Models of Language Evolution: Part II

Matilde Marcolli

MAT1509HS: Mathematical and Computational Linguistics

University of Toronto, Winter 2019, T 4-6 and W 4, BA6180

Main Reference

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

Multiple Language Models

- $\mathcal{H} = \{\mathcal{G}_1, \ldots, \mathcal{G}_n\}$ a space with n grammars
- corresponding languages $\mathcal{L}_1, \ldots, \mathcal{L}_n \subseteq \mathfrak{A}^\star$ on alphabet $\mathfrak{A}$
- probability measure $\mathbb{P}$ on $\mathfrak{A}^\star$ according to which sentences are drawn in data set $\mathcal{D}$
- speakers of language $\mathcal{L}_i$ draw sentences according to probability $\mathbb{P}_i$ with support on $\mathcal{L}_i \subset \mathfrak{A}^\star$ (positive examples)
- distribution $\mathbb{P}$ is a weighted combination

$$\mathbb{P} = \sum_i \alpha_i \mathbb{P}_i$$

where $\alpha_i = \alpha_{i,t}$ are the fractions of population (time/generation t) that speak $\mathcal{L}_i$, with $\sum_i \alpha_i = 1$

- learning algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{H}$ computable function

- probabilistic convergence: grammar $\mathcal{G}_i$ is learnable if

$$\lim_{m\to\infty} \mathbb{P}(\mathcal{A}(\tau_m) = \mathcal{G}_i) = 1$$

when τ_m examples in $\mathcal{L}_i$ drawn according to $\mathbb{P}_i$

Population Dynamics

- State Space: $\mathcal{S}$ = space of all possible linguistic compositions of the population

- identify states $s \in \mathcal{S}$ with possible probability distributions $P = (P_{\mathcal{G}})$ on $\mathcal{H}$ ...identify with previous $\alpha = (\alpha_i)$

- Example: in 3-parameter model (with 8 possible grammars) have

$$\mathcal{S} = \{P = (P_i)_{i=1,\dots,8} \,|\, P_i \geq 0,\ \sum_i P_i = 1\} = \Delta_7$$

- then distribution on $\mathfrak{A}^\star$ is $\mathbb{P}(x) = \sum_i P_i \, \mathbb{P}_i(x)$

• finite sample: probability of formulating a certain hypothesis after a sample of size m

$$p_m(\mathcal{G}_i) := \mathbb{P}(\mathcal{A}(\tau_m) = \mathcal{G}_i)$$

• limiting sample: limiting behavior for sample size $m \to \infty$

$$p(\mathcal{G}_i) := \lim_{m\to\infty} \mathbb{P}(\mathcal{A}(\tau_m) = \mathcal{G}_i)$$

• given at generation/time t a distribution $P_t \in \mathcal{S}$ of speakers of the different languages get recursion relation

$$P_{t+1} = F(P_t)$$

• take $P_{t+1,i} = p_m(\mathcal{G}_i)$ in finite sample case, or $P_{t+1,i} = p(\mathcal{G}_i)$ in limiting sample case (P_t here determines $\mathbb{P}$)

Markov Chain Model

- specify grammars $\mathcal{G}_i$ through their syntactic parameters ($n = 2^N$ number of possible settings of parameters)
- trigger learning algorithm as Markov Chain with 2^N nodes
- $T =$ transition matrix of the Markov Chain
- then probabilities $p_m(\mathcal{G}_i) = \mathbb{P}(\mathcal{A}(\tau_m) = \mathcal{G}_i)$

$$p_m(\mathcal{G}_i) = (2^{-N}\mathbf{1}_{2^N} T^m)_i$$

i-th component of vector obtained by applying (on the right) matrix T^m to normalized row vector $2^{-N}\mathbf{1}_{2^N}$ (all components 2^{-N})

