |_{0.36}$ model does still differ both in width and height, as shown by its lower standard deviation, $\sigma = 0.120$, in comparison to the Ethnologue value of $\sigma = 0.148$. Comparing the overlap in area of n_s for each of the three curves further illustrates the difference in fit, with the $PM\ |_{0.36}$ sharing 89% of its area with the Ethnologue distribution. The PDFs identified for both the $PM\ |_{0.28}$ and the $TM\ |_{0.57}$ had less than 70% overlap with the Ethnologue curves.

A common statistic used to quantify the difference between two probability distributions is the Kullback-Leibler divergence (KLD). This measure is defined as the information lost when approximating one distribution from another, as described by Kullback et al. in [31], meaning a KLD lower score in Table 1 indicates a better approximation of the Ethnologue distribution. As per[31], Kl divergence can be calculated as the following, for $p_i, q_{,i}$ elements of length k discrete PDFs P, Q;

$$KLD(P,Q) = \sum_{i=1}^{k} p_i log\left(\frac{p_i}{q_i}\right). \tag{8}$$

The calculated value of the KLD between Ethnologue and both of the PM n_s curves is significantly lower than with TM, showing both are better estimators for the actual distribution. Other divergence measures were also used to quantify this similarity. The Jensen-Shannon Di-

vergence (JSD) is a symmetric measure, as opposed to the more unstable KLD, which still provides information on how well each distribution approximates the other. As per [32], calculating JSD begins with defining M, a new discrete PDF that takes the element-wise average of each distribution,

$$M = \{\frac{p_1 + q_1}{2}, \frac{p_2 + q_2}{2}, \ldots, \frac{p_k + q_k}{2}\}. \tag{9}$$

The JSD of the two distributions is then defined as the average KLD between each PDF and M, formalised as

$$JSD(P, Q) = \frac{1}{2}\left(KLD(P, M) + KLD(Q, M)\right). \tag{10}$$

Here the observed JSD value was calculated between each estimated PDF and the Ethnologue dataset. The lowest value was for the PM at $T = 0.36$, again indicating it was closest to the real distribution. With the sum of this evidence, it could be concluded that the PM at $T = 0.36$ was the best approximation of the global n_s distribution.

	TM (T=0.57)	PM (T=0.28)	PM (T=0.36)	Ethnologue
Mode	0.520	0.526	0.426	0.435
μ	0.534	0.519	0.429	0.439
σ	0.138	0.168	0.120	0.148
Overlap	0.633	0.700	0.890	1
KLD	0.131	0.061	0.059	0
JSD	0.151	0.099	0.087	0

Table 1: A quantitative comparison of three distributions to the Ethnologue dataset. These were the TM at $T = 0.57$, and the PM at $T = 0.28$ and $T = 0.36$. These results are for the log-normalised PDF's of N, the number of speakers

3.4 Further Population Analysis

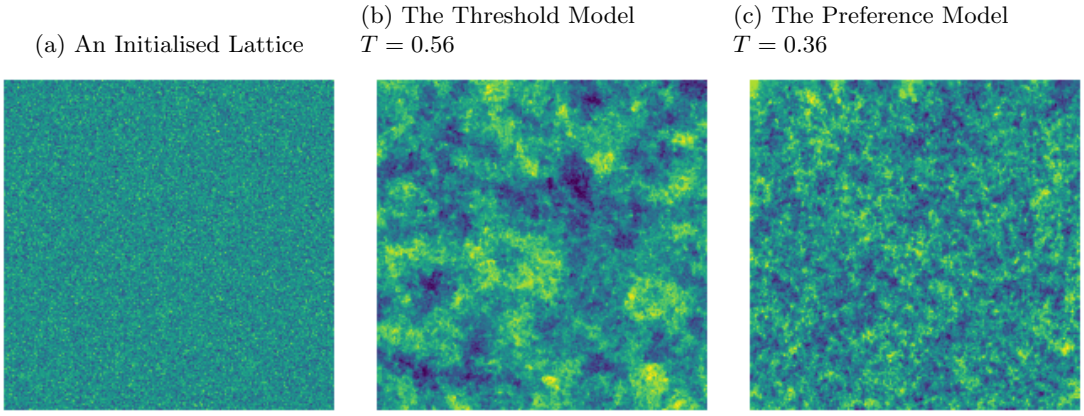

Figure 11: (a) An uninitialised lattice shown next to examples of settled state of (b) the TM at T_c and (c) the PM at $T = 0.36$. This was evaluated for $L = 300$ and with 2×10^9 time-steps

