Matheus de Carvalho Proença

Bayesian language change models and a general framework

Porto Alegre, Brasil

2013

Matheus de Carvalho Proença

Bayesian language change models and a general framework

Bachelor Thesis presented in partial fulfillment of the requirements for the degree of Computer Engineer

Universidade Federal do Rio Grande do Sul

Informatics Institute

Computer Engineering

Supervisor: Aline Villavicencio, PhD

Co-supervisor: Marco Aurelio Pires Idiart, PhD

Porto Alegre, Brasil

2013

Acknowledgements

Foremost, I would like to express my gratitude to my advisor, Aline Villavicencio, who showed above average patience and dedication when I most needed it, and was able not only to guide and assist me during the development of this work with professionalism and excellence, but was also successful in inspiring me to discover a whole new study field.

Also of extreme importance to the accomplishment of this work, I am thankful to Rodrigo Wilkens, that often showed me the brightest path to achieve my tasks, and surely made them more interesting by discussing their different aspects.

I would like to show my appreciation for Robert Berwick not only for his work on the subject, but for his demonstrations of interest in my work. They surely inspired me.

Moreover, I would like to say thanks to Dan Dediu. Your help has been essential to correctly yield this work's expected results.

Last but not least, I would like to thank my family and friends for the support.

I appreciate it.

This project received the support of CNPq projects 551964/2011-1, 478222/2011-4 and 482520/2012-4.

Abstract

Language change consists in the variation of linguistic features through time. Previous work was able to account for influences of language acquisition on those changes, establishing the dynamical system grounds behind language change. This work provides a general framework able to uniformly integrate several existing approaches to the subject and via a configurable simulator it enables the assessment of different configurations of that framework.

Key-words: Natural Language Processing, Language Evolution and Change, Bayesian Agents.

Resumo

Mudança de linguagem consiste na variação das características linguísticas ao passar do tempo. Trabalhos anteriores tiveram sucesso em contabilizar influências da aquisição da linguagem nessas mudanças, estabelecendo os princípios do sistema dinâmico por trás da mudança de linguagem. Esse trabalho fornece um framework capaz de integrar uniformemente diversas abordagens existentes e através de um simulador configurável ele permite avaliar diferentes configurações desse framework.

Palavras-chaves: Processamento de Linguagem Natural, Evolução e Mudança de Linguagem, Agentes Bayesianos.

List of Figures

List of Tables

Contents

1 INTRODUCTION

The study of language change is a core subject on cognitive sciences. The understanding of its fundamental principles can enlighten the comprehension of how human cognition works. To provide support to that kind of study, computational evaluations have been used for a long time to assess mathematical models accounting for the evolution and change of different linguistic features.

Several approaches have been proposed to model language evolution and change. Initially, the main belief was that those linguistic changes were a result of our innate capacities. Different languages were stated as instances of a Universal Grammar, imposing different restrictions over the possible linguistic features, mainly due to our genetic inheritance (CHOMSKY, 1969).

More recently, it has become clear that language acquisition and social interaction can have a major role in the construction of human languages (LIGHTFOOT, 1991; NIYOGI; BERWICK, 1995; KIRBY, 2001). Lightfoot (1991) showed that the parameters of that Universal Grammar were defined through cultural interactions, and suggested that language acquisition was an important piece of that puzzle. Niyogi & Berwick (1995) were successful in demonstrating the role of language acquisition on language evolution dynamics, showing that small imperfections of individual learning patterns could precipitate population characteristics after several generations. Kirby (2001) analysed a simple population setup, verifying that several complex linguistic features could occur due to cultural exposure alone. This population setup was named *iterated learning* and was explored thoroughly in the literature (BRIGHTON, 2002; BRISCOE et al., 2002; KIRBY; HURFORD, 2002; SMITH; KIRBY; BRIGHTON, 2003; KIRBY; SMITH; BRIGHTON, 2004; GRIFFITHS; KALISH, 2005; DEDIU, 2009; NIYOGI; BERWICK, 2009; TRIJP, 2011; SWARUP; GASSER, 2009). Furthermore, different population setups (KIRBY, 2001; NIYOGI; BERWICK, 2009; NOWAK et al., 1999), and different individual learning algorithms (DEDIU, 2008; GRIFFITHS; KALISH, 2007; OLIPHANT; BATALI, 1997; SAKAS; FODOR, 2001; LIGHTFOOT, 1999; NOWAK et al., 1999; SWARUP; GASSER, 2010) were proposed.

Among all those language evolution modelling features, different behaviours may be verified. However, a direct comparison between approaches is not always possible given different data sets and assumptions made in each of these works. Therefore there is a strong need for a unified framework that incorporates them and allows them to be uniformly compared.

1.1 Motivation and Objectives

This Computer Engineering Bachelor Thesis fits in the subject of the Natural Language Processing, a major field interfacing Computer Science, Linguistics and Cognitive Sciences in which computer techniques allow to assess different scenarios produced by different models of natural language.

This work addresses an important topic in modelling language evolution and change and its main objective is to provide a computational framework for some of the main models suggested by the literature, in particular Bayesian approaches to language evolution and change. This unified framework may be used to investigate some of the essential phenomena behind the dynamic systems ruling language evolution and change.

1.2 Document Structure

This document is structured as follows.

Chapter 2 contains a brief review of the literature. The selected papers' rationale are presented, exposing the main ideas and the related case studies when appropriate.

Chapter 3 presents the unified framework for language change models and highlights how that framework relates to the main models, giving examples of which setups relate to the published papers. Furthermore, this chapter brings up unpublished similarities between the models using the framework.

Chapter 4 describes how the individual models were executed and their results compared to the original published data.

Finally, Chapter 5 displays the conclusion and future work.

2 BACKGROUND

2.1 From language acquisition to language change

Niyogi & Berwick (1995) describe how the language acquisition mechanism has influences on the distribution of different grammars in a population, inferring that language acquisition is a major driver of language change.

The logical problem of language change is that if children acquire their grammars from their teachers[1] without errors, grammatical change could never arise in a population setting, since the very same grammar would to be transmitted from one generation to another indefinitely.

However, as children are exposed to a finite number of examples of their teachers grammars, even with error free learning methods, the probability the children will acquire a grammar that is different from their teachers' is not zero.

Children will attempt to learn the previous generation's language. From those children, only some amount will actually converge to that grammar, with the other group converging to different instances. The following generation will use that mixed grammar set as source to their own learning, having mixed examples to learn from. Over all the following generations the language state will evolve accordingly, forming a linguistic dynamic system.

To model those dynamics, a discrete time system is defined, where there are several different generations, each generation with several individuals. The system is defined in way that in any given generation each individuals learns from a single individual in the previous generation (NIYOGI; BERWICK, 1995; NIYOGI; BERWICK, 1997; NIYOGI; BERWICK, 1998). Furthermore, it is defined:

- A class of grammars $\mathcal{G}$, from which each individual chooses a target;
- A set of expressions $L_g \subseteq \Sigma^*$ generated by $g \in \mathcal{G}$;
- A probability distribution P_g over L_g, with which a speaker of g produces utterances. An utterances $s \in L_g$ is, therefore, produced with probability $P_g(s)$;
- A learning algorithm $\mathcal{A}$ that an individual will use to hypothesize a grammar with some given linguistic data. For instance, an individual exposed to a n-tuple of ex-

[1] Teachers here refers to the children's source of language data. Practically speaking, the learners acquire their grammars from several different sources, e.g. their parents, other children, other caretakers, etc.

pressions $S_n = (s_1, \ldots, s_n) \in (L_g)^n$ will acquire the grammar $g = \mathcal{A}(S_n)$, defined by the map:

$$\mathcal{A} : \bigcup_{i=1}^{\infty} (\Sigma^*)^i \to \mathcal{G} \tag{2.1}$$

- A probability distribution $P^{(t)}$ over $\mathcal{G}$, such that in generation t, a speaker of $g \in \mathcal{G}$ can be found with probability $P^{(t)}(g)$;

To evaluate how the model effectively generates scenarios according to existent phenomena, Niyogi & Berwick (1998) define a setup to reproduce the linguistic results published by Galves & Galves (1995). They expose the loss of proclitic constructions in the Portuguese spoken in Portugal from 1800 to modern times.

From the 16th to the 19th century both proclisis and enclisis were possible in root affirmative sentences with non-quantified subjects:

1. **Paulo a ama.**
2. **Paulo ama-a.**

During the 19th century, the first kind of sentence ceased to be used with root affirmative sentences with non-quantified subjects, and the second one became the sole option. That did not, however, affect Wh-subject sentences, where proclisis is still used, as in 3.

3. **Quem a ama?**

The language spoken before the 19th century is henceforward named *Classical Portuguese* (CP) and the one after the 19th century *Modern European Portuguese* (EP).

To apply the model to this specific case, Niyogi & Berwick (1998) describe three stress contours, c_i, each referring to a type of production:

- c_1, to the first kind of production, as in sentence 1;
- c_2, to the second one, as in sentence 2;
- c_3, to the third one, as in sentence 3.

Also, two grammars are defined:

- Classical Portuguese, G_{CP}, where all three stress contours happen;
- European Portuguese, G_{EP}, where c_1 does not occur.

Afterwards, the work further defines the production probabilities, as given in table 1.

Stress Contour	CP	EP
c_1	p	0
c_2	$1-2p$	$1-q$
c_3	p	q

Table 1: Production probabilities per grammar

To study the evolution of the dynamical system, the population distribution of the grammars is as follows:

- The proportion of G_{CP} speakers in a generation i is given by α_i;
- The proportion of G_{EP} speakers in a generation i is given by $1-\alpha_i$.

The learning algorithm is the Maximum Likelihood Method: It chooses between G_{CP} and G_{EP} by selecting the grammar that maximizes the probability of generating the given data.

Therefore, all the modelling pattern is defined:

- The class of grammars $\mathcal{G} = \{G_{CP}, G_{EP}\}$;
- The set of expressions, represented by the stress contours, $L_g = \{c_1, c_2, c_3\}$;
- The probability distribution over $L_{G_{CP}}$, $P_{G_{CP}} = [p, 1-2p, p]$;
- The probability distribution over $L_{G_{EP}}$, $P_{G_{EP}} = [0, 1-q, q]$;
- The probability distribution over $\mathcal{G}$, $P^{(i)} = [\alpha_i, 1-\alpha_i]$;
- The learning algorithm defined as the Maximum Likelihood Method.

To analyse the individual learning algorithm to infer population dynamics, given a set of linguistic data $S_n = \{s_1, \ldots, s_n\}$, we need to calculate the likelihoods $P(S_n|G_{CP})$ and $P(S_n|G_{EP})$. Assuming that the linguistic data set was drawn using independent and identically distributed (i.i.d.) random variables:

$$P(S_n|G_k) = \prod_{i=1}^{n} P(s_i|G_k) \tag{2.2}$$

The likelihoods are, therefore, defined by the equations 2.3 and 2.4, given that the linguistic data set has a draws of c_1, b draws of c_3 and $n-a-b$ draws of c_2:

$$P(S_n|G_{CP}) = \prod_{i=1}^{n} P(s_i|G_{CP}) = p^a(1-2p)^{(n-a-b)}p^b \tag{2.3}$$

$$P(S_n|G_{EP}) = \prod_{i=1}^{n} P(s_i|G_{EP}) = 0^a(1-q)^{(n-a-b)}q^b \tag{2.4}$$

The learning algorithms define that a learner chooses G_{EP} if its associated likelihood is greater than the one associate with G_{CP}:

$$P(S_n|G_{EP}) > P(S_n|G_{CP}) \Leftrightarrow (1-q)^{(n-b)}q^b > (1-2p)^{(n-b)}p^b \tag{2.5}$$

Verifying the evolutionary dynamics of the system, the conclusions are that language learning has some accountability in language change over time.

2.2 Language Change Models: Iterated Learning

Kirby (2001) showed that an initially unstructured communication system, mapping words to some meanings, may eventually converge to a fully compositional[2] syntactic mapping through cultural iteration.

Given two characteristics of human language: (i) the human-unique (OLIPHANT, 1999) structure-preserving nature of the evolution of the mappings between signals and meanings, both in syntactic (as *compositionality*) and morphological (as *regularity*) terms; and (ii) the *stable irregularity* usually found in existing historical records; Kirby (2001) defines the Iterated Learning Model.

Iterated Learning is defined as a sequence of agents where each agent learns with a single teacher, and afterwards act as a teacher, speaking to a single agent. This process is repeated several times.

In his setup, Kirby (2001) furthermore describes the Induction and the Invention Algorithm. The first describes a parser that will update the agent's internal representation of the language, and the second will use that representation to generate output utterance strings. Both algorithms use no biological evolution, natural selection or measures of success in communicating to achieve its results.

The setup is then used in a simulation in which a meaning was represented by a pair of signals (a, b). The initial signals produced by the Invention Algorithm were randomly associated to meanings . The results show that, after only 30 generations, the language represented already has clear stable compositionality. To show the rise of irregularity on the system, he runs the setup again, this time with a non-uniform distribution over meanings.

The given setup is of major importance to the understanding of language evolution and change phenomena, since it shows the importance of pure cultural transmission on

[2] Compositionality is the characteristic of language that states that the meaning of a complex expression is determined by the meaning of its constituent expressions and the rules used to link them.

that subject, since its learning and production algorithms show no biological or individual biases.

