Iterated Bayesian Learning and Portuguese Clitics
Qualifying Paper II

Catherine Lai

April 29, 2008

1 Introduction

Diversity of language is a key part of our understanding of natural languages now and from the past. This diversity goes hand in hand with language change. Change is pervasive at every linguistic level. However, the space of existing languages does not appear to be unconstrained. In the modern generative tradition, this is governed by Universal Grammar (UG) (see, for example, Kroch (2000)). A number of computational models have been proposed to probe the nature of language change and role learning has within this (eg. Kirby (2001); Kirby et al. (2007); Nowak et al. (1999); Niyogi and Berwick (1998); Griffiths and Kalish (2005)). All of these studies can be grouped under the umbrella of iterated learning. Within these computational models, variation arises primarily from the individual learning mechanism and the varying input coming from the rest of the adult population.

Different studies embrace different approaches to the problem. However, all of these approaches require the learner to select a grammar from a finite data set. However, the majority of computational models of language change have generally been frequentist in nature. In particular, Maximum Likelihood Estimation (MLE) is usually employed as the decision making criterion. In this context, the learner chooses the grammar which makes the input data most likely. In actuality, this amounts to assuming that grammars have specific and probabilities associated with particular syntactic constructions. Grammar acquisition is then reduced to matching observed frequencies to these grammar intrinsic probabilities. Whatever uneasiness one might have about embedding probabilities so deeply in the grammar we can immediately note another problem associated with this sort of estimation. That is, MLE requires us to conflate several factors: innate (prior) biases, social and communicative factors, and random noise.

A natural candidate for solving such problems that has the ability of separating out these factors is *Bayesian decision theory* (c.f. Berger (1985)). That is, we set the learner the goal of choosing the grammar that will maximize their *expected utility*. This differs from the MLE based approaches mentioned above in that it takes a *subjectivist* view towards to the problem. Strong arguments have been made for the philosophical advantage of the Bayesian approach (Efron, 1986; Jaynes, 2003). The main benefit from our point of view is that the utility function explicitly takes into account issues of communicability and processing/production difficulties. This allows us to, at least theoretically, separate out such factors in a principled fashion.

This paper considers these different approaches with respect to the case of clitic position change in European Portuguese. In the end, I will argue that, while MLE is the bread and butter of

frequentist statistics, it does not appear to be the right tool for this job. Other Bayesian approaches to language change have been proposed. However even within general Bayesian frameworks, MLE is still implicitly employed (cf. Griffiths and Kalish (2005) Dowman et al. (2006) and Briscoe (2000)). They, thus, fall prey to the same difficulties that frequentist approaches do.

As such, this paper proposes a new model of Bayesian iterated learning that tries to address some of these issues. As we will, see Bayesian decision theory requires new technology, in the form of priors and utility functions, that require careful construction and consideration. The model presented here is by no means the final solution but a first attempt to lay out how this framework can be applied to the logical problem of language change. In order to motivate and empirically ground this study, the models are considered in view of the case of Portuguese clitic placement.

The Portuguese data is reviewed in Section 2. The rest of the paper is structured as follows. Section 3 reviews several models of iterated learning. Having discussed the problems with these approaches, Section 4 presents the a Bayesian decision theory model for the Portuguese data. Section 5 presents the results of several simulations based on that model. Section 6 concludes the paper.

2 Portuguese clitic change

Classical Portuguese (ClP, 16th to mid-19th century) allowed direct object clitics to appear as both enclitics and proclitics, as shown in the following examples.

(1) a. *Paolo ama-a.*
Paolo loves-her.
'Paolo loves her'

b. *Paolo a ama*

In fact, Galves et al. (2005) have found proclisis to be strongly preferred form in ClP. However, in Modern European Portuguese (EP) only the enclitic form is allowed in affirmative sentences with topics, adjuncts or referential subjects (as above). Otherwise, proclitic forms are obligatory in other syntactic contexts. In particular, proclisis is the only option in in wh-questions, in negated sentences, or sentences with quantified subjects.

(2) a. *Quem a ama.*
who her loves
'Who loves her?'

b. **Quem ama-a.*

(3) a. *Alguém a ama.*
Someone her loves
'Someone loves her'

b. **Alguém ama-a.*

We can note that ClP and EP are the same with respect to *obligatory* proclitic contexts. Also, Enclisis is mandatory in both ClP and EP in V1 contexts.

According to corpus studies (Galves et al., 2005) there was a sharp rise in enclisis in the early to mid 18th century. This change led to the current state of EP in which there is no longer a

choice between proclisis and enclisis. A prosody based analysis of this change is given by Galves and Galves (1995). On a syntactic level, the change in grammar boils down to to the weak/strong setting of the Arg feature of Comp. In that analysis, Galves and Galves argued that the change was driven by vowel reduction, resulting in a change of stress patterns.

This analysis, of course, is not the last word on this phenomena (c.f. Galves (2003), Costa and Duarte (2002)). However, the fact that there is only one feature change posited in this analysis makes it ripe for a computational treatment. Galves and Galves (1995) provide a model based on the Thermodynamic Framework from statistical mechanics. In fact, this data and analysis has been used as a case study for computational language change in Niyogi and Berwick (1998) and Niyogi (2006). There appears to be enough here to highlight the differences (and deficiencies) in current computational models of language change. The next section reviews these attempts to model this change. It also reviews other models of language change with respect to this problem.

3 Modelling Portuguese Clitic Change

3.1 Iterated Learning

The computational approaches to language change reviewed below are for the most part couched in the Iterated Learning (IL) model (Kirby, 2001; Smith et al., 2003). There, an IL model consists of sequence of agents through time. Agent A_i has a hypothesis h_i which represents their linguistic competence. A_i produces utterances U_i (linguistic performance) which becomes the primary linguistic data for A_{i+1}.

Hence, there are two main components to iterated learning: language acquisition and language production. Language acquisition treats the problem of how the an agent develops, or chooses, a linguistic hypothesis. For our purposes this hypothesis will be a grammar. Syntactic change is usually framed in terms of changes of the parameter settings that make up a natural language grammar. So, it may be possible to restrict the hypothesis space to choices between parameter settings. Indeed, this is the account suggested by Galves and Galves (1995) mentioned above for Portuguese.

Once the agent has acquired a grammar, it proceeds to produce data for the next generation. As we will see below, the production component can have varying degrees of dependence on the acquisition procedure. In general, the question of how to implement either component is wide open. However, this general framework provides a good starting point for examining the asymptotic behavior and population dynamics of language change models. A good place to start with respect to the Portuguese data is with the computational model accompanying the analysis in Galves and Galves (1995).

