Statistical physics of language maps in the USA

J. Burridge,1,2 B. Vaux,2 M. Gnacik,1 and Y. Grudeva1

1School of Mathematics and Physics, University of Portsmouth, Portsmouth PO1 3HF, United Kingdom
2Faculty of Modern and Medieval Languages, University of Cambridge, Cambridge CB3 9DA, United Kingdom

(Received 22 November 2018; revised manuscript received 5 February 2019; published 14 March 2019)

Spatial linguistic surveys often reveal well-defined geographical zones where certain linguistic forms are dominant over their alternatives. It has been suggested that these patterns may be understood by analogy with coarsening in models of two-dimensional physical systems. Here we investigate this connection by comparing data from the Cambridge Online Survey of World Englishes to the behavior of a generalized zero temperature Potts model with long-range interactions. The relative displacements of linguistically similar population centers reveal enhanced east-west affinity. Cluster analysis reveals three distinct linguistic zones. We find that when the interaction kernel is made anisotropic by stretching along the east-west axis, the model can reproduce the three linguistic zones for all interaction parameters tested. The model results are consistent with a view held by some linguists that, in the USA, language use is, or has been, exchanged or transmitted to a greater extent along the east-west axis than the north-south.

DOI: 10.1103/PhysRevE.99.032305

I. INTRODUCTION

All people display linguistic idiosyncrasies [1]. These might be different words for the same action or object, syntactic differences, or systematic variations in pronunciation. Speakers’ geographical origins can often be inferred from their use of language, because people from the same region typically have many linguistic features in common. For example, native English speakers from western Canada typically call a multistory car park a “parkade,” athletic shoes worn as casual footwear “runners,” and small houses in the countryside for weekend retreats during the summer months “cabins” [2]. A collection of particularly consistent and distinctive pronunciations may be called an accent, or if vocabulary and grammar are also distinctive, a dialect [1]. The earliest known study of geographical language variation was carried out in 1876 by Georg Wenker, who asked 50,000 schoolmasters from locations across Germany to transcribe a list of sentences into the local dialect [3]. Modern computers and the creation of the internet have dramatically improved data collection and analysis [4–12], and social media has provided a new source of linguistic data [13–15]. Modelling linguistic evolution has also emerged as a subfield of statistical physics where ideas and techniques employed to relate the macroscopic behavior of physical systems to their microscopic components have been applied [16–24]. However, there is a need to develop mathematical models which provide a scientific understanding of how human-level processes [25] give rise to the observed geographical distributions and language dynamics [22,23,25,26].

It has recently been proposed [22,23] that observed geographical boundaries between linguistic features (“isoglosses”) are analogous to domain walls in physical systems [21,27,28], straightening over time, and being repelled by population centers. This process is a linguistic analog of physical phase-ordering [27], which also accounts for country shape and population distribution, and leads to predictability in the geographical distribution of language use [21–23]. In Refs. [22,23], linguistic domain walls were described by an analog of the time-dependent Ginzburg–Landau equation [27], which in the physical context provides a coarse-grained description of phase ordering in the Ising and Potts models. In this paper we further explore the hypothesis that geographical linguistic patterns may be understood as a generalized phase ordering process. However, we depart from the continuous-space Ginzburg–Landau description of language evolution [22,27], and instead compare the long-run (steady-state) behavior of a simple low-temperature generalized Potts model, to a large modern dialect survey of the United States. This departure is motivated in part by the population distribution of the United States, which is concentrated in cities, separated by sparsely populated rural areas. This settlement pattern lends itself to a discrete model where population centers form the nodes of an interacting network, each analogous to a single “spin” in the Potts model.

II. SURVEY DATA

The Cambridge Online Survey of World Englishes (COSWE) [7], initiated in 2007, consists of geographically located responses to thirty-one different questions. For example, question five is:

What word(s) do you use in casual speech to address a group of two or more people?

