Models of Language Evolution: Part IV
Linguistic Coherence as Emergent Property

Matilde Marcolli

CS101: Mathematical and Computational Linguistics

Winter 2015

Main Reference

- Partha Niyogi, *The computational nature of language learning and evolution*, MIT Press, 2006.

Population Dynamics Model

• following previous model: languages $\mu_1, \dots, \mu_n$ and linguistic evolution in population modelled by ODE

$$\dot{x}_j = \sum_i x_i \, f_i \, Q_{ij} - \phi \, x_j$$

$x_j = \alpha_j$ proportion of individuals speaking language μ_j

• matrix Q measure fidelity of language map (how much deviation from teacher to learner)

• $f_i =$ fitness

$$f_i = \sum_j x_j \, F(\mu_i, \mu_j)$$

Assumptions

- assuming as before that
  - $Q_{ii} = q$ and $Q_{ij} = \frac{1-q}{n-1}$ for $i \neq j$
  - $F(\mu_i, \mu_i) = 1$ and $F(\mu_i, \mu_j) = a$ for all $i \neq j$
  - $f_i = (1 - a)x_i + a + f_0$

Threshold behavior depending on parameter q

- for q small only stable critical point is uniform distribution: all $x_j = 1/n$
- *bifurcation* at some $q = q_1$: two new critical points $r_\pm$

- one-grammar solutions emerge where the majority of population speaks one of the languages

Without fitness

- Note: same equation with $f_i = f_0$ (without fitness function)
- would have $\phi = f_0 \sum_j x_j = f_0$
- equation would be

$$\dot{x}_j = f_0 \sum_i x_i Q_{ij} - f_0 x_j$$

becomes a linear system of ODE

- only equilibrium solution at $x_j = 1/n$, uniform distribution
- no bifurcation and no emergent behavior creating language coherence: those are effects of the presence of the fitness function

Social Learning

• this model was based on assumption that learner takes input only from one teacher (with the possibility of errors in reproduction encoded in Q_{ij})

• consider again other scenario where learner's input is coming from the entire population

• given n languages $\mathcal{L}_1, \ldots, \mathcal{L}_n$ assume a set of expressions is especially useful for language acquisition (triggers, cues, ...)

• this gives subsets $C_i \subseteq \mathcal{L}_i$; assume $C_i \cap C_j = \emptyset$ (these are unambiguous cues)

• speakers of $\mathcal{L}_i$ produce sentences randomly with distribution $\mathbb{P}_i$ and likelihood of producing a cue is

$$a_i = \mathbb{P}_i(C_i)$$

• simplifying assumption: all $a_i = a$ same

Case of two languages

- proportions $\alpha, 1-\alpha$ of speakers: function of time $x_1(t) = \alpha(t)$, $x_2(t) = 1 - \alpha(t)$
- cue-frequency based batch learner: $m = k_1 + k_2 + k_3$
  - k_1 sentences in input that are in C_1
  - k_2 in C_2
  - k_3 are not cues
- probability of $k_1 > k_2$

$$f_{1,a,m}(x_1, x_2) = \sum \binom{m}{k_1 k_2 k_3} (a x_1(t))^{k_1} (a x_2(t))^{k_2} (1-a)^{k_3}$$

sum over (k_1, k_2, k_3) with $m = k_1 + k_2 + k_3$ and $k_1 > k_2$

- probability $f_{2,a,m}$ of $k_1 < k_2$, same with sum over (k_1, k_2, k_3) with $m = k_1 + k_2 + k_3$ and $k_1 > k_2$

- symmetric assumption $a_i = a$ gives $f_{2,a,m}(x_1, x_2) = f_{1,a,m}(x_2, x_1)$
- probability after m inputs of learner acquiring $\mathcal{L}_1$

$$f_1 + \frac{1}{2}(1 - f_2 - f_1)$$

(if no cues received at all: $1/2$ chance of one language or other)

- population dynamics equation

$$x_1(t+1) = \frac{1}{2}(1 + f_{1,a,m}(x_1(t), x_2(t)) - f_{2,a,m}(x_1(t), x_2(t))$$

- a fixed point at $x_1 = x_2 = 1/2$: uniform distribution of population among the two languages