3.2 The Galves Batch Model

Galves and Galves (1995) situate their computational model of clitic change, the Galves Batch Model (GBM), in statistical mechanics framework. The basic idea is that prosody weighs-in on grammar selection during acquisition. In particular, stress patterns are learned before syntax. These stress patterns admit weights proportional to the amount of exposure the learner has to them in the pre-syntactic stage. So, on the one hand, in ClP proclisis appears in both affirmative sentences (4) and, for example, wh-questions (6), so these are lumped together for purposes of

stress contour weighting. On the other hand, matrix clauses of form in (5) form their own stress contour class.

(4) Paolo a ama (a_1)

(5) Paolo ama-a (a_2)

(6) Quem a ama (a_3)

For convenience, let a_1, a_2, a_3 refer to constructions of the forms in (4), (5) and (6) respectively. That is, proclitics in affirmative sentences, enclitics in affirmative sentences, and constructions with obligatory proclitics.

Each construction type given a probability proportional to the stress contour it is associated with it. For example, assume clauses of type a_1 and a_3 have weight p and clauses of type a_2 have weight q. For simplifying purposes, all other types of clauses are ignored. Then, we have the following probabilities:

$$\mathbf{P}(a_1|G_{ClP}) = \mathbf{P}(a_3|G_{ClP}) = p/(2p+q) \text{ and}$$
$$\mathbf{P}(a_1|G_{EP}) = 0,$$

where the former refers to sentences of the form (4) and (6). The ungrammaticality of the proclitic construction is reflected in the latter. The rest of the probability mass is spread between the other two clause types. Each grammar can then be thought of as a model with where parameters (the probabilities associated for each clause type) are derived from prosodic information. Grammar selection is then treated as a maximum likelihood estimation problem. That is, the learner selects the grammar that is most likely given the data seen by the learner *given grammar models* with parameters set concretely by the stress induced weights.

For simplicity Galves and Galves (1995) assume that the grammars for ClP (G_{ClP}) and EP (G_{EP}) differ only with respect to clitic placement. Proclisis is banned in affirmative contexts in EP. So, if the learner hears clauses of type (4) even once, G_{ClP} will be selected instead of G_{EP}. Hence, the authors derive the probability of the learner acquiring G_{ClP} as the probability that clause type a_1 occurs at least once in n samples. As seen above, this probability is tied to the stress-induced weightings. So, this then ties the change in clitic distribution to attested changes in the phonological system.

3.3 Galves Batch Learning as a Markov Process

Galves and Galves (1995) use the GBM model to find a theoretical critical point at which learners start acquiring G_{EP} and this grammar appears to take over. The asymptotic behavior of this system is further explored in Niyogi (2006) (based on Niyogi and Berwick (1998)). Niyogi implements the GBM as part of a iterated learning style Markov process. He assumes that at each generation some proportion of the population, α say, have G_{EP} (thus, $1-\alpha$ have G_{ClP}). The proportion that acquire G_{EP} at the next time step depends on α and the learning mechanism which is once against framed as an MLE problem.

Niyogi re-implements the GBM but abstracts away the role of prosody in the language change. In this formulation, G_{ClP} clauses involving proclisis have the same probability:

$$\mathbf{P}(a_1|G_{ClP}) = \mathbf{P}(a_3|G_{ClP}) = p,\ p \in [0,1],$$
$$\mathbf{P}(a_2|G_{ClP}) = 1 - 2p.$$

Similarly for EP

$$\mathbf{P}(a_1|G_{EP}) = 0,$$
$$\mathbf{P}(a_3|G_{EP}) = q,\ q \in [0,1] \text{ and}$$
$$\mathbf{P}(x_3|G_{EP}) = 1 - q.$$

Note, unlike the original algorithm presented by Galves et al., these probabilities hold across the board for *all* speakers of a particular grammar.

A language transition matrix derived purely from the parameters p, q and n, the number of samples available to the learner. That is, we treat the system as a Markov process with only two states: G_{EP} and G_{ClP}. As in the original GBM, this implementation assumes that learners are intolerant to noise and that a single sample of a_1 is enough to cause the learner to choose G_{ClP}. Note, this technology assumes that populations are infinite. This is implicit in the fact that the populations are represented as proportions.

Niyogi explores how this system evolves when varying p, q, n and the initial proportion of speakers of G_{EP}, α . He finds that populations consisting entirely of G_{EP} speakers remain G_{EP} speakers. However, certain parameter settings (eg. $np > 1$) induce equilibria where mixes of G_{ClP} and G_{EP} speakers stably co-exist. There are various combinations of p, q, n which will result the a G_{ClP} population completely moving to G_{EP} population (e.g. $np < q$).

Niyogi's implementation of GBM is very clean in terms of the formalization. However, the cleanness of the parameterization leads to other problems. Learners may still acquire G_{ClP} even though they do not see any instances of variational proclisis (a_1) in the data. This happens if there are *too many* instances of the type a_3 (obligatory proclisis). It is tempting to infer that this is due to the presence of proclisis in a_3. However, it is only really an artifact of the model. This would still happen if the syntax of the a_3 type was totally devoid of clitics.

An alternative Batch Subset Algorithm gets around this problem. This decision rule has G_{EP} rejected if there is even one occurence of a_1. Thus, the MLE component for the other syntactic types is removed. Niyogi show that the behaviour of this system depends on the sample size. Thus the change from G_{ClP} to G_{EP} could arise from there being a drop in the rate of clitic occurrences in general. This result is independent of q, the probability of a_3 in G_{EP}. However, the assumption still remains that p, the probability of a_1 in G_{ClP} is help constant through all speakers and times.

In the original GBM, the probabilities associated with a syntactic construction change with each generation. This leads to the eventual dying out of G_{ClP} after the external shock from the phonological system. Niyogi's interpretations are purely population based instead. Probabilities of syntactic types (e.g. p, q above) are intrinsic to the language here. So, a learner who acquires G_{ClP} will continue to use very high rate of variational proclisis (a_1) in spite of being surrounded by G_{EP} speakers. In fact, this deeply embedded into the asymptotic analysis of the model and accounts for the unstable equilibria in the model. However, this is counterintuitive. We would expect the desire to communicate would pressure speakers of G_{ClP} to lower the rate of variational proclisis in the face of multitudes of G_{EP} speakers. In summary, it appears the problems encountered by Niyogi's

implementation stem from the abstraction away from Galves and Galves' external (prosodic) input and the stronger parameterization of the grammars.