The most popular answers to this particular question are you guys (35%) and y’all (15%). The survey currently contains $8.28 \times 10^4$ responses world-wide, with approximately $5.8 \times 10^4$ located in the eastern half of the United States. This region
of high population density, stretching from the desert states to the Atlantic Coast, has a wide variety of local linguistic terminology and is the focus of our study. The westernmost cities in our study are San Antonio in the South (Texas) and Fargo in the North (Minnesota/North Dakota border). The land to the west of our study area is very sparsely populated (< 4 people per km²) and we approximate it as empty in our model: We treat our study area as a closed system in a linguistic sense. Further justification for this approach is provided by the USA population density map in Fig. 1. From this we see that our study area forms a self contained region of high population density bordered by desert, water (the Great Lakes, the Atlantic and the Gulf of Mexico), and country boundaries (Canada and Mexico).

Maps of the raw survey data [7] reveal a patchwork of geographical regions with distinctive language use. However, within any given region linguistic choices are never uniform across all respondents. For example, in Philadelphia, the dominant local term for a submarine sandwich is “hoagie;” but it is not universally adopted (57% of people use it in Philadelphia, compared to 8% in the eastern USA as a whole). To characterize local average linguistic behavior, we begin by clustering the locations of survey respondents using the mean shift algorithm [29], which locates the peaks of the kernel density estimate of the population distribution. Using a bandwidth of 50 km generates clusters with centroids corresponding to the locations of all significant population settlements within the eastern United States. In regions of low population density which are significantly more than 50 km away from any major settlement we find a large number of small, evenly spaced clusters, each containing only a handful of survey responses (< 20). To ensure that each cluster has sufficient data to provide a reliable linguistic sample, we repeatedly join the smallest cluster to its nearest neighbor. We set this minimum sample size to be 20, which is achieved by repeating our joining process until we have \( N = 300 \) nodes (Fig. 2). At each node \( i \) we define an average frequency vector \( \mathbf{f}_i = (f_{i1}^Q, f_{i2}^Q, \ldots, f_{iRQ}^Q) \) for each question, where \( f_{ik}^Q \) is the relative frequency of the \( k \)th response to the \( Q \)th question, and \( R_Q \) is the number of different responses for question \( Q \). Each node occupies a point in the linguistic space for each question, the probability simplex

\[
C_{iQ} = \{ \mathbf{f}_i^Q \in \mathbb{R}^{RQ} | f_{i1}^Q + f_{i2}^Q + \ldots + f_{iRQ}^Q = 1, f_{ik}^Q \geq 0 \},
\]

which can be of high dimension—the COSWE survey database [7], for example, contains more than 800 distinct families of lexical responses to Question 8, “What do you call the gooey or dry matter that collects in the corners of your eyes, especially while you are sleeping?”, with the most common being (eye) boogers, sleep, (eye) gunk, and (eye) crusties.

For each node, the combined response frequencies for all questions, concatenated into a single vector \( \mathbf{f}_i = (\mathbf{f}_i^1, \mathbf{f}_i^2, \ldots, \mathbf{f}_i^D) \), where \( n \) is the number of questions, defines a point in a high-dimensional aggregated linguistic space. To visualize the distribution of population nodes in this space we perform a principle components analysis [30] to find a low-dimensional representation of the frequency data which captures as much of its variation as possible. This is achieved by first finding a normalized linear combination of the response frequencies,

\[
z_{i1} = \mathbf{v}_1 \cdot \mathbf{f}_i \quad \text{such that} \quad |\mathbf{v}_1| = 1,
\]

which maximizes the variance of the set \( \{z_{i1}\}_{i=1}^N \). We refer to \( \mathbf{v}_1 \) as the first component loading vector. Subsequent unit loading vectors \( \mathbf{v}_k \) are chosen to maximize the variance of \( \mathbf{v}_k \cdot \mathbf{f}_i \) over all nodes subject to the condition that \( \mathbf{v}_k \cdot \mathbf{v}_j = 0 \) for all \( j < k \). If we compute only \( D \) loading vectors, then the coordinates \( \mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iD}) \) constitute a \( D \)-dimensional representation of the data which explains the maximum possible amount of the variation present in the high-dimensional data (that is, the variance of the \( \mathbf{z}_i \) is as close as possible to the variance of the \( \mathbf{f}_i \)). The result of this analysis in the case \( D = 2 \) is shown in Fig. 3.
The position of each black dot in Fig. 3 represents the linguistic state of a population node, and we note that nodes exhibit a significant degree of clustering. To analyze these clusters we define the linguistic distance (aggregated over all questions) between nodes $i$ and $j$

$$l_{ij} = |f_i - f_j|$$  \hspace{1cm} (3)

