Bayesian Models of Language Evolution & Change

And a proposal for a modeling database

Dan Dediu
The Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands
Dan.Dediu@mpi.nl

MOLE Workshop
EvoLang 2010
Utrecht, NL
14 Apr 2010
Overview

User's (genetics & language) point of view

Part I: Bayesian Models in Language Evolution
- Summary
- Advantages & disadvantages
- Can we fix them?

Part II: A database of models
- Rationale
- Examples & suggestions
Bayesian Models

Summary

• $H$ – the universe of all possible languages
• $h \in H$ – a language/hypothesis
• $A$ – a Bayesian agent
• $0 \leq p(h) \leq 1$ – the agent's “subjective” probability distribution (Press, 2003) across languages

Example: $H = \{L_1, L_2\}$, $p: H \rightarrow [0,1]$, $p(L_1) = p_1$, $p(L_2) = p_2 = 1-p_1$

• $A$ uses $p$ to produce or comprehend language
Bayesian Models

Summary

- The “real” Bayesian bit: language acquisition/learning/updating
- Bayes' rule:

\[ p(h|d) = \frac{p_{obs}(d|h) \cdot p(h)}{\sum_{h' \in H} p_{obs}(d|h') \cdot p(h')} \]

- where:
- \( d \) – the data, embodying A's experience with the language(s)
  – usually a set of utterances
Bayesian Models

Summary

- $p(h|d)$ – the **posterior** (updated) probability that hypothesis $h$ holds after exposure to $d$
- $p_{\text{obs}}(d|h)$ – the **likelihood** of observing $d$ if $h$ true
- $p(h)$ – the **prior** probability of $h$ before A “saw” $d$
  - can represent:
    - the outcome (posterior probability) of a previous learning round or
    - the result of “innate” predispositions
Bayesian Models

Summary

• Now, how is $A$ to use the distribution across languages, $p$, it has acquired to actually do stuff?

• General idea:
  - pick a **single** language $h_w$ out of the whole $H$ somehow using $p$
  - use $h_w$ to “communicate” (usually, just “speak”)

• Two major strategies:
  - **sampling**: pick $h_w$ from $H$ proportional to its $p(h_w)$
  - **maximizer**: pick $h_w$ having the highest $p(h_w)$
Bayesian Models

Summary

- **Example:**
  - \( p(L_1) = p_1 = 0.2 \) & \( p(L_2) = p_2 = 1 - p_1 = 0.8 \)
  - **maximizer:** always pick \( L_2 \)
  - **sampler:** pick \( L_2 \) about 80% of the time and \( L_1 \) the remaining 20%

- **“Intermediate” strategies** (Kirby et al., 2007):
  - pick \( h_w \) from \( H \) proportional to its \( p(h_w) \)
  - \( r = 1 \) → sampler, \( r \to \infty \) → maximizer
Bayesian Models

Summary

• Iteration:
  – current generation agent(s) $A$ use their $h_w$ to produce language data $d$
  – $d$ is fed into the next generations' agents which use it to arrive at their own $h_w$

• The social context usually is:
  – homogeneous single chains of agents
  – homogeneous populations with single teachers
  – more complex/realistic settings
Bayesian Models

Main results

- Chains:
  - sampler: converges to prior (Gibbs sampler)
  - maximizer: complex but influenced by prior (rank; Expectation-Maximization)
- Sampler is not ESS, invasion by maximizers
- Heterogeneous chains of pairs
  - complex behavior
  - no simple & clear rules?
Issues & possible fixes
The “acquisitionist” assumption

- The **acquirer** is the **locus of language change**:
  - re**interpretation** of the linguistic data
- However, the acquirer is probably **not** the only (or even the most important) locus of change
- **Competent language users** drive change: Croft, 2000; Enfield & Levinson, 2006; Ostler, 2005
- **Fixes**:
  - dynamic selection/modification of $h_w$
Issues & possible fixes
The nature of the data and hypotheses space

- **H** and **d**: tend to embody a simplistic “linguistics”:
  - “words” (forms or form-meaning mappings)
  - abstract “rules”

- But language is embedded in a rich context →
  - dialects, sociolects, registers = meaningful variation
  - underspecification, pragmatics, inference...

- **Fixes:**
  - make **H** and **d** richer (contextual, dynamic) →
    hierarchical models seem promising
  - “real” social dynamics/population structure
Issues & possible fixes

The “problem” of asymptotic behavior

- Heated debate concerning sampler vs maximizer
- Motivated by their asymptotic behavior → how “free” the cultural process really is from the prior?
- To be relevant to language → assume that present day typology is related to asymptote
- Deeper assumption: enough time and weak enough phylogenetic “inertia”
  - might not be warranted a priori for all features
  - some rates are very slow (typology, cognacy)
- Potentially non-issue?
Issues & possible fixes

What is the prior? What are the biases?