3.4 The Trigger Learning Algorithm

Niyogi also tries out Portuguese clitic placement with respect to the Trigger Learning Algorithm (TLA) of Gibson and Wexler (1994). In contrast to the GBM, the TLA is an online algorithm. Niyogi and Berwick (1996) elegantly formalizes this as a Markov process. The state space consists of possible grammars. The initial state is chosen at random. The learner stays in that state (i.e. with that grammar) until they are presented with an utterance that cannot be parsed with the current grammar (a trigger). At this point, another state is chosen at random and if the current utterance can be parsed in this state then the learner moves into that state. Nigoyi derives transition probabilities for the TLA with respect to Portuguese. It follows that G_{ClP} cannot be eliminated, in this model, if $\alpha_0 < 1/2$ (proportion of G_{EP} speakers). In fact, the system attains a stationary distribution with $\alpha_\infty \in [0, 1/2]$.[1] Since Portuguese did in fact evolve into from having a majority of ClP speakers to population of EP, this is clearly a problem.

The TLA is its memoryless. This means that changes in state only depend on the current sample. The online nature of this algorithm is attractive given children to appear to learn in increments. However, this means the learner will acquire whatever grammar is associated with the state they are in when they reach the end of the critical period for acquisition, regardless of the data they observed before this. Learners disregard all previous evidence for a particular grammar when presented with one utterance that does not fit that grammar. That is, this algorithm is not robust to noise. Theoretically, the learner will acquire the wrong grammar if they are presented with an erroneous utterance just before the end of the critical period. This reinforces the fact that models of change need to be able to deal reasonably with noise in the data.

3.5 Bayesian Updating

Briscoe (2002) presents a language acquisition model based on Bayesian updating. This is an online algorithm, Briscoe's approach is something like the TLA with memory. However, the goal here is to use all the available evidence in grammar selection in a robust fashion. Each (binary) parameter setting is associated with a probability. A parameter setting is active ('on') if its probability is greater than some threshold (e.g. 0.5 in Briscoe's implementation). The initial (prior) probabilities are left to UG to decide (as well as other cognitive factors). This presents an initial hypothesis. As each utterance is presented, the learner updates their hypothesis by changing probabilities for parameter settings. If the utterance is parseable under the current parameter setting, their associated probabilities are increased. If not, the acquisition procedure randomly flips the parameter settings of n parameters. If this results in a successful parse, then probabilities for those flipped parameter settings are reinforced. At the end of the critical period, the grammar acquired corresponds to the parameters that are active.

The updating in this acquisition processes is theoretically based on Bayes rule.

$$\mathbf{P}(h|d) = \frac{\mathbf{P}(d|h)\mathbf{P}(h)}{\mathbf{P}(d)}$$

[1] NB: Nigoyi uses α represent the proportion of G_{ClP} speakers. I am doing the opposite to fit with the analysis at the end

Where h is the hypothesized grammar and d is the observed data. Briscoe theoretically chooses the Maximum a-posteriori (MAP) parameter setting to reinforce. That is, the grammar the maximizes the posterior probability given the full data set. This is really just the grammar that maximizes likelihood of the data weighted by the grammar's prior probability. Note, in this setting, if we assume inputs to be independent, then online updating is equivalent to batch updating.

The actual implementation of this procedure presented in Briscoe (2002) differs from the theory. Here, probabilities are represented as fractions. When an utterance is successfully parsed both the numerator and the denominator are incremented by one. If the current input can only be parsed by flipping a parameter, then only the denominator of the fraction associated with that parameter is incremented. This seems like a reasonably intuitively way obtaining reinforcement but it does not have a clear relationship with Bayes rule. For some parameter P, $\mathbf{P}(P)$ is just the posterior derived from the last updates. According to Bayes rule, this is updated by multiplication with $\mathbf{P}(y|P)/\mathbf{P}(y)$. In reality, after n inputs with k unparseable inputs and $P = 1$.

$$\mathbf{P}(P|t) = \frac{num(\mathbf{P}(p)) + n - k}{den(\mathbf{P}(p)) + n}$$

It is not very obvious how to relate this to the Bayesian updating. However, Pearl (2007) derives a very similar update rule assuming that the hypotheses considered by the language learner are possible means of binomial distributions associated with particular grammars. That is, the expected probability of encountering a data point from a grammar, A say. This is equated with the probability that that grammar is the 'right' one. This is a rather indirect way of encoding the fact that learners expect their input to be noisy. This sort of updating is also very similar to model presented by Yang (2001) in terms of the use of a linear reward-penalty model for grammar selection.

In any case, this sort of update rule requires further refinement in order to be applicable to cases like Portuguese clitic change. In order to use Bayes' rule we need know how to calculate the likelihood of a syntactic type in a particular grammar (just as in the GBM). We also need to know how this should be used in the updating procedure. This is relevant to our Portuguese data if we assume proclitic and enclitic (a_1, a_2) have the same semantics.

Also, it is not clear how to derive a production model from parameter weightings. In Briscoe (2002) simply samples uniformly at random from a languages utterances to produce data for the production step. Once again, this does not seem quite right in the context of the Portuguese data where we want to model the changes in semantically equivalent types.

3.6 Bayesian Iterated Learning

A somewhat more transparent implementation of Bayesian Iterated Learning (BIL) is given by Griffiths and Kalish (2005) and Kirby et al. (2007). Bayesian iterated learning presented by these authors differs from the language models above in in explicitly considering signal and meaning pairs. Since inputs are assumed to be independent, Griffiths and Kalish use a batch process in the acquisition step. Let $Y_k = \{y_1, \ldots y_n\}$ be the utterances that an agent in generation k receives. Then, let $X_k = \{x_1, \ldots, x_n\}$ be the meanings associated with those utterances. Note that a meaning x_i can have more than one utterance associated with it. Acquisition is then based on a new instantiation of Bayes' rule where the hypothesis g is a grammar for our purposes.

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

This gives a posterior distribution over possible grammars. In Griffiths and Kalish (2005) a grammar chosen by the learner by randomly sampling a grammar from the posterior distribution. Note, as in Briscoe's approach, priors over grammars are assumed to be innate to each learner and invariable across generations. The input to each generation (*production*) is generated via another distribution $q(x)$ which describes the probability of the event that the meaning x is uttered in the real world. Importantly, this is modelled as independent of a speaker's grammar. It is not the subjective probability of x being uttered estimated by the learner.