$$= \sqrt{\sum_{Q=1}^{n} (l_{ij}^Q)^2},$$  \hspace{1cm} (4)

where

$$l_{ij}^Q = |f_i^Q - f_j^Q|$$  \hspace{1cm} (5)

is the linguistic distance in the answer to question $Q$ between the two nodes. The aggregated linguistic distances allow us to use the $K$-means method [30] to divide the data into $n_c \geq 1$ linguistic clusters. We determine the optimal value of $n_c$ by maximizing the average silhouette score, $U$ [31], over all nodes $i$, where

$$U_i = \frac{b_i - a_i}{\max(a_i, b_i)} \in [-1, 1],$$  \hspace{1cm} (6)

$$U = \frac{1}{N} \sum_{i=1}^{N} U_i,$$  \hspace{1cm} (7)

with $a_i$ the average distance between $i$ and nodes in the same cluster, and $b_i$ the average distance to nodes in different clusters. For the aggregated data shown in Fig. 3 we find $n_c = 3$, and nodes are colored according to their $K$-means cluster label. For comparison, the $n_c$ values and their scores, $U$, were $(n_c = 2, s = 0.21), (3, 0.26), (4, 0.19), \ldots$ with lower scores for larger $n_c$ values. Visual inspection of the clusters in Fig. 3 also suggests that $n_c = 3$ is the appropriate choice.

Node colors in Fig. 2 show how the results of our linguistic cluster analysis appear in geographical space. This map demonstrates that geographical proximity is a powerful predictor of linguistic similarity. The linguistic clustering process took no account of geographical location, and yet the resulting clusters divide the spatial domain into distinct regions. Similar divisions appear on the level of individual questions. For example, in Fig. 4 we have performed a linguistic clustering of the responses to question 5, and we see a sharp transition between the Southern states, where groups of people are typically addressed using the expression “y’all,” and Northern states, where the term is more typically “you guys.” The breakdown of survey results in the two clusters is given in Fig. 5. In the linguistic context, domain walls of this type are known as isoglosses [1,32], and they have been mapped and studied for over a century. Domain walls are also ubiquitous in atomic level phase ordering processes, and this connection between physics and linguistics was recently reported and explored in detail in Refs. [22,23]. The work we present here is, to the authors’ knowledge, the first quantitative comparison of a phase ordering model to a large linguistic data set.

To further explore the relationship between linguistic and geographical proximity in the USA, in Fig. 6 we have constructed a network on our set of $N = 300$ nodes in which every node connects to its four closest linguistic neighbors. Note that most connections are short range compared to the system size; population centers are typically most linguistically similar to others within a few hundred kilometres with...
the striking exception of Miami, Florida (which will come as no surprise to linguists who are well aware of the strong influence on Florida of migration from northeastern US cities [33–35]).

We may view the current residence of each person (speaker) in the US as the node in a network, embedded in two-dimensional space [36]. Each speaker will be influenced by many others, through face-to-face contact [37] and by voices heard on broadcast media. Thinking of the collection of such interactions over a significant time period as edges between nodes, we may view language evolution as a dynamical process on the resulting network. Although we cannot reconstruct the detail of these links, the fact that geographical proximity is strongly related to linguistic proximity suggests that the most important connections are spatially short range. The network through which linguistic forms spread may therefore be viewed as quasi-two-dimensional, provided we take a sufficiently coarse-grained view of the system. This has geometrical implications for the evolution of language. To understand these, we first note that there is debate about whether language change may be viewed as a process of neutral evolution [16,20,38] in the sense that linguistic traits are reproduced with a probability which is proportional to their current frequency. In statistical physics, the voter model is the simplest spatial evolutionary process which is neutral in this sense. Although models belonging to this class generate regions of space where one variant dominates (they coarsen), the boundaries of these domains do not feel surface tension [39] and the speed with which spatial domains grow is logarithmic rather than algebraic in time. Nonneutral evolution requires nonlinearity in the relationship between the local frequencies of traits and the probability that they are adopted in favour of their alternatives (more popular traits need to be further emphasized). In the social context this nonlinearity arises from a tendency of individuals to conform to the majority [40]. The Ising and majority rule models [28,41] are examples in this class, both exhibiting coarsening in which domain sizes grow algebraically in time. If we believe that geographical linguistic patterns are a result of coarsening, then the existence or otherwise of social conformity [40] in language evolution will affect the rate at which patterns form, and the forms they take. In particular, if the network within which language evolves is two-dimensional, then we might expect to see similarities between the behavior of nonneutral two-dimensional physical models, such as the Ising and Potts, and linguistic maps. In the two-dimensional case linguistic boundaries may be viewed as lines (isoglosses), and by analogy with physical systems, we would expect these to feel surface tension [22,23,27].