2.2.1 Bayesian Agents and Iterated Learning

Griffiths & Kalish (2007) provide quantitative and analytic assessments about the Iterated Learning Model. To do so, they define two Bayesian Agents, the Maximum A Posteriori and the A Posteriori Sampler that are Bayesian Inference methods, that define a posterior distribution for the linguistic data exchanged in each new generation of Iterated Learning, as described below:

- S_n a set of n utterances, given as input linguistic data;
- The prior probability distribution, $P(h)$, defining the agents' biases[3];
- The posterior distribution $P(h|S_n)$, given by:

$$P(h|S_n) = \frac{P(S_n|h)P(h)}{P(S_n)} \tag{2.6}$$

where

$$P(S_n) = \sum_{h \in \mathcal{H}} P(S_n|h)P(h) \tag{2.7}$$

An agent using Maximum A Posteriori as learning method will select the hypothesis producing the biggest posterior distribution. Therefore it will choose a grammar G_w given by:

$$G_w = \arg\max_{G_i} P(G_i|S_n) \tag{2.8}$$

A given A Posteriori Sampler learner will select a grammar using a stochastic sample from the posterior distribution. In that way, grammars with bigger posterior values will be more likely selected.

Modelling Iterated Learning as a Markov Chain on data and the probability distributions hypotheses as another Markov Chain, they compare the specific dynamical systems created by Maximum A Posteriori and A Posteriori Sampler with well established statistical methods.

A Posteriori Sampler Iterated Learning corresponds to a Gibbs Sampler (GEMAN; GEMAN, 1984). The dynamics show that the Markov Chain absorbing state will correspond to the language prior probability.

[3] The prior defines how likely the agent considers the hypothesis h as true before seeing the input linguistic data. Further discussion on the interpretations of the prior is available in the work of Griffiths & Kalish (2007).

Maximum A Posteriori Iterated Learning is approximated to a Expectation Maximisation (EM) algorithm (DEMPSTER; LAIRD; RUBIN, 1977). It shows, then, that the setup still favors languages with high prior probability. The absorbing state will, however, depend on other parameters, such as the amount of noise.

2.3 Language Change Models: Social Learning

Niyogi & Berwick (2009) suggest an improvement of the Iterated Learning Model, including the fact that learners are exposed to linguistic data from multiple sources: the Social Learning (SL).

They define:

- $P^{(t)}(g)$ the probability a speaker of generation t has the grammar g;
- P_g the probability distribution over utterances of the grammar g.

Furthermore, in SL each agent learner receives linguistic data from every agent in the previous generation, i.e., data drawn from a distribution μ_t given by the equation:

$$\mu_t = \sum_{g \in \mathcal{G}} P^{(t)}(g) P_g \tag{2.9}$$

Given a learning algorithm $\mathcal{A}$, a learner will learn a grammar h with probability

$$prob[\mathcal{A}(D) = h | D \text{ drawn according to } \mu_t] \tag{2.10}$$

The proportion of speakers of h in the generation $t+1$ is also given by that probability:

$$P^{(t+1)}(h) = prob[\mathcal{A}(D) = h | D \text{ drawn according to } \mu_t] \tag{2.11}$$

To evaluate the behaviour of the proposed language change model, Niyogi & Berwick (2009) propose a choice of the learning algorithm $\mathcal{A}$, a cue learner. Given:

- Two languages L_1 and L_2;
- A set of possible utterances grouped as two kinds:
  - Utterances producing cues[4], named u_c;
  - Utterances not producing cues, named u_n.
- A target cue language[5], defined as L_1.

[4] In this example, a cue is a language characteristic that points the learner to a given language. For instance, Verb-Object constructions can be cues to a Verb-Object language.

[5] i.e., the agent will choose L_1 if it decides that it is being exposed to a cue abundant environment.

In a set $C \subseteq (L_1\ L_2)$ of n example utterances, k cue-utterances can be counted. The learner will choose L_1 if those cues occur often enough, i.e, if $\frac{k}{n} > \tau$, otherwise selecting L_2.

Niyogi & Berwick (2009) use that abstraction to assess the evolution of the proportions of Verb-Initial and Head-Initial speakers on a population.

The SL model substantially differs from IL, in that:

1. The iterated map $s_{t+1} = f(s_t)$ is generically nonlinear;
2. As an outcome of that condition, and as parameters change continuously, bifurcations are possible;
3. For the same reasons, multiple stable states are possible, allowing language stability.

2.4 Language Change Models: A Genetic Approach

Dediu (2009) proposed several approaches to assess the importance of the genetic endowment together with cultural transmission on language change.

Namely, it uses Bayesian Inference learning algorithms together with three different cultural transmission configurations:

1. **Chains of single agents**, where a chain of single learners communicate;
2. **Chains of pairs of agents**, where a chain with two agents in each iteration communicate;
3. **Complex populations**, a setup modelling the agent's lifespan, migration and generation overlapping.

To evaluate the proposed method, an example setup is given, where agents use Maximum A Posteriori (MAP) and A Posteriori Sampler (SAM) learning algorithms, as explained in the section 2.2.1.

The setup defines a genetic endowment composed by two genes, each with two alleles. Those genes generate two linguistic features, F_1 and F_2, each with two possible values, denoted by F_i and F_i^*. To ease the notation, this linguistic features are mapped as utterances of the form $u = f_1 f_2$, where f_i corresponds to the linguistic feature F_i. Hence, four possible utterances are possible (where 1 means that * is present on the feature): 00, 01, 10 and 11. This language can be described by a vector of probabilities of the four possible utterances $\mathbf{p} = [p_{00}, p_{01}, p_{10}, p_{11}]$. Also, the agents will be exposed to linguistic data represented by the vector of occurrences of the utterances $\mathbf{n} = [n_{00}, n_{01}, n_{10}, n_{11}]$.

The likelihood function used in the learning algorithm is a multinomial distribution $\mathbf{n} \sim Multinom(\mathbf{p})$. Their prior distribution is defined using the Dirichlet distribution, $\mathbf{p} \sim Dirichlet(\alpha)$. Given that those are conjugate distributions, it can be derived that the posterior follows a Dirichlet distribution, $\mathbf{p}|\mathbf{n} \sim Dirichlet(\alpha + \mathbf{n})$.

With those definitions, Dediu (2009) concludes that the priors have considerable importance on the resulting linguistic outcomes.

2.5 Language Change Models: A Game Theoretic Approach

Nowak et al. (1999) give a completely innovative Game Theory approach. They describe a setup whose language maps words to their meanings. Initially, they describe the learners, that have two matrices $P_{n\times m}$ and $Q_{m\times n}$.

P is called active matrix. Its entries p_{ij} define the probability with which a speaking agent produces the word j referring to the meaning i. The columns of that matrix sum to one ($\sum_{j=1}^{m} p_{ij} = 1$).

Q is called passive matrix. Its entries q_{ji} define the probability with which a listening agent will link the word j to the meaning i. The rows of that matrix sum to one ($\sum_{i=1}^{n} q_{ij} = 1$).

A learner listen to other agents and constructs an association matrix, A. Its elements, a_{ij}, give how often the listener has observed other individuals referring to the meaning i using the word j. Then, it builds its active and passive matrix as below:

$$p_{ij} = \frac{a_{ij}}{\sum_{l=1}^{m} a_{il}} \tag{2.12}$$

$$q_{ij} = \frac{a_{ij}}{\sum_{l=1}^{n} a_{lj}} \tag{2.13}$$

Nowak et al. (1999) describe a measure of communication efficiency, the payoff:

$$F(L_1, L_2) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}(p_{ij}^{(1)}p_{ji}^{(2)} + p_{ij}^{(2)}p_{ji}^{(1)}) \tag{2.14}$$

Furthermore, they define the agent's payoff, i.e., its ability to convey information to all other agents in the same generation, as in the equation 2.15, that gives the payoff of the Agent I:

$$F_I = \sum_{J} F(L_I, L_J) \tag{2.15}$$

Then, they use the agents' payoffs to define three cultural transmission setups:

1. **Role Model Learning** A learner selects K Role Models from the previous generation, from whom it will sample its linguistic knowledge. The Role Models are chosen proportionally to their Payoffs;

2. **Parental Learning** Each learner selects a single parent from the previous generation. The parent is selected proportionally to its Payoff;

3. **Random Learning** The agent selects K teachers from the previous generation, randomly.

The work concludes that, from an initially random linguistic setup, a population can evolve to a common language.

3 THE FRAMEWORK

Chapter 2 presented the state of art in language evolution and change modelling. Moreover, it allows the reader to grasp some understanding of the different approaches to achieve that modelling.

To enable the comparison and analysis of the different proposed solutions for modelling of language evolution and change, a general framework is proposed. The goal of this framework is to isolate the different individual components of the individual approaches, allowing the analysis of each component separately and also the constructions of novel solutions, composed by a combination of different ones.

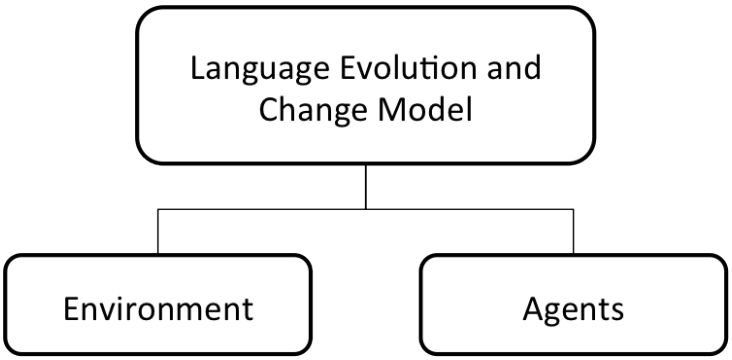


Figure 1: Proposed framework

The proposed framework, in Figure 1 divides the problem in two parts: (i) the environment and (ii) the agents, explained in the next sections.

3.1 The environment

A major part of the rationale behind modelling language change is the definition of the linguistic environment for the agents to exchange linguistic data. Changing the environment produces different data, even with identical agents (DEDIU, 2009; NIYOGI; BERWICK, 2009).

Furthermore it is possible that hearing agents give different costs or weights to linguistic data from different speakers. Therefore, the environment should also contain information about these weights.

This can be seen as a directed graph, as in Figure 2 where vertices are agents and edges are the hearer-speaker connection, with the weight as a parameter to the edge. The

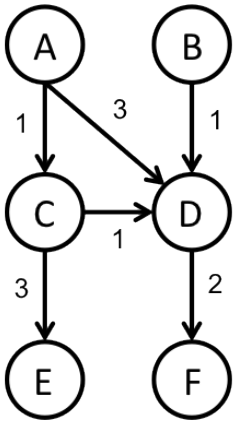


Figure 2: A sample environment

weight parameter could easily be modelled as the number of sentences communicated[1]. The graph is, therefore, modelled according to the figure 3, on which the agent A speaks n sentences to the agent B.

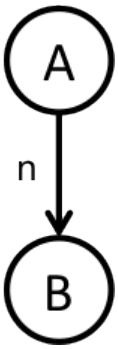


Figure 3: Environment graph logic

It should be noted that these graphs need not to be static. The information they convey is updated in the next iteration of the communication.

Dynamic graphs are useful for situations where agents change their speaking targets or change their communication weights, for instance, the parents are more important during language learning phase than later on the agent's life.

3.2 The agents

An agent is modelled as an entity receiving and producing linguistic data, as in Figure 4.

The input linguistic data is produced by the other agents in the previous iteration. For instance, in Figure 5, Agent D is exposed to the linguistic data generated by agents A, B and C in the previous iteration.

[1] This abstraction uses the hypothesis that an agent learning mechanism gives more importance to constructions heard more often.

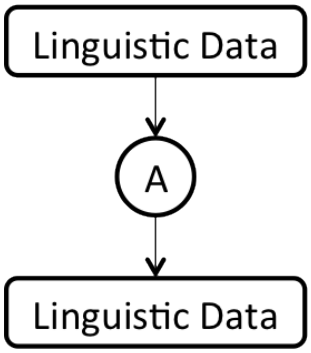


Figure 4: Agent representation

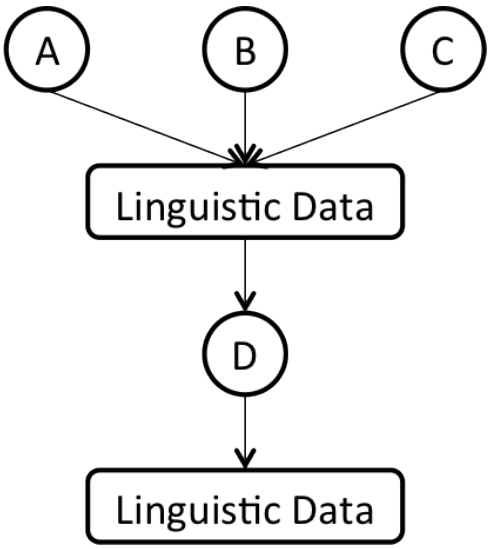


Figure 5: Input Linguistic Data formation

The agent uses this input to acquire linguistic knowledge and produce output Linguistic Data.

Linguistic Data is represented in a simplified way through storing how many times a given utterance is produced by an agent during its communication. This representation form is more often adopted quantitative models (NIYOGI; BERWICK, 2009; GRIFFITHS; KALISH, 2007; NOWAK et al., 1999; DEDIU, 2009)

3.2.1 Bayesian Agents

This work focuses on models using probabilistic methods with Bayesian Agents defined in terms of different parameters in Figure 6.