Visualising the settled states of each model at their points of log-normality offered insight into the resulting language distribution. Each language was mapped to a unique colour, allowing individual lattices to be plotted as shown in Figure 11. Visually, the TM produced more clearly

defined regions of closely aligned nodes, as shown in its larger monochrome regions. By masking over each language independently and tracking the size of contiguous groups, clusters of homogeneous language could be identified. The mean cluster size of the TM at T_c was 5.86 nodes, in comparison to 7.03 nodes for the PM. The PM also exhibited a much greater variance in cluster size at $T = 0.36$, with a standard deviation of $\sigma \approx 11$ compared to $\sigma \approx 16$ for the TM.

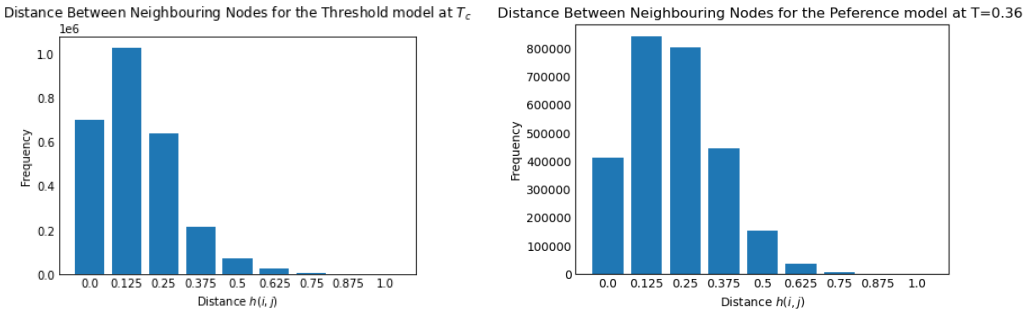

Figure 12: The distribution of the Hamming distances between adjacent nodes for each model, with $T = 0.57$ and $T = 0.36$ for the TM and PM respectively. This was evaluated for $L = 300$ and with 2×10^9 time-steps

The distance between nodes was also investigated, as shown in Figure 12. Nodes in the $PM\ |_{0.36}$ had a mean distance of ≈ 0.21 from their directly adjacent nodes, as opposed to a mean distance of ≈ 0.16 for the TM. The higher mean distance of the $PM\ |_{0.36}$ may also explain the leftwards skew of its PDF, as seen in Figure 10. This higher distance reduces the number of large languages present. Additionally, both models contained < 10 nodes with neighbours with completely inverse spin vectors. Given the extreme rarity of such nodes, it is assumed that these are simply due to the stochastic effects of thermal fluctuations. Both models also share a modal distance of 0.125 between adjacent nodes, although this is a much greater proportion for the TM

4 Discussion

Within this project, two separate modified Ising models were analysed. Each model represented languages with simple vectors and, using a modified Metropolis-Hastings algorithm, produced novel distributions of these vector languages. These models were simulated using custom scripts, with hundreds of simulations run in parallel on high performance computing systems. This allowed for a detailed investigation beyond the practical limits of conventional simulations. The phase behaviour of both models was then analysed, as shown in Figure 7. Here, despite the modifications made, both model retained some of the energy temperature dynamics of the classical Ising model, including strong evidence for the presence of phase transitions in both. The location of each model's critical temperature was estimated as being the location where the change in energy with respect to time was at its greatest. This location was understood to be the point in which both models were moving the fastest between their modes of operation and was therefore an intuitive location for the critical temperature of each model. These values are at $T = 0.57$ and $T = 0.28$ for the TM and PM respectively. This conclusion is supported by analysis of the variance in energy either side of the identified critical points. As seen in

Figure 6, both models clearly show more stable behaviour in their higher temperature regimes. Despite a large difference in value of T_c, the energy of both models at their critical temperature is relatively close, suggesting this transition occurs at states with similar disorder.