The main concern for the authors is to look at the asymptotic behavior induced by iterated learning via Markovian techniques. They show analytically that the stationary distribution of the Markov process converges to the prior fed into the system. This result is at least counterintertuitive as it says that language change is really just a random walk through the prior distribution over possible languages.

Convergence to the prior also is at odds with other (brute force) simulations that had found emergence of language properties such as regularity (Kirby, 2001; Smith et al., 2003). In Kirby et al. (2007) the learner instead acquires the MAP grammar rather than sampling from the posterior distribution. With this selection mechanism, divergence from the prior is possible. Kirby et al. (2007) also look at selection mechanisms that are intermediate between MAP and sampling from the posterior. They test biases between holistic and regular languages. They find that regular languages can be overrepresented with respect to the prior. According to Kirby et al. prior bias is most mitigated when the MAP criterion is used for grammar selection. We might also expect biases towards specific parameter settings to be overridden with enough cultural pressure. Griffiths and Kalish (2007) shows that sampling from the posterior is a form of Gibbs sampling. On the other hand, iterated learning with MAP selection corresponds to a variant of the Expectation Maximization algorithm.

3.7 Bayesian Iterated Learning applied to Portuguese Clitics

The addition of meaning into this model what really differentiates this approach from the others above. We can see this by noting that the MAP criterion is the maximum likelihood grammar weighted by the prior. If we choose the prior to by uniform then we indeed just have the GBM. Now, we can use this conditioning to reduce the parameters required.

Note, whenever we have a one to one mapping between a meaning x and a type y then $\mathbf{P}(y|x, G) = 1 - \epsilon$ if the grammar G admits x (ϵ is an error term). This happens in our Portuguese data in cases of obligatory proclisis in ClP and EP, and the obligatory enclisis in EP. So, we do not need to consider the contribution of obligatory proclisis to calculate the MLE (or MAP) grammar. Also, we only need to parameterize the error term to do the the likelihood computation for G_{EP}.

If we only care about finding MLE (or MAP) grammar, and taking probabilities from Nigoyi's implementation of GBM (e.g. $\mathbf{P}(a_1|x_1) = p$), then we have the following.

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

where a and b are the number of time proclisis (a_1) and enclisis (a_2) were observed respectively with the meaning of an affirmative declarative (x_1).

In general, this approach allow us to focus only on types that show variation. The prior notwithstanding[2], 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. That is, $\mathbf{P}(a_1|x_1) = p$ is still assumed to be an innate part of the grammar.

3.8 Model Comparison and Requirements

Nigoyi claims his implementation shows that learners will eventually choose EP over ClP under these conditions. The case is the same for Bayesian iterated learning (except with a highly skewed prior). Rates of enclisis are known to be very low in ClP (around 5% (Galves and Paixao de Sousa, 2005)). If we can assume that the production grammar is the same as the perception grammar. We expect ClP speakers to generate mostly proclitic utterances. The change in relative frequencies of enclitic and proclitic constructions is with respect to the population as a whole. That is, due solely to proportions of EP and ClP speakers in the population. Thus, the change properties are sensitive to how the languages are parameterized, rather than to the actual usage preferences of the population. If we assume this parameterization is a part of the definition of the grammar then it is not clear why ClP should have exhibited the stability it did prior to the change.

This exposes the gap between the GBM, Nigoyi's implementation of the GBM and Bayesian iterated learning. Recall, in the GBM maximum likelihood estimation is done with respect to probabilities *derived from prosody acquisition* of individuals. This is more in tune with the pure simulation approach to iterated learning of, for example, Kirby (2001). In this sense, the original GBM is also closer to Briscoe's procedure, as frequencies of types observed in the critical period directly contribute to categorical parameter setting.

We have seen above that embedding of probabilities into the definition of the grammar causes artifacts to arise in Nigoyi's implementation. It makes sense to instead put the probabilities of different construction types, like variational enclisis and proclisis, back into the acquisition procedure. One way to go about this would be to do some estimation of type frequency and then create a decision procedure that chooses the grammar. The next step is estimate production probabilities for meanings that have more than one possible construction type associated with it. Estimating this requires some understanding of the usage of different types and how this changes over time.

When we formalize this, we would like to incorporate some formal notion of why this probability estimation procedure is important to the learner. At least part of this should come from the fact that the learner wishes to communicate effectively. That is, we want to incorporate the intuitive idea that using rare forms when frequent forms exists is disfavored. Moreover, it seems desirable to model the fact that some forms are harder to produce (and process) than others. This is exactly what Galves and Galves propose when they argue for prosody driven change in clitic placement in Portuguese. In general, we want to formulate iterated learning of grammars in a manner that can take these factors into account.

[2]Recall, we can always choose this to be uniform, although I will argue later that this is not appropriate

4 Bayesian Learning Redux

The rest of this paper presents a new model of Bayesian iterated learning based on Bayesian Decision Theory and the requirements outlined above. The appropriateness of the Bayesian approach to statistics has been well debated (Efron, 1986). Two of the approaches described above were situated in Bayesian statistics. However, all the approaches above relied (at least theoretically) on maximum likelihood estimation. This makes, for example, Niyogi's approach much more similar to that of Griffiths and Kalish than might be expected.

One of the key advantages of maximum likelihood estimation for *statisticians* is that it is a convenient and reliable way of summarizing and exploring new data (Efron, 1982). It also does not require the statistician to incorporate subjective judgements into statistical models. In short, frequentist statistics is a lot easier than its Bayesian counterpart.

A maximum likelihood estimator will converge to the corresponding statistic as the amount of data seen goes to infinity (although it can be biased in the finite). This makes it very attractive for *objective* inquiry. However, it seems that the decision problem that learners face is more subjective. As mentioned above, the desired model would be one that incorporates the idea that learners choose a grammar that they believe will be most useful for them. That is, they make decisions based on expected utility.

The following sections frame the Portuguese clitic problem in terms of Bayesian decision theory. In this model, the decision rule will be based on the estimation of two parameters: the proportion of G_{EP} speakers α, and p the rate of enclisis of G_{ClP} speakers. To evaluate the decision rule we need to define the likelihood of the data with respect to G_{EP} and G_{ClP} (the likelihood function). We also need to define the prior. In this case we will define the prior so that the higher the rate of enclisis, the higher the proportion of G_{EP} speakers is expected. The final component of the decision rule is the utility function. This encodes the fact choosing one grammar over the other has implications for communicating with the rest of the speech community. We also have to define production probabilities in order to embed this decision rule into an iterated learning setting. After all this machinery is set up, Section 5 will give the results of a simulation based on this process.