1. **Learning Method** The agent uses a learning method from Bayesian Decision Theory, as Maximum Likelihood, A Posteriori Sampler, Maximum A Posteriori, Maximum Expected Utility, etc. to estimate the hypotheses;

2. **Hypotheses** The agent's internal representation of the language in its environment

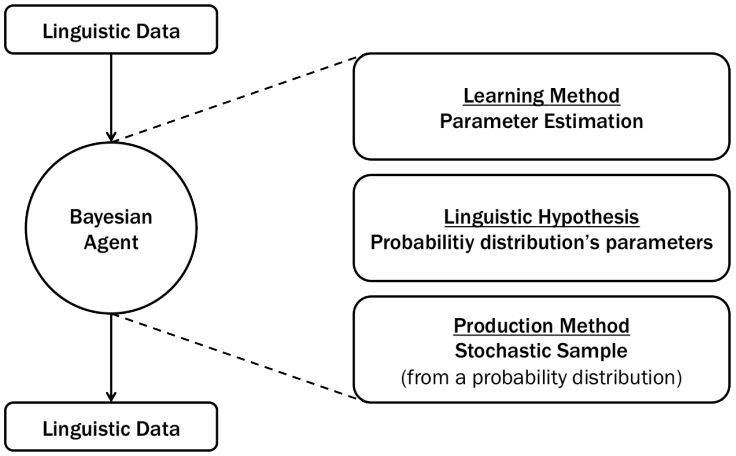


Figure 6: Bayesian Agent

is defined as a vector of probability distribution's parameters used to configure the distributions in order to produce linguistic data;

3. **Production Method** The agent generates its output Linguistic Data by drawing stochastic samples from probability distributions using its hypotheses as parameters.

3.3 Evaluation

This section evaluates the proposed framework in terms of how closely it models the approaches of Chapter 2. Similarities and differences between the individual features of those models become more clear as they are separated into the aspects covered by the framework.

3.3.1 Iterated Learning Environment

The environment described in section 2.2 is given by a set of agents distributed in generations, and a dynamical procedure to define the graph's edges, in such way that only one generation is speaking at a given iteration, and each agent speaks only once through the entire experiment. An example structure is given Figure 7.

3.3.2 Social Learning Environment

Niyogi & Berwick (2009) describe the Social Learning environment, as described in Section 2.3. This approach uses a structure graph identical to Iterated Learning, differing on the edges setup. To better illustrate that change, a sample iteration is given in Figure 8.

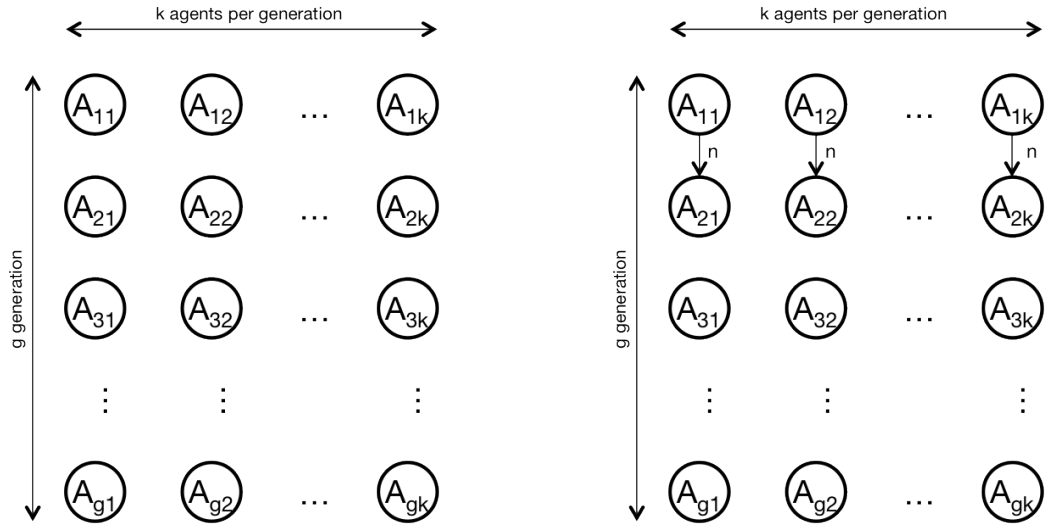


Figure 7: Iterated Learning - Structure Graph and Iteration 1

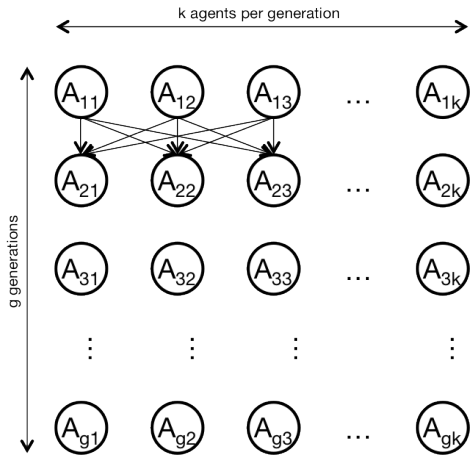


Figure 8: Social Learning - Iteration 1 (the parameter n on the edges is understood)

3.3.3 Portuguese Clitics Evaluation

Niyogi & Berwick (1998) define a setup to evaluate the proportion of Classical Portuguese speakers on a population, as a function of p, q and n, as described in Section 2.1.

The defined model uses an environment that can be mapped as Iterated Learning. Its agents are described below:

1. **Learning Method:** As described in Section 2.1, the Maximum Likelihood learning method is used. However, that method can be regarded as a Maximum A Posteriori, as follows.

A grammar chosen by a Maximum Likelihood method will be chosen according to:

$$G_w = \arg\max_{G_i \in \mathcal{H}} P(S_n|G_i) \tag{3.1}$$

A grammar chosen by a Maximum A Posterior, however, will follow:

$$G_w = \arg\max_{G_i \in \mathcal{H}} \frac{P(S_n|G_i)P(G_i)}{P(S_n)} \tag{3.2}$$

If we select a prior in such a way it follows a uniform distribution ($G_i \sim \mathcal{U}_\mathcal{H}$) over all possible grammars, the equations 3.2 and 3.1 are equivalent, since now $P(G_i) = \frac{1}{Card(\mathcal{H})} = \kappa$:

$$G_w = \arg\max_{G_i \in \mathcal{H}} \frac{P(S_n|G_i)\kappa}{\sum_{G_i \in \mathcal{H}}[\kappa P(S_n|G_i)]} = \arg\max_{G_i \in \mathcal{H}} \frac{P(S_n|G_i)\kappa}{\kappa \sum_{G_i \in \mathcal{H}}[P(S_n|G_i)]} \tag{3.3}$$

Since the likelihood function is a probability distribution, $\sum_{G_i \in \mathcal{H}}[P(S_n|G_i)] = 1$:

$$G_w = \arg\max_{G_i \in \mathcal{H}} P(S_n|G_i) \tag{3.4}$$

2. **Hypotheses:** The agent's hypotheses are defined a probability vector, as defined is the Table 1: $h_{CP} = [p, 1-2p, p]$ for Classical Portuguese speakers, and $h_{EP} = [0, 1-q, q]$ for European Portuguese speakers;

3. **Production Method:** To draw samples using those probability vectors, a Multinomial distribution must be employed.

3.3.4 Verb-Object Languages Evaluation

Niyogi & Berwick (2009) use Social Learning to assess the evolution of the proportions of Verb-Initial and Head-Initial speakers on a given population as a function of p and τ, as described in Section2.3. That abstraction defines an agent:

1. **Learning Algorithm:** The Cue-Based learning algorithm is used, as defined in Section 2.3;

2. **Hypotheses:** The agent uses a probability vector as hypothesis, defined as $p = [p(u_c), p(u_n)]$. Given p the probability a cue language speaker will produce a cue, the vector instances are $p_1 = [p, 1-p]$ for cue language speakers and $p_2 = [0, 1]$ for non cue languages speakers;

3. **Production Method:** To draw samples using those probability vectors, a Multinomial distribution must be employed.

3.3.5 Genetic Approach

Dediu (2009) proposed different setups to assess the importance of the genetic endowment on language change, as described in Section 2.4. Three environments were designed:

1. **Chains of single agents** is clearly mapped as Iterated Learning;
2. **Chains of pairs of agents** can be mapped as a limited case of Social Learning, with two agents per generation;
3. **Complex Populations** is a novel approach, that can also be mapped as a graph under the proposed framework.

To simulate his results, he also proposes an agent configuration:

1. **Learning Algorithm:** As explained in Section 2.4, the Maximum A Posteriori and the A Posteriori Sampler are used, with Dirichlet distribution as prior and Multinomial distribution as likelihood;
2. **Hypotheses:** The vector p in its definition is used as the agent's hypothesis;
3. **Production Method:** The approach supposes the learning algorithm has access to the agent's producing method to define its Likelihood distribution (DEDIU, 2009). Therefore, the learning algorithm's likelihood distribution, the Multinomial, is used as production method.

3.3.6 Game Theoretic Approach

Nowak et al. (1999) shows how stable linguistic states can arrise form initially random language basis, as described in Section 2.5. To that end they propose three cultural transmission environments, that can be modelled under the proposed framework as follows:

1. **Role Model Learning:** At each iteration, k agents are created. Then, each select K Role Models from the previous generation. Since a given Role Model cannot serve twice as teacher to the same agent, the stochastic process selecting it is without replacement. Furthermore, their parenthood probabilities are proportional to their payoffs (i.e. not uniform) requiring a non-central stochastic process. The Wallenius' Non-central Hypergeometric Distribution (WALLENIUS, 1963) is an appropriated distribution to model that process;
2. **Parental Learning:** This environment can be modelled as Role Model Learning with $K = 1$;

3. **Random Learning:** In this environment, the creation process is similar to Role Model, with the exception that the agents in the previous generation are chosen disregarding their payoffs, i.e., with an uniform distribution.

Furthermore, they define an agent:

1. **Learning Algorithm:** The learning algorithm described in Section 2.5 can be regarded as a Maximum Likelihood method, as follows:

   As originally described, A represents the agent's input Linguistic Data. As Multinomial Distribution's Maximum Likelihood Estimator is equivalent to the normalisation of the sample vector,for a data sample vector $y = [y_1, y_2, \ldots, y_n]$, the Maximum Likelihood Estimator for the parameter p_i of a Multinomial Distribution is given by:

   $$p_i = \frac{y_i}{\sum_{j=1}^{n} y_j} \tag{3.5}$$

   The game theory learning corresponds to the Maximum Likelihood Learning which in turn, as explained in Section 3.3.3, is equivalent to the Maximum A Posteriori learner with a uniform distribution as prior;

2. **Hypotheses:** For a learning agent only matrix P is important, since Q will be used to assess the agent's communication ability to form the environments, and only P will be used to generate linguistic data. Each column defines a probability vector that the agent assigns to a signal meaning pair[2]. The P matrix defines, therefore, the agent's hypotheses;

3. **Production Method:** To draw samples using P probability vectors, a Multinomial distribution must be employed.

3.4 Summary

The described approaches can be configured under the proposed abstraction as in the table below. Furthermore, those results can be laid on a plan, as in Figure 9.

Model	Environment	Learning
Kirby '01	IL	Non-Bayesian
Niyogi '98	IL	MAP[3]
Niyogi '09	SL	Cue-Based
Dediu '09 - Chain of Single Agents	IL	MAP

[2] It should be noted that every meaning will be used in a conversation transaction. (NOWAK et al., 1999)

[3] With prior given by $h \sim Uniform$

Dediu '09 - Chain of Single Agents	IL	SAM
Dediu '09 - Chain of Pairs of Agents	SL[4]	MAP
Dediu '09 - Chain of Pairs of Agents	SL[4]	SAM
Dediu '09 - Complex Populations	Complex Populations	MAP
Dediu '09 - Complex Populations	Complex Populations	SAM
Nowak '99 - Random Learning	Random	MAP[3]
Nowak '99 - Parental Learning	Role Model[5]	MAP[3]
Nowak '99 - Role Model Learning	Role Model	MAP[3]

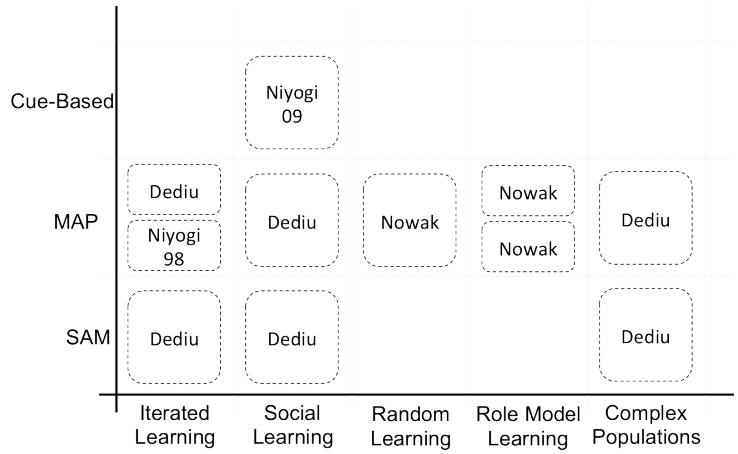


Figure 9: Analytical Results - Plan

[4] With two agents per generation.
[5] With $K = 1$.

4 EXPERIMENTAL RESULTS

To verify the correctness of both the framework described in Chapter 3 and the implementation described in Appendix A, the simulation results were compared against published results.

