# Bayesian Decision Theory, Iterated Learning and Portuguese Clitics

## Psychocomputational Models of Human Language Acquisition

Catherine Lai

Department of Linguistics
University of Pennsylvania

July 23, 2008

# Computational Models of Language Change

*Some questions:*

- How can we deal with individual and population variation in models of language change?
- Where does instability come from in these models?
- How do we use all these frequency counts to choose a grammar?

*Some Frameworks:*

- Iterated learning: (Kirby, 2001; Kirby et al., 2007)
- Dynamical systems: (Mitchener and Nowak, 2003; Nowak et al., 1999).
- Social learning: (Niyogi and Berwick, 1998; Yang, 2001)

# How do you make a decision?

*The decision rule through which a grammar is selected is crucial!*

- Are learners *just* trying to fit the probability distribution of the input data to a predefined model? This is basically what Maximum Likelihood Estimation (MLE) allows you to do.
- MLE requires us to conflate several factors: innate biases (priors), social and communicative factors, and random noise.
- If we view language learning as a problem involving *beliefs* and factors outside pure point estimation, the Bayesian view becomes very attractive.
- However, even within general Bayesian frameworks, MLE is still often implicitly employed (cf. MAP estimation Griffiths and Kalish (2005); Dowman et al. (2006); Briscoe (2000))

What does it mean to be a Bayesian?

# Outline

- Portuguese clitic data
- Models and requirements
- Bayesian decision theory

# Portuguese Direct Object Clitics

(1) $a_1$ Paolo a ama (affirmative, proclisis)

$a_2$ Paolo ama-a (affirmative, enclisis)

$a_3$ Quem a ama (obligatory proclisis)

- Affirmative sentences with topics, adjuncts or referential subjects:
  - Classical Portuguese (ClP, 16th to mid-19th century): allowed direct object enclitics and proclitics (preferred Galves et al. (2005a)). ($a_1$, $a_2$)
  - Modern European Portuguese (EP): obligatory enclisis. ($a_2$)
- Proclitic forms are obligatory in other syntactic contexts. ($a_3$)

*Note: we're treating EP as a subset of ClP.*

# Change!

- According to corpus studies (Galves et al., 2005a), there was a sharp rise in enclisis in the early to mid 18th century.
- Galves and Galves (1995): this syntactic change was driven by change in stress patterns. (although see Galves (2003); Costa and Duarte (2002)).
- *Acquisition question:* How does the learner do parameter setting?
- *Production question:* What sort of data will the learner produce for the next generation?

# The Galves Batch Model

- Galves and Galves (1995): Construction types are given a probability proportional to the stress contour associated with it.
- Clauses of type $a_1$ and $a_3$ (proclisis) have weight $p$ and clauses of type $a_2$ (enclisis) have weight $q$.

$$\mathbf{P}(a_1|G_{CIP}) = p/(2p+q) \tag{2}$$
$$\mathbf{P}(a_1|G_{EP}) = 0 \tag{3}$$

- Grammar selection via Maximum Likelihood Estimation (MLE).
- The probability of the learner acquiring $G_{CIP}$ as the probability that clause type 1 occurs at least once in $n$ samples (the critical period).

# Batch Learning as Markov Process

- Niyogi and Berwick (1998); Niyogi (2006): re-implement the GBM but take more of a a population level view.
- $\alpha_t$ = proportion of the population with $G_{EP}$ at time $t$.
- $\alpha_{t+1}$ depends on $\alpha_t$ and the learning mechanism (MLE). (Markov process with two states)

$$
\begin{aligned}
\mathbf{P}(a_1|G_{CIP}) = \mathbf{P}(a_3|G_{CIP}) &= p \text{ for some } p \in [0,1], \\
\mathbf{P}(a_2|G_{CIP}) &= 1 - 2p. \\
\mathbf{P}(a_1|G_{EP}) &= 0, \\
\mathbf{P}(a_2|G_{EP}) &= q, \text{ for some } q \in [0,1], \\
\mathbf{P}(a_3|G_{EP}) &= 1 - q,
\end{aligned}
$$

- $p$ and $q$ are *production probabilities* encoded in the grammar. These hold across the board for *all* speakers of a particular grammar.

