# Models of Language Evolution

Matilde Marcolli
CS101: Mathematical and Computational Linguistics

Winter 2015

## Main Reference

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

## From Language Acquisition to Language Evolution

- models of language acquisition behind transmission mechanism (how language gets transmitted to next generation of learners)
- perfect language acquisition implies perfect language transmission... but for evolution need *imperfect* transmission
- phonological, morphological, syntactic and semantic changes are observed
- points of view imported from evolutionary biology: population dynamics, genetic drift, dynamical systems models
- language learning at *individual level* versus *population level*

• Learning algorithm: computable function $\mathcal{A} : \mathcal{D} \to \mathcal{H}$ from primary linguistic data to a space of grammars

1. Grammatical Theory: determines $\mathcal{H}$
2. Acquisition Model: determines $\mathcal{A}$

Possibilities for change in the language transmission

1. The data $\mathcal{D}$ are changed
2. The data $\mathcal{D}$ are insufficient

• First case: presence of mixed population of speakers of different languages (not all data $\mathcal{D}$ consistent with same language)

• Second case: after finite $\tau_m$ algorithm gives hypothesis $\mathfrak{h}_m = \mathcal{A}(\tau_m)$ at some distance from target

Toy Model: two competing languages

- $\mathcal{L}_1, \mathcal{L}_2 \subset \mathfrak{A}^\star$ with $\mathcal{L}_1 \cap \mathcal{L}_2 \neq \emptyset$

- sentences in $\mathcal{L}_1 \cap \mathcal{L}_2$ are ambiguous and can be parsed by both $\mathcal{G}_1$ and $\mathcal{G}_2$ grammars

- assume each individual in the population is *monolingual*

- $\alpha_t =$ percentage of population at time $t$ (or number of generations) that speaks $\mathcal{L}_1$ and $(1 - \alpha_t) =$ percentage that speaks $\mathcal{L}_2$

- $\mathbb{P}_i =$ probability distribution for sentences of $\mathcal{L}_i$

- learning algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{H} = \{\mathcal{G}_1, \mathcal{G}_2\}$ computable

- data are drawn according to a probability distribution $\mathbb{P}$ on $\mathfrak{A}^\star$

- probability that learning algorithm will guess $\mathcal{L}_1$ after $m$ inputs

$$p_m = \mathbb{P}(\mathcal{A}(\tau_m) = \mathcal{L}_1)$$

$p_m = p_m(\mathcal{A}, \mathbb{P})$ depends on learning algorithm and distribution $\mathbb{P}$

- if $\mathbb{P}$ on $\mathfrak{A}^\star$ is supported on $\mathcal{L}_1$ (so $\mathbb{P} = \mathbb{P}_1$)

$$\lim_{m\to\infty} p_m(\mathcal{A}, \mathbb{P} = \mathbb{P}_1) = 1$$

in this case $\mathcal{G}_1$ is the target grammar the learning algorithm converges to

Population Dynamics of the two languages model

• assume size $K$ of data $\tau_K$ after which linguistic hypothesis stabilizes (locking set)

• with probability $p_K(\mathcal{A}, \mathbb{P}_1)$ the language acquired will be $\mathcal{L}_1$

• with probability $1 - p_K(\mathcal{A}, \mathbb{P}_1)$ it will be $\mathcal{L}_2$

• so the new generation will have fraction $p_K(\mathcal{A}, \mathbb{P}_1)$ of speakers of language $\mathcal{L}_1$ and fraction $1 - p_K(\mathcal{A}, \mathbb{P}_1)$ of speakers of $\mathcal{L}_2$

• assume the proportion for a given generation are $\alpha$ and $1 - \alpha$

• the following generation of learners will then receive examples generated with probability distribution

$$\mathbb{P} = \alpha \mathbb{P}_1 + (1 - \alpha)\mathbb{P}_2$$

• the following generation will then result in a population of speakers with distribution $\lambda$ and $1 - \lambda$ where

$$\lambda = p_K(\mathcal{A}, \alpha \mathbb{P}_1 + (1 - \alpha)\mathbb{P}_2)$$

• this gives the recursive dependence $\lambda = \lambda(\alpha)$ in language transmission to the following generation

Assumptions made in this model

- new learners (new generation) receive input from entire community of speakers (previous generation) in proportion to the language distribution across the population

- the probabilities $\mathbb{P}_1, \mathbb{P}_2$ of drawing sentences in $\mathcal{L}_1, \mathcal{L}_2$ do not change in time

- learning algorithm constructs a single hypothesis after each input

- populations can have unlimited growth

Memoryless Learner with two languages model

• Initialize: randomly choose initial hypothesis $\mathcal{G}_1$ or $\mathcal{G}_2$

• Receive Input $s_i$: if current hypothesis parses $s_i$ get new input, if not next step

• Single-Step Hill Climbing: switch to other hypothesis (in space of two languages) and receive new input

• how does population evolve if this Trigger Learning Algorithm is used?

