# MEAN-FIELD AND MEASURE-VALUED DIFFERENTIAL EQUATION MODELS FOR LANGUAGE VARIATION AND CHANGE IN A SPATIALLY DISTRIBUTED POPULATION

W. GARRETT MITCHENER*

**Abstract.**
Although discrete formalisms have been successful at describing the sets of grammatical sentences in human languages, new tools are needed to model language variation. An individual's speech pattern can be modeled more realistically by a stochastic grammar consisting of a set of idealized grammars together with a set of usage rates. A population can then be represented as a probability measure over a space of usage rates and physical or social locations. In this article, I investigate a measure-valued differential equation for a spatially distributed population in which individuals use stochastic grammars. Under appropriate hypotheses, and assuming that children learn based on an average feature of the nearby population's speech, the asymptotic behavior of the measure dynamics are controlled by the feature's dynamics, which can significantly reduce the dimension of the model. I discuss the example of a single usage rate for choosing one of two grammatical options. If space is unstructured, then all populations tend to a stable equilibrium dominated by one option or the other. If space consists of two well-mixed compartments, then each compartment may choose a different dominant idealized grammar, but increased migration causes a bifurcation in which one idealized grammar goes extinct. If space is continuous, numerical experiments show that the measure and feature dynamics can exhibit traveling waves.

Key words: measure-valued differential equation, semilinear differential equation, mean-field model, dimension reduction, reaction-diffusion equation, traveling wave, population, migration, language variation, language change.
AMS subject classifications: 34G20 91F20 92D25

**1. Introduction.** Human language is a hybrid system. On one hand, the set of grammatical sentences in a language can in general be described by discrete formalisms, such as context sensitive grammars, minimalism [1, 42], and optimality theory [44]. Due in part to the influence of Chomsky [7, 8, 9], much of the research in linguistics has been focused on developing discrete formal descriptions of sentences as produced and understood by idealized speakers. Idealized grammatical descriptions can be formulated for phonology (the sound system), morphology (word structure), syntax (sentence structure), and semantics (meaning structure).

There are several shortcomings to this approach. First, many languages include multiple grammatical constructions for expressing certain meanings, and discrete formalisms typically include no indication of why speakers should use one or the other [26]. To give an example from [26], most English sentences can be expressed in either active or passive voice with very little difference in meaning, and there are tendencies but no hard rules for which voice to use in any given situation. Corpus studies indicate that each individual writer varies the rates at which he or she uses various constructions across manuscripts [46], and grammatical change seems to consist of a more or less steady rise in preference for one alternative at the expense of another over decades and across the population. Second, social context has significant influence on the choices made during speech [22]. An idealized grammar states that both familiar and polite forms of address are grammatical, for example, but it is difficult to state precisely when to prefer one over the other. Third, speech production is imperfect. Children in particular make a variety of interesting mistakes that mostly disappear from their speech, but not entirely [2, 48]. Second language speakers almost always maintain a noticeable level of variation from the language as spoken by typical native

*Mathematics Department, College of Charleston, Charleston SC. MitchenerG@cofc.edu. This work was supported in part by grant #0734783 from the National Science Foundation.

speakers [47]. Fourth, native speakers sometimes disagree on whether particular sentences are grammatical, and grammaticality judgments are often graded rather than binary [3]. Fifth, it is very difficult to model language acquisition and change using discrete formalisms. Such models can be made [12, 35, 5, 13, 44] but they generally do not take into account statistical properties of speech and do not describe how speakers arrive at their usage rate for grammatical variants. They can also be overly sensitive to noise in the input.

Thus, there is a growing need for more tools that address the fuzzier aspects of speech. Some mathematical tools have been developed for addressing the lexicon [39, 19, 38, 41] and syntax [3], but there is a need for tools for modeling the dynamics of grammatical change within a population of non-idealized speakers. In this article, we formulate a general class of linguistic population dynamics that relaxes frequently-used simplifying assumptions. Specifically, speakers are allowed to use arbitrary mixtures of idealized grammars. The learning process takes the entire population state as input. Spatial and social structure are included in a speaker's state, and the dynamics include the flow of individuals from one state to another.

Since the resulting class of models involves infinite-dimensional dynamical systems, we prove a series of dimension-reduction propositions: Under certain assumptions, particularly that learning takes as input only macroscopic properties of the population's speech patterns, the infinite-dimensional dynamics are asymptotically controlled by a closed system of differential equations for those macroscopic properties. These results are then applied to example models of language change, one in which space is divided into two compartments, and one in which space is modeled continuously and the change takes the form of a traveling wave.

**1.1. The modeling process.** Different mathematical models for a single phenomenon can cover a huge range of detail, and each level has its advantages and disadvantages. For example, if one needs to model a chemical dissolved in water, the most detailed framework might be quantum mechanics, representing each subatomic particle in each molecule of solvent and solute. On a coarser scale, atoms and bonds can be represented by rods and springs. One can dispense with the individual molecules altogether and use continuum mechanics. At the coarsest scale, the solution can be represented as a homogeneous volume of liquid with a particular concentration of solute.

Within linguistics, the same range of mathematical models can formulated. On one extreme, each speaker, utterance, and meaning can be represented in full detail. Coarser models might keep track of each individual's state, but abstract away the details so that speaker states are representable as a few binary bits. Even coarser, one might keep track of the number of speakers in each of a very few states, so individual agents and sentences are not represented directly. At the coarsest level, a population may be boiled down to compartments, each of which has a single bulk speech pattern. The advantage of the detailed simulations is realism, but their disadvantage is tractability: The experimenter can run computer simulations up to a certain size and collect statistics, but proving meaningful theorems about their behavior or fitting the huge number of parameters to data is not generally possible. The advantage of the coarser models is tractability, but their disadvantage is realism: One can calculate fixed points and prove stability results for a dynamical system representing an infinite population, but its representation of language may be so simplified that its applicability is in doubt. In-between models are compromises between realism and tractability.

Currently, the linguistics modeling community favors the extremes. For example, there are detailed simulations of individuals learning a language [17, 5, 29], and continuous population dynamics [28, 33, 34, 30, 31, 20, 37, 36], but the middle ground is somewhat sparse. One purpose of this project is to build models in that middle ground, analogous to the mesoscale continuum mechanics that lie between microscale quantum physics and macroscale bulk dynamics.

**1.2. Population dynamics with probability measures.** For linguistic population dynamics in the presence of non-idealized speech, a natural mathematical tool is the probability measure, which can flexibly represent the distribution of speakers as a function of possible states. Signed measures form a Banach space, of which the probability measures are a closed subset. Much of the theory of ordinary differential equations applies directly to dynamics in Banach spaces and therefore to measure-valued differential equations. However, infinite dimensional geometry can be counter-intuitive and requires careful treatment.

The use of probability measures rather than some simpler mathematical object deserves some explanation. Let us consider a simple scenario where speakers use a mixture of two idealized grammars $G_1$ and $G_2$ that are identical except for one syntactic construction. Each individual's speech pattern is represented by a real number $z \in [0, 1]$ indicating the frequency with which he or she uses $G_1$. In the limit of a large, well mixed population, the population state at time $t$ might be represented by a probability density function $u(t, \cdot)$ where $\int_A u(t, z)\, dz$ is the fraction of the population whose usage rate is in $A$. It is possible to formulate a sensible differential equation for which $u(t, \cdot)$ takes values in the space $L^1$ in which all continuous density functions reside. However, there is a disadvantage to using this space: If the dynamics can drive a language variant to extinction, it might be necessary to include a discrete feature in the dynamics, such as a population state in which all people use the old variant at rate 0. Representing distributions with mass concentrated at a point in a continuum requires an atomic probability measure, often represented by a delta function, and $L^1$ does not include such generalized functions. Since both continuous and discrete distributions are potentially necessary, it makes sense to work in the space of signed measures rather than $L^1$. Additionally, measures can represent sets of individual agents as a sum of atomic measures, as well as infinite populations using continuous densities, potentially providing a tool that can unify infinite population models as limits of finite population models.

Therefore, consider a time-dependent probability measure $u(t, dz)$, where $u(t, A) = \int_{z \in A} u(t, dz)$ is the fraction of the population whose usage rate is in $A$. The population dynamics are then given by

$$\frac{\partial u(t, dz)}{\partial t} = Q(u(t, \cdot), dz) - u(t, dz) \tag{1.1}$$

where $Q(\mu, A)$ is the distribution for the number of children with usage rate in $A$ given that they are learning from a population with usage frequencies distributed according to the distribution $\mu$. The $Q$ term represents the distribution of births contributing to usage frequency $z$, and the $-u$ term represents deaths. In formulating this equation, it is assumed that births and deaths occur at the same rate, that this rate is independent of language, and that time has been rescaled so that this rate has unit magnitude. This model is a step upward in complexity from [28, 33, 34, 31, 18] in which it is assumed that each child learns primarily from his or her parents.

Since (1.1) contains no partial derivative in $z$, we may interpret it as an autonomous, infinite dimensional, ordinary differential equation rather than a functional partial differential equation:

$$u'(t) = dz \mapsto Q(u(t), dz) - u(t, dz)$$

or, leaving $z$ and $t$ implicit,

$$u' = Q(u) - u. \tag{1.2}$$

With that basic model in place, we introduce a spatial variable $x$ and allow for the population to be distributed continuously or discretely in space:

$$u'(t) = (dx, dz) \mapsto Q(u(t), dx, dz) - u(t, dx, dz)$$

or with $t$, $x$ and $z$ all implicit:

$$u' = Q(u) - u. \tag{1.3}$$

What remains is to add a linear term representing spatial and linguistic flow. That is, adults are allowed to move from place to place, but with the restriction that the flow rates are expressible as a linear operator $G$:

$$u' = Q(u) - u + Gu. \tag{1.4}$$

The dynamics so formulated are deterministic, but represent random variation in language as distributions over usage rates. The development of mesoscale models with stochastic components is beyond the scope of this article, but is addressed in other projects by the author [32].

**1.3. Dimension reduction.** We would like to determine if the dynamics for $u$ might be understood in terms of some mean field simplification. For example, under the simplifying assumption of an unstructured population, speakers are indistinguishable and sentences are selected uniformly at random from all speakers. Each child effectively hears and learns from the population's mean usage rates of possible variants. This suggests that we investigate the circumstances under which a closed dynamical system can be formulated for the mean of $u$, and determine the extent to which information about the mean determines the dynamics of $u$. Such circumstances would justify replacing the infinite dimensional dynamics of $u$ with finite dimensional dynamics representing the mean speech pattern of the population. Furthermore, if the population is divided into physical patches or social classes, then the same sort of finite dimensional approximation ought to be possible within each compartment, with some additional terms indicating the migration rates among compartments.

The first step is to formalize (1.4) in Section 2, and prove that under appropriate assumptions, it has unique probability-measure-valued solutions for all forward time. We then focus on the case where learning depends only on aggregate features of the population state. Rather than limit ourselves to mean dynamics, we suppose more generally that features lie in some linear space, and that there is a linear operator that extracts the aggregate features from probability measures. A crucial assumption is that migration causes features to flow: The feature extraction operator needs to be interchangeable with the migration rate operator using a feature flow rate operator.

The next step is to prove that the existence of certain stable structures within the feature dynamics implies the existence of parallel stable structures within the

the full measure dynamics. If the entire feature phase space is filled by the basins of attraction of these stable structures, which is true for generic one- and two-dimensional dynamical systems, then the asymptotic behavior of the full measure dynamics is completely accounted for by the parallel structures.

With the general mathematical machinery developed, we will turn in Section 3 to the specific example of the dynamics of the usage rate of $G_1$ as opposed to $G_2$, under the assumption that children learn from the mean usage rate. If the population is unstructured, the dynamics are simple: Generically, the population converges to a stable equilibrium dominated by $G_1$ or $G_2$. However, if the population is split into compartments, then each compartment may choose a different dominant grammar, and an increase in the migration between the compartments can lead to a bifurcation that eliminates one of the grammars. If the population is distributed on a continuous space, then (1.4) may be related to a reaction-diffusion equation, and numerical experiments show that it can exhibit the traveling waves characteristic of such equations. Each of these examples is connected to instances of language variation and change in the linguistics literature.

Although the model is described as if the spatial variable represents physical space, it could just as well be interpreted as social space, representing ethnicity, economic class, or any combination. The migration process is then interpreted as including social mobility.

## 2. Mathematical machinery.

**2.1. Notation, assumptions, and fundamental results.** To begin, here are some assumptions and notation that will be used throughout.

DEFINITION 2.1. *$\Omega$ is a locally compact Hausdorff space of states that individuals may be in. Associated with $\Omega$ is the Borel $\sigma$-algebra $\mathcal{B}\Omega$ of measurable subsets of $\Omega$. All measures considered in this paper will be finite regular Borel measures.*

For example, if the population is unstructured and there are two alternative speech patterns whose usage frequency may vary, then $\Omega = [0, 1]$. If the population has $k$ patches and two alternative speech patterns, then $\Omega = \{1, 2, \ldots, k\} \times [0, 1]$ to indicate an individual's location and speech pattern.

DEFINITION 2.2. *$\mathcal{M}$ is the Banach space of bounded measures on $(\Omega, \mathcal{B}\Omega)$ with the total variation norm,*

$$\|\mu\|_{TV} = \sup_{\text{partitions } \{F_j\} \text{ of } \Omega} \sum_j |\mu F_j| \,.$$

The measure of a set $A \in \mathcal{B}\Omega$ under $\mu \in \mathcal{M}$ will be denoted $\mu A$, and the integral of a function $f$ over a set $A$ with respect to $\mu$ will be denoted

$$\int_{x \in A} f(x)\, \mu(dx).$$

Equations satisfied by measures will sometimes be expressed in differential notation, as in

$$\nu(dx) = \int_{s \in A} f(s)\, \mu(ds, dx) - \kappa(dx).$$

Delta measures, also called point measures or delta functions, will be denoted

$$\delta_s X = \begin{cases} 1 & \text{if } s \in X \\ 0 & \text{if } s \notin X \end{cases}$$

and

$$\int_{x\in A} f(x)\,\delta_s(dx) = f(s).$$

Measures are partially ordered; in this paper, the primary use of that order is the notion of a *positive* measure: $\mu \geq 0$ means that for each $A \in \mathcal{B}\Omega$, $\mu A \in \mathbb{R}$ and $\mu A \geq 0$. Additionally, a measure can be *strictly positive:* $\mu > 0$ means that $\mu \geq 0$ and $\mu \neq 0$.