# And so...

- Learners may still acquire $G_{CIP}$ even though they do not see any instances of variational proclisis ($a_1$)! That is, if there are *too many* instances of the type $a_3$ (obligatory proclisis).
- Is because there is proclisis in $a_3$? No! This would still happen if the syntax of the $a_3$ type was totally devoid of clitics.
- Also, a learner who acquires $G_{CIP}$ will continue to use a (possibly) very high rate of variational proclisis ($p$) in spite of being surrounded by $G_{EP}$ speakers.
- Shouldn't we expect that the desire to communicate would pressure speakers of $G_{CIP}$ to lower the rate of variational proclisis in the face of multitudes of $G_{EP}$ speakers?
- How do we deal with noise? (c.f. Briscoe (2002)) What about biases? How about being Bayesian?

# Bayesian Iterated Learning

Signal/meaning pairs (Griffiths and Kalish, 2005)

- $(Y_k, X_k) = \{(y_1, x_1) \dots (y_n, x_n)\}$: (utterance, meaning) pairs received by agent in generation $k$. ($y \rightarrow x$ is many to one).
- This allows us to focus only on types that show variation.
- Grammar selection is based on the posterior ($g$ is the hypothesized grammar),

$$\mathbf{P}(g|X_k, Y_k) = \frac{\mathbf{P}(Y_k|X_k, g)\mathbf{P}(g)}{\mathbf{P}(Y_k|X_k)},$$

  Priors over grammars are assumed to be innate and invariable across generations.
- Also, add an error term to account for random noise.
- Griffiths and Kalish (2005); Kirby et al. (2007) show analytically that convergence to the prior depends on the selection mechanism (MAP, sampling from the posterior, etc.)

# BIL and Portuguese

- BIL $\approx$ Griffiths and Kalish (2005) does not take into account variation in the community (!). However, in general IL allows more than one agent in a generation Kirby and Hurford (2002).
  $\Rightarrow$ BIL is like the previous models except for the priors.
- For Portuguese, we don't have to consider cases of obligatory proclisis ($a_3$) since they do not differentiate the two grammars.
- However, $\mathbf{P}(a_1|x_1) = p$ is still seems to be an innate part of the grammar with MAP estimation.

# What would we like in the model?

- Frameworks are frameworks – they still need articulation.
- We would like to incorporate some formal notion of why frequency estimation is important to the learner.
- At least part of this should come from the fact that the learner wishes to communicate effectively with a variety of speakers.
- For example, we want to incorporate the intuitive idea that using rare forms when frequent forms exists may be disfavored.
- Also forms that are harder to produce (and process) should be disfavoured (c.f. the prosody argument).
- The decision problem that learners face is subjective – learners choose a grammar that they believe will be most useful for them. That is, they make decisions based on *expected utility*.

# The Components of the Bayesian Decision Rule

Bayesian decision rule: maximize the expected utility of taking: action $a$ from decision set $\Theta$ with respect to the possible values of $\theta$ and the observed values of $y$. That is,

$$\hat{a} = \text{argmax}_a \int_\Theta U(a, \theta)\mathbf{P}(\theta|y)d\theta$$

- The likelihood function
- The prior
- The utility function
- The decision rule
- The production distribution

# The Parameter setting problem

- Things the learner doesn't know but would like to (parameters, $\theta$):
  - $\alpha =$ proportion of $G_{EP}$ speakers [syntactic parameter 'ON']
  - $p =$ rate of enclisis of $G_{CIP}$ speakers
- The only evidence the learner has for any given parameter is the count of inputs that support the parameter setting and a count of those that oppose it (observations, $y$).
- The task of the learner is to use these frequency counts to evaluate what is the best grammar for them (decision set, $\Theta$).

# The Likelihood Function

- Treat the data as $N$ (independent) Bernoulli trials: $S_N = \{(y_i, x_i)\}_{i=1}^{N}$.
- Let, $k$ be the number of cases that were parseable with parameter setting on. e.g. enclitics.
- The likelihood function is:

$$\mathbf{P}(S_N|\alpha, p) = \binom{N}{k}[(1-\alpha)p+\alpha]^k[(1-\alpha)(1-p)]^{N-k}$$

- $\alpha, p$ are dummies here, they aren't part of the grammar.
- **Note:** $G_{EP}$ is a subset of $G_{CIP}$ so does not present any counter-evidence for $G_{CIP}$ in this model.

# The Prior

Prior beliefs of the learner about possible combinations of $\alpha$ and $p$:

- If $\alpha = 1$ then the population is entirely made up of $G_{EP}$ speakers, the value of $p$ is irrelevant as it only applies to $G_{CIP}$ speakers.
- The simplest hypothesis is that $\alpha = 1$, $p = 1$ is a maximum. i.e. before being wiped out, $G_{CIP}$ speakers would have increasingly used enclitic constructions to fit with the rest of the population.
- Similarly, if $\alpha = 0$ then the population would most likely be using proclitic construction a large proportion of the time. So, maxima around $\alpha = 0$, $p = 0.05$ (Galves et al., 2005b).