- if number of inputs m small: only fixed point (stable)
- for larger m other fixed points appear (one language becomes dominant)
- for larger m uniform solution $x_1 = x_2 = 1/2$ becomes unstable
- the value of m where bifurcation occurs is a function of parameter a
- can also keep m fixed and vary a:
  - a close to zero: only $x_1 = x_2 = 1/2$ (stable fixed point)
  - bifurcation when a grows: new stable fixed points and $x_1 = x_2 = 1/2$ becomes unstable
  - bifurcation occurs at a value of a dependent on m

Stability of $x_1 = x_2 = 1/2$: more details

- derivative at the fixed point

$$f'_{1,a,m}(1/2,1/2) = \sum_{k_1>k_2} \binom{m}{k_1\, k_2\, k_3} a^{m-k_3}(1-a)^{k_3}(k_1-k_2)(\frac{1}{2})^{k_1+k_2-1}$$

similar for $f'_{2,a,m}$

- $f'_{1,a,m}(1/2,1/2)|_{a=0} = 0$ so by continuity for small a have

$$|f'_{1,a,m}(1/2,1/2)| < 1$$

stability while in this range

- also see that when $a = 1$, for sufficiently large m have $f'_{1,a,m}(1/2,1/2)|_{a=1} > 1$ so in between will cross value 1: where bifurcation occurs

- emergence of linguistic coherence in the population

Case of n languages

• learner is exposed to a mixture of languages form the environment

• learner scans incoming data for cues and chooses the language from which largest number of cues is received

• if multiple languages with same number of cues: pick one among them randomly

• same simplifying assumption as before $\mathbb{P}_i(C_i) = a$ same for all languages

Algorithm

1. Count cues
   - k_i = number of cues in C_i out of m inputs
   - k_{n+1} = number of non-cues (in any of the languages)
   - $m = k_1 + \cdots + k_n + k_{n+1}$
2. Find maximal languages: languages $\mathcal{L}_i$ with $k_i = \max_j k_j$: $\mathcal{I}$ = set of indices of $\mathcal{L}_i$ maximal
3. Choose language: if $|\mathcal{I}| = 1$ choose that language; if $|\mathcal{I}| > 1$ choose one language randomly in the set $\mathcal{I}$ with probability $1/|\mathcal{I}|$
4. of naive version: just choose a language randomly among all n with probability $1/n$

Population Dynamics in this model

- $\mathbb{P} = \sum_i x_i(t)\mathbb{P}_i$ probability with which input is generated
- $p_i = p_i(t) = a x_i(t)$ probability of receiving a cue from language $\mathcal{L}_i$; $p_{n+1} = 1 - a$
- probability of receiving (strictly) more cues from language $\mathcal{L}_1$ than from any other

$$F_{1,m,a}(x_1, \ldots, x_n) = \sum \binom{m}{k_1 \cdots k_{n+1}} p_1^{k_1} \cdots p_n^{k_n} p_{n+1}^{k_{n+1}}$$

sum over all $(k_1, \ldots, k_{n+1})$ with $m = k_1 + \cdots + k_{n+1}$ and $k_1 > k_j$ for all $j \neq 1$

- similar for other languages with symmetry

$$F_{i,m,a}(\cdots, x_i, \cdots, x_j \cdots) = F_{j,m,a}(\cdots, x_j, \cdots, x_i \cdots)$$

- in this model, probability that learner will choose $\mathcal{L}_i$ after m input data

$$f_{i,m,a}(x_1, \dots, x_n) = F_{i,m,a}(x_1, \dots, x_n) + (1 - \sum_{j=1}^{n} F_{j,m,a}(x_1, \dots, x_n))\frac{1}{n}$$

(with naive version of choice in the cue-less case)

- Recursion relation for population distribution in next generation

$$x_i(t+1) = f_{i,m,a}(x_1(t), \dots, x_n(t))$$

Fixed Points

- $f = (f_{i,m,a})_{i=1}^{n}$ continuous map $f : \Delta_{n-1} \to \Delta_{n-1}$
- Results
  1. f has finite number of fixed points: at most $m2^n$
  2. for small m only fixed point is $(\frac{1}{n}, \ldots, \frac{1}{n})$, stable
  3. for fixed (sufficiently large) m number of fixed points varies with a: small a only one fixed point (uniform distribution); as a increases bifurcation: other fixed points arise
  4. large values of $a \sim 1$: uniform distribution no longer stable, only the fixed points with one dominant language are

Language Learning and Statistical Physics

• these bifurcations and emergence of linguistic coherence reminiscent of behavior of Ising model and spin glass systems in Statistical Physics

• an ensemble of interacting components

• degree of interaction governed by a thermodynamic parameter $\beta \sim 1/T$ inverse temperature

• these systems often exhibit *phase transitions* between different regimes, at some critical temperature $T = T_c$ (different states of matter, loss of magnetization, etc.)