The *mass* of a measure $\mu$ is given by the measure of the whole space under $\mu$, $\mu\Omega$. Note that for positive measures, $\mu\Omega = \|\mu\|_{TV}$.

Definition 2.3. *$\mathcal{M}^\times$ is the subset of $\mathcal{M}$ of strictly positive measures,*

$$\mathcal{M}^\times = \{\mu \in \mathcal{M} \mid \mu > 0\}. \tag{2.1}$$

Definition 2.4. *$\mathcal{P}$ is the set of probability measures on $\Omega$.*

$$\mathcal{P} = \{\mu \in \mathcal{M}^\times \mid \mu\Omega = 1\} \tag{2.2}$$

*and for each $\mu \in \mathcal{P}$, $\|\mu\|_{TV} = \mu\Omega = 1$.*

Assumption 2.5. *$Q : \mathcal{P} \to \mathcal{P}$ is the learning function. It must satisfy a Lipschitz condition on $\mathcal{P}$: There is a constant $L > 0$ such that*

$$\forall \mu_1, \mu_2 \in \mathcal{P} \quad \|Q(\mu_1) - Q(\mu_2)\|_{TV} < L\,\|\mu_1 - \mu_2\|_{TV}\,. \tag{2.3}$$

The $Q$ function takes as input a probability measure that represents how individuals in the population are distributed over their possible states. Its output is a probability measure that represents the distribution of the state of a random person born into that environment. The Lipschitz condition guarantees that it is continuous, plus a bit smoother.

Definition 2.6. *A function $f : \mathcal{M} \to \mathcal{M}$ is said to* respect positivity *if $\mu \in \mathcal{M}^\times$ implies $f(\mu) \in \mathcal{M}^\times$.*

Assumption 2.7. *$G : \mathcal{M} \to \mathcal{M}$ is a bounded linear operator such that for all $\mu \in \mathcal{M}$,*

$$(G\mu)\Omega = 0, \tag{2.4}$$

*and $M(t) = e^{tG}$ is a time-dependent bounded linear operator that respects positivity for all $t \geq 0$.*

The $M$ process represents part of the population flow, specifically, $M(t)\mu$ gives the state of a population initially in state $\mu$ after experiencing migration (but not birth and death) for a time $t$. The migration rate operator $G$ represents instantaneous flow due to migration and will be used to incorporate these effects into the population dynamics. The constraint $(G\mu)\Omega = 0$ means that the net flow over the entire population is zero, even for an un-normalized measure $\mu$, so that the population is self-contained and should not grow due to migration. More formally, an immediate consequence of this constraint is that $M(t)$ preserves the mass of measures, as is seen from the power series

$$\begin{aligned}(M(t)\mu)\Omega &= \left(\mu + tG\mu + \frac{t^2}{2}G^2\mu + \dots\right)\Omega \\ &= \mu\Omega\end{aligned}$$

because in each term past the first, $(G^n\mu)\Omega = (G(G^{n-1}\mu))\Omega = 0$. From this and the assumption that $M(t)$ respects positivity, it follows that for all $t \geq 0$, $M(t) : \mathcal{P} \to \mathcal{P}$.

For a single-compartment population with no migration, $G = 0$ and $M(t) = I$. It is possible to generalize $M$ to a semigroup and let $G$ be its generator, but there is no need for such generality in the examples in this paper. Instead, we will eventually assume a somewhat more specific form for $M$ and $G$.

ASSUMPTION 2.8. *$u : [0, \infty) \to \mathcal{P}$ is the time-dependent population state. Sampling the population at time $t$ produces random a element of $\Omega$ distributed according to $u(t)$. The dynamics of $u$ are*

$$u' = Q(u) - u + Gu, \quad u(t_0) = u_0. \tag{2.5}$$

As a technicality, there is no theoretical difficulty dealing with derivatives or integrals of a Banach-space valued function of a real variable. See for example, chapter III of [10]. Furthermore, bounded linear operators may be exchanged with derivatives and integrals in $t$, which justifies operations such as

$$\left( \int g(t)\, dt \right) S = \int (g(t)S)\ dt$$

where $S \in \mathcal{B}\Omega$ and $g : \mathbb{R} \to \mathcal{M}$.

The various assumptions on $Q$ and $G$ are required to prove that solutions to (2.5) are well defined. A general result, Theorem 5.1 from section VI.5 of [27], will be applied. The assumptions so far immediately satisfy the following hypotheses of this theorem:

- $\mathcal{P}$ is closed.
- $A(u) = Q(u) - u + Gu$ is continuous, $A : \mathcal{P} \to \mathcal{M}$.
- For each $u \in \mathcal{P}$, $\|A(u)\|_{TV} \leq 2 + \|G\|$, so $A$ maps bounded sets to bounded sets.

The theorem requires confirmation of two other hypotheses. First is a condition that says $A$ does not drive the dynamics off $\mathcal{P}$.

PROPOSITION 2.9. *For all $u \in \mathcal{P}$,*

$$\liminf_{h \to 0^+} \frac{1}{h} d\left(u + hA(u); \mathcal{P}\right) = 0$$

*where $d(x; D)$ is the distance from the point $x$ to the set $D$, as in*

$$d(x; D) = \inf\{\|x - y\|_{TV} \mid y \in D\}.$$

*Proof.* We will need the expansions

$$e^{h(G-I)} = I + h(G - I) + \frac{1}{2}h^2(G - I)^2 + \dots$$
$$e^{-h} = 1 - h + \frac{1}{2}h^2 + \dots$$

transformed into

$$I + h(G - I) = e^{h(G-I)} + O\left(h^2\right)$$
$$h = 1 - e^{-h} + O\left(h^2\right)$$

The expression of interest is

$$\begin{aligned} u + hA(u) &= (I + h(G - I))u + hQ(u) \\ &= e^{h(G-I)}u + (1 - e^{-h})Q(u) + O\left(h^2\right) \\ &= \overbrace{e^{-h}(e^{hG}u) + (1 - e^{-h})Q(u)}^{y_0} + O\left(h^2\right) \end{aligned}$$

Since $Q(u) \in \mathcal{P}$ and $e^{hG}u = M(h)u \in \mathcal{P}$, the dominant term $y_0$ is a convex combination of probability measures, so it is also a probability measure. Therefore, $d(u + hA(u); \mathcal{P}) \leq \|u + hA(u) - y_0\|_{TV} = O\left(h^2\right)$. □

Second is a bound on a sort of one-sided Gâteau differential.

PROPOSITION 2.10. *For each $u$ and $v \in \mathcal{P}$,*

$$\lim_{h \to 0^+} \frac{1}{h} \left( \|u - v\|_{TV} - \|u - v - h(A(u) - A(v))\|_{TV} \right) \leq (L + 1 + \|G\|) \|u - v\|_{TV}$$

*Proof.* The limit exists thanks to the convexity of $\|\cdot\|_{TV}$ and a monotonicity argument, as in § II.5 of [27]. Thus we only need to show a bound.

Let $r = \|u - v\|_{TV} - \|u - v - h(A(u) - A(v))\|_{TV}$. Using the triangle inequality,

$$\begin{aligned} |r| &\leq \|h(A(u) - A(v))\|_{TV} \\ &\leq h \left( \|Q(u) - Q(v)\|_{TV} + \|u - v\|_{TV} + \|Gu - Gv\|_{TV} \right) \\ &\leq h(L + 1 + \|G\|) \|u - v\|_{TV} \end{aligned}$$

□

PROPOSITION 2.11. *Every initial value problem* (2.5) *has a unique solution $u : [0, \infty) \to \mathcal{P}$. If $u_1$ and $u_2$ satisfy the differential equation, then for all $t \geq 0$*

$$\|u_1(t) - u_2(t)\|_{TV} \leq e^{(L+1+\|G\|)t} \|u_1(0) - u_2(0)\|_{TV}$$

*Proof.* This follows from Theorem 5.1 from § VI.5 of [27]. □

**2.2. Feature dynamics and dimension reduction.** Under the simplifying assumption the population is unstructured and well mixed, we consider learning functions $Q$ that depend on $u$ only through its mean. More generally, consider a bounded linear operator $T$ taking a population state $u$ to those features that are relevant to learning, and suppose that $Q(u) = q(Tu)$. This yields the differential equation

$$u' = q(Tu) - u + Gu \tag{2.6}$$

where as in Proposition 2.11, there is a unique solution $u : [0, \infty) \to \mathcal{P}$ for each initial condition $u(0) = u_0 \in \mathcal{P}$. The features $m = Tu$ also satisfy a differential equation, derived by applying $T$ to both sides of (2.6):

$$Tu' = q(Tu) - Tu + TGu$$

By making appropriate assumptions about how $T$ interacts with $G$ and introducing a related feature flow rate operator $H$, the $u$ dynamical system yields a closed dynamical system for features $m = Tu$,

$$m' = q(m) - m + Hm$$

Thus, with appropriate assumptions, the infinite dimensional $u$ dynamics can be reduced to much lower dimensional $m$ dynamics, and many features of the $u$ dynamics are controlled by the underlying $m$ dynamics. This section proves several results about the extent to which the asymptotic dynamics of $m$ determine the asymptotic dynamics of $u$.

We need the following additional definitions and assumptions.

ASSUMPTION 2.12. *$\mathcal{Y}$ is a Banach space representing interesting features of elements of $\mathcal{M}$. Its norm will be denoted $\|\cdot\|_{\mathcal{Y}}$*

ASSUMPTION 2.13. *$T : \mathcal{M} \to \mathcal{Y}$ is a bounded linear operator that extracts aggregate features from a probability density.*

For example, $T$ could be the mean usage rate, any moment of the usage rate, or any tuple of moments, in which case $\mathcal{Y} = \mathbb{R}^n$. The assumption that $T$ is linear allows us to swap it with $d/dt$, which is important in connecting the full measure-valued population dynamics with the feature-valued population dynamics. Typically, $T$ will be many-to-one and specified on $\mathcal{P}$, but it generalizes to all of $\mathcal{M}$ by linearity.

ASSUMPTION 2.14. *In this section, we consider the case where the learning function has the form $Q(u) = q(Tu)$. Formally, $q : \mathcal{Q} \to \mathcal{P}$ is a Lipschitz-continuous function defined on some closed and bounded subset $\mathcal{Q} \subset \mathcal{Y}$. We require for each $u \in \mathcal{P}$ that $Tu \in \mathcal{Q}$ so that $q \circ T : \mathcal{P} \to \mathcal{P}$ is well defined. We also require that for each $m \in \mathcal{Q}$ there exists at least one $u \in \mathcal{P}$ such that $Tu = m$.*

ASSUMPTION 2.15. *$H : \mathcal{Y} \to \mathcal{Y}$ is a bounded linear operator representing feature flow rates. If $m \in \mathcal{Q}$ then $Hm \in \mathcal{Q}$. It is related to $T$ and $G$ by the identity*

$$TG = HT. \tag{2.7}$$

*We also assume that $\|G\| < 1$ and $\|H\| < 1$.*

The norm constraints on $G$ and $H$ ensure that $G - I$ and $H - I$ are non-singular, and that

$$T(G - I)^{-1} = (H - I)^{-1}T \tag{2.8}$$

Using these assumptions in conjunction with power series for functions of the operators $G$ and $H$ gives, for example,

$$Te^{(G-I)t} = e^{(H-I)t}T. \tag{2.9}$$

The existence of $H$ allows the feature operator $T$ to be swapped with the migration rate operator $G$. It represents the effects of migration on features rather than on the population distribution.

ASSUMPTION 2.16. *$m : [0, \infty) \to \mathcal{Y}$ is the time-dependent vector of features representing the population state. The dynamics of $m$ are derived from the dynamics of $u$ by applying $T$ to both sides of (2.6):*

$$\begin{aligned} T(u') &= Tq(Tu) - Tu + TGu \\ (Tu)' &= Tq(Tu) - Tu + HTu \end{aligned}$$

*Setting $m = Tu$ yields the closed dynamical system*

$$m' = Tq(m) - m + Hm. \tag{2.10}$$

PROPOSITION 2.17. *For every initial condition $m(t_0) = m_0$ in $\mathcal{Q}$ there is a unique solution to* (2.10)*, defined for all $t \geq t_0$.*

*Proof.* Since $q$ is Lipschitz continuous, the standard Picard-Lindelöf theorem guarantees the existence of unique solutions to initial value problems on finite time intervals. (See for example Chapter VI of [27].)

Given $m_0$, define $u_0 = q(m_0)$. From Proposition 2.11, there is a unique solution $u(t)$ for (2.6) defined for all $t \geq t_0$. This gives a solution $m(t) = Tu(t)$ for (2.10) defined for all $t \geq t_0$. Since the solution for $m$ is unique on every finite time interval, it follows that $m(t) = Tu(t)$ is the unique solution for $m$ defined for all $t$. □

It turns out that the asymptotic dynamics of $m$ determine much about the asymptotic dynamics of $u$. First, we show that fixed points for one correspond exactly to fixed points for the other.