- assuming starting with uniform distribution $2^{-N}\mathbf{1}_{2^N} \in \mathcal{S}$
- limiting distribution $p(\mathcal{G}_i)$ with T_∞

Interpreting limiting behavior

- seen in case of Markov Chain Model, if non-unique closed class in Markov Chain decomposition, limiting T_∞ matrix can have initial states with different probabilities of reaching different targets
- Example of matrix T_∞ in 3-parameter model seen before

$$T_\infty = \begin{pmatrix} & 2/5 & & 3/5 & \\ & 1 & & & \\ & 2/5 & & 3/5 & \\ & 1 & & & \\ & & & 1 & \\ & & & 1 & \\ & & & 1 & \\ & & & 1 & \end{pmatrix}$$

starting at s_1 or s_3 will reach s_5 with probability $3/5$ and s_2 with probability $2/5$

- interpret these probabilities as limiting composition of speakers population

3-parameter model with *homogeneous* initial population

- assume initial population consists only of speakers of one of the languages $\mathcal{L}_i$ (that is, initial P has $P_i = 1$ and all other $P_j = 0$)
- after a number of generations observe drift towards other languages: population no longer homogeneous (or remains homogeneous, depending on initial state)
- at next generation

$$p_m(\mathcal{G}_i) = (PT^m)_i \quad \text{or} \quad p(\mathcal{G}_i) = (PT_\infty)_i$$

with P the initial distribution

- then this gives new $P = (P_i)$ and recompute p_m and p with this P for next generation, etc.

Shortcomings of the model

• simulations in case of 3-parameter model show: +V2-languages remain stable; -V2-languages all drift towards +V2-languages

• contrary to what known from historical linguistics: languages tend to lose the +V2-parameter (switching to -V2): both English and French historically switched from +V2 to -V2

• model relies on many ad hoc assumptions about the state space, the learning algorithm, the probability distributions... some of these are clearly not realistic/adequate to actual language evolution

• Problem: identify which assumptions are inadequate and modify them and improve the model

Can model the S-shape?

• An observation of Historical Linguistics: S-shape of Language Change

• a certain change in a language begins very gradually, then proceeds at a much faster rate, then tails off slowly again