Language Evolution in Locally Connected Societies

- two possible languages: $\{\mathcal{L}_0, \mathcal{L}_1\} = \{0, 1\}$
- Graph G representing linguistic agents and their interaction
  - each vertex $v \in V(G)$ has an associated random variable $X_v(t)$
  - $X_v(t) \in \{0, 1\}$: language of agent occupying position v
  - $X_v(t+1)$ language occupying same position at next step (generation)
  - $\mathbb{P}(X_v(t+1) = 1) = g_{a,m}(\mu_v(t))$

$$\mu_v = \frac{1}{\mathrm{val}(v)} \left(\sum_{e \in E(G): \partial(e) = \{v, v'\}} X_{v'}(t) \right)$$

- nearest neighbor interaction considered only

- as before assuming $a = \mathbb{P}_i(C_i)$ same for both languages
- a possible choice for the function $g_{a,m} : [0,1] \to [0,1]$:

$$g(x) = \frac{1}{2} + \frac{1}{2}(f_{1,a,m}(x, 1-x) - f_{1,a,m}(1-x, x))$$

with $f_{1,a,m}$ as before counting probability of set of cues $k_1 > k_2$

$$f_{1,a,m}(x, 1-x) = \sum \binom{m}{k_1 k_2 k_3} (ax)^{k_1} (a(1-x))^{k_2} (1-a)^{k_3}$$

sum over (k_1, k_2, k_3) with $m = k_1 + k_2 + k_3$ and $k_1 > k_2$

• study evolution of

$$\alpha_G(t) = \frac{1}{\#V(G)} \sum_{v \in V(G)} X_v(t)$$

average number of $\mathcal{L}_1$-speakers at time/generation t

• for a complete graph have all language users connected to all others: recover model in which learning from whole community

• can consider asymptotic behaviors when size of graph becomes large $\#V(G) = N \to \infty$

• can simplify the geometry making special assumptions on the graph: e.g. a square lattice

The Ising Model of spin systems on a graph G

- configurations of spins $s : V(G) \to \{\pm 1\}$
- magnetic field B and correlation strength J: Hamiltonian

$$H(s) = -J \sum_{e \in E(G): \partial(e) = \{v, v'\}} s_v s_{v'} - B \sum_{v \in V(G)} s_v$$

- first term measures degree of alignment of nearby spins
- second term measures alignment of spins with direction of magnetic field

Equilibrium Probability Distribution

- Partition Function $Z_G(\beta)$

$$Z_G(\beta) = \sum_{s:V(G)\to\{\pm 1\}} \exp(-\beta H(s))$$

- Probability distribution on the configuration space: Gibbs measure

$$\mathbb{P}_{G,\beta}(s) = \frac{e^{-\beta H(s)}}{Z_G(\beta)}$$

- low energy states weight most
- at low temperature (large β): ground state dominates; at higher temperature (β small) higher energy states also contribute

Average Spin Magnetization

$$M_G(\beta) = \frac{1}{\#V(G)} \sum_{s:V(G)\to\{\pm 1\}} \sum_{v\in V(G)} s_v \, \mathbb{P}(s)$$

- Free energy $F_G(\beta, B) = \log Z_G(\beta, B)$

$$M_G(\beta) = \frac{1}{\#V(G)} \frac{1}{\beta} \left(\frac{\partial F_G(\beta, B)}{\partial B} \right) |_{B=0}$$

- *thermodynamic limit*: $\#V(G) = N \to \infty$

$$m(\beta) = \lim_{\#V(G)\to\infty} M_G(\beta)$$

- in these thermodynamic limits need to fix a way in which the geometry of the graph grows: if it is a lattice, just grow size N of lattice; for other kinds of graphs, fix how smaller graphs embedded in larger graphs