PROPOSITION 2.18. *If $\bar{u} \in \mathcal{P}$ is a fixed point of* (2.6)*, then $\bar{m}$, defined as*

$$\bar{m} = T\bar{u}$$

*is a fixed point of* (2.10)*.*

*Proof.* Given that

$$q(T\bar{u}) + (G - I)\bar{u} = 0,$$

apply $T$ to both sides to derive

$$Tq(\bar{m}) + (H - I)\bar{m} = 0.$$

□

PROPOSITION 2.19. *For each fixed point $\bar{m}$ of* (2.10) *there is a unique fixed point $\bar{u}$ of* (2.6) *such that $T\bar{u} = \bar{m}$, and $\bar{u}$ is given by*

$$\bar{u} = -(G - I)^{-1}q(\bar{m}).$$

*Proof.* The hypothesis that $\bar{m}$ is a fixed point gives

$$\bar{m} = -(H - I)^{-1}Tq(\bar{m}).$$

The definition of $\bar{u}$ implies that

$$T\bar{u} = -T(G - I)^{-1}q(\bar{m}) = -(H - I)^{-1}Tq(\bar{m}) = \bar{m}$$

and that

$$q(\bar{m}) = -(G - I)\bar{u}.$$

With this information, the right hand side of (2.6) evaluated at $\bar{u}$ gives

$$q(T\bar{u}) + (G - I)\bar{u} = q(\bar{m}) + (G - I)\bar{u} = 0.$$

To prove uniqueness, let $\bar{v}$ be another fixed point with $T\bar{v} = \bar{m}$. Since $0 = q(\bar{m}) + (G - I)\bar{v}$, the non-singularity of $G - I$ implies that $\bar{v} = \bar{u}$. □

The basins of attraction of asymptotically stable features are also in parallel, as the next two propositions prove.

PROPOSITION 2.20. *Let $m_1$ and $m_2$ be solutions to (2.10), and let $u_1$ and $u_2$ be solutions to (2.6) with $Tu_1(0) = m_1(0)$ and $Tu_2(0) = m_2(0)$. Suppose*

$$\lim_{t\to\infty} \|u_1(t) - u_2(t)\|_{TV} = 0.$$

*Then*

$$\lim_{t\to\infty} \|m_1(t) - m_2(t)\|_{\mathcal{Y}} = 0.$$

*Proof.* Applying $T$ to (2.6) and using (2.7) shows that

$$(Tu_1)' = Tq(Tu_1) - Tu_1 + HTu_1$$

which means that $Tu_1$ solves (2.10). Since initial value problems under (2.10) have unique solutions, it follows that for all $t \geq 0$, $m_1(t) = Tu_1(t)$. Similarly, $m_2(t) = Tu_2(t)$.

Therefore,

$$\|m_1(t) - m_2(t)\|_{\mathcal{Y}} \leq \|T\| \, \|u_1(t) - u_2(t)\|_{TV}$$

and since the right hand side converges to 0, so does the left. □

PROPOSITION 2.21. *Let $u_1, u_2 : [0,\infty) \to \mathcal{P}$ be solutions to (2.6). Let $m_1 = Tu_1$ and $m_2 = Tu_2$. Note that $m_1$ and $m_2$ satisfy (2.10). Suppose*

$$\lim_{t\to\infty} \|m_1(t) - m_2(t)\|_{\mathcal{Y}} = 0.$$

*Then*

$$\lim_{t\to\infty} \|u_1(t) - u_2(t)\|_{TV} = 0.$$

*Proof.* If we regard $m$ as known, we may think of the $u$ dynamics as a linear differential equation in $u$ with a non-homogeneous term containing $m$, that is,

$$u'(t) - (G - I)u(t) = q(m(t)).$$

We multiply by the integrating factor $e^{-(G-I)t}$ and rearrange the terms to find

$$u(t) = e^{(G-I)t}u(0) + \int_0^t e^{(G-I)(t-s)} q(m(s)) \, ds. \tag{2.11}$$

Applying this to $u_1$ and $u_2$ and taking the difference gives

$$\begin{aligned} \|u_2(t) - u_1(t)\|_{TV} \leq{}& \left\|e^{(G-I)t}\right\| \|u_2(0) - u_1(0)\|_{TV} \\ &+ \int_0^t \left\|e^{(G-I)(t-s)}(q(m_2(s)) - q(m_1(s)))\right\|_{TV} ds \end{aligned}$$

The assumption that $\|G\| < 1$ is key here, since it guarantees that $e^{(G-I)t}$ is shrinking as $t$ increases:

$$\begin{aligned} \left\|e^{(G-I)t}\right\| &= \left\|e^{-It}e^{Gt}\right\| \\ &= e^{-t}\left\|e^{Gt}\right\| \\ &\leq e^{-(1-\|G\|)t} \end{aligned}$$

We need to split the integral. Given $\varepsilon > 0$, and a Lipschitz constant $L$ for $q$, let $\tau$ be sufficiently large that for each $s \geq \tau$,

$$\|m_2(s) - m_1(s)\|_{\mathcal{Y}} < \frac{\varepsilon(1 - \|G\|)}{L}, \tag{2.12}$$

Introducing the positive constants

$$\begin{aligned} C_1 &= \|u_2(t_0) - u_1(t_0)\|_{TV} \\ C_2 &= \int_0^\tau \left\| e^{-(G-I)s}(q(m_2(s)) - q(m_1(s))) \right\|_{TV} ds \\ C_3 &= C_1 + C_2 \end{aligned}$$

it follows that

$$\begin{aligned} \|u_2(t) - u_1(t)\|_{TV} &\leq \left\| e^{(G-I)t} \right\| \|u_2(0) - u_1(0)\|_{TV} \\ &\quad + \left\| e^{(G-I)t} \right\| \int_0^\tau \left\| e^{-(G-I)s} \left( q(m_2(s)) - q(m_1(s)) \right) \right\|_{TV} ds \\ &\quad + \int_\tau^t \left\| e^{(G-I)(t-s)} \right\| \|q(m_2(s)) - q(m_1(s))\|_{TV} ds \\ &\leq e^{-(1-\|G\|)t}(C_1 + C_2) \\ &\quad + \int_\tau^t e^{-(1-\|G\|)(t-s)} L \|m_2(s) - m_1(s)\|_{\mathcal{Y}} ds \\ &\leq e^{-(1-\|G\|)t} C_3 + L \cdot \frac{\varepsilon(1 - \|G\|)}{L} \int_\tau^t e^{-(1-\|G\|)(t-s)} ds \\ &\leq e^{-(1-\|G\|)t} C_3 + \varepsilon \left( 1 - e^{-(1-\|G\|)(t-\tau)} \right) \end{aligned}$$

Letting $t \to \infty$,

$$\begin{aligned} \limsup_{t\to\infty} \|u_2(t) - u_1(t)\|_{TV} &\leq \limsup_{t\to\infty} e^{-(1-\|G\|)t} C_3 + \varepsilon(1 - e^{-(1-\|G\|)(t-\tau)}) \\ &\leq \varepsilon. \end{aligned}$$

Since the final inequality holds for arbitrary $\varepsilon > 0$, it follows that $\|u_2(t) - u_1(t)\|_{TV} \to 0$ as $t \to \infty$. □

COROLLARY 2.22. *Let $u$ be a solution to* (2.6)*, and let $m = Tu$. Suppose $\bar{m} \in \mathcal{Q}$ and $m(t) \to \bar{m}$ as $t \to \infty$. Then $u(t) \to -(G - I)^{-1} q(\bar{m})$ as $t \to \infty$.*

*Proof.* Observe that $\bar{m}$ must be a fixed point of (2.10). Then $\bar{u} = -(G - I)^{-1} q(\bar{m})$ is a fixed point of (2.6) by Proposition 2.19. The conclusion follows from Proposition 2.21. □

Another consequence of Propositions 2.20 and 2.21 is that if two $u$ trajectories are different but initially map to the same feature $m_0$, then the difference between the two trajectories shrinks to zero. In other words, only information derived from the features persists.

COROLLARY 2.23. *Suppose $u_{01}$ and $u_{02}$ are probability measures with $Tu_{01} = Tu_{02} = m_0$. Let $u_1$ and $u_2$ be the solutions of* (2.6) *with initial conditions $u_{01}$ and $u_{02}$ respectively. Then*

$$\lim_{t\to\infty} \|u_1(t) - u_2(t)\|_{TV} = 0.$$

In addition to fixed points and their basins of attraction, limit cycles and their basins of attraction exist in parallel.

PROPOSITION 2.24. *Let $u$ be a solution to (2.6), and let $m = Tu$. Suppose $m$ converges to a limit cycle as $t \to \infty$. Then $u$ converges to a limit cycle as $t \to \infty$ and this limit cycle is unique.*

*Proof.* The limit cycle $\tilde{u}$ of (2.6) may be recovered from the limit cycle $\tilde{m}$ and its period $\tau$. Convergence of $u$ with $\tilde{u}$ then follows from Proposition 2.21.

To begin, if $\tilde{u}_0$ is any element of $\mathcal{P}$ that happens to satisfy $T\tilde{u}_0 = \tilde{m}(0)$, then the solution to (2.6) starting at $\tilde{u}_0$ will satisfy $T\tilde{u}(t) = \tilde{m}(t)$ and converge with $u$, but there is no guarantee that a generic $\tilde{u}$ is actually a limit cycle. Only one such $\tilde{u}_0$ will work.

Using an integrating factor of $e^{-(H-I)t}$ with (2.10), it follows that

$$\tilde{m}(t) = e^{(H-I)t}\tilde{m}(0) + \int_0^t e^{(H-I)(t-s)}Tq(\tilde{m}(s))\,ds.$$

Since $\tilde{m}(\tau) = \tilde{m}(0)$,

$$\tilde{m}(0) = \left(I - e^{(H-I)\tau}\right)^{-1} \int_0^\tau e^{(H-I)(\tau-s)}Tq(\tilde{m}(s))\,ds.$$

Note that since $\left\|e^{(H-I)\tau}\right\| = e^{-\tau}\left\|e^{H\tau}\right\| \leq e^{-(1-\|H\|)\tau}$ and $\|H\| < 1$, we know that $\left\|e^{(H-I)\tau}\right\| < 1$. Therefore, the operator $I - e^{(H-I)\tau}$ is non-singular, so its inverse is well defined. Similarly, the operator $I - e^{(G-I)\tau}$ is non-singular because $\|G\| < 1$. Using (2.9), we can swap the $T$ all the way to the left to get

$$\tilde{m}(0) = T\left(\overbrace{\left(I - e^{(G-I)\tau}\right)^{-1} \int_0^\tau e^{(G-I)(\tau-s)}q(\tilde{m}(s))\,ds}^{\tilde{u}_0}\right).$$

Using the $\tilde{u}_0$ so defined as the initial condition, the solution $\tilde{u}$ to (2.6) satisfies

$$\begin{aligned}
\tilde{u}(\tau) &= e^{(G-I)\tau}\tilde{u}_0 + \int_0^\tau e^{(G-I)(\tau-s)}q(T\tilde{u}(s))\,ds \\
&= e^{(G-I)\tau}\tilde{u}_0 + \int_0^\tau e^{(G-I)(\tau-s)}q(\tilde{m}(s))\,ds \\
&= e^{(G-I)\tau}\tilde{u}_0 + \left(I - e^{(G-I)\tau}\right)\tilde{u}_0 \\
&= \tilde{u}_0
\end{aligned}$$

which verifies that $\tilde{u}$ is a limit cycle.

Uniqueness follows from the observation that if there were two $u$-limit cycles for $\tilde{m}(0)$, they would have to converge with each other as $t \to \infty$, which is impossible unless they coincide. □

These theorems mean that if the $m$ dynamics include a stable fixed point or stable limit cycle, a parallel stable feature and its basin of attraction must exist in the $u$ dynamics. If the $m$ dynamics lead all trajectories to converge to some fixed point or limit cycle, then the $u$ dynamics lead all trajectories to converge to parallel fixed

points or limit cycles. If $\mathcal{Y}$ is one- or two-dimensional, then this argument will hold for generic $m$ dynamics [15, §1.9].

Intuitively, the formula (2.11) means that the initial condition is forgotten exponentially rapidly, and at any given $t$, the measure $u(t)$ is dominated by a weighted average of recent values of $q(m(t))$. If at some $t_0$ the value of $m(t_0)$ is known exactly but the corresponding $u(t_0)$ is unknown, then the value of $u(t)$ may be approximated for $t > t_0$ by choosing any initial condition $v_0$ with $Tv_0 = m(t_0)$ and using the approximation

$$u(t) \approx v(t) = e^{(G-I)(t-t_0)}v(t_0) + \int_{t_0}^{t} e^{(G-I)(t-s)}q(m(s))\,ds. \tag{2.13}$$

The convergence of $v$ with $u$ follows from Proposition 2.21 with $m_1 = m_2 = m$.

However, if $m(t_0)$ can only be approximately estimated as $m(t_0) \approx \tilde{m}_0$, and if the nearby trajectories of $m$ are unstable, then the approximation (2.13) is potentially doomed: Since there is no way to guarantee $Tv_0 = m(t_0)$, we would have to choose $v_0$ such that $Tv_0 = \tilde{m}_0$, and consider a solution $\tilde{m}(t)$ with $\tilde{m}(t_0) = \tilde{m}_0$. There is no way to guarantee that $m(t)$ and $\tilde{m}(t)$ converge, so Proposition 2.21 does not apply. Thus, sensitive dependence on initial conditions for $m$ can result in sensitive dependence on initial conditions for $u$.

**2.3. More specific migration rate operators.** In preparation for some specific instances of this family of models, we now consider a more specific form for the migration rate operator $G$.

ASSUMPTION 2.25. *$K : \mathcal{M} \to \mathcal{M}$ is a bounded linear operator that respects positivity. We require $\|K\| < 1/2$.*

ASSUMPTION 2.26. *$J : \mathcal{M} \to \mathcal{M}$ is a bounded linear operator with the property that for all $t \in \mathbb{R}$, the linear operator $e^{tJ}$ respects positivity. We require $\|J\| < 1/2$.*

The $K$ operator represents immigration (arrival) and $J$ represents emigration (departure). The overall migration rate is $G = K - J$. The norm constraints on $K$ and $J$ ensure that $\|G\| < 1$.

These operators must have several special properties that are automatically true in the case of $G = 0$. We need a conservation of population constraint so that everyone who departs one location must arrive somewhere else:

ASSUMPTION 2.27. *For each $\mu \in \mathcal{M}^\times$,*

$$(K\mu)\Omega = (J\mu)\Omega \tag{2.14}$$

*This ensures that* (2.4) *is satisfied for $G = K - J$.*

The different positivity constraints for $K$ and $J$ reflect the fact that departure rates tend to be diagonal but arrival rates tend to be diffuse. That is, some fraction of the people at each location leave in an infinitesimal time interval. Then the whole set of moving people distributes itself over the whole space. In general, any realistic $K$ will respect positivity. The $J$ operator, however, requires more care, as it is important that the mass of people leaving a location does not exceed the number present.