- It is assumed that the “real” prior (i.e., $p_0(h)$ before $A$ has seen any $d$) are the biases
- But biases contain (there can be variation in):
  - “real” prior, $p_0(h)$
  - likelihood function, $p_{obs}(h|d)$
  - language selection mechanism, $p(h|d) \rightarrow h_w$
  - ontogenetic development (implicit) genome $\rightarrow p_0(h)$
  - communicative structure & rules, etc, etc

- Fixes: explicit modeling, enlargement of “biases”
Issues & possible fixes

The “prior” and the development

- Even with the previous caveats, the “real” prior $p_0(h)$ is assuming a neat dichotomy between “innate” and “acquired”
- This dichotomy is patently wrong:
  - Genes & environment interact in complex ways
  - Both are equally required
- Fixes:
  - we must show that this modeling dichotomy is acceptable, or
  - explicitly model the development of $p_0(h)$
Issues & possible fixes

Omniscience & “pre-science”

- Ferdinand and Zuidema (2009)
  - If $p_0(h) \equiv$ learning biases then the learner must be omniscient about the possible sources of the data

- For the learning to be meaningful (result in acceptably similar language across generations):
  - The acquirer's likelihood function $p_{obs}(h|d)$ must match the actual production mechanism $p(h|.) \rightarrow d$
  - “Pre-science”: How? Learning of $p_{obs}(h|d)$?
Issues & possible fixes
Computational burden & level of analysis

• David Marr's (1982) classic 3 levels:
  – **computational** (semantic, content)
  – **algorithmic** (syntactic, form)
  – **implementational** (physical, medium)

• Probably computational?
• But what are the lower levels? What would count as a crucial experiment/falsifying data?
• What “approximations” actually do the work?
• What would be the reason for nature to “want” to simulate such a Bayesian mechanism?
Issues & possible fixes
Some practical modeling issues

• Computational:
  – *MCMC* (slow, convergence)
  – *conjugate priors* (contentious, possible artifacts)

• How well does it **scale** to
  – more realistic “languages”
  – richer context
  – more complex populations/social networks
Advantages of Bayesian Models

- Some very important advantages:
  - robust, rich & well studied mathematical & philosophical framework
  - relatively easy to understand
  - (conceptual and, possibly, practical) standardization
  - some (limited?) empirical support
  - fashionable (?)
Suggestions & Conclusions

- **Hierarchical Bayesian Models**
  - including a richer context (social, communicative)
  - explicit modeling of pre-learning development
  - must be investigated how well they behave & how natural the assumptions required are

- We have to **qualify** the generalizability and practical relevance of our results

- Empirical investigations of the appropriateness of the Bayesian assumptions

- Probably **not** the solution for every problem...
Part II: A Database for Models in Language Evolution & Change

● Overview:
  - Open access to rich descriptions of such (mathematical & computational) models
  - Source code (where possible)
  - Relevant publications & results
  - Searching & indexing
  - Comments, discussions/forum, voting

● Advantages:
  - Centralized resource → emergence of standards
  - Avoid “reinvention of the wheel”/bad new models
  - Increase speed of development of the field
NCBI, UCSC Genome Browser, HapMap, Felsenstein's Phylogeny Software

  - **Data** (GenBank, dbDNP, Entrez Nucleotide db…)
  - **Primary research** (PubMed)
  - **Summary findings** (OMIM)
  - **Software** (online, stand-alone) (BLAST)

- **UCSC Genome Browser** ([http://genome.ucsc.edu/](http://genome.ucsc.edu/))


  - Comprehensive list (400+ packages & servers)
Databases

- Clearly a **key ingredient** in the current explosive growth and theoretical & practical success of genetics/biology/biotech
- We **have to** implement the idea to promote a healthy and quick growth of our field
- **How?**
  - **Extend** Jun Wang’s *Language Evolution and Computation Bibliography* ([http://www.isrl.illinois.edu/~amag/langev/](http://www.isrl.illinois.edu/~amag/langev/))
  - **Create** a new online resource
- **Management:**
  - **Owner/Elite board** (NCBI, ...)
  - **Open** (Wikipedia)
Thank you!

Thanks to: K. Smith, J. Zuidema, S. Kirby, J. Hurford, B. Ladd, M. Christiansen, S. Levinson, N. Enfield, M. Dunn & A. Dima