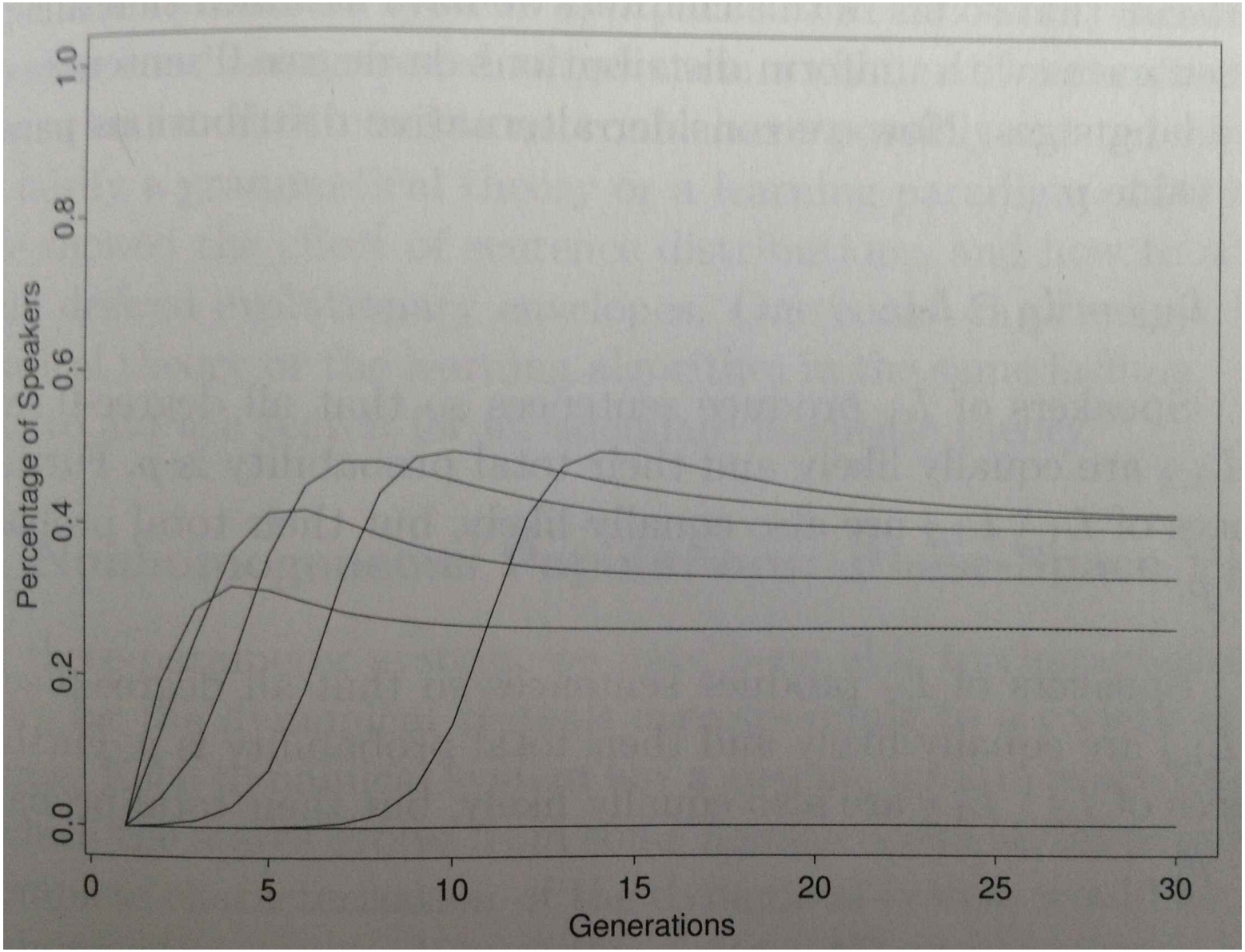
• in evolutionary biology, a logistic shaped curve governs replacement of organisms and of genetic alleles that differ in Darwinian fitness

• in the case of linguistic change, what mechanism would produce this kind of shape?

• in the model of language evolution have dependence on maturation time K and probability distribution $\mathbb{P}$ (a, b parameters in the 2-languages case; distribution P in multilingual case)

- simulations in 2-languages model finds:
  - initial rate of language change is highest when K small
  - curves *not* S-shaped: if start with homogeneous $\mathcal{L}_1$ population, percentage of $\mathcal{L}_2$ speakers with generations first grows to *above* limiting value, then *decreases*
  - slopes and location of the peak and of asymptotic value depend on K

- not clear is these are spurious phenomena due to assumptions in the model, or they say something more about the S-shape empirical observation of historical linguists


Percentage of Speakers
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
Generations

Effect of the dependence on $\mathbb{P}_i$

- probability distributions $\mathbb{P}_i$ with which sentences in $\mathcal{L}_i$ are generated
- these are the underlying parameters of the dynamical system
- seen already in 2-languages model, as dependence on a, b parameters
- if think of these parameters as *changing* in time, can affect a language change by (gradual) modification of the parameters
- can use for "reverse engineering": if know a certain change occurs, find what diachronic modification of the parameters is needed in order to affect that change, then use to model other changes

3-parameter model with *non-homogeneous* initial population

- $P = (P_i)_{i=1,\dots,8}$ probability distribution on the 8 possible grammars of the 3-parameter model
- $P \in \Delta_7$ point in 7-dimensional simplex in $\mathbb{R}^8_+$
- $T =$ transition matrix of the Markov Chain of 3-parameter model
- given P_t (with $P_0 = P$) this determines $\mathbb{P} = \mathbb{P}_t$ on $\mathfrak{A}^\star$ which determines entries of $T = T_t$
- non-homogrneous Markov Chain with $T = T_t$
- for finite-sample size m use matrix T_t^m
- recursion step: next generation distribution $P_{t+1} = P_t\, T_t^m$