4.1 Portuguese Clitic Placement Change

The first test scenario uses the case described in Section 3.3.3 to verify its results against the data published in Niyogi & Berwick (1998): To assess the proportion of Classical Portuguese and European Portuguese on a population of speakers after a given number of generations each agent would acquire the most likely the grammar, as described in Section 3.3.3.

Among the possible solutions of the inequation $P(S_n|G_{CP}) \overset{?}{>} P(S_n|G_{EP})$, the case were $p < q < 2p$ was selected for this test. In that case, the state update rule is given by the equation 4.1, that states that the proportion α_{i+1} of G_{CP} speakers in the generation $i + 1$ is:

$$\alpha_{i+1} = (1 - \alpha_i p)^n \tag{4.1}$$

Niyogi & Berwick (1998) example that if $p = 0.05$, $q = 0.02$ and $n = 4$ and if the parents were all speaking Classical Portuguese ($\alpha = 1$) then the probability with which the child would pick G_{EP} (European Portuguese) is 0.66. In this case, even if the majority of children choose the grammar of European Portuguese, the speakers of Classical Portuguese will never disappear. In fact, the fixed point is 0.11. Roughly 11 percent of the population will continue to speak Classical Portuguese. (NIYOGI; BERWICK, 1998)

The results of the simulation with those parameters, in Figure 10, are consistent with the original results reported by Niyogi & Berwick (1998).

4.2 Game Theoretic Example

The second evaluation is against the results published in Nowak et al. (1999) (a game theory based approach explained in Section 3.3.6) using Parental and Random Learning. Although the original data was unavailable the simulation results with the same measures were consistent with those published, what can be verified in Figure 11 and in Appendix C.

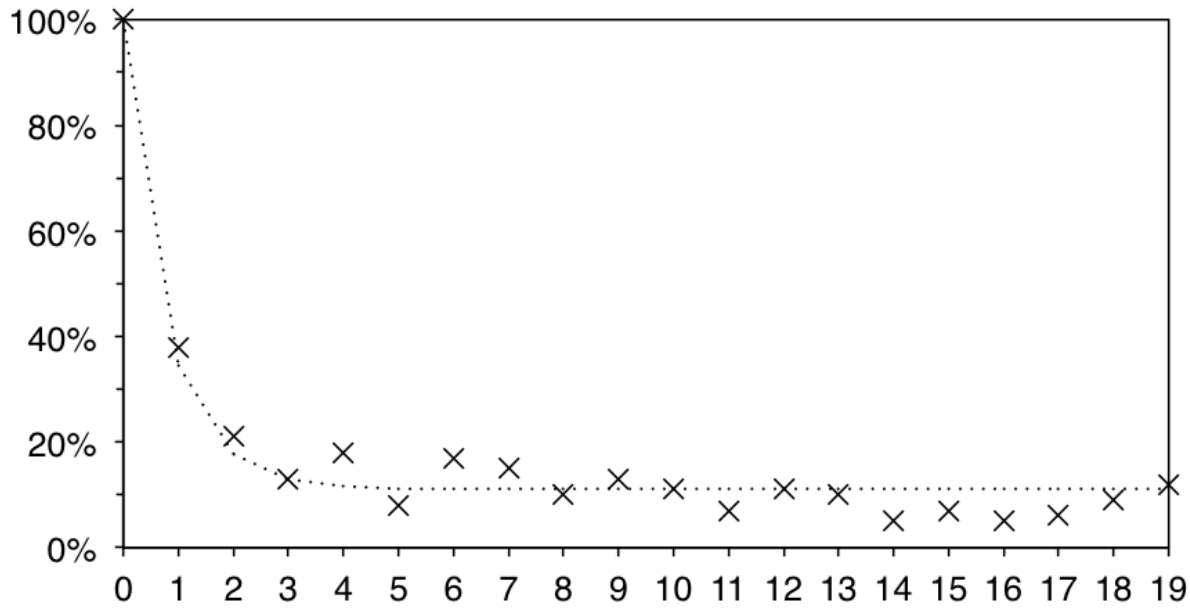
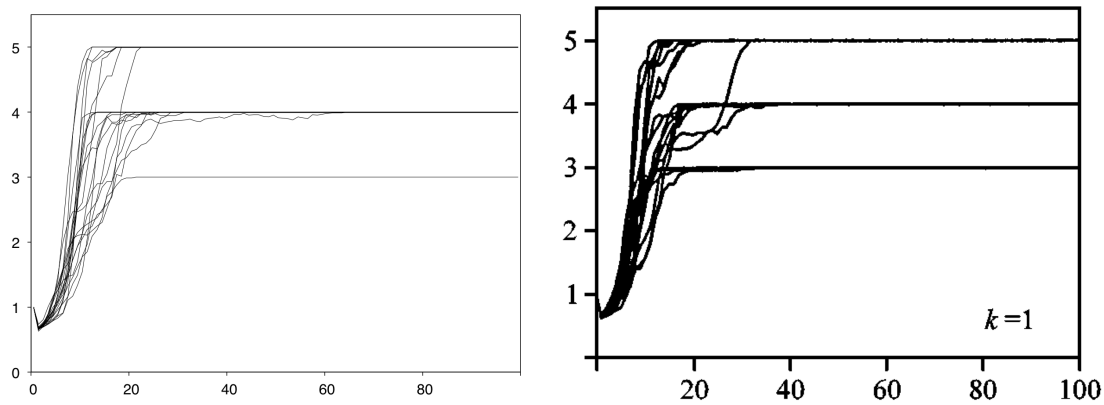


Figure 10: Portuguese Clitics Placement Change - Classical Portuguese speakers proportion (vertical axis) vs. Generation - *Dotted line displays the theoretical expected values, crosses display simulated values.*

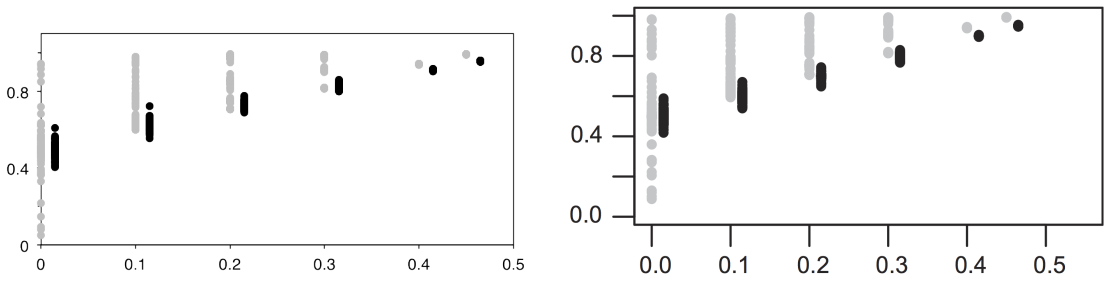


(a) Simulated (b) Published (NOWAK et al., 1999)

Figure 11: Evaluation Results - Parental Learning, $k = 1$ - Generations on horizontal axis, Normalized Payoff on vertical axis

4.3 Bayesian Inference and Iterated Learning

Dediu (2009) used Bayesian Inference learning agents with Iterated Learning to assess the importance of genetic endowments in language change, as described in Section 2.4. That approach was mapped to the proposed framework in Section 3.3.5, and the simulation suggest that the abstraction yields similar results to the published ones, as displayed in Figure 12.



(a) Simulated (b) Published (DEDIU, 2009)

Figure 12: Evaluation Results - Bayesian Inference and Iterated Learning - $mean(p_1)$ on the vertical axis, μ on the horizontal axis. **SAM** agents on black (displaced by 0.015) and **MAP** agents on grey.

5 CONCLUSIONS

In this work, a unifying framework was proposed, a simulator was built under that abstraction, and the results were validated against established models.

Given the abstract framework proposed, the results of the selected models' analysis, explaining how hey fit under the framework, and the simulation results, some conclusion may be drawn.

Firstly, it is clear that breaking monolithic models in different individual pieces allow to better understand the accountability of each phenomena. For instance, the 11 studied models can be expressed as only 5 environments and 3 learning methods.

The proposed framework is a useful tool and a valid proposal of structure defining in which pieces those models may be decomposed. Furthermore, the framework allows to build mixed approaches, using parts from initially different publications.

Finally, the simulation results show that the simulator is a correct implementation of the framework, correctly reproducing the same results as the ones originally published.

5.1 Future Work

Several expansions are foreseen for this work.

First of all, this work only models Bayesian Agents, even though the framework proposed could easily be applied to agents based on other theories.

Finally, mixed models were not produced due to the lack of comparative data for evaluation. However, the analysis of a mixed language change environment could interesting tendencies.

Bibliography

BRIGHTON, H. Compositional syntax from cultural transmission. *Artificial Life*, 2002. MIT Press, Cambridge, MA, USA, v. 8, n. 1, p. 25–54, mar. 2002. Cited on page 15.

BRISCOE, T. et al. *Linguistic Evolution through Language Acquisition*. Cambridge, UK: Cambridge University Press, 2002. Cited on page 15.

CHOMSKY, N. *Aspects of the Theory of Syntax*. Cambridge,MA: The MIT Press, 1969. Cited on page 15.

DEDIU, D. The role of genetic biases in shaping the correlations between languages and genes. *Journal of Theoretical Biology*, 2008. v. 254, n. 2, p. 400 – 407, 2008. Cited on page 15.

DEDIU, D. Genetic biasing through cultural transmission: Do simple Bayesian models of language evolution generalise? *Journal of theoretical biology*, 2009. Elsevier, v. 259, n. 3, p. 552–561, 2009. Cited 7 times on pages 15, 23, 24, 27, 29, 33, and 39.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1977. Blackwell Publishing for the Royal Statistical Society, v. 39, n. 1, p. 1–38, 1977. Cited on page 22.

GALVES, A.; GALVES, C. *A case study of prosody driven language change: From Classical to Modern European Portuguese*. Master Thesis — Universidade de São Paulo, São Paulo, Brasil, 1995. Cited on page 18.

GEMAN, S.; GEMAN, D. Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1984. IEEE, n. 6, p. 721–741, 1984. Cited on page 21.

GRIFFITHS, T. L.; KALISH, M. L. A Bayesian view of language evolution by iterated learning. In: *Proceedings of the 27th Annual Conference of the Cognitive Science Society*. Stresa,Italy: Erlbaum, 2005. p. 827–832. Cited on page 15.

GRIFFITHS, T. L.; KALISH, M. L. Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, 2007. Blackwell Publishing Ltd, v. 31, n. 3, p. 441–480, 2007. ISSN 1551-6709. Cited 4 times on pages 15, 21, 29, and 49.

KIRBY, S. Spontaneous evolution of linguistic structure – An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 2001. v. 5, n. 2, p. 102–110, abr. 2001. Cited 2 times on pages 15 and 20.

KIRBY, S.; HURFORD, J. The emergence of linguistic structure: An overview of the iterated learning model. In: CANGELOSI, A.; PARISI, D. (Ed.). *Simulating the Evolution of Language*. London: Springer Verlag, 2002. cap. 6, p. 121–148. Cited on page 15.

KIRBY, S.; SMITH, K.; BRIGHTON, H. From UG to universals: linguistic adaptation through iterated learning. *Studies in Language*, 2004. v. 28, n. 3, p. 587–607, 2004. Cited on page 15.

LIGHTFOOT, D. *How to Set Parameters: Arguments from Language Change*. Cambridge, MA: MIT Press, 1991. Cited on page 15.

LIGHTFOOT, D. *The Development of Language: Acquisition, Change, and Evolution*. Oxford: Wiley-Blackwell, 1999. Cited on page 15.

NIYOGI, P.; BERWICK, R. C. *The Logical Problem of Language Change*. Cambridge, MA, USA, 1995. Cited 2 times on pages 15 and 17.

NIYOGI, P.; BERWICK, R. C. Evolutionary consequences of language learning. *Linguistics and Philosophy*, 1997. Springer, v. 20, n. 6, p. 697–719, 1997. Cited on page 17.

NIYOGI, P.; BERWICK, R. C. The logical problem of language change: A case study of european portuguese. *Syntax*, 1998. Blackwell Publishers Ltd., v. 1, n. 2, p. 192–205, 1998. ISSN 1467-9612. Cited 4 times on pages 17, 18, 31, and 37.

NIYOGI, P.; BERWICK, R. C. The proper treatment of language acquisition and change in a population setting. *Proceedings of the National Academy of Sciences*, 2009. v. 106, n. 25, p. 10124–10129, 2009. Cited 7 times on pages 15, 22, 23, 27, 29, 30, and 32.

NOWAK, M. A.; PLOTKIN, J. B.; KRAKAUER, D. C. et al. The evolutionary language game. *Journal of Theoretical Biology*, 1999. London, New York, Academic Press., v. 200, n. 2, p. 147–162, 1999. Cited 10 times on pages 15, 24, 29, 33, 34, 37, 38, 65, 66, and 67.

OLIPHANT, M. The learning barrier: Moving from innate to learned systems of communication. *Adaptive Behavior*, 1999. v. 7, n. 3-4, p. 371–383, 1999. Cited on page 20.

OLIPHANT, M.; BATALI, J. Learning and the emergence of coordinated communication. *The newsletter of the Center for Research in Language*, 1997. v. 11, n. 1, 1997. Cited on page 15.

SAKAS, W. G.; FODOR, J. D. The strucutural triggers learner. In: BERTOLO, S. (Ed.). *Language Acquisition and Learnability*. Cambridge, UK: Cambridge University Press, 2001. cap. 5. Cited on page 15.