PROPOSITION 2.28. *$M(t) = e^{(K-J)t}$ respects positivity, as required.*

*Proof.* Let $\mu_0 \in \mathcal{M}^\times$, and let $\mu(t) = e^{(K-J)t}\mu_0$. Note that $\mu$ is the unique solution to $\mu'(t) = -J\mu + K\mu$ with initial condition $\mu_0$. This initial value problem may be treated as semilinear, so that it falls under Theorem 5.1 in section VIII.5 of [27] as follows. The operator $-J$ is the generator of the semigroup $e^{-Jt}$ on the closed set $D = \mathcal{M}^\times \cup \{0\}$. The function $K$ takes the role of the potentially nonlinear term.

The technical assumptions of this theorem are satisfied by the fact that $K$ and $J$ are bounded linear operators. Its conclusion is that the solution $\mu$ takes values in $D$.

From the discussion following (2.4), the mass of $\mu$ is constant, so $\mu$ is never the zero measure. Thus, for every $t \in [0, \infty)$, $\mu(t) \in \mathcal{M}^{\times}$. □

**2.4. Dynamics under a migration kernel and linear speech features.** To derive some more specific results, we suppose that the population inhabits some space $\mathbb{S}$, which might be a set of compartments, a subset of Euclidean space, a circle, a torus, or some combination of such spaces. We also assume that speech patterns are elements of a Banach space $\mathbb{L}$, which might represent usage rates for various idealized grammars, for example. The individual state space for this problem is $\Omega = \mathbb{S} \times \mathbb{L}$, representing the fact that each individual has a spatial location $x \in \mathbb{S}$ and a language or speech pattern $z \in \mathbb{L}$. To simplify the notation, $u(t, X, Z)$ will be synonymous with $u(t, X \times Z)$ which is the probability that an individual is within $X \times Z$ at time $t$.

We assume that $K$ and $J$ are derived from a migration kernel $\kappa$ as follows.

ASSUMPTION 2.29. *$\kappa : \mathcal{B}\mathbb{S} \times \mathbb{S} \to \mathbb{R}$ is a migration kernel such that the probability that an individual at a location $s$ moves to somewhere in $X \subset \mathbb{S}$ during the time interval $(t, t + \Delta t)$ is $\kappa(X, s)\Delta t + o(\Delta t)$ as $\Delta t \to 0$. Given $X \subset \mathbb{S}$, $\kappa(X, \cdot)$ is a positive real-valued, bounded, measurable function. Given $s \in \mathbb{S}$, $\kappa(\cdot, s)$ is a finite, positive, regular Borel measure. To simplify certain results, $\kappa$ does not incorporate the probability that individuals do not move; that is, for each $x$, $\kappa(\{x\}, x) = 0$ and $\kappa(\mathbb{S}, \cdot) \leq 1$.*

Note that $\kappa$ does not depend on the speech pattern: We make the simplifying assumption that migration is independent of speech pattern.

ASSUMPTION 2.30. *$k : \mathbb{S} \to [0, \infty)$ is the net migration rate out of a point, given by*

$$k(s) = \kappa(\mathbb{S}, s) = \int_{x \in \mathbb{S}} \kappa(dx, s),$$

*that is, the probability that an individual at $s$ moves away in the time interval $(t, t+\Delta t)$ is $k(s)\Delta t + o(\Delta t)$ as $\Delta t \to 0$. We assume that*

$$\sup_{s \in \mathbb{S}} k(s) < \frac{1}{2} \tag{2.15}$$

*to guarantee that the required bounds on the migration rate operators $K$, $J$, and $G$ hold.*

Consequently, the time until an individual leaves $s$ is approximately exponentially distributed with rate $k(s)$ and mean $1/k(s)$. The probability measure for the conditional distribution of the destination $x$ of individuals leaving $s$ given that they leave during the time interval $(t, t + \Delta t)$ is $\kappa(dx, s)/k(s) + o(1)$ as $\Delta t \to 0$. The unconditional probability measure for the next location $x$ of an individual currently at $s$ after the time interval $(t, t + \Delta t)$ is $\kappa(dx, s)\Delta t + (1 - k(s)\Delta t)\delta_s(dx) + o(\Delta t)$ as $\Delta t \to 0$.

The $K$ and $J$ operators applied to an arbitrary $\mu \in \mathcal{M}$ are therefore defined by

$$\begin{aligned}(K\mu)(dx, dz) &= \int_{s \in \mathbb{S}} \kappa(dx, s)\mu(ds, dz), \\ (J\mu)(dx, dz) &= \int_{s \in \mathbb{S}} \kappa(ds, x)\mu(dx, dz) \\ &= k(x)\mu(dx, dz)\end{aligned}$$

The total migration rate operator is

$$(G\mu)(dx, dz) = \left( \int_{s \in \mathbb{S}} \kappa(dx, s)\mu(ds, dz) \right) - k(x)\mu(dx, dz).$$

Clearly, $K$ respects positivity. To check that $e^{tJ}$ respects positivity, observe that $J$ is essentially diagonal, so that

$$\begin{aligned}(e^{tJ}\mu)(dx, dz) &= \left(1 + tk(x) + \frac{(tk(x))^2}{2!} + \dots \right) \mu(dx, dz) \\ &= e^{tk(x)}\mu(dx, dz).\end{aligned}$$

As required by the conservation constraint (2.14),

$$\begin{aligned}(K - J)\mu\Omega &= \int_{z \in \mathbb{L}} \int_{x \in \mathbb{S}} \left( \left( \int_{s \in \mathbb{S}} \kappa(dx, s)\mu(ds, dz) \right) - k(x)\mu(dx, dz) \right) \\ &= \int_{z \in \mathbb{L}} \left( \int_{s \in \mathbb{S}} k(s)\mu(ds, dz) - \int_{x \in \mathbb{S}} k(x)\mu(dx, dz) \right) \\ &= 0.\end{aligned}$$

The bound (2.15) on $k$ ensures that the norm constraints on $K$ and $J$ are satisfied:

$$\begin{aligned}\|J\mu\|_{TV} &\leq \int_{z \in \mathbb{L}} \int_{x \in \mathbb{S}} k(x) \, |\mu| \, (dx, dz) \\ &\leq \frac{1}{2} \|\mu\|_{TV}\end{aligned}$$

$$\begin{aligned}\|K\mu\|_{TV} &\leq \int_{z \in \mathbb{L}} \int_{x \in \mathbb{S}} \int_{s \in \mathbb{S}} \kappa(dx, s) \, |\mu| \, (ds, dz) \\ &\leq \int_{z \in \mathbb{L}} \int_{s \in \mathbb{S}} k(s) \, |\mu| \, (ds, dz) \\ &\leq \frac{1}{2} \|\mu\|_{TV}\end{aligned}$$

With the assumptions so far, the requirements of Section 2.1 are satisfied, so Proposition 2.11 applies and there is a unique solution to initial value problems for $u$.

To reduce the measure-valued dynamics to feature dynamics as in Section 2.2, we must verify more properties of $K$ and $J$ and make several additional assumptions. We assume that learning takes place from a local average of speech patterns. To formulate the $T$ operator, we need a representation of a physical neighborhood. This is accomplished with a spatial influence kernel:

ASSUMPTION 2.31. *$\phi : \mathbb{S} \times \mathbb{S} \to [0, \infty)$ is an influence kernel, where $\phi(x, s)$ represents the influence of speech patterns at location $s$ on a child learning at location $x$. We assume that $\phi$ is integrable and bounded. Furthermore, for each $x$, the function $\phi(x, \cdot)$ should be strictly positive in some open neighborhood around $x$, indicating that there is at least some local influence on children learning at $x$.* For notational convenience, define the operators

$$\begin{aligned}(M_0\mu)(dx) &= \int_{z \in \mathbb{L}} \mu(dx, dz) \\ (M_1\mu)(dx) &= \int_{z \in \mathbb{L}} z\mu(dx, dz)\end{aligned} \tag{2.16}$$

Note that $M_0\mu$ is a probability measure on $\mathbb{S}$ if $\mu$ is a probability measure on $\Omega$, and represents the marginal spatial distribution of the population ignoring speech patterns. In general, $\|M_0\mu\|_{TV} = \|\mu\|_{TV}$. Also, $M_1$ is an $\mathbb{L}$-valued measure on $\mathbb{S}$. (See Chapter 8 of [40] for an introduction to vector-valued measures and further references.)

To combine the mean speech patterns from all speakers in an area, we will need to weight $M_0\mu$ and $M_1\mu$ against the influence kernel $\phi$, for which the following notation will be useful,

$$(\phi \circledast \nu)(x) = \int_{s \in \mathbb{S}} \phi(x, s)\nu(ds). \tag{2.17}$$

Thus, the mean speech pattern seen by a child as a function of location $x$ and weighted by $\phi$ is

$$\frac{(\phi \circledast M_1\mu)(x)}{(\phi \circledast M_0\mu)(x)}.$$

With the goal of expressing learning as a function of this mean while maintaining the form $Q(u) = q(Tu)$, we define

$$T\mu = \begin{pmatrix} M_0\mu \\ M_1\mu \end{pmatrix}. \tag{2.18}$$

Thus, $\mathcal{Q} = T(\mathcal{P})$ is a set of pairs, the first component of which is an $\mathbb{R}$-valued probability measure $m_0$ representing the spatial distribution of the population, and the second component of which is an $\mathbb{L}$-valued measure $m_1$ representing the average speech pattern over a set of locations. We take $\mathcal{Y} \supset \mathcal{Q}$ to be the Banach space of all such pairs under the norm

$$\left\| \begin{pmatrix} m_0 \\ m_1 \end{pmatrix} \right\|_{\mathcal{Y}} = \max\left\{ \|m_0\|_{TV}, \|m_1\|_{TV} \right\} \tag{2.19}$$

For an $\mathbb{L}$-valued measure $\nu$, $|\nu|$ is the $\mathbb{R}$-valued measure

$$|\nu|\, A = \sup_{\text{partitions } \{F_j\} \text{ of } A} \sum_j \|\nu A\|_{\mathbb{L}}$$

and $\|\nu\|_{TV} = |\nu|\,\Omega$. Therefore $T : \mathcal{M} \to \mathcal{Y}$ is a bounded linear operator with $\|T\mu\|_{\mathcal{Y}} \leq \|\mu\|_{TV}$. Given a local learning function $q_{\text{loc}}(p, dz)$ that gives the probability measure representing the speech pattern $z$ of a child who learns from hearing a local average speech pattern $p$, the overall learning function is

$$q\left( \begin{pmatrix} m_0 \\ m_1 \end{pmatrix}, dx, dz \right) = q_{\text{loc}}\left( \frac{(\phi \circledast m_1)(x)}{(\phi \circledast m_0)(x)}, dz \right) m_0(dx). \tag{2.20}$$

As a technical point, we must be careful about the case $(\phi \circledast m_0)(x) = 0$, which happens only when a region around $x$ is uninhabited and no births should take place there. Since we assumed that $\phi(x, \cdot)$ is strictly positive in some open neighborhood around $x$, the only way for $(\phi \circledast m_0)(x)$ to be zero is for $m_0$ to be zero in an open neighborhood around $x$, in which case $m_0(dx) = 0$, and we adopt the convention that $q(m, dx, dz)$ should be zero for such values of $x$. Note that since $q$ is defined in

(2.20) as a product of the spatial distribution and a speech pattern measure, the birth process does not change the spatial distribution of the population.

It remains to verify (2.7) for the appropriate operator $H$. It will be useful to overload the $\circledast$ notation so that

$$(\kappa \circledast \nu)(dx) = \int_{s\in\mathbb{S}} \kappa(dx, s)\nu(ds)$$

which represents a flow rate into $dx$ given a spatial distribution $\nu$. The composition of the operators $M_0$ and $M_1$ with the migration operators $K$ and $J$ can be reversed with Fubini's theorem, and expressed using $\circledast$,

$$\begin{aligned}\big(M_1(K\mu)\big)(dx) &= \int_{z\in\mathbb{L}} z \int_{s\in\mathbb{S}} \kappa(dx, s)\mu(ds, dz)\\ &= \int_{s\in\mathbb{S}} \kappa(dx, s) \int_{z\in\mathbb{L}} z\mu(ds, dz)\\ &= (\kappa \circledast M_1\mu)(dx)\end{aligned}$$

$$\begin{aligned}\big(M_1(J\mu)\big)(dx) &= \int_{z\in\mathbb{L}} zk(x)\mu(dx, dz)\\ &= \big(k(x)M_1\mu\big)(dx)\end{aligned}$$

and similarly for $M_0$.

Treating $m$ as a two-component vector,

$$m = \begin{pmatrix} m_0 \\ m_1 \end{pmatrix}$$

and interpreting integrals accordingly, the correct choice of $H$ is

$$Hm = \kappa \circledast m - km \tag{2.21}$$

where $\kappa \circledast m$ gives the flow rate of features into a set of locations and $km$ gives the flow rate out of a set of locations.