# The Prior

As a function:

$$f(\alpha, p) = \frac{1}{c} e^{-(p-(0.95\alpha+0.05))^2}$$

where $c$ is a normalizing constant. $f(\alpha, p)$ is then just a squared Gaussian with mean $0.95\alpha + 0.05$. This is the rate of enclisis found in the Tycho Brahe corpus.

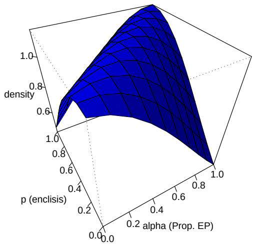



Figure: The prior density: $f(\alpha, p)$

# The Utility Function

- The learner wants to acquire the same grammar as the rest of its community.
- The learner wants to be able to play both roles of speaker and hearer successfully.
- A speaker of CIP will be able to understand both EP and CIP speakers without any penalty. However, a speaker of EP will have difficulty understanding CIP speakers. Conversely, EP speakers will be able to converse without penalty but not vice-versa.
- Assume the individual plays speaker/hearer half the time.

$$U(a, \alpha, p) = \begin{cases} -\frac{1}{2}\alpha & \text{if } a = 0 \ (G_{CIP}), \\ -\frac{1}{2}(1-\alpha) & \text{if } a = 1 \ (G_{EP}). \end{cases}$$

*This is also where we should be encoding pronounciation difficulty!*

# Utility Maximization

The learner does not actually know what $\alpha$ and $p$ are. They need to *infer* it from frequencies $k$ and $N$. Instead of trying to pin this down (or stipulate it) expected utility maximization hedges its bets. So,

$$\mathbf{E}[U(a,\alpha,p)|S_N] = \int_{[0,1]^2} U(a,\alpha,p)d\mathbf{P}(\alpha,p|S_N).$$

To find out whether the parameter should be set 'off' (and simplified), we calculate:

$$\mathbf{E}[U(0,\alpha,p)|S_N] > \mathbf{E}[U(1,\alpha,p)|S_N].$$

$$\int_{[0,1]^2} (2\alpha - 1)\mathbf{P}(S_N|\alpha,p)f(\alpha,p)d(\alpha,p) < 0$$

If this last statement is true, the learner should choose $G_{CIP}$.

# Estimating Production Rates

- Assume that production probabilities are derivable from the frequencies observed in the acquisition process.
- For a CIP speaker:

$$\mathbf{P}(a_1|x_1) = (N-k)/N,$$
$$\mathbf{P}(a_2|x_1) = k/N$$

- For an EP speaker:

$$\mathbf{P}(a_2|x_1) = 1.$$

- Let $\alpha_0$ be the proportion of $G_{EP}$ speakers observed in generation 0. Then the probability of getting the enclitic version in the first round.

$$q_0 = \mathbf{P}_{pop}(a_2|x_1, T=0) = (1-\alpha_0)p_0 + \alpha_0$$

# Over and Over...

- The proportion of speakers who will see $k$ enclitic constructions in $N$ Bernoulli trials is:

$$\binom{N}{k} q_t^c (1 - q_t)^{N-k}$$

  where $q_t$ is probability of seeing an enclitic in generation $t$.
- This proportion of speakers will then contribute enclitics with a rate of $k/N$ to the next generation, $t + 1$.

# Initial $G_{CIP}$ rate of enclisis between 60-70%

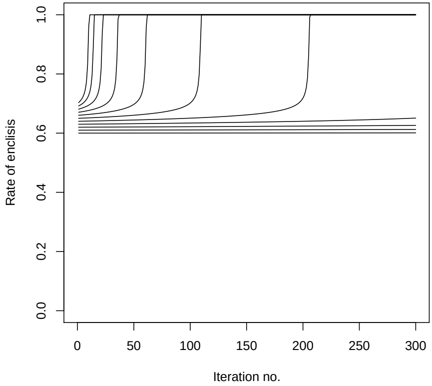



Figure: Rate of enclisis. $N = 100$, $\alpha_0 = 0$, and the initial rate of $G_{CIP}$ enclisis, $p$, ranges from 0.6 to 0.7.