SMITH, K.; KIRBY, S.; BRIGHTON, H. Iterated learning: a framework for the emergence of language. *Artificial life*, 2003. v. 9, n. 4, p. 371–86, jan. 2003. Cited on page 15.

SWARUP, S.; GASSER, L. The iterated classification game: New model of the cultural transmission of language. *Adaptive Behavior - Animals, Animats, Software Agents, Robots, Adaptive Systems*, 2009. Sage Publications, Inc., Thousand Oaks, CA, USA, v. 17, n. 3, p. 213–235, jun. 2009. Cited on page 15.

SWARUP, S.; GASSER, L. The classification game: combining supervised learning and language evolution. *Connect. Sci*, 2010. Taylor & Francis, Inc., Bristol, PA, USA, v. 22, n. 1, p. 1–24, mar. 2010. Cited on page 15.

TRIJP, R. van. Can iterated learning explain the emergence of case marking in language? In: CAUSMAECKER, P. D.; MAERVOET, J.; MESSELIS, T.; VERBEECK, K.; VERMEULEN, T. (Ed.). *Proceedings of the 23rd Benelux Conference on Artificial Intelligence (BNAIC 2011)*. Ghent: KAHO Sint-Lieven, 2011. p. 288–295. Cited on page 15.

WALLENIUS, K. T. *Biased sampling: the noncentral hypergeometric probability distribution*. Stanford, California, 1963. (Technical report, 70). Cited on page 33.

APPENDIX A – Implementation

This chapter describes the implementation of the proposed framework.

A.1 Architecture

For reasons of extensibility and portability of the framework, it was implemented in Java, an Object Oriented programming language. Furthermore, a set of customised was implemented.

A.2 Linguistic Data

The Linguistic Data exchanged by the agents at each iteration is modelled using the class `BayesianLinguisticData`, as in Figure 13.

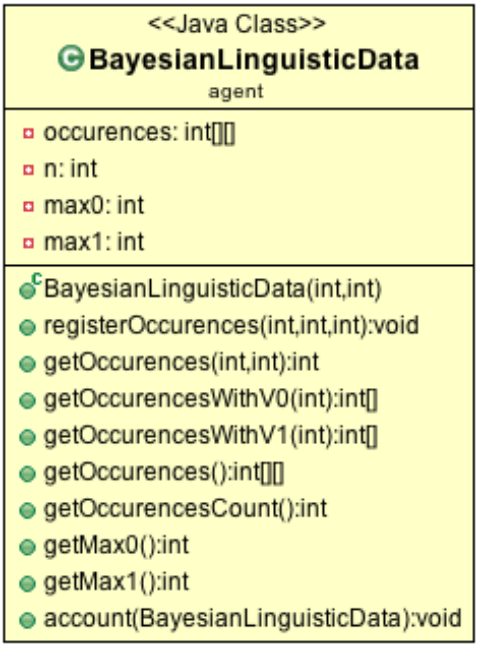


Figure 13: BayesianLinguisticData Class Diagram

An utterance is represented in the class as two integer indexes. This approach allows to account signal-meaning pair utterances, representing the signal and the meaning

as integers. Furthermore, it allows the representation of uni-dimensional language setups[1] by setting the first integer as zero.

The class then stores another integer for each utterance, representing how many times that specific utterance has been used in the communication represented by this Linguistic Data.

A.3 Agents

The agents have been implemented using four distinct classes, displayed in the Figure 14.

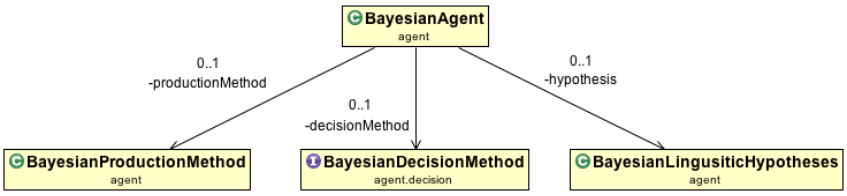


Figure 14: Agent Class Structure

The main class implementing the agent's behaviour is `BayesianAgent` abstract class. Beyond its Learning Method, Hypotheses and Production Method, the agent has a field called `memory`, where it stores its input Linguistic Data. That field is updated through the `listen` method, that takes Linguistic Data as argument. In that way, the agent can listen to several different speakers, constructing its memory.

A.3.1 Learning Method

The agent's Learning Method is modelled through the `BayesianDecisionMethod` interface. It has a single method `decide`, that takes LinguisticData as argument. Its output are Linguistic Hypotheses, i.e. the internal representation of the language the agent has learned observing the given Linguistic Data.

A.3.2 Linguistic Hypotheses

The class `BayesianLinguisticHypotheses` models the agent's Linguistic Hypotheses. It contains several probability distribution parameters.

A.3.3 Production Method

The class `BayesianProductionMethod` is responsible for producing the agent's output Linguistic Data. It stores several probability distributions. Those will be used in

[1] As, for instance, the one employed in the case described in the section 2.1.

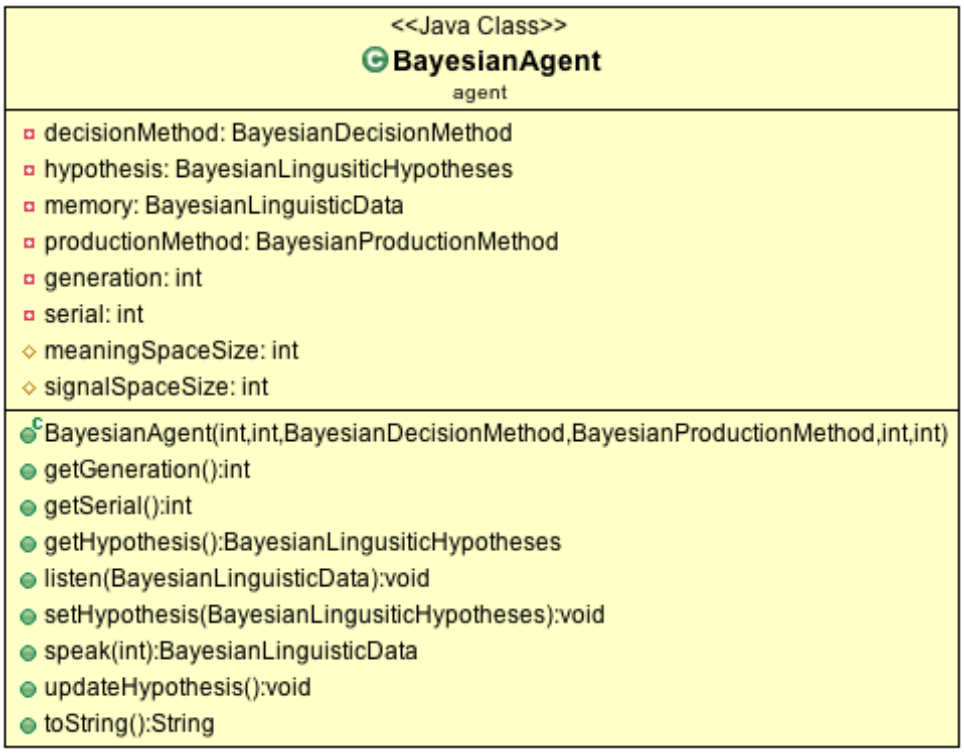


Figure 15: BayesianAgent Abstract Class Diagram

Figure 16: BayesianDecisionMethod Interface Diagram

the method `produce`, on which they will use the parameters passed as argument to draw n samples each, producing the output Linguistic Data. n is also passed as argument. The number of probability distributions is equal to the language's first dimension size.

A.3.4 Implemented Agents

To be able to run a simulation, a key parts of an Agent need to be implemented: The `BayesianLearningMethod` interface. Some models were selected and their rationale, as described in the section 3.3, was implemented as follows.

Four learning methods were implemented. Their class structure is displayed in Figure 19.

A.3.4.1 Bayesian Inference Methods

To implement the methods described on Griffiths & Kalish (2007), four classes were designed.

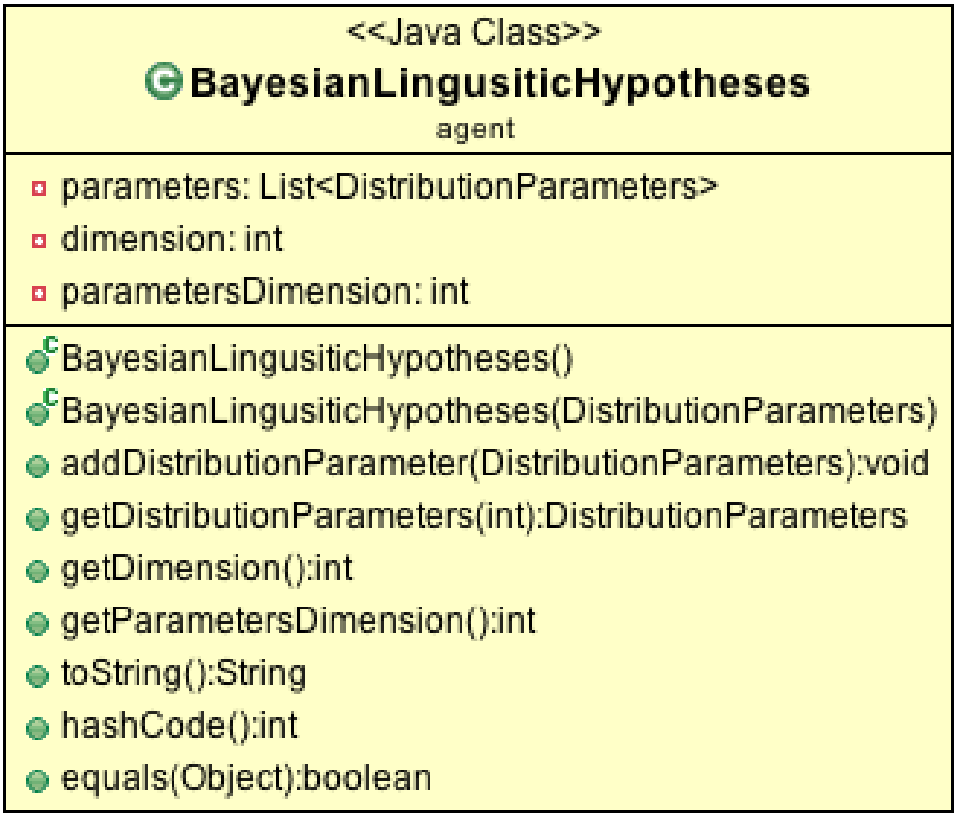


Figure 17: BayesianHypotheses Class Diagram

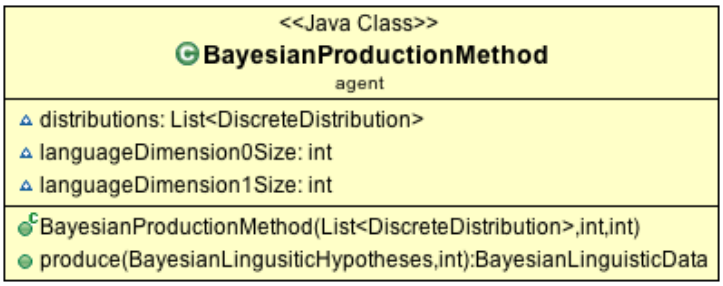


Figure 18: BayesianProductionMethod Class Diagram

The first, `BayesianInference`, contains the structure behind those learning methods. It stores the Posterior probability distribution and contains two important methods. `posteriorParameters` decides the parameters to be set on the posterior distribution, using the agent's input Linguistic Data and the Prior distribution's parameters as argument. `decide` effectively decides the output Linguistic Data distribution parameter's, using the posterior distribution parameter's as argument.