A calculation verifies that $\|H\| \le 1$. First, we find a bound on the norm of the operator $\kappa \circledast \cdot$ acting on a $\mathbb{C}$- or $\mathbb{L}$-valued measure $\nu$ on $\mathbb{S}$.

$$\begin{aligned}\|\kappa \circledast \nu\|_{TV} &= \sup_{\text{partitions } \{F_j\} \text{ of } \mathbb{S}} \sum_j \left| \int_{s\in\mathbb{S}} \kappa(F_j, s)\nu(ds) \right|\\ &\le \sup_{\text{partitions } \{F_j\} \text{ of } \mathbb{S}} \int_{s\in\mathbb{S}} \sum_j \kappa(F_j, s)\,|\nu|\,(ds)\\ &= \int_{s\in\mathbb{S}} \kappa(\mathbb{S}, s)\,|\nu|\,(ds)\\ &= \int_{s\in\mathbb{S}} k(s)\,|\nu|\,(ds)\\ &\le \|k\|_{\sup} \|\nu\|_{TV}\end{aligned}$$

Applying this result to the two components of $m \in \mathcal{Y}$ gives $\|\kappa \circledast m\|_{\mathcal{Y}} \le \|k\|_{\sup} \|m\|_{\mathcal{Y}}$, so $\|\kappa \circledast \cdot\| \le \|k\|_{\sup}$. Using the triangle inequality and the bound (2.15) on $k$, $\|H\| = \|(\kappa \circledast \cdot) - kI\| \le 2\|k\|_{\sup} < 1$.

Now the results of Section 2 apply, and the asymptotic dynamics of $u$ under (2.6) are controlled by the asymptotic dynamics of $m$ under (2.10), reproduced here with $K$ and $J$ filled in:

$$\begin{aligned}\frac{d}{dt}\begin{pmatrix}m_0(dx)\\m_1(dx)\end{pmatrix} &= \left(\int_{z\in\mathbb{L}}\begin{pmatrix}1\\z\end{pmatrix} q_{\text{loc}}\left(\frac{(\phi\circledast m_1)(x)}{(\phi\circledast m_0)(x)}, dz\right)\right) m_0(dx)\\ &\quad - \begin{pmatrix}m_0(dx)\\m_1(dx)\end{pmatrix}\\ &\quad + \left(\int_{s\in\mathbb{S}}\kappa(dx,s)m(ds)\right) - k(x)m(dx)\\ &= Tq(m) - m + \kappa\circledast m - km\end{aligned} \tag{2.22}$$

Since $q_{\text{loc}}(q, dz)$ is always a probability measure in $dz$, the upper component of the integral with respect to $z$ will always give 1. Thus the equation for $m_0$ can be simplified:

$$\frac{d}{dt}m_0(dx) = \left(\int_{s\in\mathbb{S}}\kappa(dx,s)m_0(ds)\right) - k(x)m_0(dx) \tag{2.23}$$

Equation (2.23) is linear in $m_0$ with no dependence on $m_1$. It may therefore be solved using an operator exponential, so $m_0$ can be taken as known when investigating the dynamics of $m_1$.

**3. Dynamics for a single binary choice.** We now apply the mathematical machinery of Section 2: Assume there are two grammatical options $G_1$ and $G_2$ for expressing a particular meaning, and that each individual uses $G_1$ some fraction of the time and $G_2$ the rest. Section 3.1 discusses a population with no spatial structure or migration. In Section 3.2, the population is divided into two compartments. In Section 3.3, the population is evenly distributed over a circle. For each of these examples, the measure-valued differential equation has unique solutions for each initial condition, and the dimension reduction propositions apply. We can then examine the asymptotically stable structures of the feature dynamics and conclude that the full measure-valued dynamics have parallel structures.

**3.1. A well-mixed population.** If we imagine a child learning from a reasonably large sample of the population and retaining no memory of who said which sentence, then the child will hear the $G_1$ option at approximately the average usage rate. Space consists of a single point, for which there is only a single probability measure, so there is no need to represent it. The migration operators are very simple: $K = J = 0$.

The changing population is therefore represented by a time-dependent probability measure $u(t)$ on $\Omega = [0, 1]$, and for a set $A \subset \Omega, A \in \mathcal{B}\Omega$, the measure of $A$ at time $t$, denoted $u(t)A$, is the fraction of the population that uses the $G_1$ option at a rate in $A$.

The most obvious choice of features is the mean, or in general some moment or vector of moments of the distribution. We set $\mathcal{Y} = \mathbb{R}$, $\mathcal{Q} = [0, 1]$, and

$$T\mu = \int_{z\in\Omega} z\,\mu(dz). \tag{3.1}$$

We assume that the learning function $q : \mathcal{Q} \to \mathcal{P}$ is continuously differentiable, which implies that $g = T \circ q$ is also continuously differentiable. We will focus on the feature dynamics and leave $q$ unspecified, because the dimension reduction theorems

imply that given $g$, any admissible $q$ yields measure-valued dynamics with the same asymptotic behavior. The shape of $g$ is the driving force behind the dynamics.

The feature dynamics $m' = g(m) - m$ now take place in a one-dimensional interval. Thus, the possible behaviors of $m$ are sharply limited: $m(t)$ must converge to a fixed point as $t \to \infty$. Different mean learning algorithms $g(m)$ yield different fixed point configurations.

**3.1.1. The case of mean-perfect learning.** If we suppose that learning is mean-perfect, that is, children exactly reproduce the mean usage rate of $G_1$ in the overall population, then $g(m) = m$, and the $m$ dynamics are simply $m' = 0$. Thus, the initial mean usage rate remains unchanged and the population converges to $u = q(Tu_0)$.

This learning algorithm may be appropriate for cases in which a language stably maintains multiple options for expressing a meaning. An example in English is the *dative alternation* [4]. Many verbs such as *give* that take a subject and two objects can be used with or without a preposition on the indirect object:

(3.2) John gives a book to Fred

(3.3) John gives Fred a book

This choice has been present in English for centuries, and there is no sign that either of these options is in danger of disappearing.

Other persistent alternations include *stranding* vs. *pied-piping* of certain verbal particles,

(3.4) I turned the light on — The particle *on* is stranded

(3.5) I turned on the light — The particle *on* follows the verb *turn* via pied-piping

and the choice of *that* as a complementizer or a null complementizer, known as *that-deletion*

(3.6) I know that the light is on

(3.7) I know the light is on

See [21], for example.

**3.1.2. The case of sigmoid learning.** Frequently, languages prefer to use one option almost exclusively. As an example from English, a few special verbs occur before the negative word *not* or its contraction *-n't* and before the subject of a question:

(3.8) I can see the other side.

(3.9) He can't see the other side.

(3.10) Can you see the other side?

However, most verbs are left in a lower position in the syntactic tree and cannot appear before *not* or in inverted questions. A syntactic process called *do-support* inserts the auxiliary verb *do* in negative statements and questions:

(3.11) I like mowing grass.

(3.12) *He likes not mowing grass.

(3.13) He doesn't like mowing grass.

(3.14) *Like you mowing grass?

(3.15) Do you like mowing grass?

(The * indicates an ungrammatical sentence.) Old English had a syntactic process called *verb raising* that raised all verbs to a higher position. In a verb-raising grammar main verbs appear before negation and in inverted questions and there is no need for the auxiliary *do*. Old English used verb raising almost exclusively, but over the centuries the grammar changed. Modern English uses *do*-support almost exclusively.

Despite this change, both grammars were more or less stable for centuries. To model this mutual exclusion, a one-dimensional phase portrait must contain two stable fixed points close to the extremes, separated by an unstable fixed point. This implies that $g(m)$ is sigmoid shaped. That is, $g$ is smooth, strictly monotone, and bounded, with one inflection point, but $g$ is not necessarily an exponential sigmoid as in $f(x) = 1/(1+e^{-x})$. Furthermore, there must be three solutions to $g(m) = m$ to generate the correct number of fixed points. See Figure 3.1.

This one-dimensional model is not capable of representing language change. Every population tends to one of the stable fixed points and stays there. Even given a fairly large perturbation, trajectories in this model tend to return to their original equilibrium because the stable fixed points are well away from the boundary point separating their basins of attraction. Propositions 2.18, 2.19, 2.20, 2.21 and their corollaries apply. Therefore, the measure-valued dynamics are also constrained to exhibit two stable fixed points for any admissible learning function $q$. A more complex, inherently higher-dimensional model is required to model transitions between grammars.

**3.2. A single binary choice with two regions.** The mathematical machinery developed so far also works if the population is divided into compartments. Consider the simplest case, a linguistic population with two regions which will be called north and south. As before, we assume that there are two idealized grammars, that children learn from a sample of sentences spoken by people in their native region, that they effectively learn from the mean speech pattern of their native region, and that such learning takes place under a sigmoid learning function as in Section 3.1.2. In addition, people move from one region to another. This model of was analyzed heuristically in [30] assuming idealized speech, but it lies within the current rigorous framework and there is no longer any need to assume that each individual's speech pattern is limited to strictly $G_1$ or $G_2$. Instead, the dimension reduction results from Section 2.2 allow us to formulate the same two-dimensional dynamical system on the foundation of measure dynamics. We will relate the behavior of this reduced system to a change in English syntax.

Each individual is in either the north ($\mathtt{N}$) or the south ($\mathtt{S}$), and may be characterized by a usage rate for $G_1$ between 0 and 1. Thus, $\mathbb{S} = \{\mathtt{N}, \mathtt{S}\}$ and $\mathbb{L} = [0, 1]$. A measure $\nu$ on $\mathbb{S}$ is a linear combination of delta measures on the two points of $\mathbb{S}$, or essentially a pair of numbers $(\nu_N, \nu_S)$:

$$\nu(dx) = \nu_N \delta_{\mathtt{N}}(dx) + \nu_S \delta_{\mathtt{S}}(dx)$$

Integration against such a measure is just a sum.

$$\int_{x \in \mathbb{S}} f(x)\nu(dx) = f(\mathtt{N})\nu_N + f(\mathtt{S})\nu_S.$$

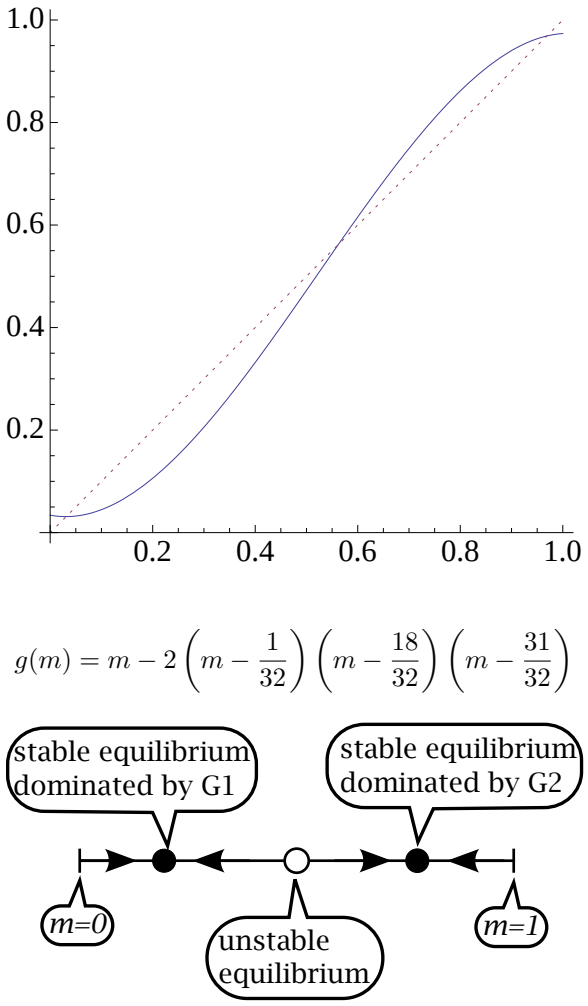



FIG. 3.1. *If the mean learning function $g(m)$ is a sigmoid as shown in the graph, the phase portrait for the $m$ dynamics includes two stable fixed points separated by one unstable fixed point. Generically, the population will tend toward one of the stable fixed points thereby settling into a state where one of the two grammars is used almost exclusively. The specific function given here is used as an example throughout.*

A measure $\mu$ on $\Omega$ is essentially pairs of measures $(\mu_N, \mu_S)$ on $\mathbb{L}$:

$$\mu(dx, dz) = \mu_N(dz)\delta_{\mathtt{N}}(dx) + \mu_S(dz)\delta_{\mathtt{S}}(dx).$$

This suggests that measures should be written as row vectors and elements of $\Omega$ should be written as column vectors with indices `N` and `S` instead of 1 and 2, as in

$$\begin{aligned}\mu(X \times Z) &= \mu\big((\{\mathtt{N}\} \times Z_N) \cup (\{\mathtt{S}\} \times Z_S)\big) \\ &= \begin{pmatrix}\mu_N & \mu_S\end{pmatrix}\begin{pmatrix}Z_N \\ Z_S\end{pmatrix} \\ &= \mu_N Z_N + \mu_S Z_S\end{aligned}$$

Thus, the migration kernel may be represented as a matrix that acts on the right

$$\kappa = \begin{pmatrix} 0 & \eta_N \\ \eta_S & 0 \end{pmatrix}.$$

The migration operators are

$$K = \begin{pmatrix} 0 & \eta_N \\ \eta_S & 0 \end{pmatrix}, \quad J = \begin{pmatrix} \eta_N & 0 \\ 0 & \eta_S \end{pmatrix} \tag{3.16}$$

with the understanding that the action of a matrix $A$ on a measure $\mu$ is multiplication on the right

$$\begin{pmatrix} \mu_N & \mu_S \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} a_{11}\mu_N + a_{21}\mu_S & a_{12}\mu_N + a_{22}\mu_S \end{pmatrix} \tag{3.17}$$

With the migration operators handled, we turn our attention to the feature operator $T$:

$$T\mu = \begin{pmatrix} \int_0^1 \mu_N(dz) & \int_0^1 \mu_S(dz) \\ \int_0^1 z\,\mu_N(dz) & \int_0^1 z\,\mu_S(dz) \end{pmatrix} \tag{3.18}$$

That is, $T\mu$ is the mass and mean of the north and south submeasures of $\mu$. The corresponding $H$ operator is the matrix $K - J$ with its normal right action on $\mathbb{R}^2$. The feature-based learning function $q$ is defined by applying a local learning function $q_{\text{loc}} : [0,1] \to \mathcal{P}$ to the mean usage rate in each region. Because the individual submeasures do not in general have unit mass, we must divide by their masses when applying $q_{\text{loc}}$.

$$q \begin{pmatrix} m_{N0} & m_{S0} \\ m_{N1} & m_{S1} \end{pmatrix} = \begin{pmatrix} m_{N0}\, q_{\text{loc}}\left(\frac{m_{N1}}{m_{N0}}\right) & m_{S0}\, q_{\text{loc}}\left(\frac{m_{S1}}{m_{S0}}\right) \end{pmatrix} \tag{3.19}$$

It should be understood that if $m_{N0} = 0$, then the left entry should be 0, and similarly for the right entry. These conditions cover the cases when one of the regions is empty and no births should take place there. To unify (3.19) with (2.20), note that since $\mathbb{S} = \{\mathtt{N}, \mathtt{S}\}$ and children learn only from others in their native region, the influence kernel $\phi$ may be represented as the identity matrix.