From Fig. 6 we observe that in addition to being typically short range, the distribution of connections is not isotropic: a disproportionate number of edges appear to run closer to the east-west direction than to north-south. To explore this effect, in Fig. 7 we have plotted the relative geographical coordinates of the five nearest linguistic neighbors. Also shown in Fig. 7 are the eigenvectors of the gyration tensor

$$T = \begin{pmatrix} \langle \Delta x^2 \rangle & \langle \Delta x \Delta y \rangle \\ \langle \Delta x \Delta y \rangle & \langle \Delta y^2 \rangle \end{pmatrix}.$$  (8)
where $\Delta x$, $\Delta y$ are the relative displacements in the east-west and north-south directions respectively and $\langle \cdot \rangle$ denotes the average over these displacements. The eigenvectors of $T$ are closely aligned with the north-south and east-west directions, and the ratio of the roots of the corresponding eigenvalues is 1.40. There is, therefore, strong anisotropy in the geographical distribution of linguistic near-neighbors. In other words, the drop in cultural and linguistic affinity between population settlements appears to decline more slowly with east-west displacements than north-south. To verify that this anisotropy is not just an artefact of the placement of the nodes, in Fig. 8 we have plotted the relative geographical coordinates of the five nearest geographical neighbors of every node. In this case we find very little anisotropy: the ratio of the roots of the eigenvalues of the gyration tensor is very close to one (1.08). If we consider the displacements between all pairs of nodes we find a ratio of 1.17, with axis of elongation lying approximately midway between horizontal and vertical. We therefore conclude that linguistic anisotropy is not a consequence of node placement.

It is possible that linguistic anisotropy is a historical artefact of the west-moving colonisation of the continent [42,43], leading to disproportionately strong east-west cultural identification, or it may be due to the existence of more extensive east-west oriented physical routes of communication (e.g., roads, air flights), or even a combination of both these things. Whatever its origin, we explore the effect of enhanced east west connectivity on linguistic maps below.

### III. THE MODEL

Our aim is to test the extent to which spatial linguistic patterns can be explained by minimal statistical physics models of conformity–driven evolution. Since we are interested in the behavior of a network, in the current paper we revert to the simplest possible model whose steady states are a discrete analog of those generated by the continuous space memory models defined in Refs. [22,23]. Our model is a generalization of the $q$-state Potts model [44], described below. Our aim is not to capture the full stochastic evolution of the frequency vectors $f^0_j$ (to be considered in further work), but to model the discrete patterns which emerge from the survey data after clustering. We note that there exists a large body of work exploring the processes which drive linguistic change [1,32,45–48], and factors such as gender, networks (social and physical), social status, and cultural identities all play important roles.

Our model is defined as follows. At time $t$, each node $i$ of the network is in a discrete state $s_i(t) \in \{1, 2, \ldots, q\}$. We relate this discrete state to our continuous observational data by viewing $s_i$ as the label of the linguistic cluster assigned to $f^0_j$ by our clustering method of choice. We assume that nodes evolve so as to maximize conformity within their linguistic neighborhood. We define this neighborhood using a discrete version of the interaction kernel defined in Ref. [22]. Letting $r_i$ be the location of node $i$, we first define a raw kernel $\phi(r_i - r_j)$ giving the influence of node $i$ on node $j$ in the absence of variations in the populations of nodes. We then define the normalized population weighted influence of node $j$ on node $i$ to be

$$J_{ij} = \frac{\phi(r_i - r_j) P_j}{\sum_k \phi(r_i - r_k) P_k} \quad (9)$$

According to this definition, the influence of a node scales in proportion to its population $P_j$: If two nodes are equidistant from a speaker, then she is twice as likely to converse with another speaker from a node with twice the population (see Fig. 9). We note that the idea of modeling linguistic change using population-based measures of influence was introduced by the sociolinguist Peter Trudgill in his gravity model [1,46], which predicts how changes jump from one settlement to another. Here we take a different approach by defining dynamics which seek to minimize a global nonconformity function, analogous to the Hamiltonian of the $q$-state Potts model.
first define the indicator function that two nodes belong to different clusters:

\[ \Delta_{ij}(t) = \begin{cases} 0 & \text{if } s_i(t) = s_j(t) \\ 1 & \text{otherwise}. \end{cases} \tag{10} \]