The second, `MultinomialDirichlet`, implements the `posteriorParameters` method, using a Dirichlet distribution as Prior, and a Multinomial distribution as likelihood. Since those are conjugate distributions, there as is closed formula to that operation. Specifically,

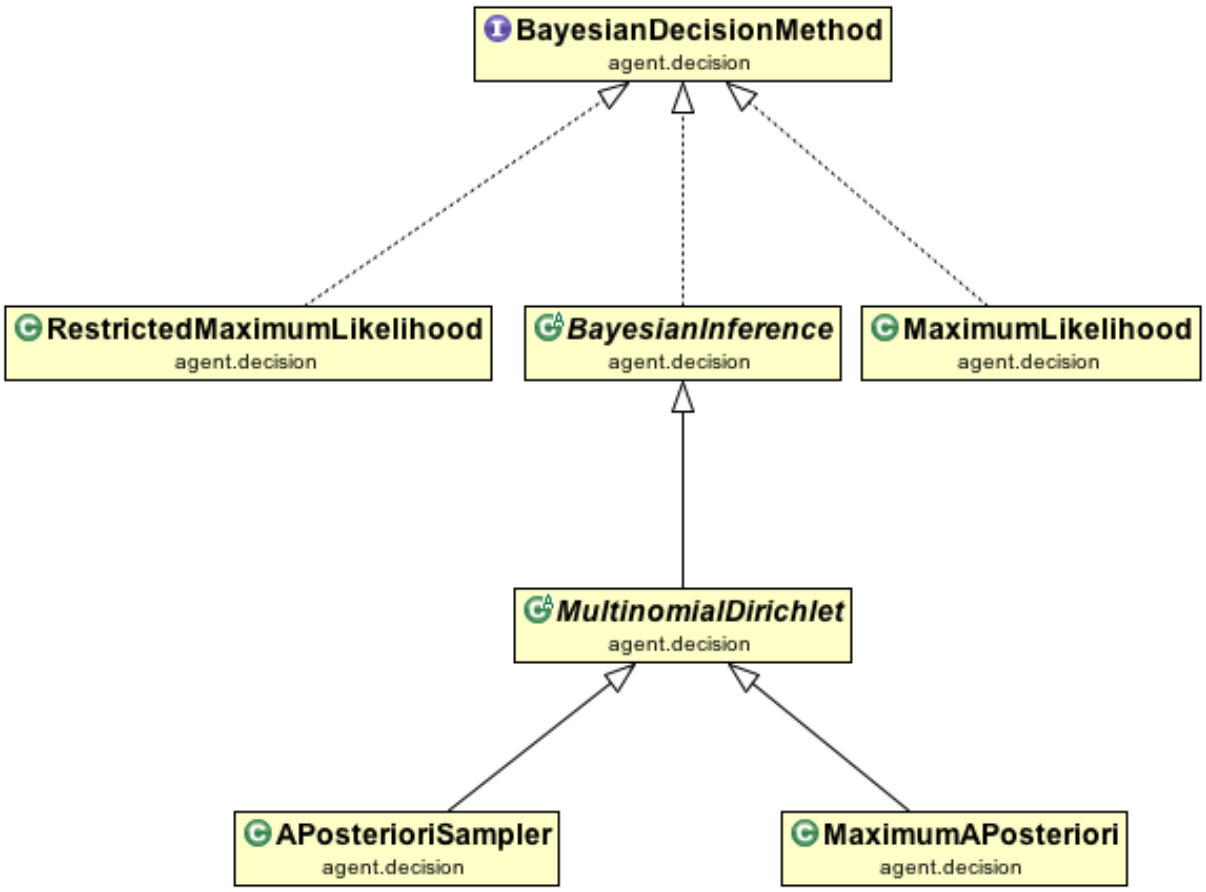


Figure 19: Learning Methods Class Structure

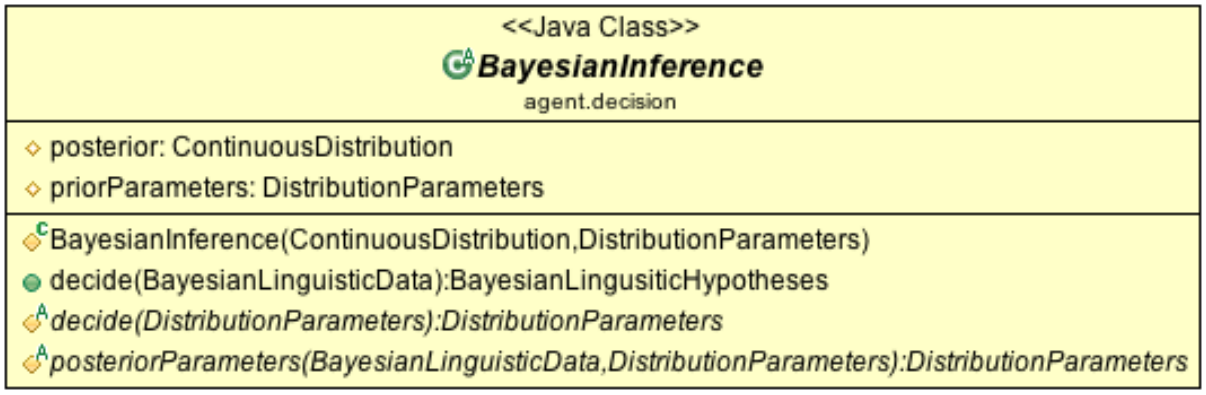


Figure 20: BayesianInference Abstract Class Diagram

given:

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)} \tag{A.1}$$

If $h \sim Dirichlet(\alpha)$ and $d \sim Multinomial(h)$, then $h|d \sim Dirichlet(\alpha+d)$. Therefore, the method returns Dirichlet parameters defined as the sum of the prior parameters with the frequency of the utterances on the input Linguistic Data.

The third class, `MaximumAPosteriori` defines the `decide` method. It defines that the output Linguistic Data's distribution's parameters are the ones that maximize the Posterior Distribution. Therefore, it returns the mode of the Distribution.

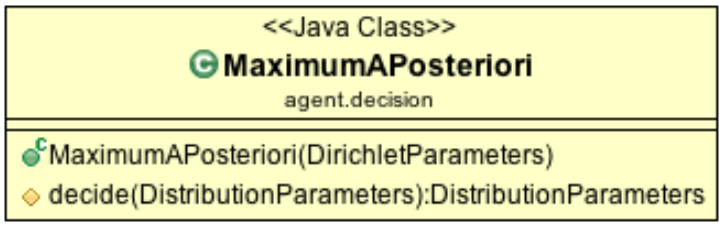


Figure 21: MaximumAPosteriori Class Diagram

Finally, the fourth class, `APosterioriSampler`, also defining the `decide` method, does so by simply drawing a stochastic sample from the posterior distribution.

Figure 22: APosterioriSampler Class Diagram

A.3.4.2 Maximum Likelihood

As explained in section 3.3.3, Maximum Likelihood Learning Method can be regarded as Maximum A Posteriori with an Uniform distribution as Prior.

However, to use that approach without restricting the choice on the Likelihood distribution, a numerical approach would be necessary. This happens because in order to analytically construct the Posterior distribution on Bayesian Inference methods, the use of conjugate distributions is necessary[2].

[2] For instance, the Beta distribution is the conjugate prior to Bernoulli and Binomial Likelihoods, as is Dirichlet to Categorical and Multinomial.

Even though the Uniform distribution is somewhat general, and can be regarded as specific cases of other distributions[3], it still limits the choice of the Likelihood distribution to the conjugate of those Priors.

For those reasons, the Maximum Likelihood method was implemented under a different class inheritance, saving the choices of Likelihood distributions.

The class `MaximumLikelihood` implementing the named Learning Method, does so by using distribution probabilities' Maximum Likelihood Estimator.

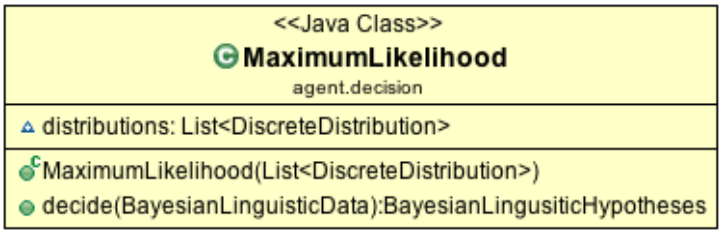


Figure 23: MaximumLikelihood Class Diagram

A.3.4.3 Maximum Likelihood with Restricted Parameter Space

The Maximum Likelihood Learning method with constraints on the Parameter space is implemented by the `RestrictedMaximumLikelihood` class. It uses a distribution probability likelihood function to select a language among a set of possible target languages.

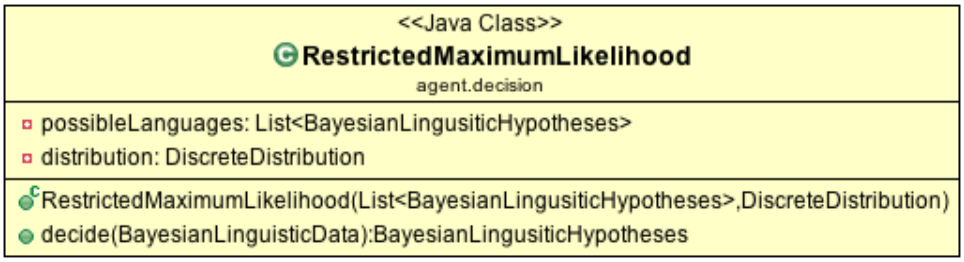


Figure 24: RestrictedMaximumLikelihood Class Diagram

A.4 Environment

The environment is mainly represented by the abstract class `Environment`. It contains a graph[4] and two important methods. The first, `iterate` iterates the environment,calling `updateEnvironment` to put its structure to date, and producing the communication between the agents, following the edges. `updateEnvironment` is an abstract

[3] For instance, using Beta distributions with $\alpha = \beta = 1$, or using Dirichlet distribution with $\forall \alpha_i, \alpha_i = 1$.
[4] Whose data structures are given by the `JGraphT` library.

method that is responsible to make any changes to the Environment structure between iterations.

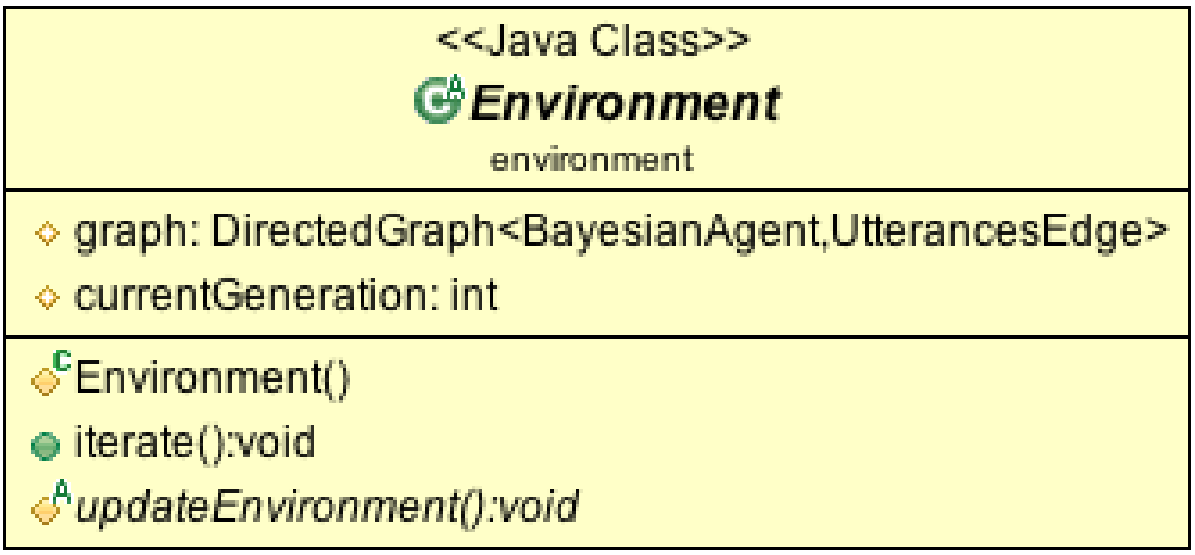


Figure 25: Environment Abstract Class Diagram

The environment graph edges are constructed using the `UtterancesEdge` class. It is solely a wrapper to the integer parameter of the edges, representing the number of sentences exchanged.

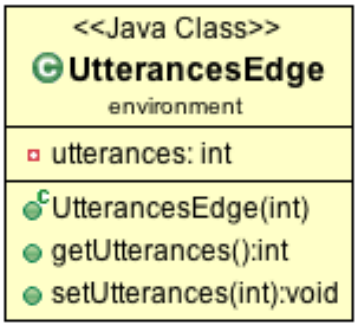


Figure 26: UtterancesEdge Class Diagram

A.4.1 Implemented Environments

Four distinct environments were implemented, using a structure composed by eight classes, as displayed in the figure 27.

All the implemented environments share the inheritance to `NonOverlappingGenerations` class, modelling environments with well-defined, separated generations. It defines the `updateEnvironment`, that calls `setEdges`, that will update the edge relationship between the agents, `storeGenerationHypothesis`, for storage of the agents metrics, and `initFirstGenerationHypothesis`, for initialising the environment.