- set values (given $\mathbb{P}_1$ and $\mathbb{P}_2$):

$$a = \mathbb{P}_1(\mathcal{L}_1 \cap \mathcal{L}_2), \quad 1 - a = \mathbb{P}_1(\mathcal{L}_1 \smallsetminus \mathcal{L}_2)$$

$$b = \mathbb{P}_2(\mathcal{L}_1 \cap \mathcal{L}_2), \quad 1 - b = \mathbb{P}_2(\mathcal{L}_2 \smallsetminus \mathcal{L}_1)$$

- $a$ and $b$ are the probabilities of users of languages $\mathcal{L}_1$ and $\mathcal{L}_2$ of generating ambiguous sentences
- assume a very short "maturation time": $K = 2$
- Result: the ratio $\alpha_{t+1}$ of the $t+1$-st generation satisfies

$$\alpha_{t+1} = A\,\alpha_t^2 + B\,\alpha_t + C$$

$$A = \frac{1}{2}((1-b)^2 - (1-a)^2), \quad B = b(1-b) + (1-a), \quad C = \frac{b^2}{2}$$

Explanation:

- start with $\alpha_t$ proportion of $\mathcal{L}_1$-users
- compute probability of learner acquiring $\mathcal{L}_1$ in two steps ($K = 2$)
- probabilities for a random example:
  - in $\mathcal{L}_1 \setminus \mathcal{L}_2$ with probability $\alpha_t(1 - a)$
  - in $\mathcal{L}_1 \cap \mathcal{L}_2$ with probability $\alpha_t a + (1 - \alpha_t)b$
  - in $\mathcal{L}_2 \setminus \mathcal{L}_1$ with probability $(1 - \alpha_t)(1 - b)$
- also probability $1/2$ of choosing $\mathcal{L}_1$ as initial hypothesis

- if started with $\mathcal{L}_1$, to have $\mathcal{L}_1$ after two steps:
  - either $\mathcal{L}_1$ retained in both steps
  - or switch from $\mathcal{L}_1$ to $\mathcal{L}_2$ at next step and back from $\mathcal{L}_2$ to $\mathcal{L}_1$ at second step
- first case happens with probability $\alpha_t + (1 - \alpha_t)b$
- second case happens with probability $\alpha_t(1 - a)(1 - \alpha_t)(1 - b)$

- if started with $\mathcal{L}_2$, to have $\mathcal{L}_1$ in two steps:
  - either switch to $\mathcal{L}_1$ at first step and retain $\mathcal{L}_1$ at second
  - or retain $\mathcal{L}_2$ at first and switch to $\mathcal{L}_1$ at second
- the first case happens with probability $\alpha_t(1-a)(\alpha_t+(1-\alpha_t)b)$
- the second case happens with probability $((1-\alpha_t)+\alpha_t a)\alpha_t(1-a)$
- putting all these possibilities together gives the right counting

## Long term behavior

- if $a = b$ simple exponential growth
- for $a \neq b$ behavior similar to *logistic map*: in particular it has a regime with *chaotic behavior*
- the chaotic regime is avoided because of the constraints $a, b \leq 1$
- the fact that the recursion is a *quadratic* function reflects the choice $K = 2$
- for higher values of $K$ would get higher order polynomials

