

Models of Language Evolution: Part III

Matilde Marcolli

CS101: Mathematical and Computational Linguistics

Winter 2015

Main Reference

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

Language Evolution and Fitness

- Language evolution modeled on ideas from biological evolution
 - a reproductive process: **learning algorithm** (individuals of new generation produce language using linguistic input data from previous generation)
 - transmission errors: **multilingual models and population dynamics**
 - **fitness test?**
- in biological evolution **reproductive fitness** drives evolution by natural selection
- **Problem**: is there a good fitness function in linguistics?
- develop a notion of **communicative efficiency**

Origins of Language

- how did language arise, from our pre-human ancestors?
- like everything else in the biological world: by Darwinian evolution by natural selection
- evolutionary advantage in terms of reproductive fitness
- communicative efficiency provides biological fitness
- related idea: *coherence* (homogeneous linguistic population) is an *emergent phenomenon* resulting from the behavior of interacting individual linguistic agents
- passage to a coherent state resulting from a bifurcation in the dynamics

Mutual Intelligibility

- \mathcal{L}_1 and \mathcal{L}_2 two languages: want to define a **communicative fitness** or **mutual intelligibility** function $F(\mathcal{L}_1, \mathcal{L}_2)$
- how to improve (or maximize) $F(\mathcal{L}_1, \mathcal{L}_2)$ depends on *ambiguities* in the languages and on learning algorithms
- want an algorithm that identifies $\arg \max_{\mathcal{L}'} F(\mathcal{L}, \mathcal{L}')$ or approximates it arbitrarily well
- if \mathcal{L} has ambiguities $\arg \max_{\mathcal{L}'} F(\mathcal{L}, \mathcal{L}')$ need not be $\mathcal{L}' = \mathcal{L}$

Language as Probabilistic Association between Form and Meaning

- language as an association matrix linking referents to signals
- M referents and N signals: $A = (A_{ij})$ an $M \times N$ -matrix, values of entries A_{ij} = strength of the association
- this matrix drives both *production* and *comprehension* (assigning signals to meanings and meanings to signals)
- also used as framework to model “communication in the animal and the machine”
- set of all possible signals is the set of all strings \mathcal{A}^* ; set of all possible meanings is set of strings over a *semantic alphabet*: infinite matrices

- \mathcal{S} = set of signals; \mathcal{M} = set of meanings (finite or countable)
- assume $\mathcal{S} = \mathfrak{A}_1^*$ and $\mathcal{M} = \mathfrak{A}_2^*$ (linguistic and semantic alphabets)
- **Communication System**: probability measure μ on $\mathcal{S} \times \mathcal{M}$
- generalizes the (normalized) “association matrix” to infinite case
- **Encoding Matrix** (production): $s_i \in \mathcal{S}$ and $m_j \in \mathcal{M}$

$$P_{ij} = \mu(s_i | m_j) = \frac{\mu(s_i, m_j)}{\sum_k \mu(s_k, m_j)}$$

set equal zero if denominator sum is zero

- **Decoding Matrix** (comprehension):

$$Q_{ij} = \mu(m_i | s_j) = \frac{\mu(s_j, m_i)}{\sum_k \mu(s_j, m_k)}$$

and zero if denominator sum is zero

- in other models P and Q not required to come from the same measure μ , but this is a way to ensure *consistency* between active and passive usage of a language
- Note that μ determines P and Q , but these matrices don't determine μ : can consider equivalence classes of measures μ that determine same P, Q
- **Useful signals**: (actually used in production or comprehension)

$$\mathcal{S}_\mu = \{s \in \mathcal{S} \mid \exists m \in \mathcal{M}, \mu(s, m) > 0\}$$

- this should correspond in formal language theory to the set of well formed (grammatical) sentences
- **Expressible meanings**: (can be expressed within the language)

$$\mathcal{M}_\mu = \{m \in \mathcal{M} \mid \exists s \in \mathcal{S}, \mu(s, m) > 0\}$$