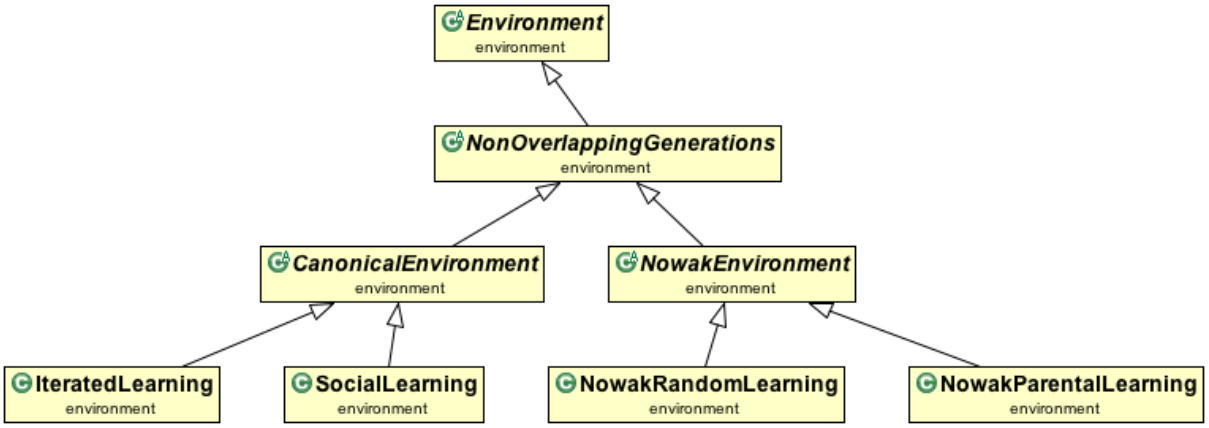


Figure 27: Environments Class Structure

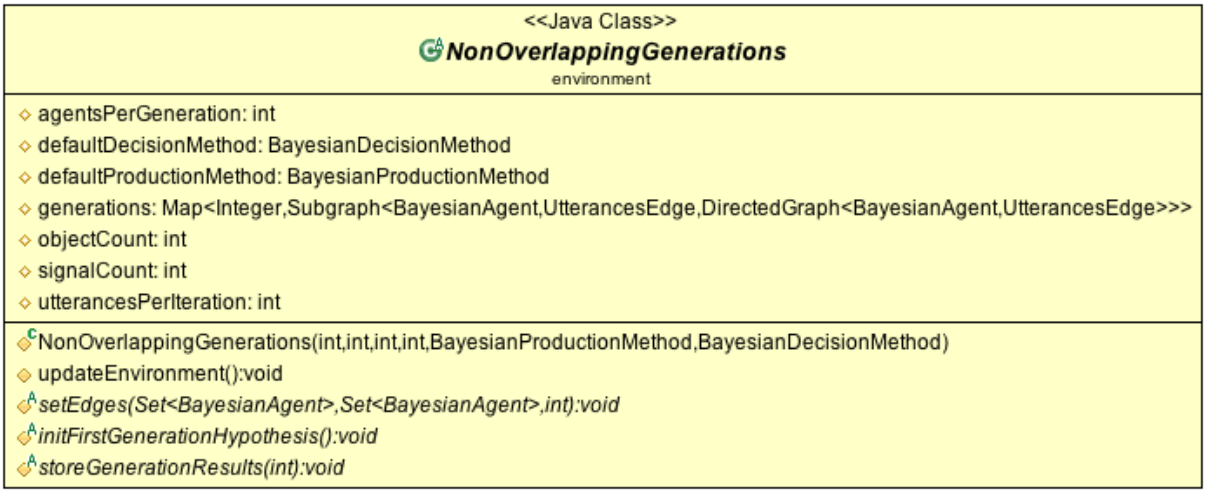


Figure 28: NonOverlappingGenerations Abstract Class Diagram

A.4.1.1 Canonical Environments

Some common features among Iterated Learning and Social Learning were implemented on the `CanonicalEnviroment` abstract class. It implementes the initialisation method using Hypotheses passed as arguments on the constructor, and the agent's metrics storage, saving every agent Hypotheses to further analysis.

A.4.1.2 Iterated Learning

Iterated Learning environments are constructed using the `IteratedLearning` class, that implements the `setEdges` method as required, associating a parent agent to each child agent.

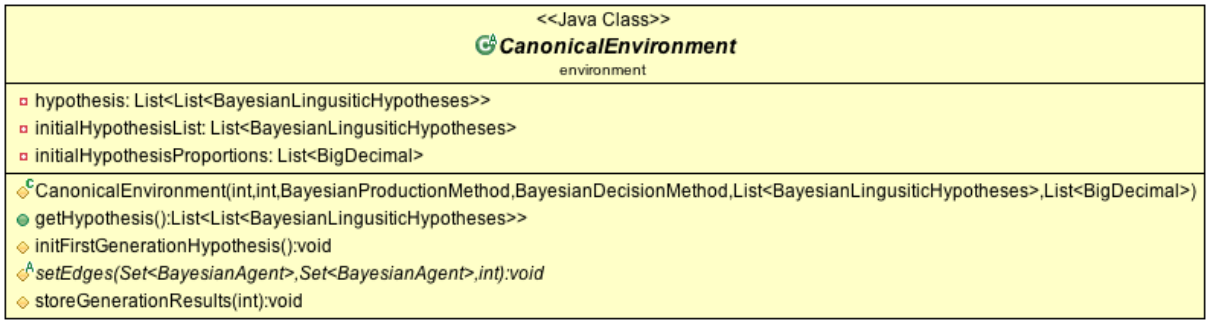


Figure 29: CanonicalEnviroment Abstract Class Diagram

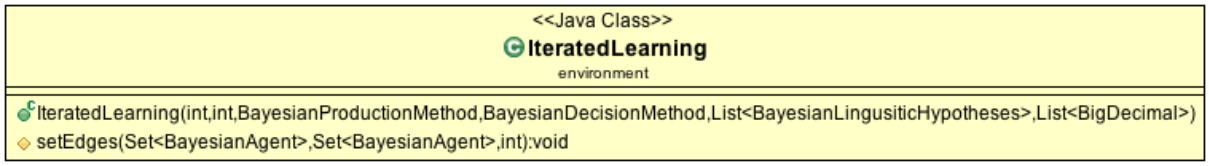


Figure 30: IteratedLearning Class Diagram

A.4.1.3 Social Learning

Social Learning is implemented on the `SocialLearning` class, implementing the `setEdges` in way that a child agent is connected to every parent agent.

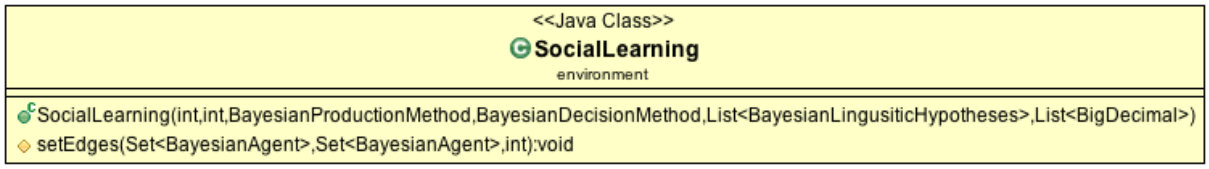


Figure 31: SocialLearning Class Diagram

A.4.1.4 Nowak Environments

Parental Learning and Random Learning have several common features, implemented on `NowakEnvironment` abstract class. It the defines the initialisation method, defining random initial hypotheses. Furthermore it stores the generations' resulting payoff.

A.4.1.5 Random Learning

Random Learning is implemented using the `NowakRandomLearning` class. It implements the `setEdges` method using random parenthood probabilities.

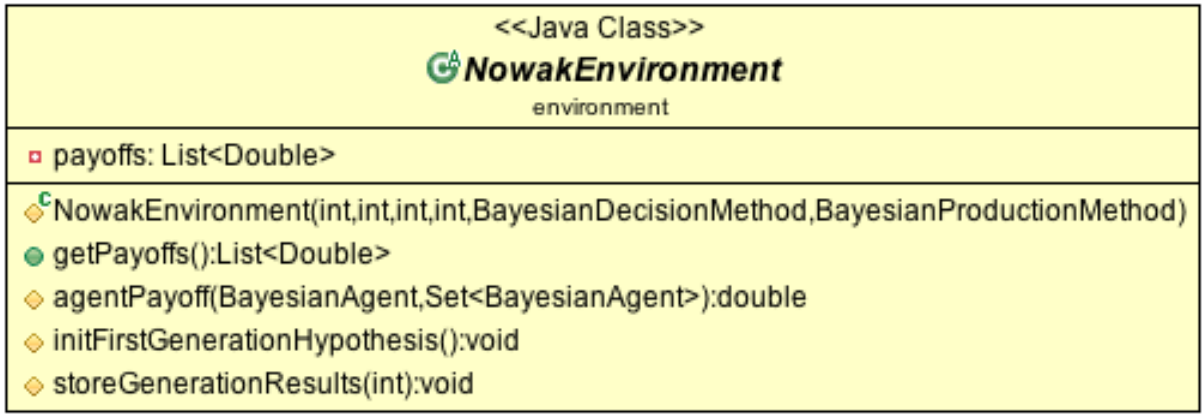


Figure 32: Nowak Abstract Class Diagram

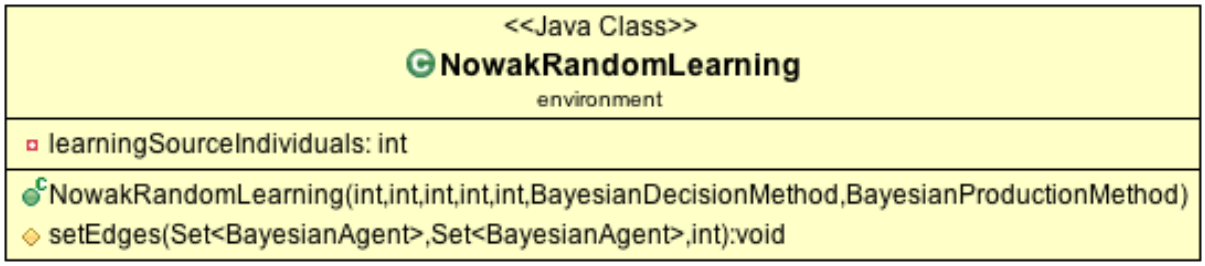


Figure 33: NowakRandomLearning Class Diagram

A.4.1.6 Parental Learning

Parental Learning, which implements a `setEdges` method using the agent's payoff as weight to its parenthood probability, is implemented with the `NowakParentalLearning` class.

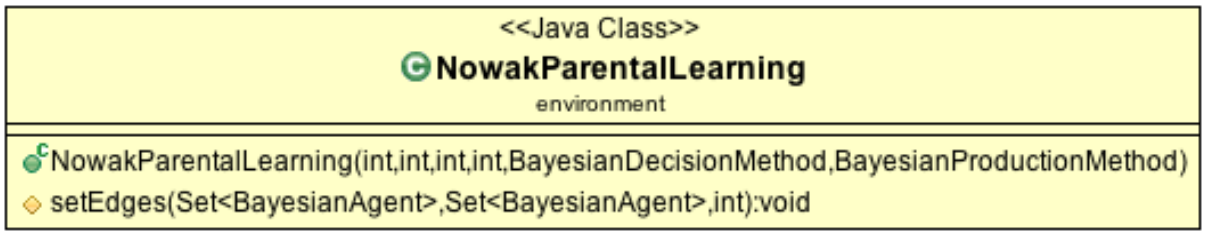


Figure 34: NowakParentalLearning Class Diagram

A.5 Statistical Module

The statistical features modelled consist on probability distributions and its parameters. Those are modelled by three interfaces: (i) `ContinuousDistribution`, (ii)

`DiscreteDistribution` and (iii) `DistributionParameters`. Each distribution implements the methods `mode`, `sample`, `likelihood` and `maximumLikelihoodEstimator`.

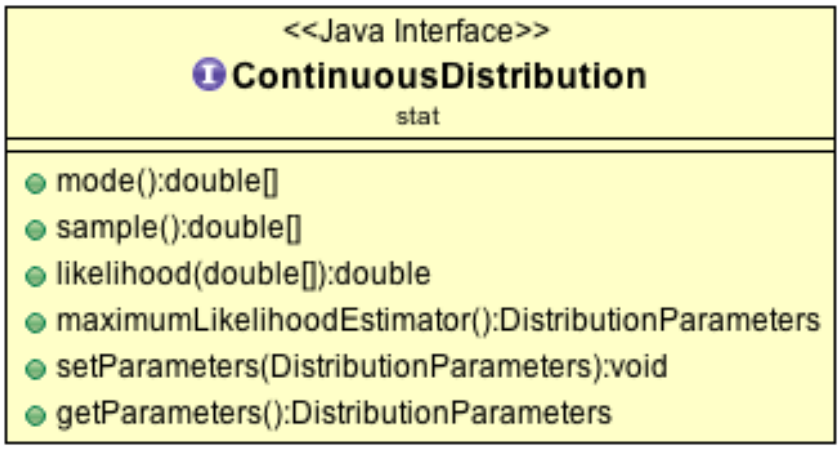


Figure 35: ContinuousDistribution Interface Diagram

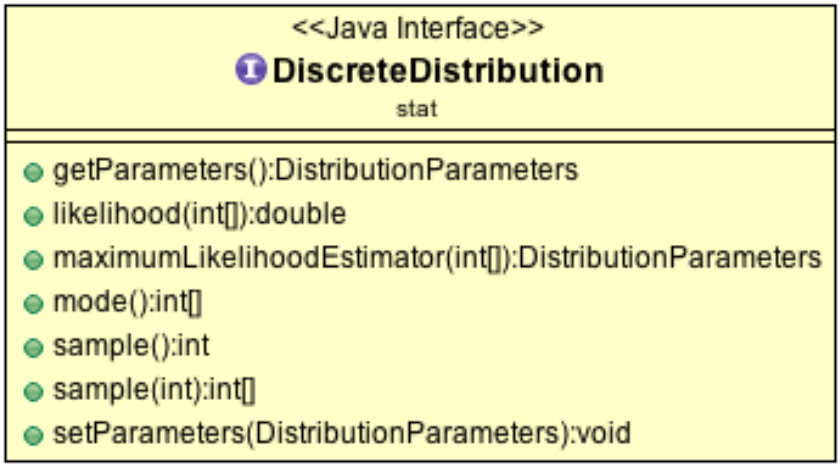


Figure 36: DiscreteDistribution Interface Diagram

A.5.1 Implemented Features

Two distributions, and its corresponding parameters, were implemented: (i) Multinomial and (ii) Dirichlet, according to the class structure in the figure 38.