4.1 The Bayesian Decision Rule

The general form of a Bayesian decision is choosing action $\hat{a}$ from decision set Θ given an unknown parameter θ for which one has observed data y. The goal is to maximize the expected utility of taking action a with respect to the possible values of θ and the known values of y. That is,

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

We can formulate the problem for the parameter setting as follows. Now, let

$$\mathbf{P}(\text{parameter is on}) = \alpha.$$

We can interpret α as the proportion of speakers in the population who have the parameter set on. For the Portuguese data α is the proportion of G_{EP} speakers in the population. As above, we assume there is only one parameter setting differentiating G_{EP} and G_{ClP}. The only evidence the learner has for this is frequency counts of the different syntactic constructions under consideration. For any given parameter we have a count of inputs that support the parameter setting and a count of those that oppose it. The task of the learner is to use these frequency counts to evaluate what is the best grammar for them.

4.2 The Likelihood Function

In order to calculate expected utility of choosing a particular parameter setting, we want to find the likelihood of the data given that we have this much evidence for the parameter being 'on'.

Let, $G \rightarrow \{G_{ClP}, G_{EP}\}$ be a random variable that representing the grammar of the speaker contributing input to the current learner. Let $p = \mathbf{P}(a_2|x_1, G = G_{ClP})$. That is, the probability that a speaker of G_{ClP} utters an enclitic construction (a_2) in the variational case (x_1). Let S_N be the N pieces of input the learner receives. Then, what we want to evaluate is $\mathbf{P}(S_N|\alpha, p)$. That is, the probability of seeing S_N given that α of the population are G_{EP} speakers and p is the rate at which G_{ClP} speakers are producing enclitics. We can keep the simplifying assumption that each data point is independent. So we can calculate $\mathbf{P}(S_N|\alpha, p)$ as N Bernoulli trials. Let, k be the number of cases that were parseable with this parameter setting. Now, the likelihood function of seeing data S_N can be derived follows.

$$\begin{aligned}
\mathbf{P}(S_N|\alpha, p) &= \binom{N}{k}\mathbf{P}(a_2|x_1)^k\mathbf{P}(a_1|x_1)^{N-k}. \\
&= \binom{N}{k}[\mathbf{P}(G = G_{ClP})\mathbf{P}(a_2|x_1, G = G_{ClP}) + \mathbf{P}(G = G_{EP})]^k \\
&\quad \times [\mathbf{P}(G = G_{ClP})\mathbf{P}(a_1|x_1, G = G_{ClP}))]^{N-k}. \\
&= \binom{N}{k}[(1-\alpha)p + (1-\alpha)]^k[(1-\alpha)(1-p)]^{N-k}
\end{aligned}$$

Now let $a \in \{0, 1\}$ be the possible outcomes of the learner's decision. That is, $a = 1$ means that the learner sets the parameter on (G_{EP}) and $a = 0$ means that it is set off (G_{ClP}). Note, G_{EP} is a subset of G_{ClP} so does not present any counter-evidence for G_{ClP} in this model. Unlike the GBM we do not necessarily reject G_{EP} upon hearing a single affirmative proclitic construction.

4.3 The Prior

We also need to consider the prior beliefs of the learner about possible combinations of α and p, $\mathbf{P}(\alpha, p)$. There are two points to be to addressed. First, if $\alpha = 1$ then population is entirely made up of G_{EP} speakers. Then, the value of p is irrelevant as it only applies to G_{ClP} speakers. The simplest hypothesis is that $p = 1$ is a maximum when $\alpha = 1$. That is, before being wiped out G_{ClP} would have been using enclitic constructions in increasing number to fit with the rest of the population. Second, if $\alpha = 0$ then we have a population of G_{ClP} speakers and so the rate of enclisis should reflect the stable rate observed for Classical Portuguese. That is, the population would most likely be using proclitic construction a large proportion of the time. So, the maxima of when $\alpha = 0$ should be $p = 0.05$ to fit with the figures from Galves and Paixao de Sousa (2005).

In this manner, we assume a prior with density $f(\alpha, p)$ as shown in Figure 1. With form

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

where c is a normalizing constant:

$$c = \int_{[0,1]^2} f(\alpha, p)d(\alpha, p).$$

$f(\alpha, p)$ is then just a squared Gaussian with mean $0.95\alpha + 0.05$.[3] This is a simplification of matters. It allows us to express skepticism at the suggestion that remnant speakers of G_{ClP} would hold out with high proportions of proclisis. However, clearly more could be put into this prior.

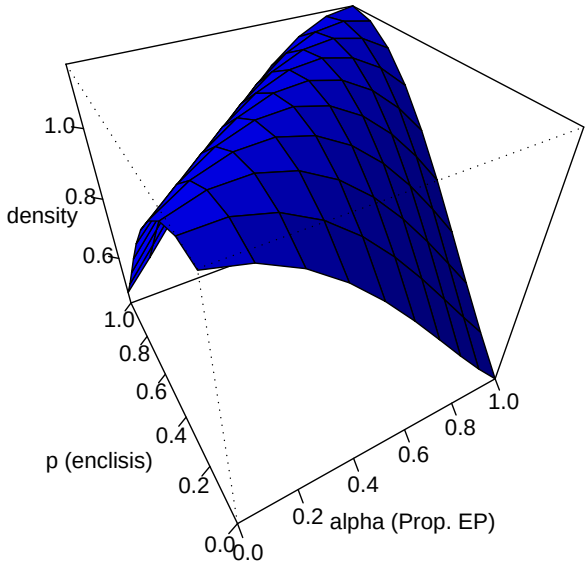

Figure 1: The prior function, $\mathbf{P}(\alpha, p)$

4.4 The Utility Function

At this stage, we want to set out a utility function that captures two facts. First, the learner wants to acquire the same grammar as the rest of its community. Second, the learner wants to be able to play both roles of speaker and hearer successfully. Since EP is a subset of ClP, a speaker of ClP will be able to understand both EP and ClP speakers without any penalty. However, a speaker of EP will have difficulty understanding ClP speakers. Conversely, EP speakers will be able to converse without penalty with ClP speakers but not vice-versa. Now, assume that an individual plays speaker and hearer roles half of the time. We can now set up the utility function as follows.