We call this the indicator of nonconformity between nodes. The total nonconformity of the entire system is then

\[ H(t) = \sum_{i,j} J_{ij} \Delta_{ij}(t). \tag{11} \]

We note that the influence numbers \( J_{ij} \) are not symmetric in their indices because larger nodes will exert greater linguistic influence on their smaller neighbors than these neighbors exert in return.

Conformity driven dynamics are then implemented using the Metropolis algorithm [44]: we randomly select a node \( i \), and then propose a new state, \( s'_i \), selected uniformly at random from the set of alternatives to the current state \( s_i(t) \). We let \( \Delta H \) be the change in total nonconformity that would result from the proposed change, and then we accept this change with probability

\[ \Pr(s_i(t) \rightarrow s'_i) = \begin{cases} 1 & \text{if } \Delta H < 0 \\ e^{-\beta \Delta H} & \text{if } \Delta H \geq 0, \end{cases} \tag{12} \]

where \( \beta \), the classical inverse temperature, controls the level of noise in the dynamics. In the zero noise (zero temperature) limit \( \beta \rightarrow \infty \), only those changes which increase conformity are allowed. For simplicity we consider the zero temperature limit from here on. Although these dynamics are minimal and coarse grained, they capture an important aspect of spatially distributed social ordering phenomena, namely, that even if the individual behavior of agents tends to lead toward social conformity, the system as a whole may exhibit regionalism because the stochastic process of conforming takes place at small spatial scales. The system can therefore become “stuck” in a suboptimal global state, because no single change of individual state can increase conformity.

To allow for anisotropy in our interaction kernel, we define the anisotropic distance \( d_{ij} \) between nodes as

\[ d_{ij}(r_i - r_j) = \sqrt{\frac{(x_i - x_j)^2}{A^2} + \frac{(y_i - y_j)^2}{1^2}}, \tag{13} \]

where \( A \) measures the extent to which the east-west components of geographical displacements are effectively shrunk by enhanced connectivity (cultural or physical). Setting \( A > 1 \) is equivalent to squashing the system in the east-west direction. Given the anisotropic distance, we define our raw interaction kernel \( \phi \) to be a truncated Cauchy distribution,

\[ \phi(r_i - r_j) = \begin{cases} \left(1 + \left(\frac{d_{ij}}{\gamma}\right)^2\right)^{-1} & \text{if } d_{ij} < R \\ 0 & \text{otherwise}, \end{cases} \tag{14} \]

where \( \Delta r_{ij} = |r_i - r_j| \) is the distance between nodes \( i \) and \( j \). We refer to \( \gamma \) as the interaction range and \( R \) as the cutoff. Our choice of \( \phi \) is in part guided by experimental data which suggest that human displacements collectively follow a truncated power law but are individually highly repetitive and predictable [49,50], with considerable heterogeneity within the population. We also note that the choice of a long-range algebraic kernel as opposed to a short-range exponentially decaying kernel is more consistent with our linguistic proximity network (Fig. 6), which contains some links which stretch hundreds of kilometres. The large distance cutoff may also be justified on purely theoretical grounds since without it, inverse square-law interactions preclude the possibility of stable domain walls in phase ordering models [51].

IV. RESULTS

A. Influence of initial conditions

Since we have no information regarding the early linguistic state of our population nodes, we initialize the system with each node in a random state \( s_i(0) \in [1, \ldots, q] \). We note that the number of possible initial conditions is very large indeed \( >10^{90} \)—significantly larger than the number of atoms in the universe. This approach may be justified on the basis that although these initial conditions almost certainly do not reflect reality, the ordering process causes very large numbers of different initial states to converge to a much smaller subset of equilibrium configurations, so it should not matter a great deal how the system began. We quantify this assertion by defining a simple measure of the number of distinctive long-run configurations generated by the model. These stable states are attractors of the dynamics in the sense that each one is the final destination of very large numbers of different randomized initial states.