Communication

- two systems μ_1 and μ_2
- probability of a meaning being successfully communicated from μ_1 to μ_2 or from μ_2 to μ_1 (with σ distribution on \mathcal{M})

$$\mathbb{P}(1 \rightarrow 2) = \sum_i \sigma(m_i) \sum_j \mu_1(s_j | m_i) \mu_2(m_i | s_j)$$

$$\mathbb{P}(2 \rightarrow 1) = \sum_i \sigma(m_i) \sum_j \mu_2(s_j | m_i) \mu_1(m_i | s_j)$$

- **Communicability** $F(\mu_1, \mu_2) = \frac{1}{2}(\mathbb{P}(1 \rightarrow 2) + \mathbb{P}(2 \rightarrow 1))$

$$F(\mu_1, \mu_2) = \frac{1}{2}(\text{Tr}(P^{(1)} \Lambda^\tau Q^{(2)}) + \text{Tr}(P^{(2)} \Lambda^\tau Q^{(1)}))$$

$\Lambda =$ diagonal matrix entries $\Lambda_{ii} = \sigma(m_i)$

- $F(\mu_1, \mu_2)$ probability of understanding each other in two way communication: $0 \leq F(\mu_1, \mu_2) = F(\mu_2, \mu_1) \leq 1$
- $F(\mu, \mu)$ communicability between linguistic agents with same language: $0 < F(\mu, \mu) \leq 1$
- Note: if had $\mathcal{S} = \mathcal{M}$ and μ supported on diagonal, P and Q would be identity and $F(\mu, \mu) = 1$
- Role of distribution σ on \mathcal{M} : not marginal of μ , but determined by “external world”, which meanings are more likely to be communicated in a given context, F may be larger or smaller depending on this external context (two linguistic agents may communicate better or worse in different contexts)

Best Response

- suppose one language given $\mu = \mu_0$
- Want to maximize communicability

$$F(\mu_0, \mu_*) = \sup_{\mu} F(\mu_0, \mu)$$

- Algorithm that approaches the best response μ_*
- construct a family of languages μ_ϵ such that $F(\mu_0, \mu_\epsilon)$ gets arbitrarily close to $\sup_{\mu} F(\mu_0, \mu)$ when $\epsilon \rightarrow 0$

Finite Languages: simplified model

Assume:

- 1 Languages are finite with μ an $M \times N$ matrix
- 2 The distribution σ is uniform $1/M$
- 3 The measure μ_0 has *unique maximum property*: for all $s \in \mathcal{S}$ there is a unique $m = m(s) \in \mathcal{M}$ and for $m \in \mathcal{M}$ a unique $s = s(m) \in \mathcal{S}$ with

$$\mu_0(s | m(s)) = \max_{m \in \mathcal{M}} \mu_0(s | m), \quad \mu_0(s(m) | m) = \max_{s \in \mathcal{S}} \mu_0(s | m)$$

Best Decoder

- Find a matrix Q_* with

$$\sum_{ij} \mu_0(s_i | m_j) Q_{*,ij} = \max_Q \sum_{ij} \mu_0(s_i | m_j) Q_{ij}$$

maximize over non-negative row-stochastic matrices

- this is given by

$$Q_{*,ij} = \begin{cases} 1 & \mu_0(s_i | m_j) = \max_k \mu_0(s_i | m_k) \\ 0 & \text{otherwise} \end{cases}$$

Best Encoder

- Find a matrix P_* with

$$\sum_{ij} P_{*,ij} \mu_0(m_j | s_i) = \max_P \sum_{ij} P_{ij} \mu_0(m_j | s_i)$$

maximize over non-negative column-stochastic matrices

- this is given by

$$P_{*,ij} = \begin{cases} 1 & \mu_0(m_j | s_i) = \max_k \mu_0(m_j | s_k) \\ 0 & \text{otherwise} \end{cases}$$