*Change doesn't take off from* $p = 0.05$*!*

# More interactions?

- The stability of $G_{CIP}$ is really assumed by the model via the prior.
- Crucially, the simulation above still does not incorporate the effects of simultaneous change in other modules of language (e.g. phonology).
- Production changes? We could define a new decision problem that estimates production probabilities.

# Conclusion

Take home points:

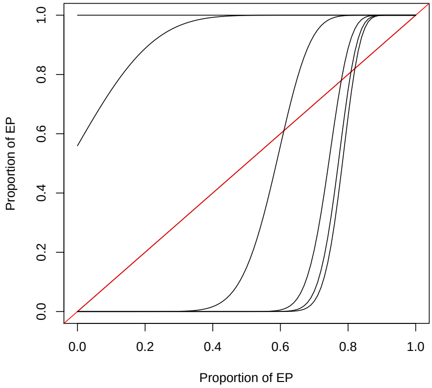
- This model articulates the general social learning model Niyogi (2006): learners learn from an (infinite) population.
- The decision procedure was presented as a utility maximizing decision rule where the learner estimates population frequencies in order to maximize communicability.
- Ideally we would look at a change in progress where we could do better estimation of the prior and utility functions.

*Thanks!*

Especially to: Charles Yang, Andrew Clausen, and Ling 575/506-ers!

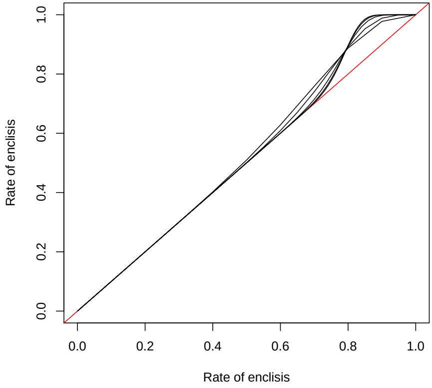
# Simulation: 300 iterations

Figure: Transition diagram: Proportion of $G_{EP}$ speakers. Different curves represent different $G_{CIP}$ enclisis rates ($p$): 0.05, 0.1, 0.2, 0.5, 0.8, and 1. $n = 100$ and $\alpha = 0$.



# Different input sizes

Figure: Transition diagram: Overall rates of enclisis. Different curves represent different input sizes *n*: $n = 10, 20, 50, 80, 100$. $\alpha = 0$



# Initial $G_{CIP}$ rate of enclisis between 60-70%

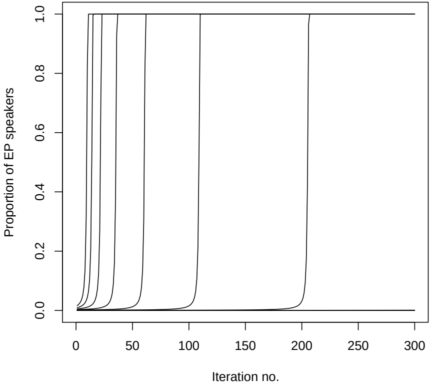



Figure: Proportions of $G_{EP}$ speakers. $n = 100$, $\alpha_0 = 0$, and the initial rate of $G_{CIP}$ enclisis, $p$, ranges from 0.6 to 0.7.

# Portuguese BIL Example

- If there is a 1-1 mapping between a meaning $x$ and a type $y$ then

$$\mathbf{P}(y|x, G) = 1 - \epsilon,$$

  if $G$ admits $x$ ($\epsilon$ is an error term).
- Let frequencies for input be: $a_1 = a, a_2 = b, a_3 = c$.
- Let, $x_1, x_3$ be the meanings associate with $a_1$ and $a_3$ respectively.
- We do not need to consider the contribution of obligatory proclisis to calculate the MLE (or MAP) grammar.

$$
\begin{aligned}
\mathbf{P}(G|Y_k, X_k) &\propto \mathbf{P}(Y_k|X_k, G)\mathbf{P}(G) \\
&= \prod_{i=1}^{k} \mathbf{P}(y_i|x_i, G)\mathbf{P}(x_i) \\
&= \mathbf{P}(a_1)^a \mathbf{P}(a_2)^b \mathbf{P}(a_3)^c \;\; \mathbf{P}(a_1|x_1, G)^{a'} \mathbf{P}(a_1|x_3, G)^{a''} \\
&\quad \mathbf{P}(a_2|x_1, G)^{b'} \mathbf{P}(a_2|x_3, G)^{b''} \\
&\quad \mathbf{P}(a_3|x_1, G)^{c'} \mathbf{P}(a_3|x_3, G)^{c''} \mathbf{P}(G)
\end{aligned}
$$

Where $a = a' + a''$ and similarly for the other frequency counts. Proclisis in affirmative sentences is simply given the error probability, $\epsilon$, in $G_{EP}$.