The results of Section 2 apply, so we may focus our attention on the feature matrix $m$,

$$m = \begin{pmatrix} m_{N0} & m_{S0} \\ m_{N1} & m_{S1} \end{pmatrix} = \begin{pmatrix} \int_0^1 \mu_N(dz) & \int_0^1 \mu_S(dz) \\ \int_0^1 z\,\mu_N(dz) & \int_0^1 z\,\mu_S(dz) \end{pmatrix} \tag{3.20}$$

The dynamics of $m$ in this case simplify to

$$\begin{aligned} m'_{N0} &= \eta_S m_{S0} - \eta_N m_{N0} \\ m'_{S0} &= \eta_N m_{N0} - \eta_S m_{S0} \\ m'_{N1} &= m_{N0}\left(\int_0^1 z\, q_{\text{loc}}\left(\frac{m_{N1}}{m_{N0}}, dz\right)\right) + \eta_S m_{S1} - \eta_N m_{N1} - m_{N1} \\ m'_{S1} &= m_{S0}\left(\int_0^1 z\, q_{\text{loc}}\left(\frac{m_{S1}}{m_{S0}}, dz\right)\right) + \eta_N m_{N1} - \eta_S m_{S1} - m_{S1} \end{aligned} \tag{3.21}$$

The equations for the masses are uncoupled from the others, as in (2.23). Since the feature matrix also satisfies $m_{N0} + m_{S0} = 1$, we may eliminate one of those two variables entirely, as knowing one determines the other. We eliminate $m_{S0}$, thus,

$$m'_{N0} = \eta_S(1 - m_{N0}) - \eta_N m_{N0} = \eta_S - (\eta_N + \eta_S)m_{N0}. \tag{3.22}$$

This equation implies that $m_{N0}$ and $m_{S0}$ converge exponentially fast to equilibrium values, representing the long-term behavior of the migration process alone:

$$\begin{aligned} m_{N0} \to \bar{m}_{N0} &= \frac{\eta_S}{\eta_N + \eta_S} \\ m_{S0} \to \bar{m}_{S0} &= \frac{\eta_N}{\eta_N + \eta_S} \qquad \text{as } t \to \infty. \end{aligned} \tag{3.23}$$

It is advantageous at this point to introduce new variables representing the mean usage rates of the two regions

$$\begin{aligned} x_N &= \frac{m_{N1}}{m_{N0}} \\ x_S &= \frac{m_{S1}}{m_{S0}} \end{aligned} \tag{3.24}$$

and derive the following dynamics for them from (3.21).

$$\begin{aligned} x'_N &= g(x_N) + \eta_S \frac{m_{S0}}{m_{N0}} x_S - \eta_N x_N - x_N \\ x'_S &= g(x_S) + \eta_N \frac{m_{N0}}{m_{S0}} x_N - \eta_S x_S - x_S \\ &\text{where } g(p) = \int_0^1 z \, q_{\text{loc}}(p, dz) \end{aligned} \tag{3.25}$$

Since $m_{N0}$ and $m_{S0}$ each flow toward a unique fixed value, the variables $x_N$ and $x_S$ are ultimately controlled by simple two dimensional dynamics,

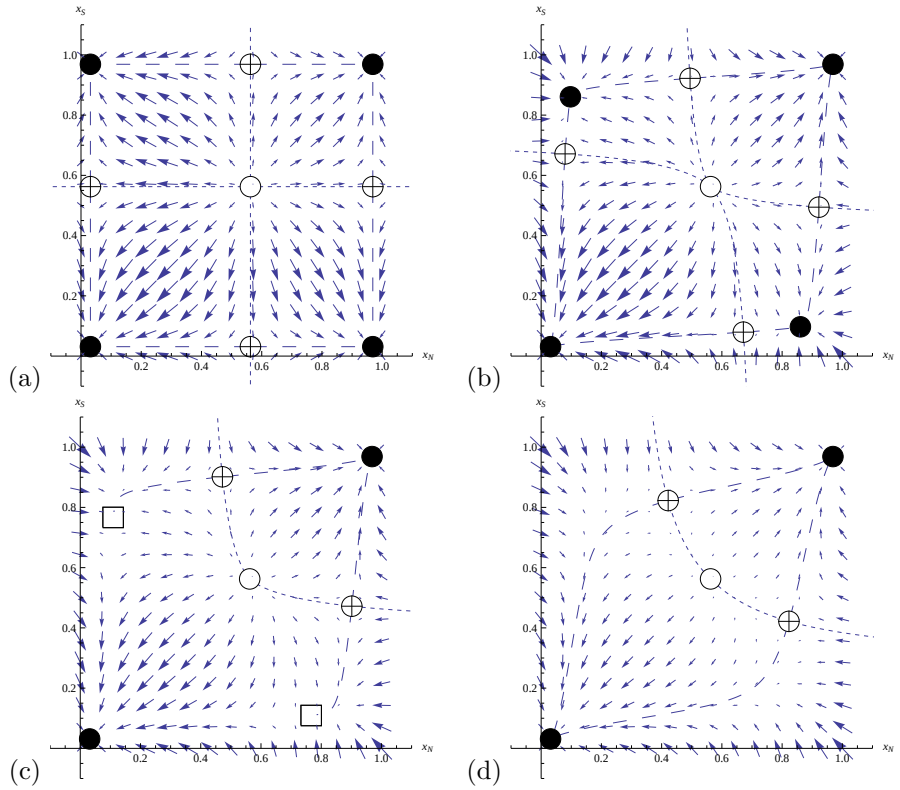
$$\begin{aligned} x'_N &= g(x_N) - x_N + \eta_N(x_S - x_N) \\ x'_S &= g(x_S) - x_S + \eta_S(x_N - x_S) \end{aligned} \tag{3.26}$$

For the rest of this section, we will assume that $g$ is a sigmoid, as in Figure 3.1.

If no mixing at all occurs, that is $\eta_N = \eta_S = 0$, then $x_N$ and $x_S$ uncouple completely. Intuitively, each region picks a dominant language independently of the other. The population as a whole can stably maintain both $G_1$ and $G_2$ through split states in which one region is dominated by $G_1$ and the other by $G_2$. The resulting $(x_N, x_S)$ phase portrait has four stable fixed points separated by a variety of unstable fixed points, as in Figure 3.2(a).

If $\eta_N$ and $\eta_S$ are sufficiently large, then the two regions mix strongly with each other and effectively become a single region. This yields phase portraits as in Figure 3.2(d) where the stable population states require both regions to be dominated by the same grammar.

In between, there are intermediate states and a pair of bifurcations representing the loss of the stable split states, as in Figure 3.2(c). These bifurcations may be interpreted as a model of language change through contact between dialects: Two initially separate populations maintain different dialects, but as contact between the regions increases, they effectively become unified and one dialect or the other disappears.




Key: ∘ = unstable fixed point, • = stable fixed point, ⊕ = saddle. Thick lines are the stable manifolds of saddles. Thin lines are the unstable manifolds of saddles.

FIG. 3.2. *Phase portraits for* (3.26) *with* $\eta_N = \eta_S = \eta$. *In (a),* $\eta = 0$. *As* $\eta$ *increases in (b), (c), and (d), the fixed points shift until bifurcations wipe out the stable split states. In (c), both stable split states bifurcate at the same value of* $\eta$ *because of the symmetric choice of* $\eta_N = \eta_S = \eta$.

We may even set the migration parameters $\eta_N = \eta_S = \eta$, put $\eta$ into motion as a function of $t$, and visualize the change as a time trace as in Figure 3.3. We start the population near the split state closest to $(1, 0)$, which represents a population whose northern region uses $G_1$ and whose southern region uses $G_2$. At first, the population tracks the stable split state as it shifts, and maintains both grammars. Once $\eta$ is large enough, the bifurcation occurs and the stable split state vanishes. Then the population converges quickly to the single-language fixed point near $(0, 0)$, and $G_1$ becomes essentially extinct.

As a final detail, since all trajectories tend to a fixed point in this model, the dimension reduction theorems guarantee that the full measure-valued dynamical system has the same asymptotic behavior for any admissible learning function $q$: All trajectories converge to a steady-state probability measure. For small values of $\eta$, there are stable split states, but as $\eta$ increases, a bifurcation eliminates those split-state fixed points.

**3.2.1. The loss of V2 in English.** There is a change in word order in Middle English that is thought to have been caused by contact between different grammars. Middle English can be divided into two or more regional dialects. Initially, all had some form of the *verb-second* or V2 rule, still present in modern German, which moved

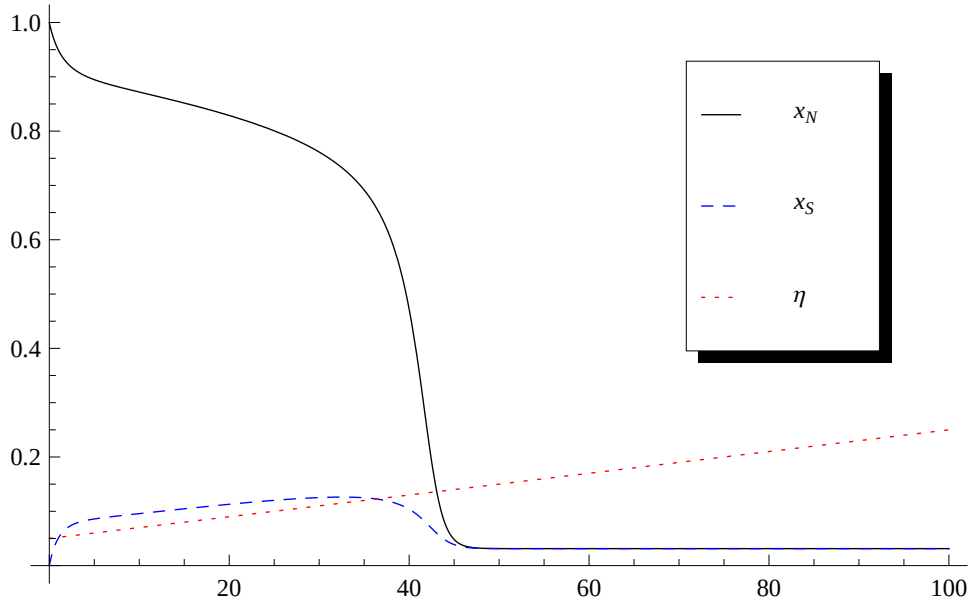



FIG. 3.3. *Values of $x_N$ (solid) and $x_S$ (dotted) as functions of time, starting from $x_N \approx 1$ and $x_S \approx 0$, where $\eta_N = \eta_S = \eta$ and $\eta$ (dashed) increases linearly as a function of time. The population state $(x_N, x_S)$ tracks the stable fixed point representing a split state until the bifurcation annihilates it around $t = 65$. Then the population converges quickly to the single-language fixed point near $(0, 0)$ representing the extinction of $G_1$.*

the finite verb to the front of the sentence, and a topic in front of that. Northern Middle English already had a different form of V2 than southern Middle English because of contact with Norse-speaking settlers. Apparently, increased contact between the northern and southern dialects led to the development of a non-V2 word order similar to Modern English [25, 30]. Although this scenario is somewhat more complicated than the two-grammar choice studied here, the two-grammar dynamics and bifurcation still give some insight into how contact can lead to language change.

**3.3. A single binary choice in continuous space.** As an alternative to a compartment model, we will consider in this section a population spatially distributed over a circle, so $\mathbb{S}$ is the interval $[0, 1]$ with periodic boundary points. As before, an individual's speech pattern is represented by a usage rate between 0 and 1. Initially, the model is in integral form, but it can be related to a reaction-diffusion equation, and we will investigate the possibility of traveling wave solutions, which represent the spread of a language change. The traveling wave can be related to a phonology change taking place in Pennsylvania.

To simplify this example, we assume that $m_0$ is fixed at a uniform distribution,

$$m_0(dx) = dx$$

and omit explicit dependence on $m_0$ where possible. The migration kernel $\kappa(ds, x)$ is assumed to have a peak with mirror symmetry centered at $x$ when viewed as a function of $s$. In particular, for each $x$,

$$\int_{s\in\mathbb{S}} (s - x)\kappa(ds, x) = 0.$$

Thus, the population effectively diffuses through space.

We assume that the mean of the learning distribution $q$ has a smooth sigmoid-shaped density function $g(p)$ as in Figure 3.1,

$$g(p) = \int_{z\in[0,1]} z\, q_{\text{loc}}(p, dz) \tag{3.27}$$

Additionally, $\phi(x,\cdot)$ must also have a peak with mirror symmetry centered at $x$, indicating that the greatest influence is from nearby speakers. Since we will be holding the spatial distribution $m_0$ constant and uniform, $\phi \circledast m_0$ is constant. The assumed form of $\phi$ means that $\phi \circledast \mu$ is a scaled smoothing of $\mu$.