Constrain relating them needs to be satisfied: $\exists \mu_*$

$$\mu_*(s_i | m_j) = P_{*,ij}, \quad \mu_*(m_j | s_i) = Q_{*,ij}$$

Problem: this does not always work

Approximations

- define $P_{ij}^0 = \mu_0(s_i | m_j)$ and $Q_{ij}^0 = \mu_0(m_j | s_i)$
- **Result:** for μ_0 finite with unique max and σ uniform

$$\sup_{\mu} F(\mu_0, \mu) = \frac{1}{2M} \text{Tr}(P^0 \tau Q_{\star} + P_{\star} \tau Q^0)$$

- this follows from two properties:
 - $F(\mu_0, \mu) \leq \frac{1}{2M} \text{Tr}(P^0 \tau Q_{\star} + P_{\star} \tau Q^0)$ for all μ
 - $\forall \epsilon \exists \mu_{\epsilon}$ with

$$\lim_{\epsilon \rightarrow 0} \left(\frac{1}{2M} \text{Tr}(P^0 \tau Q_{\star} + P_{\star} \tau Q^0) - F(\mu_0, \mu_{\epsilon}) \right) = 0$$

- first property true by definition of best decoder and best encoder

Construction of the measures μ_ϵ with

$$\lim_{\epsilon \rightarrow 0} \mu_\epsilon(s_i | m_j) = P_{*,ij} \quad \lim_{\epsilon \rightarrow 0} \mu_\epsilon(m_j | s_i) = Q_{*,ij}$$

- Auxiliary matrix X :

$$X_{ij} = \begin{cases} 1 & P_{*,ij} + Q_{*,ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

- form a **Graph**: G_X
 - vertices = entries of matrix X that are = 1
 - edges = lines connecting 1 entries on the same row and on the same column

- **Fact:** if measure μ_0 has *unique maximum property* then the graph does not have loops (tree or multiconnected forest)
- then construction of μ_ϵ :
 - for each component of G_X : take each pair of vertices
 - if connected by horizontal (vertical) line: look at corresponding entries of Q_\star (or P_\star): one of them is one the other is zero
 - orient the edge from the vertex with entry 0 to the one with entry 1
 - start from one of the vertices: replace corresponding entry of X with ϵ
 - follow oriented path replace successive elements of X with ϵ^k (increasing k along *reverse* orientation of edges: unambiguous because no loops): matrix A^ϵ
 - measure μ_ϵ :

$$\mu_\epsilon(s_i, m_j) = \frac{A_{ij}^\epsilon}{\sum_{k,l} A_{kl}^\epsilon}$$

Limiting behavior of μ_ϵ

- normalize each column of A^ϵ so that sum adds up to one to get $\mu_\epsilon(s_i | m_j)$
- want to show these $\mu_\epsilon(s_i | m_j)$ converge to $P_{\star,ij}$
- each column of A^ϵ contains contains *at most one* edge of a connected component of G_X because P_\star (resp. Q_\star) has at most one 1 entry per column (resp. row) so X has at most two
- for $\epsilon \rightarrow 0$ only lowest power of ϵ dominant, others to zero faster
- dominant term is column entry where P_\star is 1: in the limit it gives the column of P_\star
- argument for $\mu_\epsilon(m_j | s_i) = Q_{\star,ij}$ is similar using rows

More general cases

- dropping all assumptions of unique maximum for μ_0 , uniform distribution σ , and finite N and M :
 - N. Komarova, P. Niyogi, *Optimizing the mutual intelligibility of mutual agents in a shared world*, Artificial Intelligence Journal, 154 (2004) 1–42.

Learning

- in this model: trying to communicate with an agent whose language is μ : best response strategy is trying to approximate μ_* constructing some μ_ϵ (while μ_* itself need not exist)
- but measure μ is unknown to learner: two possible scenarios
 - **full information**: can sample μ directly for (meaning,sentence) pairs; then strategy is sample μ as accurately as possible, construct P_* and Q_* and from those μ_ϵ
 - **partial information**: meaning is not directly accessible, only sentences are, and a feedback response on whether interpretation of sentence by learner is correct