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

[3] This is the rate of enclisis found in the Tycho Brahe corpus. We really just want to represent that enclisis is available in G_{ClP} but not widely used

That is, if the learner chooses G_{ClP} they will only potentially pay from speaker point of view. This loss will be proportional to the size of G_{EP} population. Similarly, choosing G_{EP} will potentially incur a loss from the hearer side proportional to the size of the G_{ClP} population.[4]

4.5 The Decision Rule

Now, the decision task for the learner is then to choose the grammar that will maximize their expected utility. This is the crucial difference between this approach and the MAP based decision mechanisms reviewed above. To evaluate expected utility, we need to abstract away from α and p and instead integrate the utility function with respect to the posterior $\mathbf{P}(\alpha, p|S_N)$. This represents the fact that the learner does not actually know what α and p are. They need to infer it from frequencies N and k. Instead of trying to pin this down (or stipulate it) expected utility maximization hedges its bets. Thus we integrate over all possible values of $(\alpha, p) \in [0, 1]^2$.

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

Now, we can simplify this using the following instantiation of Bayes rule:

$$\mathbf{P}(\alpha, p|S_N) = \frac{\mathbf{P}(\alpha, p)\mathbf{P}(S_N|\alpha, p)}{\mathbf{P}(S_N)}.$$

So,

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

Where $f(p, a)$ is the density of the prior $\mathbf{P}(\alpha, p)$ as defined above. To find out whether the parameter should be set 'off', we calculate:

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

That is (disregarding constants),

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

This leaves us with the following concrete decision rule for this case,

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

[4] This is the place where production difficulties should come in, but this will have to wait for further development.

4.6 The Production Step

To investigate language change in an iterated learning fashion, we need to describe a production step. For the Portuguese data, this primarily means dealing with meanings with more than one construction type. As in Briscoe's model, this does not fall out immediately from decision procedure above. In fact, it seems reasonable to assume that *production* probabilities are malleable through a lifespan.[5] However, as a first step we might assume that production probabilities are derivable from the frequencies observed in the acquisition process.

This, of course, is dependent on whether the grammar licenses the construction. So after acquisition, a ClP speaker would have $\mathbf{P}(a_1|x_1) = (N-k)/N$ and $\mathbf{P}(a_2|x_1) = k/N$ On the other hand, an EP speaker would have $\mathbf{P}(a_2|x_1) = 1$. In a population analysis, we can estimate these from the G_{ClP}/G_{EP} proportions and the production frequencies of the first (parent) generation. For example, let the random variable $T = t_0^\infty$ represent the t-th generation. Let k_0 and α_0 be the proportion of variational enclitics and the proportion of G_{EP} speakers observed in generation 0 respectively. Then we have the following proportion of input data as enclitics in the first generation:

$$q_0 = \mathbf{P}_{pop}(a_2|x_1, T=0) = (1-\alpha_0)\frac{k_0}{N} + \alpha_0.$$

Iterated learning proceed as follows. The first generation of learners samples its input from the binomial distribution centered at q_0.[6] We can determine the proportion of speakers who will see $K = c$ enclitic constructions as the probability of seeing c successes in n Bernoulli trials:

$$\mathbf{P}(K = c|\alpha) = \binom{N}{c} q^c (1-q)^{N-c}.$$

The proportion of learners that will acquire G_{ClP} or G_{EP} is determinable from the decision procedure above. This proportion of speakers will then contribute enclitics with a rate of c/N to the next generation. Let $p_{2,t}^{cp}$ and $p_{2,t}^{ep}$ be the probability of enclisis (a_2) coming from G_{ClP} and G_{EP} speakers in generation t respectively. For G_{ClP} we have different proportions learners of G_{ClP} who see c enclitics in the learning phase who will produce enclitics a rate of c/N. The proportion contributed by $p_{2,t}^{ep}$ is the entire proportion of non G_{ClP} speakers for that generation.

$$p_{2,t}^{cp} = \sum_{c=0}^{N} I_c^{ClP} \binom{N}{c} q_t^c (1-q_t)^{N-c} (c/N).$$

$$p_{2,t}^{ep} = 1 - \sum_{c=0}^{N} I_c^{ClP} \binom{N}{c} q_t^c (1-q_t)^{N-c}.$$

Where $I_c^{ClP} = 1$ if the decision procedure chooses G_{ClP} when c enclitic constructions are observed. The probability of generation $t+1$ observing an enclitic is then:

$$q_{t+1} = \mathbf{P}_{pop}(a_2|x_1, T=t+1) = p_{2,t}^{cp} + p_{2,t}^{ep}$$

[5] We might want to make inferences on set windows of time, for example.

[6] Notice the similarity to Pearl (2007)!

This leaves asymptotic behaviour still very much in the hands of initial conditions. Perhaps a better way to proceed would be to think of the decision rule above as a rule for deciding which grammar to use, and hence which construction. This would fit better with models of speaker alignment. For now, however, lets just see what this approximation leads us to.

5 A Simulation

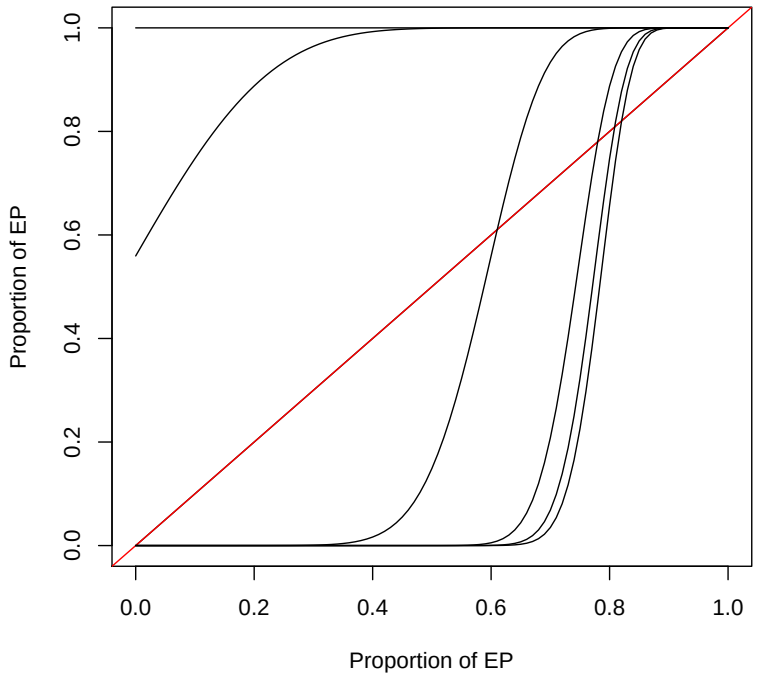

Figure 2: Transition diagram: Proportion of G_{EP} speakers. Different curves represent different G_{ClP} enclisis rates (p): 0.05 (rightmost), 0.1, 0.2, 0.5, 0.8 (leftmost), and 1. $N = 100$ and $\alpha = 0$.