Result: for an arbitrary $K$

$$\alpha_{t+1} = \frac{B + \frac{1}{2}(A-B)(1-A-B)^K}{A+B}$$

$$A = (1-\alpha_t)(1-b), \qquad B = \alpha_t(1-a)$$

Explanation: Markov Chain with two states describing the TLA

- Transition matrix $T$

$$T_{12} = (1-\alpha_t)(1-b) = A, \qquad T_{21} = \alpha_t(1-a) = B$$

and $T_{11} = 1 - T_{12}$ and $T_{22} = 1 - T_{21}$

- after $m$ examples moved by transition matrix $T^m$

- probability of acquiring language $\mathcal{L}_1$ after $m$ examples is

$$\frac{1}{2}(T_{11}^m + T_{21}^m)$$

- recursively have $T^m = TT^{m-1}$

$$T_{11}^m = (1-A)T_{11}^{m-1} + BT_{12}^{m-1}$$

$$T_{11}^m = \frac{B}{A+B} + \frac{A(1-A-B)^m}{A+B}$$

- similarly obtain inductively

$$T_{21}^m = \frac{B}{A+B} + \frac{B(1-A-B)^m}{A+B}$$

- putting these together gives succession rule at $m = K$

## Population behavior in the model

- the function $f(\alpha) = f_{a,b,K}(\alpha)$

$$f_{a,b,K}(\alpha) = \frac{B(\alpha) + \frac{1}{2}(A(\alpha) - B(\alpha))(1 - A(\alpha) - B(\alpha))^K}{A(\alpha) + B(\alpha)}$$

$$A(\alpha) = (1-\alpha)(1-b), \qquad B(\alpha) = \alpha(1-a)$$

- only one stable fixed point in $\alpha \in [0,1]$ interval
- $f(0) = b^K/2$ and $f(1) = a^K/2$, $f$ continuous, find only one $\alpha = f(\alpha)$ and can check at that point $|f'(\alpha)| < 1$
- if $a = b = 1/2$ fixed point is at $\alpha = 1/2$ (population converges to this mix from all initial conditions)

$$f_{\frac{1}{2},\frac{1}{2},K}(\alpha) = \alpha(1 - b^K) + \frac{b^K}{2}$$

• if $a \neq b$ with $a > b$: fixed point close to $\alpha = 0$: most population speaks $\mathcal{L}_2$

• if $a \neq b$ with $a < b$: fixed point close to $\alpha = 1$: most population speaks $\mathcal{L}_1$

• transition of the fixed point from a value close to zero to a value close to one very sharp for small values of $a, b$, more gradual for larger values of $a, b$ (close to one)

## Limiting behavior when $K \to \infty$

- limiting function and recursion

$$f_{a,b,\infty}(\alpha) = \frac{\alpha(1-a)}{\alpha(1-a) + (1-\alpha)(1-b)}$$

$$f'_{a,b,\infty}(\alpha) = \frac{(1-a)(1-b)}{((1-b) + \alpha(b-a))^2}$$

$$\alpha_{t+1} = f_{a,b,\infty}(\alpha_t)$$

- if $a = b$ just have $\alpha_{t+1} = \alpha_t$ population preserved, no change
- fixed point behavior: if $a > b$ two fixed points $\alpha = f_{a,b,\infty}(\alpha)$ at $\alpha = 0$ (unstable) and $\alpha = 1$ (stable)
- if $a < b$ same two fixed points but with switched stability

## Batch Error-Based Learner in the two languages model

- still memoryless learner, but replace trigger learning algorithm (TLA) with batch error-based

- learner waits until all set of $K$ samples collected before choosing a hypothesis, then pick the one that best fits the entire set $\tau_K = (s_1, \ldots, s_K)$

- for each $\mathcal{L}_i$ error-measure

$$e(\mathcal{L}_i) = \frac{k_i}{K}$$

with $k_i$ = number of sentences in $\tau_K$ that cannot be parsed by $\mathcal{L}_i$

- then hypothesis is chosen as

$$\mathcal{A}(\tau_K) = \arg\min_i e(\mathcal{L}_i)$$

## Procedure

- Group together sentences in $\tau_K = (s_1, \ldots, s_K)$ into
  1. $n_1$ sentences in $\mathcal{L}_1 \setminus \mathcal{L}_2$
  2. $n_2$ sentences in $\mathcal{L}_1 \cap \mathcal{L}_2$
  3. $n_3$ sentences in $\mathcal{L}_2 \setminus \mathcal{L}_1$

with $n_1 + n_2 + n_3 = K$

- Choose $\mathcal{L}_1$ if $n_1 > n_3$; choose $\mathcal{L}_2$ if $n_3 > n_1$
- if $n_1 = n_3$ deterministic or randomized way of choosing either $\mathcal{L}_i$
- Example: choose $\mathcal{L}_1$ if $n_1 \geq n_3$