Stability Analysis

- given distributions $\mathbb{P}_i$ used in computing transition matrix T

$$T_{ij} = \mathbb{P}(s_i \to s_j) = \sum_{x \in \mathfrak{A}^\star \,:\, \mathcal{A}(s_i, x) = s_j} \mathbb{P}(x)$$

$\mathcal{A}(s, x)$ determines algorithm's next hypothesis, given current state of Markov Chain s and input x

- let T_i be transition matrix when $\mathbb{P} = \mathbb{P}_i$ (target language is $\mathcal{L}_i$)
- data $\mathcal{D}$ all coming from $\mathcal{L}_i$ drawn according to $\mathbb{P}_i$ (instead of mixture of languages with combination $\mathbb{P}$)
- at generation/time t new distribution will be

$$\mathbb{P} = \mathbb{P}_t = \sum_i P_{t,i}\, \mathbb{P}_i,$$

where P_t distribution of languages at generation t

- get T_t transition matrix by

$$(T_t)_{ab} = \sum_{x \in \mathfrak{A}^\star : \mathcal{A}(s_a, x) = s_b} \sum_i P_{t,i}\, \mathbb{P}_i(x) = \sum_i P_{t,i}\ (T_i)_{ab}$$

- fixed points are solutions of $P_{t+1} = P_t$
- finite-sample case size m: $P = (P_i)$ satisfying

$$P = \frac{1}{8} \mathbf{1}_8 (\sum_i P_i T_i)^m$$

- limiting case: can show it becomes solution $P = (P_i)$ of

$$P = \mathbf{1}_8 (I_{8\times 8} - \sum_i P_i T_i + \mathbf{1}_{8\times 8})^{-1}$$

where $I_{N\times N}$ identity and $\mathbf{1}_{N\times N}$ matrix with all entries 1

- assuming initial distribution $1/8\, \mathbf{1}_8$

Multilingual Learners

• realistically, leaners in a multilingual population do not zoom in on a unique grammar, but become *multilingual* themselves (even if in previous generation each individual speaks only one language)

• a more realistic model should take this into account

• instead of learning algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{H}$ consider as a map

$$\mathcal{A} : \mathcal{D} \to \mathcal{M}(\mathcal{H})$$

with $\mathcal{M}(\mathcal{H}) =$ set of probability measures on $\mathcal{H}$

• if $\mathcal{H}$ has n elements, identify $\mathcal{M}(\mathcal{H}) = \Delta_{n-1}$ simplex

Bilingualism

- learning algorithm $\mathcal{A} : \mathcal{D} \to [0, 1]$
- a bilingual speaker (languages $\mathcal{L}_1$ and $\mathcal{L}_2$) produces sentences $x \in \mathfrak{A}^\star$ according to probability

$$\mathbb{P}_\lambda(x) = \lambda \mathbb{P}_1(x) + (1 - \lambda)\mathbb{P}_2(x)$$

with $\mathbb{P}_i$ supported on $\mathcal{L}_i \subset \mathfrak{A}^\star$

- for a single speaker a particular value λ is fixed
- over the entire population it varies according to a distribution $P(\lambda)$ over $[0, 1]$