We first define a simple method for measuring the similarity of two maps. Two clusterings of nodes are equivalent if they can be transformed into one another by permuting cluster labels. To compare maps we must find the permutation of labels which maximizes the number of nodes which have the same label in both maps. This can be achieved using the Hungarian algorithm [52], and results in a similarity score, \( S(m_1, m_2) \), giving the fraction of nodes which belong to the same cluster in maps \( m_1 \) and \( m_2 \). We may then generate a set, \( D \), of distinctive maps by repeatedly generating a new map, \( m \), and adding it to \( D \) only if \( S(m, m_j) < 90\% \) for all \( m_j \in D \). For a small interaction range \(|D|\) can be very large but for the range of parameter values used in this work \(|D|\) is only a few hundred. Since all initial conditions are attracted toward this set, we argue that the current linguistic state of the system may still be significantly predictable even though we lack historical data; it does not matter where one starts, because in the end one will end up near to \( D \). This is our justification for considering the behavior of the model aggregated over a large number of entirely random initial states when comparing to aggregated survey data.

B. Comparison with empirical data

Using linguistic distances, \( r_{ij}^D \) [Eq. (5)], we can apply the K-means method to cluster nodes according to their responses to individual survey questions. We may then apply the silhouette method [31] to each question [see Eqs. (6) and (7)], to find the optimal number of clusters in each case. When clustering the responses to individual questions we use the symbol \( q \) to denote the number of clusters, emphasizing that we are viewing cluster labels as states in our \( q \)-state Potts model. Silhouette analysis reveals that in all but one case the optimal
number is either $q = 2$ or $q = 3$, with $q = 2$ being optimal in 21 out of 31 cases. However, visual inspection of PCA plots for each question indicates that in some of the $q = 2$ cases it might be more appropriate to view their results as a single cluster. We therefore explore the behavior of the two- and three-state versions of our model over a large number of simulation runs, using equal proportions of each $q$ value.

To make our predictions we generate 750 Potts maps for each $q \in \{2, 3\}$ using $3 \times 10^4$ Metropolis updates, sufficient to reach a near-equilibrium state, each starting from a different randomly selected initial condition. We think of each of these simulation runs as the response to a single fictitious survey question, and we compare the aggregated simulation results to the aggregated survey clustering, as described below.

By comparing the equilibrium behavior of our model to empirical data, we are implicitly assuming that the current aggregated geographical distribution is a near-equilibrium state. In reality, language is constantly evolving [1,45] with new linguistic variants being created and older variants coming back into fashion, possibly leading to significant changes in the geographical distributions of individual linguistic forms. Within the Potts model framework, such changes correspond to the temporary application of a (possibly localized) external field creating a bias in favor of a particular variant. Our equilibrium assumption remains valid provided that the frequency and duration of such events is sufficiently low, so most variables will be close to equilibrium at any given moment. We do not rule out the possibility that the main linguistic zones in the USA will change in the future, but according to our model, this would require a change in the structure of interactions between population centers.

Letting $s_k^i$ be the final state of the node $i$ in map $k$ we define the aggregated linguistic state of node $i$ to be the vector of states for all simulated maps

$$s_i = (s_1^i, s_2^i, \ldots).$$

The first two principle components of these vectors, obtained from simulations, are shown in Fig. 10 in the case where $R = 500$ km and $\gamma = 120$ km. To compare the model to our survey data we use $K$ means to assign each node to one of three clusters based purely on its aggregated linguistic state, and then plot these clusters geographically in Fig. 11. Visual inspection reveals a close match to the survey map in Fig. 2, confirmed by a similarity score of $S(\text{model, data}) = 90.67\%$.

To test the robustness of this result to variations in interaction kernel parameters we repeat our simulations and analysis for $R \in \{400, 500, 600\}$ and $\gamma \in \{90, 100, \ldots, 180\}$. The results in the isotropic case ($A = 1$) are shown in Table I, where we see that a similarity score over 80% is achieved in 13 out of 30 cases. These high scoring maps all exhibit approximately the same pattern illustrated in Fig. 11. In cases when this pattern fails to emerge the system selects one of a number of alternative states, and the score drops considerably. When interaction range is excessively low, nodes become increasingly independent, and linguistic clusters either do not form, or are very small.