Ising Model on a 2-dimensional lattice

- $\exists$ critical temperature $T = T_c$ where phase transition occurs
- for $T > T_c$ equilibrium state has $m(T) = 0$ (computed with respect to the equilibrium Gibbs measure $\mathbb{P}_{G,\beta}$
- demagnetization: on average as many up as down spins
- for $T < T_c$ have $m(T) > 0$: spontaneous magnetization
- Warning: beware of thermodynamic limits!
- a lot of technical problems in these spin glass models go into how one takes these limits where the size N of the graph $N \to \infty$ (even for simple geometries like lattice case)

A Spin Glass model of Language Learning

• a multilingual society = a graph G

• linguistic agents = vertices of the graph $V(G)$

• which agents interact with which others = edges $E(G)$

• possible languages = spin states
- Ising models $\{\pm 1\}$: two languages model
- Potts models $\{1, \ldots, q\}$: many languages model

• distribution of population across different languages = average magnetization

• previous analysis for "input from whole society" = mean field theory for case of Ising model on complete graph

Syntactic Parameters and Ising/Potts Models

• a different view on how to use spin glass models for language evolution

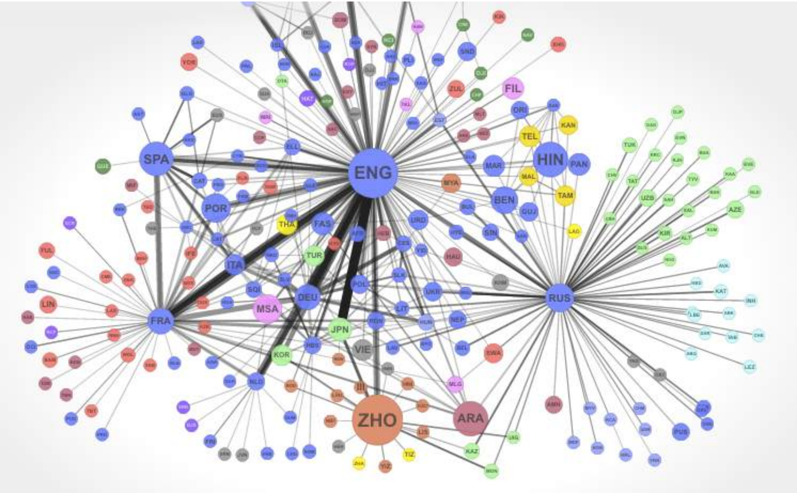
• characterize set of $n = 2^N$ languages $\mathcal{L}_i$ by binary strings of N syntactic parameters (Ising model)

• or by ternary strings (Potts model) if take values ± 1 for parameters that are set and 0 for parameters that are not defined in a certain language

• a system of n interacting languages = graph G with $n = \#V(G)$

• languages $\mathcal{L}_i$ = vertices of the graph (though of as, for instance, the language that occupies a certain geographic area)

• languages that have interaction with each other = edges $E(G)$ (geographical proximity, or high volume of exchange for other reasons)



graph of language interaction (detail) from Global Language Network of MIT Medialab, with interaction strengths J_e on edges based on number of book translations

• if only one syntactic parameter, would have an Ising model on the graph G: configurations $s : V(G) \to \{\pm 1\}$ set the parameter at all the locations on the graph

• variable interaction energies along edges (some pairs of languages interact more than others) • magnetic field B and correlation strength J: Hamiltonian

$$H(s) = -\sum_{e\in E(G):\partial(e)=\{v,v'\}} \sum_{i=1}^{N} J_e\, s_{v,i}\, s_{v',i}$$

• if N parameters, configurations

$$\underline{s} = (s_1, \ldots, s_N) : V(G) \to \{\pm 1\}^N$$

• if all N parameters are independent, then it would be like having N non-interacting copies of a Ising model on the same graph G (or N independent choices of an initial state in an Ising model on G)

- an interesting problem in this model is the entailment of parameters: it is known that flipping certain syntactic parameters causes others to flip as well
- so in addition to the edge interaction, instead of alignment with external magnetic field term in H

$$-B \sum_{v \in V(G)} s_v$$

should have a term that favors alignment of entailed parameters

- set of parameters $\mathcal{P}$ with $N = \#\mathcal{P}$; subset $\mathcal{E} \subset \mathcal{P} \times \mathcal{P}$ of entailments: pairs of parameters (Π, Π') such that flippIng Π causes Π' to flip as well
- term in Hamiltonian favoring alignments of entailed parameters

$$-B \sum_{v \in V(G)} \sum_{(\Pi, \Pi') \in \mathcal{E}} s_{v,\Pi} \, s_{v,\Pi'}$$