Figure 4 shows the results of a simulation varying the initial condition k_0. Each simulation ran for 300 iterations starting with a population of ClP speakers. Transition diagrams for various enclisis rates (G_{ClP} speakers) is shown in Figure 2. This shows the expected proportion of G_{EP} speakers in the next time step given an existing proportion of G_{EP} speakers. If the rate of enclisis for the G_{ClP} population is 5%, then only fairly large surge G_{EP} speakers in one generation (over 60% of the population) will successfully trigger a move to G_{EP}. If the rate of G_{ClP} enclisis is high however, for example 80%, the move to G_{EP} is inevitable even for a pure G_{ClP} population. Transition diagrams for overall rates of enclisis are shown in Figure 3. This shows that, in this model, enclisis rates are fairly stable. This suggests even reasonably high rates of enclisis (e.g. 50%) will not drive any change to G_{EP}.

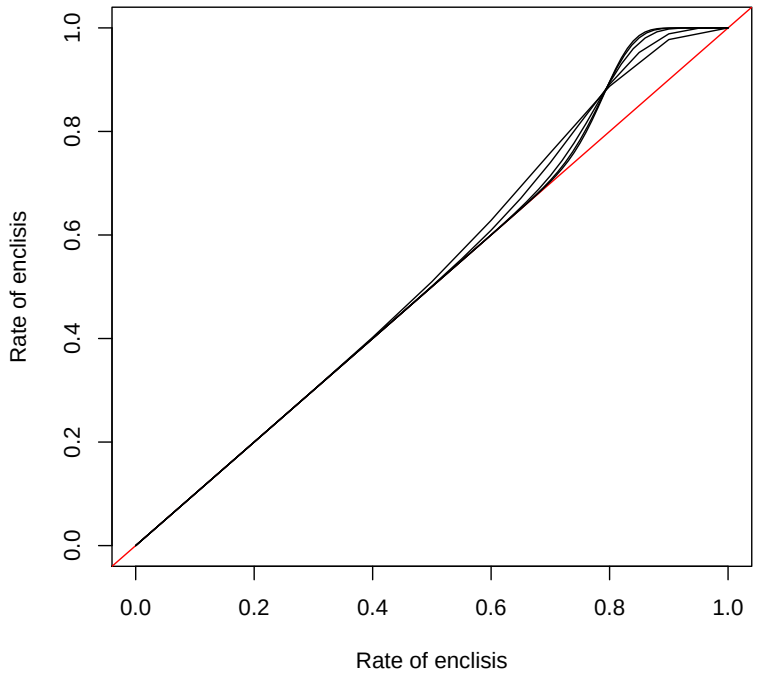

Figure 3: Transition diagram: Overall rates of enclisis. Different curves represent different input sizes N: $N = 10, 20, 50, 80, 100$. $\alpha = 0$

Figure 4 shows the results of iterated learning with input size $N = 100$, $\alpha_0 = 0$ (initially all G_{ClP}) and initial levels of G_{ClP} enclisis varying between 60% and 70%. Figure 5 shows the corresponding rates of G_{ClP} enclisis through the simulation. The jump seems to occur between $k = 63$ and $k = 64$. The former predicts that the population will stay purely ClP speakers. The latter, however, shows that sharp rise in the rate of enclisis at around 200 generations. If the initial rate of enclisis is higher than this, then converge on EP is inevitable in this system.

According to the actual frequencies observed in the Tycho Brahe corpus (Galves and Paixao de Sousa, 2005), rates of enclisis were around 0.05 in the classical period. According to the model above, this would not lead an increase in enclisis. Indeed, this is assumed by the prior, so the stability of G_{ClP} is really assumed by the model. Crucially, the the simulation above still does not incorporate the effects of simultaneous change in other modules of language (e.g. phonology). It is possible that an external shock could be enough to move the rate of enclisis to the point where increase becomes inevitable. Studies of the Portuguese data suggest that this was actually the case.

The model also ignores changes in production after the acquisition period. To account for these factors, we could define a new decision problem that estimates production probabilities. This would

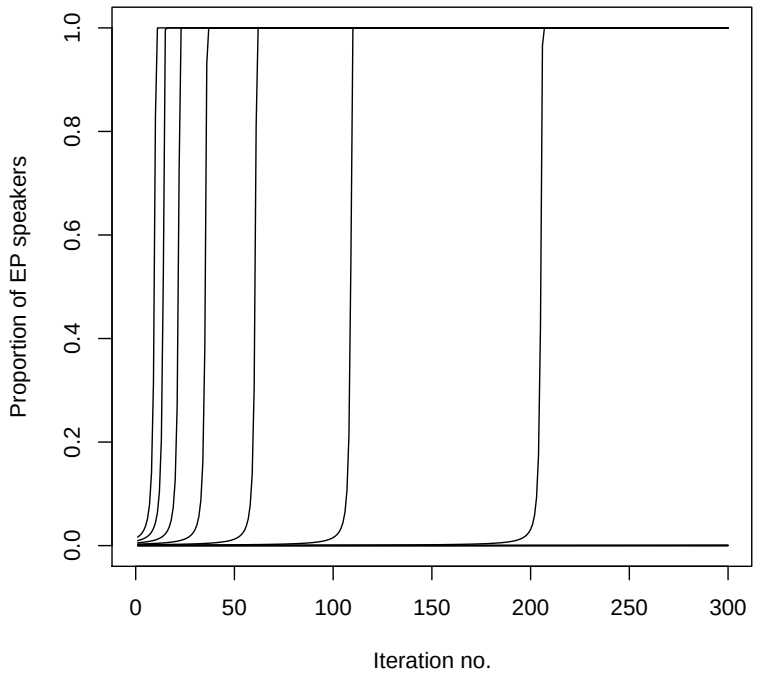

Figure 4: Iterated Learning: Proportions of G_{EP} speakers. $N = 100$, $\alpha_0 = 0$, and the initial rate of G_{ClP} enclisis, p, ranges from 0.6 to 0.7 (fastest convergence to $\alpha = 1$).

give weights to constructions derived from other modules of the language. This process would run in parallel and beyond grammar acquisition.