**3.3.1. Connection to a reaction-diffusion equation.** Continuing from Section 2.4, it is natural to consider cases where $m_1$ has a smooth time-dependent density $w$ with respect to Lebesgue measure,

$$m_1(dx) = w(x)dx$$

and use asymptotic arguments to relate the $m_1$ dynamics to a reaction-diffusion equation in $w$. Because of the assumptions on $\phi$, the local average usage rate passed to $q_{\text{loc}}$ is nearly an identity transformation,

$$\frac{(\phi \circledast m_1)(x)}{(\phi \circledast m_0)(x)} \approx w(x)$$

This means that the learning term in (2.22) is a sort of local average of a sigmoid,

$$\begin{aligned} G(w,x) &= \int_{z\in[0,1]} z\, q_{\text{loc}}\left(\frac{(\phi \circledast m_1)(x)}{(\phi \circledast m_0)(x)}, dz\right) \\ &\approx g(w(x)). \end{aligned} \tag{3.28}$$

Using a series for $w$ about $x$ and dropping terms of higher order than quadratic, the $w$ dynamics derived from the $m_1$ dynamics of (2.22) become (suppressing the explicit dependence of $w$ on $t$)

$$\begin{aligned} \partial_t w(x) &= G(w,x) - w(x) \\ &\quad + \int_{s\in\mathbb{S}} \left(w(x) - \partial_x w(x)(s-x) + \frac{\partial_x^2 w(x)}{2}(s-x)^2\right)\kappa(ds,x) \\ &\quad - k(x)w(x) + \cdots \end{aligned} \tag{3.29}$$

Integrating term by term,

$$\begin{aligned} \partial_t w(x) &= G(w,x) - w(x) \\ &\quad + w(x)\left(\int_{s\in\mathbb{S}} \kappa(ds,x)\right) - \partial_x w(x)\left(\int_{s\in\mathbb{S}} (s-x)\kappa(ds,x)\right) \\ &\quad + \frac{1}{2}\partial_x^2 w(x)\left(\int_{s\in\mathbb{S}} (s-x)^2\kappa(ds,x)\right) - k(x)w(x) + \cdots \end{aligned}$$

the first integral $\int_s \kappa(ds,x) = k(x)$ cancels out. The second integral $\int_s (s-x)\kappa(ds,x) = 0$ because $\kappa$ is assumed to have mirror symmetry about $x$. The remaining integral

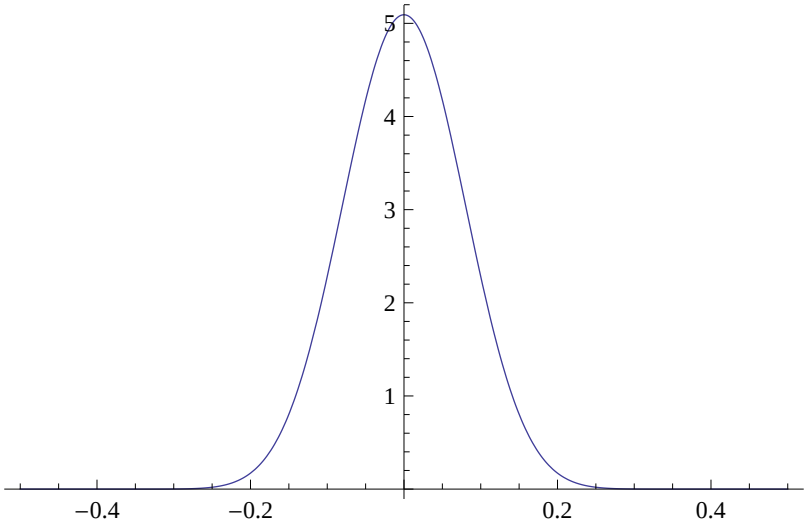



FIG. 3.4. *Graph of* $\phi(x, 0)$.

$\int_s (s-x)^2 \kappa(ds, x)$ is some function we will denote $\sigma^2(x)$ that represents the space-dependent variance of $\kappa$. Dropping the remaining terms, we are left with a reaction-diffusion partial differential equation [11, 14],

$$\partial_t w(x) = G(w, x) - w(x) + \frac{\sigma^2(x)}{2} \partial_x^2 w(x). \tag{3.30}$$

If as supposed $G(w, x) \approx g(w(x))$, then $G(w, x) - w(x)$ will be roughly a cubic-shaped function of $w(x)$, and this equation will have traveling wave solutions [11, Section 4.2]. The interpretation of such a solution is that a language change can begin at one point and propagate throughout the space.

With this connection to reaction-diffusion equations, it is reasonable to suppose that the measure dynamics may exhibit solutions typical of reaction-diffusion equations, such as traveling waves, diffusive Turing instabilities, and pattern formation, depending on the specific choice of $q$, $K$, and $J$.

**3.3.2. Some numerical results.** We now examine some pictures of the measure-valued dynamics at work. To keep the numerics simple, the choice of $\mathbb{S}$ is a circle $S^1$ represented as the interval $[0, 1]$ with periodic boundary conditions. The feature extraction operator $T$ takes a space-dependent probability measure to its space-dependent mean, with an influence kernel

$$\phi(x, s) = \frac{128}{6435} \left(1 + \cos 2\pi(x - s)\right)^8$$

representing the assumption that children learn from the average speech patterns of nearby speakers. See Figure 3.4. The $1 + \cos$ structure creates positive function with a bump around $x$. The power 8 narrows the bump. The constant factor $128/6435$ ensures that for each $x$, the total influence is 1, that is $(\phi \circledast m_0)(x) = \int_{\mathbb{S}} \phi(x, s) ds = 1$. This normalization is convenient but not strictly necessary because $\phi$ is always used in a quotient as in (2.20).

The migration kernel is similar:

$$\kappa(ds, x) = \frac{16}{6435} \left(1 + \cos 2\pi(x - s)\right)^8 ds$$

with

$$k(x) \equiv \frac{1}{8}.$$

The population remains uniformly distributed in space and only the speech patterns change.

The learning function $q_{\text{loc}}(p)$ is defined to be a $\beta$-distribution with mean $g(p)$ and variance $p^2(1-p)^2$, where $g$ is the polynomial sigmoid function depicted in Figure 3.1. This choice of the variance is so that the $\beta$-distributions have no singularities in their densities, which seems to improve numerical stability. The choice of the $\beta$-distribution is so that we have a realistic specific example for this demonstration. Any single-mode family of distributions supported on $[0, 1]$ and determined smoothly by the mean usage rate would work as well.

The calculations are performed by a Mathematica notebook. At each time step, $u(t, x, z)$ is represented by samples for $(x, z) \in [0, 1] \times [0, 1]$ based on the function's value at each point on a 64 by 65 grid. Each step in the numerical method is an Euler step in $t$ with step size 0.1, followed by a normalization step that ensures $\int_0^1 u(t, x, z)dz = 1$ for every $x$ on the grid. Integrals are computed using the trapezoid rule. Since these figures are for demonstration purposes and no numerical instability is apparent, there is no need at this point for more sophisticated numerical methods.

The results of the $u$ dynamics are shown in Figure 3.5. The results of Section 2 apply, so the asymptotic behavior of $u$ reflects the asymptotic behavior of the feature dynamics. Since $m_0$ is fixed, the interesting feature is the location-dependent mean usage rate $m_1$. The corresponding $m_1$ dynamics are shown in Figure 3.6. There are two spatially uniform steady states given by $\bar{m}_1 =$ a solution to $g(m) = m$. These represent the states where everyone everywhere strongly prefers one idealized grammar over the other.

The initial condition in these figures represents a population where half of the population, centered about $x = 1/2$, prefers $G_1$ and the other half prefers $G_2$. Near the boundaries at $1/4$ and $3/4$, the two mix due to migration. Since the sigmoid function is slightly asymmetric in favor of $G_2$, the preference for $G_2$ tends to spread. Two waves develop, travel toward each other, and meet in the middle, resulting in the disappearance of $G_1$. Since the $m_1$ dynamics show that the initial condition is attracted to that steady state, the full $u$ dynamics must be attracted to a parallel steady state.

The $m_1$ computation is substantially faster than the full measure dynamics. The good agreement of the $u$ and $m_1$ calculations shows that the dimension-reduction results are of practical as well as theoretical value.

It is also possible to do numerical experiments with the reaction-diffusion equation (3.30). Using Mathematica's built-in numerical solver and the approximation $G(w, x) \approx g(w(x))$, we obtain the pictures in Figure 3.7. This calculation is even faster than the $m_1$ dynamics. Compared to the $u$ and $m_1$ dynamics, the $w$ dynamics are qualitatively the same: Two waves meet in the middle and $G_1$ goes extinct. However, the waves take about 5 times as long to disappear, which suggests that the approximation (3.28) for $G(w, x)$ is too crude.

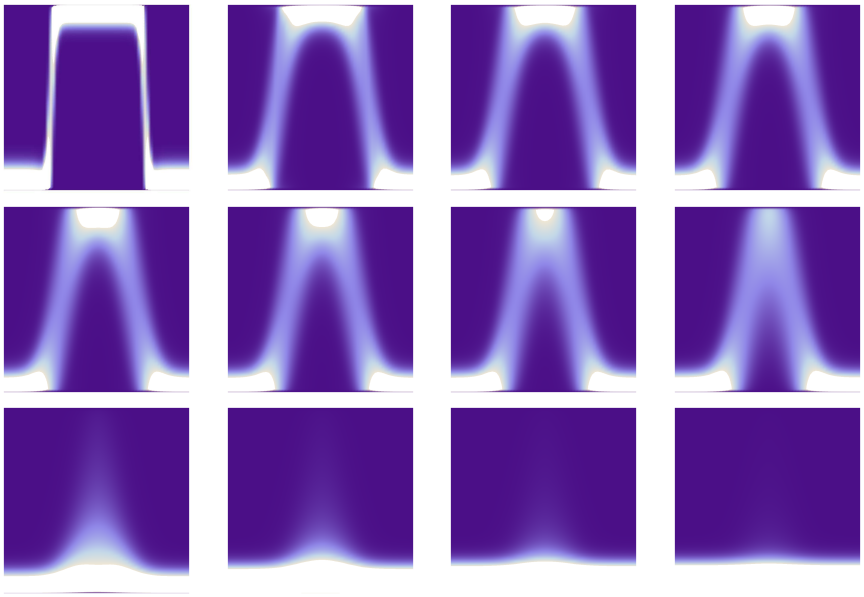


FIG. 3.5. *Density plots of $u(t, x, z)$ for $t = 0, 2, 4, \ldots$ reading left to right and top to bottom. In each plot, the lighter colors represent higher values of $u$ and the darker colors represent values close to 0. The horizontal axis is $x \in [0, 1]$ and the vertical axis is $z \in [0, 1]$.*

**3.3.3. An example of a traveling wave from the linguistics literature.** A well-studied feature of speech in western Pennsylvania is the so-called *low back merger,* in which the low back vowels as in *cot* and *caught* are no longer distinguished. Data collected in 1940 and 1988, as displayed in [16], indicates that the region in which the vowels are merged is growing. The eastern boundary is moving to the east, and part of the wave seems to have stopped at the Susquehanna river and the Pennsylvania German region. Other parts of the boundary coincide "with a weakness in the face-to-face oral communication network, as indicated by population density and traffic patterns" [16, p. 171–172]. This barrier indicates that the process by which the change spreads is spatial and local (although [16] discusses a second region in eastern Pennsylvania where the merger seems to have arisen more or less independently). Furthermore, there is inherent asymmetry in the acquisition of the vowel system with the merger compared to the vowel system without it: In the presence of speakers that use the merged system frequently, a child will have less evidence that the vowels are distinct, making the acquisition and use of the non-merged vowel system more unlikely.

The model described here in Section 3.3 agrees qualitatively with the linguistic data. A combination of local influence and migration, in conjunction with asymmetric learning tendencies, suffice to create a traveling wave in which one grammar disappears in favor of the other.

**4. Discussion and conclusion.** We began by considering the dynamics of a population of individuals whose speech patterns are variable. Each individual's state, including a location and a speech patten, is represented as an element of a set $\Omega$. The population is represented as a probability measure on $\Omega$, and learning, birth, and migration cause it to evolve deterministically. If children learn primarily from some

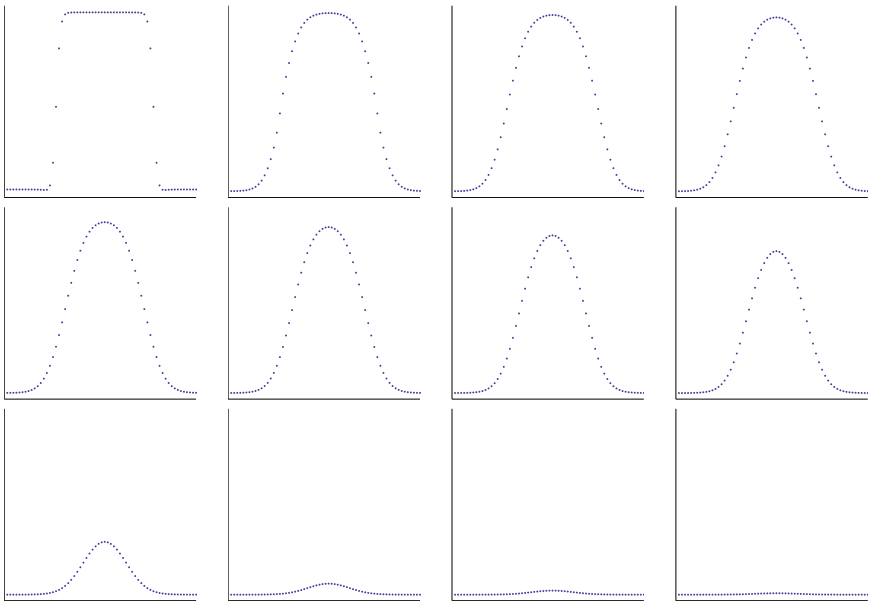


FIG. 3.6. *Plots of $m_1(t,x)$ for $t = 0, 2, 4, \ldots$ reading left to right and top to bottom. The horizontal axis is $x \in [0,1]$ and the vertical axis is $m \in [0,1]$.*

average feature of the population, then it is tempting to formulate the dynamics in some simpler Banach space rather than try to deal with the measure dynamics. However, there is some question as to whether the resulting mean-field feature dynamics accurately represent the original measure dynamics.