To investigate the effects of interaction anisotropy, we stretch the interaction kernel in the east-west direction by set-

![Fig. 10. Principle components analysis of aggregated Potts results with $R = 500$ km and $\gamma = 120$ km. The figure shows a typical output, where variations occur between simulation runs.](image)

![Fig. 11. Cluster analysis of aggregated Potts results with $R = 500$ km and $\gamma = 120$ km.](image)
TABLE II. Percentage similarity scores between model and data for anisotropic interaction kernel with $A = 1.15$ (east-west displacements shrunk).

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>90</th>
<th>100</th>
<th>110</th>
<th>120</th>
<th>130</th>
<th>140</th>
<th>150</th>
<th>160</th>
<th>170</th>
<th>180</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{1,\gamma}$</td>
<td>400</td>
<td>0.82</td>
<td>0.89</td>
<td>0.90</td>
<td>0.89</td>
<td>0.87</td>
<td>0.87</td>
<td>0.89</td>
<td>0.89</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.89</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>600</td>
<td>0.88</td>
<td>0.85</td>
<td>0.85</td>
<td>0.84</td>
<td>0.83</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.82</td>
</tr>
</tbody>
</table>

In this case we obtain a robust match between model and survey maps for all parameter values. We note that our model is systematically unable to generate the linguistic connection between the southern tip of Florida and the Northeast: the connection is a result of unusually high migratory flows between these two locations [34] which are not captured by our interaction kernel. We estimate that this discrepancy reduces similarity scores by $\approx 2\%$ on the basis that three out of our three hundred nodes are systematically misclassified. Our analysis of the linguistic proximity network (Figs. 6 and 7) suggested a greater east-west affinity, and our results indicate that such an affinity causes the observed linguistic zones to appear with high probability in our model. That fact that this pattern is very robust with respect to the choice of interaction parameters suggests that the observed large scale patterns may be explained by the distribution of people, and by system shape, rather than the fine detail of their interactions. In fact, it is likely that regional differences in connectivity and population density would make the interaction range vary by location. Comparing maps generated for individual survey questions to the set of simulated maps, we find the best-match between survey and simulated maps, averaged over all survey questions, is $\approx 80\%$. If we stretch our kernel in the opposite direction by setting $A = 0.96$, we find no strong matches to our survey maps (Table III).

As an alternative to dedicated linguistic surveys, social media platforms such as Twitter may be used to generate very large datasets of geotagged text which may be analyzed to discover geographical variations in language use [13,14]. A recent analysis of 924 million tweets generated by 6.6 million users in the USA over one year used hierarchical clustering to divide the country into distinct linguistic zones. For example, the five most distinctive clusters are shown in Fig. 12, along with the aggregated result of our two- and three-state Potts models, divided into five clusters using $K$-means. We note the close match, and also that clusters appear to have densely populated areas at their heart with boundaries lying in less densely populated areas. These features were predicted by the

TABLE III. Percentage similarity scores between model and data for anisotropic interaction kernel with $A = 0.96$ (east-west displacements increased).

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>90</th>
<th>100</th>
<th>110</th>
<th>120</th>
<th>130</th>
<th>140</th>
<th>150</th>
<th>160</th>
<th>170</th>
<th>180</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{1,\gamma}$</td>
<td>400</td>
<td>0.63</td>
<td>0.63</td>
<td>0.6</td>
<td>0.6</td>
<td>0.58</td>
<td>0.63</td>
<td>0.6</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.61</td>
<td>0.61</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.63</td>
<td>0.64</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>600</td>
<td>0.62</td>
<td>0.62</td>
<td>0.63</td>
<td>0.62</td>
<td>0.66</td>
<td>0.63</td>
<td>0.64</td>
<td>0.66</td>
<td>0.66</td>
</tr>
</tbody>
</table>

memory based surface tension models [22,23] upon which the current paper builds.