Clearly there are also clearly many details that need to be teased out. For example, we also need a more linguistically realistic view of the prior. However, this is not really something that should prevent a Bayesian approach to this problem. Moreover, the Bayesian framework provides a clearer framework for dealing with these hard but necessary components such as the prior.

6 Conclusion

This paper has reviewed a number of approaches to iterated learning. These have been viewed in the light of the history of Portuguese clitic placement. The loss of proclisis in affirmative contexts has been analyzed as the change in a single parameter setting by Galves and Galves (1995). This makes it a convenient testing ground for models of language change. The approaches reviewed generally relied on maximum likelihood estimation. That is, matching of observed frequencies to parameterized models of language. In Niyogi and Berwick (1998), the probabilities of licit construc-

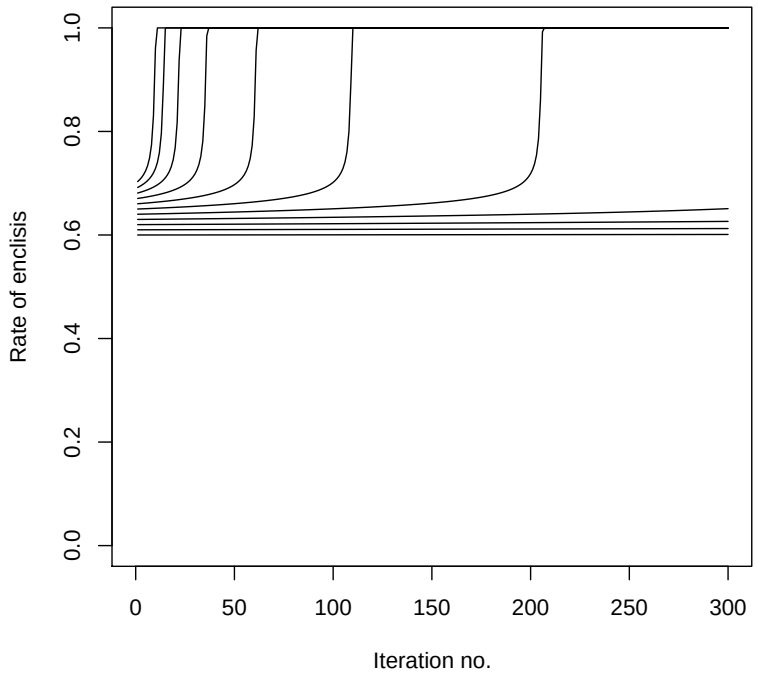

Figure 5: Iterated Learning: Rate of enclisis. $n = 100$, $\alpha_0 = 0$, and the initial rate of G_{ClP} enclisis, k/N, ranges from 0.6 to 0.7 (fastest convergence to $p = 1$.

tions actually *defined* the language alongside actual categorical parameter settings. The problems arising from this were mitigated slightly by using a Bayesian approach to iterated. However, the BIL procedures described by Kirby et al. (2007), for example, still rely on point estimation (via the MAP criterion) than true Bayesian learning.

In general, it seems that pure frequency fitting is not quite the right approach to this problem. In fact, the outcome is categorical - the learner chooses a discrete grammar.[7] This sort of problem falls fairly solidly in the realm of decision theory. The acquisition procedure of Briscoe (2002) reaches for this by putting more emphasis on modelling the changing beliefs of the learner given the input data. However, his implementation does not have a clear relationship with Bayes rule. Moreover, all of the approaches reviewed above lack a notion of utility. It is certainly treated as crucial in Bayesian decision theory.

Once we decide to change the task from estimation to decision making, the case for a Bayesian approach becomes much stronger. Bayesian decision theory has been shown to subsume frequentist decision theory. Whether or not this works in practice depends on how able we are at meeting the

[7]Or possibly a set of grammars (Kroch, 2000).

demands of a subjectivist theory. As a first step, I have outlined new version of Bayesian iterated learning. Here, the acquisition procedure is an attempt to capture Briscoe's procedure in a more transparently Bayesian way. It also incorporates the idea of utility into the procedure. However, in order to complete the IL production step, we need to reformulate production probabilities in a more sensible manner.

The simple model presented assumed agents in IL simply used the frequencies from the acquisition phase. This is clearly not enough. This needs a better grounding in what we know about production frequencies change during a lifespan. It also needs to be able to incorporate input from other modules. Only then will be able to test claims, as that of Galves and Galves, that syntactic change can be driven by prosody for example. The next obvious step is to incorporate an measure of production difficulty into the utility function. This parameter can then be estimated with respect to the corpus data as presented by Galves and Paixao de Sousa (2005). However, the correct way of fitting the output of the model to the data is not obvious and requires more attention. It is also clear that a more careful linguistic attention needs to be devoted to the development of the prior.

There is evidently a lot more work to be done. Both in differentiating between prosodic change and true syntactic change, and in developing a better model of testing this. Hopefully the Bayesian approach to the latter can help the former.

References

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer.

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.

Efron, B. (1982). Maximum Likelihood and Decision Theory. *The Annals of Statistics*, 10(2):340–356.

Efron, B. (1986). Why Isn't Everyone a Bayesian? *The American Statistician*, 40(1):1–5.

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. (2005). 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. and Paixao de Sousa, M. C. (2005). 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.

Gibson, E. and Wexler, K. (1994). Triggers. *Linguistic inquiry*, 25(3):407–454.

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.*

Griffiths, T. and Kalish, M. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, 31:441–480.

Jaynes, E. (2003). *Probability Theory: The Logic of Science.* Cambridge University Press.

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. *Proc Natl Acad Sci US A*.

Kroch, A. (2000). Syntactic change. In *The Handbook of Contemporary Syntactic Theory*, pages 699–729. Blackwell Publishing.

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

Niyogi, P. and Berwick, R. (1996). A language learning model for finite parameter spaces. *Cognition*, 61(161-193):122.

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.

Pearl, L. (2007). *Necessary Bias in Natural Language Learning.* PhD thesis, University of Maryland.

Smith, K., Kirby, S., and Brighton, H. (2003). Iterated learning: a framework for the emergence of language. *Artificial Life*, 9(4):371–386.

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