• probability of a sentence $x \in \mathfrak{A}^\star$ being produced, when input comes from entire population

$$\mathbb{P}(x) = \int_0^1 \mathbb{P}_\lambda(x)\, P(\lambda)\, d\lambda$$

$$\mathbb{P}(x) = \mathbb{P}_1(x) \int_0^1 \lambda\, P(\lambda)\, d\lambda + \mathbb{P}_2(x) \int_0^1 (1-\lambda)\, P(\lambda)\, d\lambda$$

$$\mathbb{P}(x) = \mathbb{E}_P(\lambda)\, \mathbb{P}_1(x) + (1 - \mathbb{E}_P(\lambda))\, \mathbb{P}_2(x)$$

with expectation value

$$\mathbb{E}_P(\lambda) = \int_0^1 \lambda\, P(\lambda)\, d\lambda$$

• so input from population with probabilities $\mathbb{P}_\lambda$ distributed according to $P(\lambda)$ same as input from single speaker with probability $\mathbb{P}_{\mathbb{E}_P(\lambda)}$

- learner (next generation) will also acquire bilingualism with a factor λ which is deduced (via the learning algorithm) from the incoming data
- again subdivide m input data sentences into groups
  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 $n_1 + n_2 + n_3 = m$
- three possible learning algorithms
  1. $\mathcal{A}_1$ sets $\lambda = \frac{n_1}{n_1+n_3}$ (ignoring ambiguous sentences)
  2. $\mathcal{A}_1$ sets $\lambda = \frac{n_1}{n_1+n_2+n_3}$ (ambiguous sentences read as in $\mathcal{L}_2$)
  3. $\mathcal{A}_3$ sets $\lambda = \frac{n_1+n_2}{n_1+n_2+n_3}$ (ambiguous sentences read as in $\mathcal{L}_1$)

- at time/generation t set $u_t := \mathbb{E}_P(\lambda)$ average over population in that generation
- recursive relation u_{t+1} from u_t in terms of parameters

$$a = \mathbb{P}_1(\mathcal{L}_1 \cap \mathcal{L}_2) \qquad b = \mathbb{P}_2(\mathcal{L}_1 \cap \mathcal{L}_2)$$

different weight put on the ambiguous sentences by the probability distributions $\mathbb{P}_i$ of the two languages

- Result: recursion for the three algorithms

$$\begin{aligned}
\mathcal{A}_1 : \quad u_{t+1} &= \frac{u_t(1-a)}{u_t(1-a) + (1-u_t)(1-b)} \\
\mathcal{A}_2 : \quad u_{t+1} &= u_t(1-a) \\
\mathcal{A}_3 : \quad u_{t+1} &= u_t(1-b) + b
\end{aligned}$$

Explanation

- case of $\mathcal{A}_2$:

$$u_{t+1} = \mathbb{E}(\frac{n_1}{n_1+n_2+n_3}) = \frac{\mathbb{E}(n_1)}{m} = \mathbb{P}_1(\mathcal{L}_1 \setminus \mathcal{L}_2)\mathbb{E}_P(\lambda) = (1-a)u_t$$

- case of $\mathcal{A}_3$:

$$u_{t+1} = \mathbb{E}(\frac{n_1+n_2}{n_1+n_2+n_3}) = \frac{\mathbb{E}(n_1)}{m} + \frac{\mathbb{E}(n_2)}{m}$$

$$= u_t(1-a) + u_t a + (1-u_t)b$$

- case of $\mathcal{A}_1$:

$$u_{t+1} = \mathbb{E}(\frac{n_1}{n_1+n_3}) = \sum \binom{m}{n_1 n_2 n_3} \alpha^{n_1} \beta^{n_2} \gamma^{n_3} \frac{n_1}{n_1+n_3}$$

with $\alpha = (1-a)u_t$, $\beta = au_t + b(1-u_t)$, $\gamma = (1-b)(1-u_t)$

$$u_{t+1} = \sum_{k=0}^{m} \sum_{n_1=0}^{k} \binom{k}{n_1} \binom{m}{k} \alpha^{n_1} \beta^{m-k} \gamma^{k-n_1} \frac{n_1}{k}$$

$$= \sum_{k=0}^{m} \binom{m}{k} \beta^{m-k} (1-\beta)^k \frac{\alpha}{1-\beta} = \frac{\alpha}{1-\beta}$$

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