The proofs in this paper demonstrate that for a general class of non-linear learning algorithms and linear migration processes, the measure dynamics can indeed be reduced to a dynamical subsystem in an appropriate Banach space without losing any essential information. If the feature dynamics are simple enough that all trajectories converge to some limit cycle or fixed point, then the same is true of the original measure dynamics, and a single observation of the state of the feature dynamics suffices to give an approximation to the corresponding state of the measure dynamics that improves exponentially as time advances. However, if the feature dynamics are chaotic, then sensitive dependence on initial conditions makes it practically impossible to recover the state of the measure dynamics from a single observation of the feature dynamics.

We explored a scenario in which children must learn rates at which to use two alternative grammatical constructions to communicate a single meaning. We assume that children learn only from the mean usage rates. In the case of a well-mixed single-compartment population, the one-dimensional mean-field dynamics have two stable fixed points, one for each alternative construction. An unstable fixed point separates their basins of attraction. For a population with two compartments, children learn only from the mean usage rate within their native compartment. The resulting feature dynamics are two dimensional. If there is very low migration between the regions, then stable split states are possible, in which one compartment prefers one grammatical option and the other compartment prefers the other. As the migration rate increases, the compartments effectively merge. Bifurcations take place that annihilate the stable

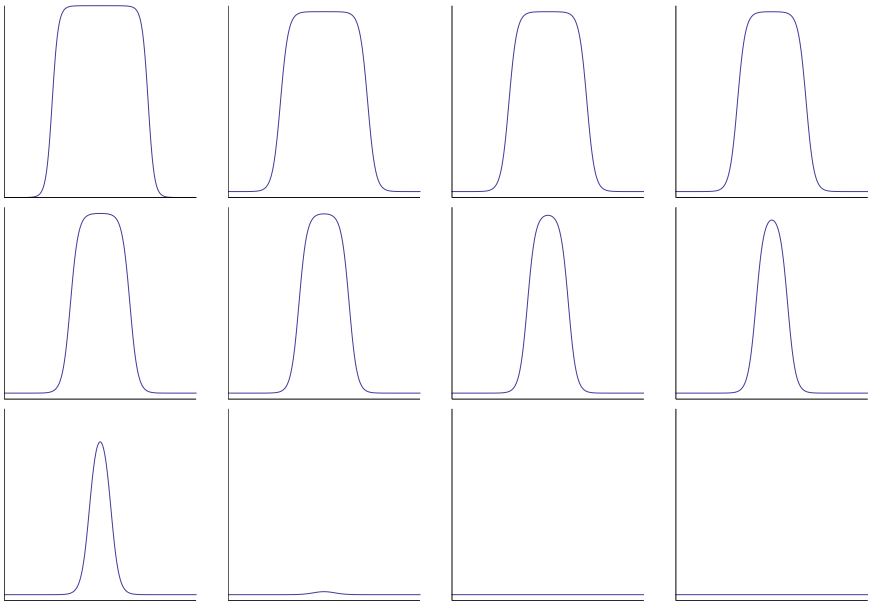


FIG. 3.7. *Plots of $w(t,x)$ for $t = 0, 10, 20, \ldots$ reading left to right and top to bottom. The horizontal axis is $x \in [0,1]$ and the vertical axis is $w \in [0,1]$. Note the difference in time scale compared to Figures 3.5 and 3.6.*

split states, and the entire population ends up with a single preferred choice. This scenario suggests a mechanism by which increased contact among dialects can lead to the extinction of one of them, as seems to have happened in the loss of the verb-second property of Middle English. The mathematical framework for the two-compartment population dynamics could be extended to deal with many compartments.

In addition, we considered the same grammatical scenario with a continuous representation of space. We assume that children learn from other individuals based on their proximity, so the dominant influence is a mean speech pattern weighted by a spatial influence kernel. After verifying that the general mathematical machinery applies, we used an asymptotic argument to relate the feature dynamics to a reaction-diffusion equation, which suggests that the feature and measure dynamics might have traveling wave solutions. A numerical experiment further supports the existence of such solutions. The traveling wave seen in the numerical experiment is associated with the disappearance of one grammar in favor of the other, which agrees qualitatively with observations by Herold of a change in the vowel system spreading eastward in western Pennsylvania [16].

Throughout, the spatial component of $\Omega$ has been interpreted as being literally spatial. However, it is certainly possible to include social or economic status using exactly the same mathematics. All that changes is the interpretation: compartments might represent social classes, and a continuous scale could represent wealth. Migration would then include social and economic mobility.

Similarly, the elements of $\mathbb{L}$ in the examples have all been usage rates of idealized grammars. More generally, any Banach-space-valued feature of speech could be used, for example, the frequencies of vowel formants [22, 23].

There are several shortcomings of this general modeling framework and the specific examples. These provide opportunities for further research.

First, the feature dynamics assume that the population is large enough that children see only mean features. A more realistic model would directly take into account the discrete nature of human populations and the fact that children learn from a large but finite number of sentences spoken primarily by individuals that they have spatial and social proximity to. See for example [5, 12, 29, 44] for learning models based on simulated sentences. It should be possible to prove that under certain hypotheses, discrete finite population models in the framework of Section 2 converge to continuous infinite population models.

Second, we have assumed throughout that migration is independent of language, which simplifies the mathematics but is unrealistic. People tend to form social and economic neighborhoods within cities, for example, and language is correlated with these factors. Adults can change their speech patterns as they age and move among social classes. People also tend to sort themselves spatially into culturally and linguistically homogeneous clusters. To model these effects, the framework discussed here would have to be adapted to allow $K$ and $J$ to depend on $x$ and $u$, thereby introducing additional nonlinearities into the $u$ dynamics.

Third, we have assumed that each individual's speech pattern is drawn from an unconditional probability distribution, but people are known to change their speech pattern as they age and within the social context of each conversation. To account for social context, an additional parameter, say $c$, would have to be added to $u$, so that $u(t, x, c)$ is a probability measure on speech patterns indicating how someone located at $x$ speaks in social context $c$.

An important feature of one- and two-dimensional deterministic continuous dynamical systems is that generically all trajectories converge to a fixed point or a limit cycle. This means that such models of language change are doomed to be "single shot," meaning they are only able to mimic a single instance of a language change in one direction. Some external force is required to push the model to repeat or reverse the change. For example, the traveling wave in Section 3.3 models the merger of two vowels. Given that vowels can merge, that known languages have been experiencing phonemic change for centuries, and that languages exist with as few as three phonemically distinct vowels, it is paradoxical that any language has more than a few vowels. The resolution is that there are competing forces that can cause vowels to split, such as is happening with the short $a$ vowel in some northern dialects of English in the United States [24]. To model a fluctuating vowel system in which vowels merge, split, and shift would require a higher dimensional representation of individual speech patterns, plus some source of random fluctuations. That would give room for a merger in one part of the vowel inventory to be followed by a split elsewhere, preventing a total collapse, and eventually a restoration of a lost distinction.

In summary, the framework outlined in this article puts certain mean-field models of language variation and change on a more secure mathematical footing as reductions of measure-valued dynamical systems. Specifically, it provides a way to eliminate the simplifying assumption that individuals use a single idealized grammar. It also provides for spatially and socially distributed dynamics. Further studies could include relaxing the assumption that the learning and spatial dynamics are independent, and the incorporation of age structure. These possibilities will be addressed in future articles.

## REFERENCES

[1] David Adger, *Core Syntax: A minimalist approach*, Oxford University Press, Oxford, 2003.
[2] Misha Becker, *Learning verbs without arguments: The problem of raising verbs*, Journal of Psycholiguistic Research, 34 (2005), pp. 173–199.
[3] Rens Bod, Jennifer Hay, and Stefanie Jannedy, eds., *Probabilistic Linguistics*, MIT Press, Cambridge, MA, 2003.
[4] Joan Bresnan and Tatiana Nikitina, *The gradience of the dative alternation*, in Uyechi and Wee [45], ch. 13, pp. 161–184.
[5] E. J. Briscoe, *Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device*, Language, 76 (2000), pp. 245–296.
[6] Angelo Cangelosi and Dominico Parisi, eds., *Simulating the Evolution of Language*, Springer-Verlag, New York, 2002.
[7] Noam Chomsky, *Aspects of the Theory of Syntax*, MIT Press, Cambridge, MA, 1965.
[8] ———, *Language and Mind*, Harcourt Brace Jovanovich, New York, 1972.
[9] ———, *Language and Problems of Knowledge*, MIT Press, Cambridge, MA, 1988.
[10] Nelson Dunford and Jacob T. Schwartz, *Linear Operators Part I: General Theory*, vol. VII of Pure and Applied Mathematics, Interscience Publishers, Inc., New York, 1958.
[11] Lawrence C. Evans, *Partial Differential Equations*, American Mathematical Society, Providence, Rhode Island, 1998.
[12] E. Gibson and Kenneth Wexler, *Triggers*, Linguistic Inquiry, 25 (1994), pp. 407–454.
[13] E. Mark Gold, *Language identification in the limit*, Information and Control, 10 (1967), pp. 447–474.
[14] Peter Grindrod, *Patterns and Waves: The theory and applications of reaction-diffusion equations*, Clarendon Press, Oxford, 1991.
[15] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1990.
[16] Ruth Herold, *Solving the actuation problem: Merger and immigration in eastern pennsylvania*, Language Variation and Change, 9 (1997), pp. 165–189.
[17] Simon Kirby and James R. Hurford, *The emergence of structure: An overview of the iterated learning model*, in Cangelosi and Parisi [6], pp. 121–148.
[18] Natalia L. Komarova, Partha Niyogi, and Martin A. Nowak, *The evolutionary dynamics of grammar acquisition*, Journal of Theoretical Biology, 209 (2001), pp. 43–59.
[19] Natalia L. Komarova and Martin A. Nowak, *The evolutionary dynamics of the lexical matrix*, Bulletin of Mathematical Biology, 63 (2001), pp. 451–485.
[20] Anthony Kroch, *Reflexes of grammar in patterns of language change*, Language Variation and Change, 1 (1989), pp. 199–244.
[21] Anthony Kroch and Kathy Small, *Grammatical ideology and its effect on speech*, in Linguistic Variation: Models and Methods, David Sankoff, ed., Academic Press, New York, 1978.
[22] William Labov, *Principles of Linguistic Change: Internal Factors*, vol. 1, Blackwell, Malden, MA, 1994.
[23] ———, *Principles of Linguistic Change: Social Factors*, vol. 2, Blackwell, Malden, MA, 2001.
[24] ———, *Transmission and diffusion*, Language, 83 (2007), pp. 344–387.
[25] David Lightfoot, *The Development of Language: Acquisition, Changes and Evolution*, Blackwell, Malden, MA, 1999.
[26] Christopher D. Manning, *Probabilistic syntax*, in Bod et al. [3], ch. 8.
[27] Robert H. Martin, Jr., *Nonlinear Operators and Differential Equations in Banach Spaces*, John Wiley & Sons, New York, 1976.
[28] W. Garrett Mitchener, *Bifurcation analysis of the fully symmetric language dynamical equation*, Journal of Mathematical Biology, 46 (2003), pp. 265–285.
[29] ———, *Simulating language change in the presence of non-idealized syntax*, in Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition, New Brunswick, NJ, 2005, Association for Computational Linguistics, pp. 10–19.
[30] ———, *A mathematical model of the loss of verb-second in Middle English*, in Ritt et al. [43]. Proceedings of the 13th International Conference on English Historical Linguistics.
[31] ———, *Game dynamics with learning and evolution of universal grammar*, Bulletin of Mathematical Biology, 69 (2007), pp. 1093–1118.
[32] ———, *A stochastic model of language change through social structure and prediction-driven instability*. Submitted, 2009.
[33] W. Garrett Mitchener and Martin A. Nowak, *Competitive exclusion and coexistence of universal grammars*, Bulletin of Mathematical Biology, 65 (2003), pp. 67–93.
[34] ———, *Chaos and language*, Proceedings of the Royal Society of London, Biological Sciences, 271 (2004), pp. 701–704.

[35] Partha Niyogi, *The Informational Complexity of Learning*, Kluwer Academic Publishers, Boston, 1998.
[36] ———, *The Computational Nature of Language Learning and Evolution*, MIT Press, Cambridge, MA, 2006.
[37] Martin A. Nowak, Natalia L. Komarova, and Partha Niyogi, *Evolution of universal grammar*, Science, 291 (2001), pp. 114–118.
[38] Martin A. Nowak, D. C. Krakauer, and A. Dress, *An error limit for the evolution of language*, Proceedings of the Royal Society of London, Series B, 266 (1999), pp. 2131–2136.
[39] Martin A. Nowak, Joshua Plotkin, and David C. Krakauer, *The evolutionary language game*, Journal of Theoretical Biology, 200 (1999), pp. 147–162.
[40] E. Pap, ed., *Handbook of Measure Theory*, vol. I, Elsevier, Amsterdam, 2002.
[41] Joshua Plotkin and Martin A. Nowak, *Language evolution and information theory*, Journal of Theoretical Biology, 205 (2000), pp. 147–159.
[42] Andrew Radford, *Minimalist Syntax: Exploring the structure of English*, Cambridge University Press, Cambridge, UK, 2004.
[43] Nikolaus Ritt, Herbert Schendl, Christiane Dalton-Puffer, and Dieter Kastovsky, eds., *Medieval English and its Heritage: Structure, meaning and mechanisms of change*, vol. 16 of Studies in English Medieval Language and Literature, Peter Lang, Frankfurt, 2006. Proceedings of the 13th International Conference on English Historical Linguistics.
[44] Bruce Tesar and Paul Smolensky, *Learnability in Optimality Theory*, MIT Press, Cambridge, MA, 2000.
[45] Linda Uyechi and Lian Hee Wee, eds., *Reality Exploration and Discovery: Pattern Interaction in Language and Life*, CSLI Publiscations, Stanford, 2007.
[46] Anthony Warner, *Why DO dove: Evidence for register variation in Early Modern English negatives*, Language Variation and Change, 17 (2005), pp. 257–280.
[47] Lydia White, *Universal Grammar and Second Language Acquisition*, vol. 1 of Language Acquisition & Language Disorders, John Benjamins Publishing Company, Philadelphia, 1989.
[48] Charles D. Yang, *Knowledge and Learning in Natural Language*, Oxford University Press, Oxford, 2002.