When investigating each model's language distribution, unexpected behaviour was observed. Whilst log-normal behaviour seemed likely at the critical temperature hypothesised for the TM, within the PM no such behaviour was observed at the critical point. Instead, the model only produces approximately log-normal distributions in the region of $T = 0.36$, significantly higher than its proposed critical temperature. This may be due to the approximate nature of the system's phase transition - the change in dynamics within the classical Ising model is also gradual and does not occur instantaneously at the critical temperature. The author conjectures this log-normality occurs as an intermediate stage between the complex structures formed below T_c and the completely random distribution of languages seen above this value. At temperatures beyond these values, both systems provided insufficient evidence to reject either hypothesis concerning the nature of each distribution, possibly as log-normal curves can be well approximated as normal for low enough values of σ .

In order to compare the results of both models with real-world data, multiple data transformation were necessary in order to account for the vast differences in scale. These transformations consist of taking the logarithm of the observed speaker counts for each language, then normalising each distribution into the range $(0, 1)$. As each distribution has seemingly log-normal characteristics, the distributions of their logarithms should be expected to approximately follow the characteristic bell shape of a Gaussian distribution. However, the logarithm operation does loose information, as it constrains any comparisons into only describing relative properties. The following normalisation of the logarithm is essential to allow for visual and statistical comparisons, and does not change the shape of the curve produced. In practice, this normalisation makes each comparison relative only to the largest language in each dataset, as they each contain population 1 languages and as such the minimum of the log transformed data was already mapped to 0.

With the effects of these transformations in mind, a comparison of the models n_s curves to the real-world is shown in Figure 10. In particular, the PM produces results with a strong similarity to that of the Ethnologue dataset at $T = 0.36$. Both these distributions share a mode and mean of $N \approx 0.43$. Considering the previously mentioned values, this corresponds to both having an average language size proportional to the largest language to the power of 0.43 . Other measures further quantified this similarity between the two scaled distributions, as shown in Table 1. The calculated shared area (overlap) of $\approx 0.89\%$ reflects the already visually apparent similarity in shape between the two. Further still, calculating the divergence of each model from the Ethnologue dataset further supports the conclusion $PM\ |_{0.36}$ is a significantly better approximation of the true distribution when compared to either the PM and TM at their respective critical temperatures. The reason for the improved quality of fit at $T = 0.36$ instead of at $T_c = 0.28$ is not fully understood, however I believe this is caused by the smooth nature of the phase transition seen within the Ising model.

As shown in Figure 11, each model also successfully produced lattices containing contiguous clusters of individual languages. At the edge of these regions, some approximation of continuity in language can also be observed. Most adjacent nodes were ≤ 1 spin state away from one another, with neighbours of directly opposite spins occurring very rarely within either the TM or the PM. This near continuity at the boundary between languages could be observed both in these visualisations and within Figure 12. These fuzzy boundaries between languages are a property also observed in real-world language and were successfully replicated by both models. This is a clear advantage of the spin vector language representation, as established distance metrics can be used to quantify the nature of the language continua seen.

Despite the less accurate fit of the TM to the Ethnologue dataset, the results it provides still give valuable insight into the applicability of modified Ising models to language problems in general. The Threshold relation proposed was derived from the author's personal hypotheses about the role of language in communication and still produced the characteristic log-normality observed in real-world data. The value for this threshold was also selected based upon the poorly defined boundary between languages and dialects; it is in some sense arbitrary. Future tuning of this parameter may produce more significant results.

Ultimately, the successful reproduction of some characteristics in both models illustrates the potential of the framework established by Houghton [1] in modelling language dynamics. The ability to implement hypotheses about the interactions of these spin language speakers as simple mathematical relationships has great potential for further experimentation. This is assisted by the large body of literature surrounding the dynamics and implementations of the Ising model. Simulations of these models used well documented algorithms, familiar to researchers outside classical linguistics. As such, the modified Ising model has much room to apply new hypotheses and modifications.

4.1 Comparison to Related Models

The modified Ising models presented here bear some obvious comparisons with the work of Stauffer et al. [12]. Both achieved this using $2D$ lattices, each populated by nodes assigned with a vector language representation. Importantly, the lattice nodes in Stauffer's model contain populations of size > 1, allowing the model to represent significantly larger populations. Larger lattice sizes and longer language vectors are used, allowing for a more complete representation of the scale of global language distribution. However, this work sits paradigmatically opposed with the models presented in this project and much of the existing literature. No consideration is made of agent-agent interaction, with languages instead mutating at some constant rate and only propagating into unoccupied nodes. Similarly, no consideration is made of the closeness in language space between nodes, instead these language vectors serve more as simple labels of each node. The model presented is also never numerically compared to the real language distribution seen globally. Attempts to modify the model beyond the presence of mutating bit strings were also unsuccessful, suggesting this model fails to generalise to alternate hypotheses than those initially outlined. With these considerations, the author finds this model is a weaker

alternative to the models presented here, even without a way of qualitative comparison.