Result: population dynamics $\alpha_{t+1} = f_{a,b,K}(\alpha_t)$ with

$$f_{a,b,K}(\alpha) = \sum \binom{K}{n_1\, n_2\, n_3} p_1(\alpha)^{n_1} p_2(\alpha)^{n_2} p_3(\alpha)^{n_3}$$

with sum over $(n_1, n_2, n_3) \in \mathbb{Z}_+^3$ with $n_1 + n_2 + n_3 = K$ and $n_1 \geq n_3$

$$p_1(\alpha) = \alpha(1-a), \quad p_2(\alpha) = \alpha a + (1-\alpha) b, \quad p_3(\alpha) = (1-\alpha)(1-b)$$

Properties of Dynamics

- $b = 1 \;\Rightarrow\; p_3(\alpha) = 0$; $a = 1 \;\Rightarrow\; p_1(\alpha) = 0$
- have $1 - a = \mathbb{P}_1(\mathcal{L}_1 \smallsetminus \mathcal{L}_2)$ and $1 - b = \mathbb{P}_2(\mathcal{L}_2 \smallsetminus \mathcal{L}_1)$
- so $b = 1$ implies $n_3 = 0$ and $a = 1$ implies $n_1 = 0$
- so for $b = 1$ always $n_1 \geq n_3$ so always $\mathcal{L}_1$

- for $a = 1$ have $n_1 \geq n_3$ only when $n_3 = 0$ so get

$$f_{1,b,K}(\alpha) = (1 - (1 - \alpha)(1 - b))^K$$

- then $\alpha = 0$ not a fixed point but $\alpha = 1$ is fixed
- stability of $\alpha = 1$ fixed point depends on $K$ and $b$: stability iff

$$b > 1 - \frac{1}{K}$$

- when passes to unstable, *bifurcation* occurs and new (stable) fixed point appears in interior of interval $(0, 1)$
- when $a \neq 1$ and $b \neq 1$: for most values $\alpha = 1$ stable fixed point, and two fixed points $\alpha_1 < \alpha_2$ in $(0, 1)$, first stable, second unstable

## Asymptotic Behavior when $K \to \infty$

- assume $K \to \infty$ with $\frac{n_1}{K} \to p_1$ and $\frac{n_3}{K} \to p_3$
- then if $p_1 > p_3$ have $\alpha(1-a) > (1-\alpha)(1-b)$ and $\alpha_t \to 1$
- when $K = \infty$ have $\alpha = 0$ and $\alpha = 1$ stable fixed points and unstable

$$\alpha = \frac{1-b}{(1-b)+(1-a)}$$

- Note: asymmetry of behavior when $n_1 = n_3$ (choosing $\mathcal{L}_1$) becomes less and less noticeable in the large $K$ limit

## Cue-Based Learner in the two languages model

- learner examines data for indications of how to set linguistic parameters
- a set $C \subset \mathcal{L}_1 \smallsetminus \mathcal{L}_2$ of examples that are cues to target being $\mathcal{L}_1$
- if elements from $C$ occur sufficiently frequently in $\mathcal{D}$ learning algorithm chooses $\mathcal{L}_1$, if not it chooses $\mathcal{L}_2$
- learner receives $K$ samples input $\tau_K = (s_1, \dots, s_K)$
- $k/K =$ fraction of the input that is in the cue set
- probability that a user of language $\mathcal{L}_1$ produces a cue: $p = \mathbb{P}_1(C)$
- probability that learner receives a cue as input $= \alpha_t\, p$
- threshold $\mathfrak{t}$ with $k/K > \mathfrak{t}$: achieved with probability

$$\sum \binom{K}{i} (p\alpha_t)^i (1 - p\alpha_t)^{K-i}$$

where sum is over $i$ in the range $K\mathfrak{t} \leq i \leq K$

## Population Dynamics with cue-based learner

- recursion relation for the fractions of population speaking the two languages:

$$\alpha_{t+1} = f_{p,K}(\alpha_t) = \sum_{K\mathrm{t} \leq i \leq K} \binom{K}{i} (p\alpha_t)^i (1 - p\alpha_t)^{K-i}$$

- when $p = 0$ cues never produced, only stable equilibrium is $\alpha = 0$ (reached in one step)
- for $p$ small, $\alpha = 0$ stays unique stable fixed point
- as $p$ increases *bifurcation* occurs:
  - two new fixed points arise $\alpha_1 < \alpha_2$
  - $\alpha = 0$ remains stable; $\alpha_1$ is unstable; $\alpha_2$ is stable
- at $p = 1$ stable fixed points $\alpha = 0$ and $\alpha = 1$ and one unstable fixed point in between