Learning with full information

- **Event** E_{ij} : sentence s_i is produced to communicate meaning m_j
- probability of event E_{ij} is $\sigma(m_j) \mu(s_i | m_j)$
- for large n events drawn uniformly randomly frequencies k_{ij}/n approximate probability
- can estimate the $\sigma(m_j) \mu(s_i | m_j)$ using sampling frequencies k_{ij}/n
- use estimated $\sigma(m_j) \mu(s_i | m_j)$ to compute P_* , Q_* , μ_ϵ

Learning with Partial Information

- learner guesses meaning without direct access to it: if correct guess know meaning, if not only have negative information, asymmetric
- suppose guess meaning uniformly randomly among $M = \#\mathcal{M}$ possible meanings: guess m_j with probability $1/M$
- **Event** E_{ij} : sentence s_i is produced, meaning m_j is guessed, successfully
- probability of event E_{ij} is $\frac{1}{M}\sigma(m_j)\mu(s_i | m_j)$
- also can be empirically estimated from frequencies of correct guessing k_{ij}/n
- so apparently different setup leads to very similar procedure anyway

Communicative Efficiency: a phonetics example

- suppose English words are transmitted from a speaker to a receiver
- each word is a list of **phonemes**: if every phoneme is received correctly communicative efficiency would be 1
- some phonemes are notoriously difficult to distinguish in transmissions: *p* and *b* for example
- this causes ambiguities in words such as *bit* versus *pit* or *pat* versus *bat*
- since this can cause different meaning associations to sentences there is a **loss of communicative efficiency**
- subdivide the lexicon into **cohorts**: equivalence classes of words that become indistinguishable if certain phonemes are no longer distinguished

- p_i = probabilities (frequencies) of words in the original lexicon \mathcal{W}
- information content of the lexicon measured by Shannon entropy

$$S(\mathcal{W}) = - \sum_{w_i \in \mathcal{W}} p_i \log(p_i)$$

- after passing to cohorts $\mathcal{W}_{/\sim}$ probabilities $P_k = \sum_{w_i \in C_k} p_i$
- information content of set of cohorts

$$S(\mathcal{W}_{/\sim}) = - \sum_{C_k \in \mathcal{W}_{/\sim}} P_k \log(P_k)$$

- normalized **Information Loss**

$$IL(\mathcal{W}, \sim) = \frac{S(\mathcal{W}) - S(\mathcal{W}_{/\sim})}{S(\mathcal{W})}$$

measures the “functional load” in communication carried by the ability to distinguished those phonemes

Communicative Fitness

- \mathcal{H} set of n possible languages
- identify languages with measures μ_k on $\mathcal{M} \times \mathcal{S}$ (meanings and sentences/signals)
- **mutual intelligibility matrix**

$$A_{ij} = \sum_{m \in \mathcal{M}} \sigma(m) \sum_{s \in \mathcal{S}} \mu_i(s | m) \mu_j(m | s)$$

probability that a speaker of language μ_i is understood by a receiver who speaks language μ_j

- **simplified model**: assume $A_{ii} = 1$ (each language has perfect intelligibility with itself) and $A_{ij} = a$ for some $0 \leq a \leq 1$ for all pairs $i \neq j$, same for all pairs
- also assume population of constant size with every person speaking only one language
- linguistic distribution of the population: $\alpha_k \geq 0$, with $\sum_{k=1}^n \alpha_k = 1$
- **individual communicative fitness** of a speaker of language μ_k : average communicative efficiency with the rest of the population
- **mutual intelligibility** of μ_i and μ_j

$$F(\mu_i, \mu_j) = \frac{1}{2}(A_{ij} + A_{ji})$$

- so **average communicative efficiency** of a speaker of μ_i

$$f_i = f_0 + \sum_{j=1}^n F(\mu_i, \mu_j) \alpha_j$$

f_0 = background, independent of language (but dependent of how much specific environment facilitates communication)

- if everybody spoke the same language μ (assuming $\mathbb{A}_{ii} = 1$) would have $f = f_0 + 1$; if other languages are present, lower value of fitness f
- following basic rule of evolution by natural selection: assume **individuals reproduce in proportion to their fitness**
- assuming successful communication is an evolutionary advantage in the Darwinian sense
- in this model also make the assumption that children learn language from their parents and not from the entire community