4.2 Current Limitations

Due to the aforementioned limits in scale, comparing the results of each model with real-world data was a significant challenge. Even after aggregating the results across each of the 15 runs, each model only produced 1.35×10^6 samples, 3 orders of magnitude smaller than the actual global population. Similarly, the use of $n = 8$ spin vectors constrained the models to only 256 possible languages - far less than the 6520 catalogued within the Ethnologue dataset. Whilst the logarithm-normalisation operation allows for a relative comparison of the shape of each distribution, these comparisons only consider the distribution of language size in relation to the largest languages present.

4.3 Future Improvements

Initial efforts in addressing the limitations of each model could simply begin with increasing the scale of each simulation. For example, the boundary effects and resulting instability seen within Figure 4 appeared to be steadily decreasing as L increased. The chosen value of $L = 300$ was a compromise between these effects and the limitations of the $\mathcal{O}(L^2)$ complexity of each model, a balance that may have negatively effected the results obtained. Investigation into these would be computationally intensive, requiring either expanded computing resources or significant optimisation of the code used for each model. Optimisation, perhaps through moving into a higher performance language, would be a clear initial goal as this project was already highly dependent on the University of Bristol BluePebble high performance computing facilities. These boundary effects might also be rectified through the use of periodic boundaries, eliminating these edge cases entirely by defining the lattice not as a 2d grid but as a torus, avoiding this discontinuity.

Compared to increasing population, increasing the amount of languages represented would be significantly less demanding. Only length $n = 13$ language vectors are needed to represent the $\approx 7 \times 10^4$ languages catalogued globally. This improvement would also help simplify a direct comparison between each model and real-world data.

When looking to further investigate the Threshold Model specifically, an exploration of the threshold (θ) value may prove promising. The value of 0.51 chosen was motivated through existing research on the language/dialect distinction. An improved value could be found either through parameter sweeps or perhaps based upon new distinctions of intelligibility. This could be included, with minimal modification, to the existing simulation scripts.

5 Conclusion

Within this project, two separate models were investigated - each with the goal of replicating some generalisation of human language dynamics. These simple agent-based models, derived from the statistical physics, used the ordinal 2D Ising model as a framework. The inclusion of spin based language vectors created a simple representation of linguistic diversity. Both the Threshold relation proposed by the author and the Preference relation initially proposed by Conor Houghton introduced simple models of interaction driven language evolution. This project quantified the dynamics of both systems, hoping to provide an initial understanding from which further work may be done. The presence and location of the phase transitions in each model were successfully identified, supported by an analysis of the energy-temperature dynamics. Both models also successfully capture some of the characteristic log-normality of the real language distribution, although this occurred substantially above the proposed value for the critical temperature in the PM. Further work is needed to establish whether this is due to error in how T_c is evaluated or if this is a result from other aspects of model design, such as the finite boundary conditions. At the locations identified, both models produced lattices populated by distinct clusters of languages, with some approximation of continuity between these regions despite the discretisation of languages used. Comparisons to the real-world data distribution showed the Preference model at $T = 0.36$ produced the best approximation of the real-world language distribution, although both models did replicate its characteristic log-normal distribution. However, these comparisons are only relative, leaving space for future work to develop both the size and scale of each model and further these comparisons

References

[1] Conor Houghton. An ising-like model for language evolution. *arXiv preprint arXiv:2305.13916*, 2023.

[2] Noam Chomsky and George A. Miller. Introduction to the formal analysis of natural languages. *Journal of Symbolic Logic*, 33(2):299–300, 1968.

[3] Robert C Berwick et al. Evolution, brain, and the nature of language. *Trends in cognitive sciences*, 17(2):89–98, 2013.

[4] Michael Boissonneault and Paul Vogt. A systematic and interdisciplinary review of mathematical models of language competition. *Humanities and Social Sciences Communications*, 8(1):1–12, 2021.

[5] Luc Steels. A self-organizing spatial vocabulary. *Artificial life*, 2(3):319–332, 1995.