## Fixed Point Analysis (more details)

- fixed points $\alpha = f(\alpha)$ of function

$$f_{p,K}(\alpha) = \sum_{Kt \leq i \leq K} \binom{K}{i} (p\alpha)^i (1 - p\alpha)^{K-i}$$

- for all $p$ and $K$ have $f_{p,K}(0) = 0$, for stability check $|f'(0)| < 1$:

$$f'_{p,K}(\alpha) = pF'(p\alpha) \quad \text{with} \quad f_{p,K}(\alpha) = F(p\alpha)$$

$$F(\alpha) = \sum_{Kt \leq i \leq K} \binom{K}{i} \alpha^i (1 - \alpha)^{K-i}$$

- differentiate term by term gives $F'(\alpha)$:

$$\sum_{K_\mathfrak{t} \leq k \leq K-1} \binom{K}{k} \left(k\alpha^{k-1}(1-\alpha)^{K-k} - (K-k)\alpha^k(1-\alpha)^{K-k-1}\right) + K\alpha^{K-1}$$

with $K_\mathfrak{t}$ smallest integer larger than $K\mathfrak{t}$

- Expanding and grouping terms

$$F'(\alpha) = K\left(\sum_{K_\mathfrak{t} \leq k \leq K-1} \frac{(K-1)!}{(K-k)!(k-1)!}\alpha^{k-1}(1-\alpha)^{K-k}\right)$$

$$-K\left(\sum_{K_\mathfrak{t} \leq k \leq K-1} \frac{(K-1)!}{k!(K-k-1)!}\alpha^{k}(1-\alpha)^{K-k-1} - \alpha^{K-1}\right)$$

- cancellations leave

$$F'(\alpha) = K \binom{K-1}{K_t - 1} \alpha^{K_t - 1} (1-\alpha)^{K-K_t}$$

- $f'_{p,K}(0) = pF'(0) = 0$ hence stability of $\alpha = 0$
- $f_{p,K}(1) = F(p) < 1$ (for $p < 1$); since $f_{p,K}(0) = 0$ with $f'_{p,K}(0) = 0$ and continuous: even number of crossings of graph of $f_{p,K}$ and diagonal in $(0, 1]$
- if $2m$ such points $\alpha_1, \dots, \alpha_{2m}$ with slope $f'_{p,K}(\alpha_j)$ alternating larger and smaller than 1 (slope of diagonal)
- is each successive interval $(\alpha_{2j-1}, \alpha_{2j+1})$ derivative $f'_{p,K}$ changes from larger to smaller to larger than 1, so $f'_{p,K}(\alpha) - 1$ changes sign twice, so derivative $f''_{p,K}$ has zero, same for every interval $(\alpha_{2j-2}, \alpha_{2j})$

- second derivative $f''_{p,K}(\alpha) = p^2 F''(p\alpha)$
- show that $f''_{p,K}$ vanishes at most once in $(0,1) \Rightarrow$ at most two fixed points in $(0,1]$
- in fact have

$$F''(\alpha) = K \binom{K-1}{K_{\mathfrak{t}}-1} \alpha^{K_{\mathfrak{t}}-2} (1-\alpha)^{K-K_{\mathfrak{t}}-1} (K_{\mathfrak{t}} - 1 - (K-1)\alpha)$$

Limiting Behavior for $K \to \infty$ (with $k/K \to p\alpha$)

- if $p\alpha < \mathfrak{t}$ all learners choose $\mathcal{L}_2$
- if $p\alpha > \mathfrak{t}$ all learners choose $\mathcal{L}_1$
- for $p < \mathfrak{t}$ (hence $p\alpha < \mathfrak{t}$ for all $\alpha \in [0,1]$) only stable fixed point $\alpha = 0$
- for $p > \mathfrak{t}$ two stable fixed points $\alpha = 0$ and $\alpha = 1$ with basins of attraction $\alpha_0 \in [0, \mathfrak{t}/p)$ and $\alpha_0 \in (\mathfrak{t}/p, 1]$
- in this model a change from $\mathcal{L}_1$ to $\mathcal{L}_2$ (or vice versa) achieved by moving $p$ across threshold