V. VORONOI NULL MODEL

In defining our model we took account of, and made assumptions regarding, the effect of social conformity and the relative sizes of population centers. To assess the extent to which our assumptions are necessary to explain the observed geographical linguistic patterns, or whether the patterns have a more trivial origin, we now define a simple null model of linguistic clustering. The only assumption of this simpler model is that nearby nodes should be linguistically similar. A simple way to achieve this is to select $q$ nodes at random, assign each a different label, and then assign all other node labels according to which of the original $q$ nodes is closest. In this way we generate a discrete version of the Voronoi tessellation [53]. Having generated a large number of tessellations, each representing the null-map for a single fictitious survey question, we repeat our principle components and cluster analysis. The principle components plot (Fig. 13) takes the form of a circle, lacking any recognisable clusters in linguistic space. This high degree of symmetry reflects the fact that the null model treats all nodes as equivalent in size and status, and ensures only that geographically nearby nodes are close linguistically. Despite the lack of clearly identifiable clusters, we may still apply $K$-means clustering, which seeks to minimize within-cluster variation [30]. The result is a symmetrical division of points into approximately equal sized clusters, shown in Fig. 13.

There is similarity between the shapes of the main linguistic zones in our Voronoi null model and the survey map (Figs. 2 and 13). This similarity is perhaps not surprising, given that the null model captures one of our main empirical observations: that people nearby to each other tend to use similar linguistic terms. In this sense the null model captures
a crude form of linguistic or social alignment between nearby nodes. However, it lacks dynamics and any kind of natural derivation in terms of the speaker-level process of conforming to the local majority. Our Potts model, which accounts for the sizes of population centers and models social conformity through an empirically motivated interaction kernel, produces clustering in linguistic space, which is present in the survey data (Figs. 3 and 10) but not in the null model. Beyond linguistic clustering, the Potts model also produces a closer match to the survey map. We therefore suggest that plausible coarse-grained conformity-driven dynamics which account for uneven population distribution are an important ingredient in understanding the distribution of language features.

VI. DISCUSSION

We have examined the spatial distribution of linguistic features in the eastern USA and compared these distributions to a generalized Potts model defined on the network of population centers, taking account of long-range interactions, social conformity, population sizes, and interaction anisotropy. The steady states of this model are discrete analogues of those generated by the continuous space memory models defined in Refs. [22,23], where the dialect patterns of England and Italy were modeled.

In the case of isotropic interactions, our Potts model predicted shapes of linguistic zones in close (≈90%) agreement with survey data [7] for 13 out of 30 sets of interaction parameters tested. Analysis of the linguistic proximity network inferred from survey data suggested that linguistic affinity was not in fact isotropic, and introducing such anisotropy into our model produced a close agreement to the survey maps for all parameter values.

It is interesting to note a possible connection to crossing probabilities in percolation [54]. In low-temperature two-dimensional magnetic systems it is common to see stripe states formed by magnetic domain walls crossing the system. The appearance and direction of these crossings depends strongly on the aspect ratio of the system: they become increasingly common as the aspect ratio is increased, and in high ratio systems they typically run across the short axis. By shrinking east-west displacements we were effectively changing the aspect ratio of our system, making an east-west domain wall substantially more likely.

Our model indicates that without anisotropy the current population distribution could have generated a linguistic north-south split, but this distribution is only one of a number of possibilities. By including anisotropy we find that the split is almost inevitable. Our work therefore takes a step towards answering the question of whether the observed north-south linguistic divide in the USA is merely a consequence of population distribution and geography. Our results support the idea that enhanced east-west linguistic transmission has occurred in the USA. Enhanced east-west transmission could have arisen from the historic westerly colonization (migration) of people. Alternatively, the existence of better east-west transportation links, either historically or in the modern setting, could provide an explanation, and in future work we might analyze historic and modern transportation networks to test this. However, we note that the two possible explanations (migration and the quality of transportation links) may be interdependent.

ACKNOWLEDGMENTS

J.B. is grateful for the support of a Royal Society APEX award 2018–2020 (APXR1\180117). The authors are grateful to Marius Jøhndal for curating and supplying the survey data. The authors thank the anonymous referees for investing their time in reading the paper and providing much useful feedback. The individual contributions of the authors were as follows: J.B. devised the study and model, wrote simulations, performed data analysis, and wrote the manuscript. B.V. devised the original survey, provided data and linguistic expertise, and edited the manuscript. M.G. processed and analyzed data, ran simulations, and generated figures and tables. Y.G. performed exploratory simulations and data analysis, directed by J.B.