[6] Katrien Beuls and Luc Steels. Agent-based models of strategies for the emergence and evolution of grammatical agreement. *PloS one*, 8(3):e58960, 2013.

[7] Lorenzo Pucci, Pietro Gravino, and Vito DP Servedio. Modeling the emergence of a new language: Naming game with hybridization. In *International Workshop on Self-Organizing Systems*, pages 78–89. Springer, 2013.

[8] Dorota Lipowska and Adam Lipowski. Phase transition and fast agreement in the naming game with preference for multi-word agents. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(8):P08001, 2014.

[9] Barry A Cipra. An introduction to the ising model. *The American Mathematical Monthly*, 94(10):937–959, 1987.

[10] Elleard FW Heffern et al. Phase transitions in biology: from bird flocks to population dynamics. *Proceedings of the Royal Society B*, 288(1961), 2021.

[11] Daniele Marinazzo et al. Information transfer of an ising model on a brain network. *BMC Neuroscience*, 14:1–2, 2013.

[12] Dietrich Stauffer et al. Bit-strings and other modifications of viviane model for language competition. *Physica A: Statistical Mechanics and its Applications*, 376:609–616, 2007.

[13] Viviane M De Oliveira et al. Theoretical model for the evolution of the linguistic diversity. *Physica A: Statistical Mechanics and its Applications*, 361(1):361–370, 2006.

[14] Karthik Siva, Jim Tao, and Matilde Marcolli. Syntactic parameters and spin glass models of language change. *Linguist. Anal*, 41(3-4):559–608, 2017.

[15] Marcelo AF Gomes et al. Scaling relations for diversity of languages. *Physica A: Statistical Mechanics and its Applications*, 271(3-4):489–495, 1999.

[16] Robert D Greenberg. The politics of dialects among serbs, croats, and muslims in the former yugoslavia. *East European politics and societies*, 10(03):393–415, 1996.

[17] Søren Wichmann. How to distinguish languages and dialects. *Computational Linguistics*, 45(4):823–831, 2020.

[18] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

[19] Charlotte Gooskens and Wilbert Heeringa. Perceptive evaluation of levenshtein dialect distance measurements using norwegian dialect data. *Language Variation and Change*, 16:189 – 207, 10 2004.

[20] Chaoju Tang and Vincent J. van Heuven. Predicting mutual intelligibility of chinese dialects from multiple objective linguistic distance measures. *Linguistics*, 53(2):285–312, 2015.

[21] David M Eberhard, Gary F Simons, and Charles D. Fennig. Ethnologue: Languages of the world. twenty-seventh edition. dallas, texas: Sil international. `https://www.ethnologue.com/`. Accessed 10/04/2024.

[22] Chikara Furusawa, Takao Suzuki, Akiko Kashiwagi, Tetsuya Yomo, and Kunihiko Kaneko. Ubiquity of log-normal distributions in intra-cellular reaction dynamics. *Biophysics*, 1:25–31, 2005.

[23] Andrew P Allen, B-L Li, and Eric L Charnov. Population fluctuations, power laws and mixtures of lognormal distributions. *Ecology Letters*, 4(1):1–3, 2001.

[24] Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251, 2004.

[25] Peter Wittek. 14 - boosting and adiabatic quantum computing. In Peter Wittek, editor, *Quantum Machine Learning*, pages 139–151. Academic Press, Boston, 2014.

[26] Kirill P Kalinin and Natalia G Berloff. Computational complexity continuum within ising formulation of np problems. *Communications Physics*, 5(1):20, 2022.

[27] David B Hitchcock. A history of the metropolis–hastings algorithm. *The American Statistician*, 57(4):254–257, 2003.

[28] Johann Ostmeyer, Evan Berkowitz, Thomas Luu, Marcus Petschlies, and Ferenc Pittler. The ising model with hybrid monte carlo. *Computer Physics Communications*, 265:107978, 2021.

[29] A Batsidis, P Economou, and G Tzavelas. Tests of fit for a lognormal distribution. *Journal of Statistical Computation and Simulation*, 86(2):215–235, 2016.

[30] Winfried Stute, Wenceslao Gonzáles Manteiga, and Manuel Presedo Quindimil. Bootstrap based goodness-of-fit-tests. *Metrika*, 40:243–256, 1993.

[31] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[32] Frank Nielsen. On a generalization of the jensen–shannon divergence and the jensen–shannon centroid. *Entropy*, 22(2):221, 2020.