If we only care about finding MLE (or MAP) grammar, and taking probabilities from Nigoyi's implementation of GBM, then we have the following.

$$\mathbf{P}(G_{CIP}|Y_k, X_k) \propto \mathbf{P}(G_{CIP}) \frac{(p - \epsilon/2)}{(1 - p - \epsilon))^{a'}} \frac{((1 - 2p - \epsilon/2)}{(1 - p - \epsilon)^{b'}}$$

$$\mathbf{P}(G_{EP}|Y_k, X_k) \propto \mathbf{P}(G_{EP})(\epsilon/2)^{a'}(1 - \epsilon/2)^{b'}$$

- The explicit connection between meaning and types allows us to reduce the parameter space needed to evaluate the two grammars in question.
- We only need to parameterize the error term to do the the likelihood computation for $G_{EP}$. In general, it will allow us to focus only on types that show variation.
- The prior notwithstanding, this reduction in the parameter space is welcome in comparison with Nigoyi's implementation.
- However, this still suffers from over-parameterization the problems associated with MLE. $\mathbf{P}(a_1|x_1) = p$ is still assumed to be an innate part of the grammar.

Briscoe, E. J. (2002). Grammatical acquisition and linguistic selection. In Briscoe, E. J., editor, *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, chapter 9. Cambridge University Press.

Briscoe, T. (2000). Grammatical Acquisition: Inductive Bias and Coevolution of Language and the Language Acquisition Device. *Language*, 76(2):245–296.

Costa, J. and Duarte, I. (2002). Preverbal sbjects in null subject languages are not necessarily dislocated. *Journal of Portuguese Linguistics*, 2:159–176.

Dowman, M., Kirby, S., and Griffiths, T. L. (2006). Innateness and culture in the evolution of language. In *Proceedings of the 6th International Conference on the Evolution of Language*, pages 83–90.

Galves, A. and Galves, C. (1995). A case study of prosody driven language change: from classical to modern European Portuguese. *Unpublished MS, University of Sao Paolo, Sao Paolo, Brasil*.

Galves, C. (2003). Clitic-placement in the history of portuguese and the syntax-phonology interface. In *Talk given at 27th Penn Linguistics Colloquium, University of Pennsylvania, USA*.

Galves, C., Britto, H., and Paixão de Sousa, M. (2005a). The Change in Clitic Placement from Classical to Modern European Portuguese: Results from the Tycho Brahe Corpus. *Journal of Portuguese Linguistics*, pages 39–67.

Galves, C., de Sousa, P., and Clara, M. (2005b). Clitic Placement and the Position of Subjects in the History of European Portuguese. In Geerts, T., Ginneken, V., Ivo, and Jacobs, H., editors, *Romance Languages and Linguistic Theory 2003: Selected papers from 'Going Romance' 2003*, pages 97–113. John Benjamins, Amsterdam.

Griffiths, T. and Kalish, M. (2005). A Bayesian view of language evolution by iterated learning. In *Proceedings of the XXVII Annual Conference of the Cognitive Science Society*.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110.

Kirby, S., Dowman, M., and Griffiths, T. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, 104(12):5241.

Kirby, S. and Hurford, J. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In Cangelosi, A. and Parisi, D., editors, *Simulating the Evolution of Language*, chapter 6, pages 121–148. Springer-Verlag.

Matsen, F. A. and Nowak, M. A. (2004). Win-stay, lose-shift in language learning from peers. *PNAS*, 101(52):18053–18057.

Mitchener, W. and Nowak, M. (2003). Competitive exclusion and coexistence of universal grammars. *Bulletin of Mathematical Biology*, 65(1):67–93.

Niyogi, P. (2006). *The computational nature of language learning and evolution*, chapter 7. MIT Press.

Niyogi, P. and Berwick, R. (1998). The Logical Problem of Language Change: A Case Study of European Portuguese. *Syntax*, 1(2):192–205.

Nowak, M., Plotkin, J., and Krakauer, D. (1999). The evolutionary language game. *Journal of Theoretical Biology*, 200(2):147–162.

Yang, C. (2001). Internal and external forces in language change. *Language Variation and Change*, 12(03):231–250.