- further simplify the model by assuming each learner has only one teacher (literally "mother tongue")
- also allow for mistakes during language acquisition
- probability of a transition from language μ_i to language μ_j is Q_{ij} (depends on \mathbb{A} : on how close the different languages are)
- **Population Dynamics**

$$\alpha_{t+1,j} = \frac{\sum_{i=1}^n \alpha_{t,i} f_i Q_{ij}}{\sum_{k=1}^n \alpha_{t,k} f_k}$$

reproduction proportional to fitness: percentage of new generation produced by speakers of language μ_i in previous generation is $f_i \alpha_{t,i}$ (normalized by $\sum_k f_k \alpha_{t,k}$)

ODE: turn difference equation into ordinary differential equation

- normalization condition $\sum_k \alpha_k = 1$ gives $\sum_k \dot{\alpha}_{t,k} = 0$
- positivity $\alpha_k \geq 0$ becomes condition $\dot{\alpha}_{t,k}|_{\alpha_{t,k}=0} \geq 0$
- continuous time differential equation

$$\dot{\alpha}_{t,k} = \sum_{i=1}^n f_i \alpha_{t,i} Q_{ij} - \phi \alpha_{t,k}$$

with $\phi(t) = \sum_k f_k \alpha_{t,k}$ *average fitness of the population*

- for case with $A_{ii} = 1$ and $A_{ij} = a$ for $i \neq j$ fitness

$$f_i = (1 - a)\alpha_i + a + f_0$$

- *learning fidelity*: probability $\frac{1}{n} \leq q \leq 1$ of learning same language as primary teacher

$$Q_{ii} = q \quad \text{and} \quad Q_{ij} = \frac{(1 - q)}{n - 1} \quad \text{when } i \neq j$$

perfect learning $q = 1$; random guessing $q = 1/n$

- then differential equation

$$\dot{\alpha}_{t,k} = (1 - a) \left(-\alpha_{t,k}^3 + \alpha_{t,k}^2 q + \sum_{j \neq k} \alpha_{t,j}^2 \left(\frac{1 - q}{n - 1} - \alpha_{t,k} \right) \right) - \frac{(a + f_0)(1 - q)(n\alpha_{t,k} - 1)}{n - 1}$$

Equilibrium Solutions (critical points $\dot{\alpha}_{t,k} = 0$)

- each x_j root of polynomial (with $\gamma = \sum_j x_j^2$)

$$P_{a,q,n}(x) = (1-a) \left(-x^3 + x^2 q + (\gamma - x) \left(\frac{1-q}{n-1} - x \right) \right) \\ - \frac{(a+f_0)(1-q)(nx-1)}{n-1}$$

- if $x_\ell = X$ and all other $x_k = \frac{1-X}{n-1}$ (so $\sum_j x_j = 1$) then equation becomes

$$X^3 - X^2 q + \frac{(1-X)^2}{n-1} \left(X - \frac{1-q}{n-1} \right) + \frac{(a+f_0)(1-q)(nX-1)}{(1-a)(n-1)} = 0$$

- Cubic polynomial: three solutions given by $\frac{1}{n}$, and r_{\pm}

$$r_{\pm} = \frac{-(1-a)(1+(n-2)q) \mp \sqrt{D}}{2(a-1)(n-1)}$$

$$D = 4(-1-a(n-2)-f_0(n-1))(1-q)(n-1)(1-a)+(1-a)^2(1+(n-2)q)^2$$

- So in total $2n + 1$ solutions:

- 1 uniform solution $x_k = 1/n$ for all k
- 2 one $x_{\ell} = r_+$ and all other $x_k = (1 - r_+)/ (n - 1)$ (n possibilities for x_{ℓ})
- 3 one $x_{\ell} = r'_-$ and all other $x_k = (1 - r_-)/ (n - 1)$ (n possibilities for x_{ℓ})

the last two cases lead to one preferred language (and all the others with same distribution)