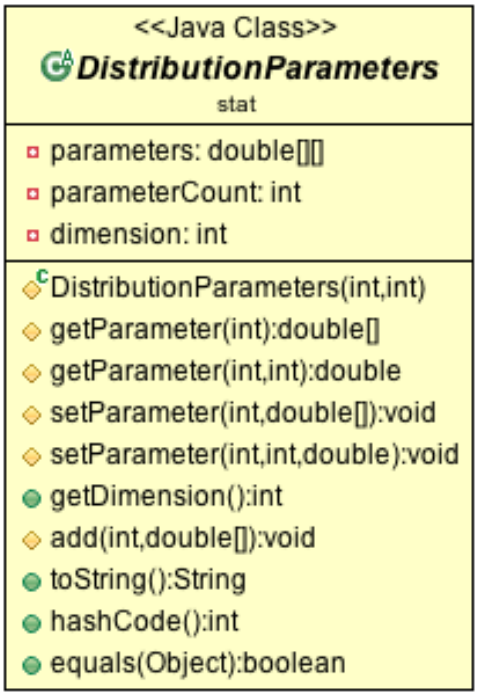


Figure 37: DistributionParameters Interface Diagram

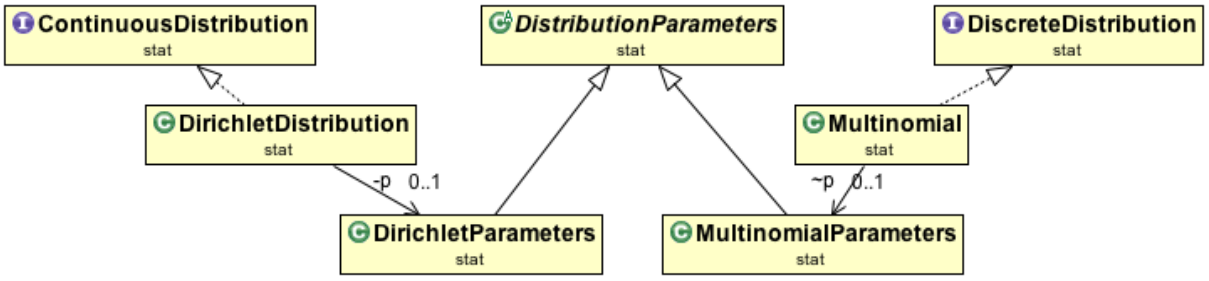


Figure 38: Distributions Class Structure

APPENDIX B – Sample Experiment

As an example, this chapter describes the procedures needed to simulate the experiment whose results are presented on Section 4.1, namely the one assessing the change on Portuguese Clitic Placement.

Firstly, the target languages (Classical and European Portuguese) are defined:

```
double[] cp = {p, 1 - 2*p, p};
double[] ep = {0, 1 - q, q};

BayesianLingusiticHypotheses classicalPortuguese =
        new BayesianLingusiticHypotheses(new MultinomialParameters(cp));
BayesianLingusiticHypotheses europeanPortuguese =
        new BayesianLingusiticHypotheses(new MultinomialParameters(ep));

List<BayesianLingusiticHypotheses> languages = new ArrayList<>(2);
languages.add(classicalPortuguese);
languages.add(europeanPortuguese);
```

The standard learning method can be defined. The Restricted Maximum Likelihood uses a set of target languages and a probability distribution to calculate its likelihoods:

```
BayesianDecisionMethod portugueseCliticsDecider =
    new RestrictedMaximumLikelihood(
                                languages,new Multinomial()
                                );
```

The standard production method is defined using a set of probability distributions and the language dimensions:

```
List<DiscreteDistribution> distributions = new ArrayList<>(1);
distributions.add(new Multinomial());

int meaningSpaceSize = 1;
int signalSpaceSize = cp.length;
BayesianProductionMethod mult = new BayesianProductionMethod(
        distributions,meaningSpaceSize,signalSpaceSize
        );
```

Then, the environment can be set, using six parameters:

1. The number of agents per generation;
2. The number of sentences exchanged at each iteration;
3. The standard production method to be used when creating agents;
4. The standard learning method to be used when creating agents;
5. A list of hypotheses defining the initial hypotheses;
6. A list of numbers defining the proportion associated with each initial hypothesis.

```
List<BayesianLingusiticHypotheses> initList = new ArrayList<>();
initList.add(classicalPortuguese);
List<BigDecimal> proportion = new ArrayList<>();
proportion.add(new BigDecimal(1.0));

IteratedLearning il = new IteratedLearning(
        agentsPerGeneration,
        n,
        mult,
        portugueseCliticsDecider,
        initList,
        proportion);
```

The experiment can then be performed, iterating the environment through the generations:

```
for(int i=0; i<generations; i++){
        il.iterate();
}
```

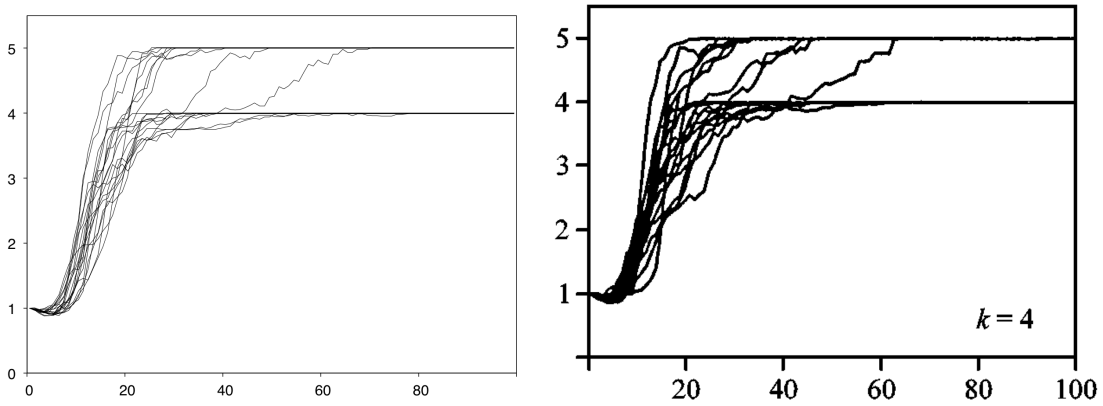
Finally, the experiment results can be accounted, verifying the proportion of Classical Portuguese speakers per generation:

```
List<List<BayesianLingusiticHypotheses>> hyp = il.getHypothesis()

List<Double> perGen = new ArrayList<>();

for (int i = 0; i < h.size(); i++) {
        List<BayesianLingusiticHypotheses> thisGeneration = hyp.get(i);
        perGen.add(0.0);

        for (int j = 0; j < thisGeneration.size(); j++) {
                if (thisGeneration.get(j).equals(classicalPortuguese))
                        perGen.set(i, perGen.get(i) + 1);
                }

                perGen.set(i, perGen.get(i) / thisGeneration.size());
```

```
        }
}

for(int gen=0; gen < hyp.size(); gen++){
        System.out.println(gen+";"+perGen.get(gen));
}
```

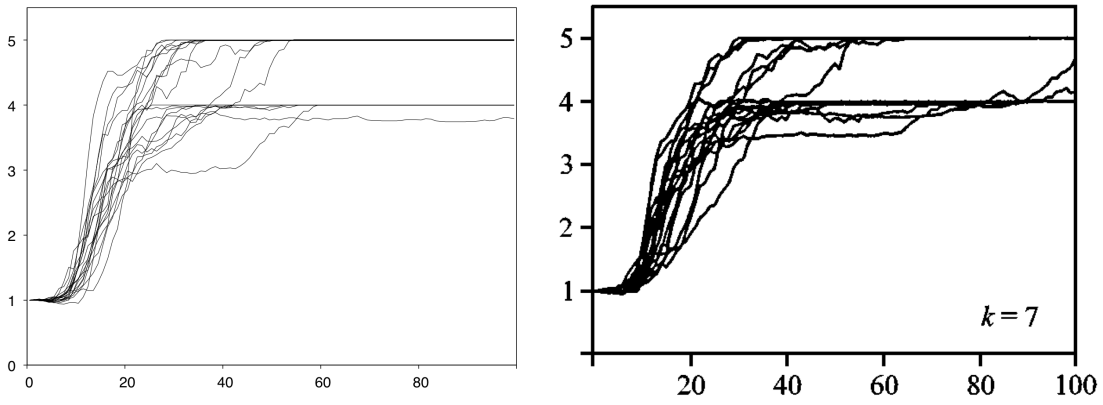
APPENDIX C – Evaluation Results - Game Theoretical Approach

This appendix contains the evaluation results, complementing the ones presented on section 4.2. It should be noted that, on the Random Learning results, the original plots contained the variable k (number of sentences exchanged) on evidence. However, the variable K (number of parents selected) is the variable parameter on those, and should therefore be the one displayed.



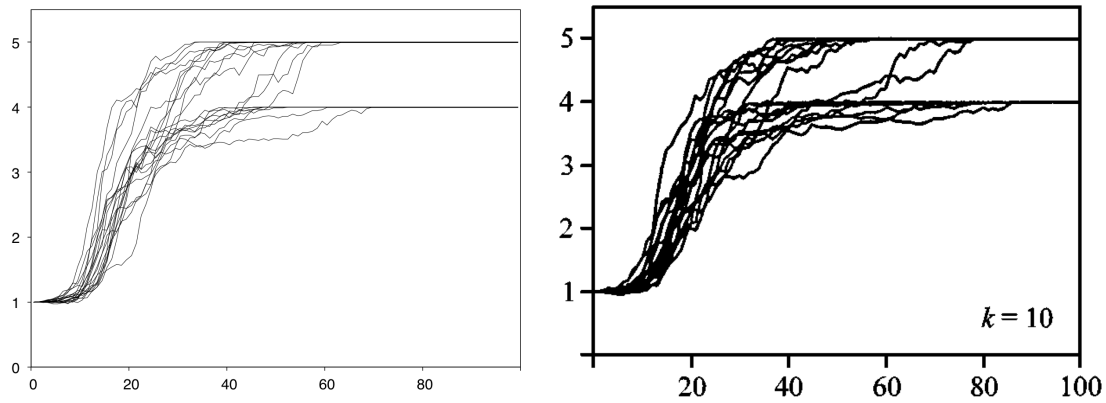
(a) Simulated (b) Published (NOWAK et al., 1999)

Figure 39: Evaluation Results - Parental Learning, $k = 4$ - Generations on horizontal axis, Normalized Payoff on vertical axis



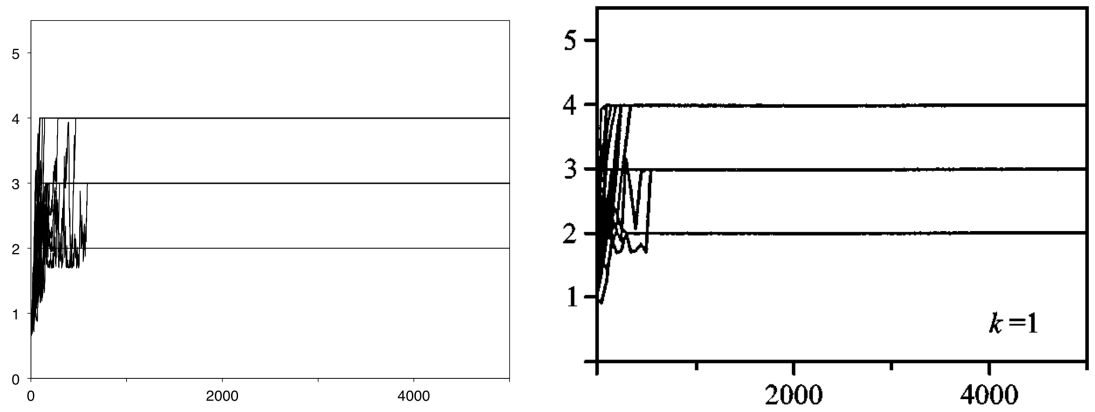
(a) Simulated (b) Published (NOWAK et al., 1999)

Figure 40: Evaluation Results - Parental Learning, $k = 7$ - Generations on horizontal axis, Normalized Payoff on vertical axis



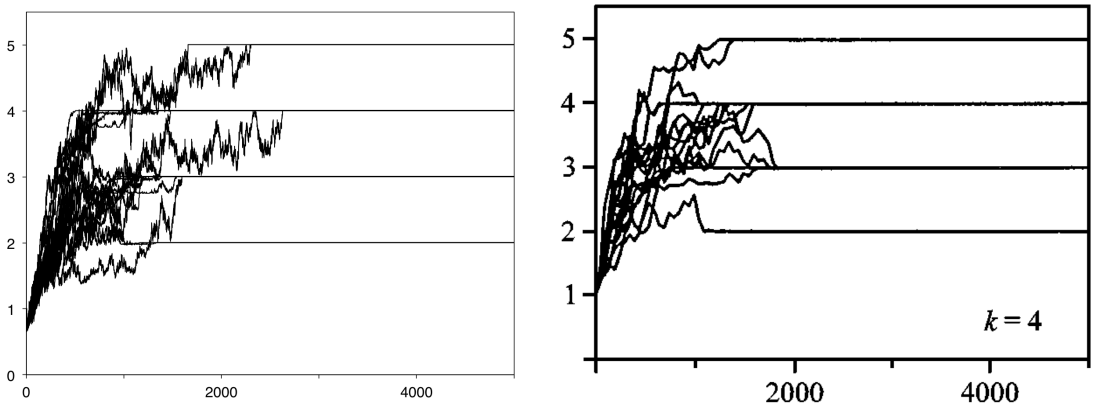
(a) Simulated (b) Published (NOWAK et al., 1999)

Figure 41: Evaluation Results - Parental Learning, $k = 10$ - Generations on horizontal axis, Normalized Payoff on vertical axis



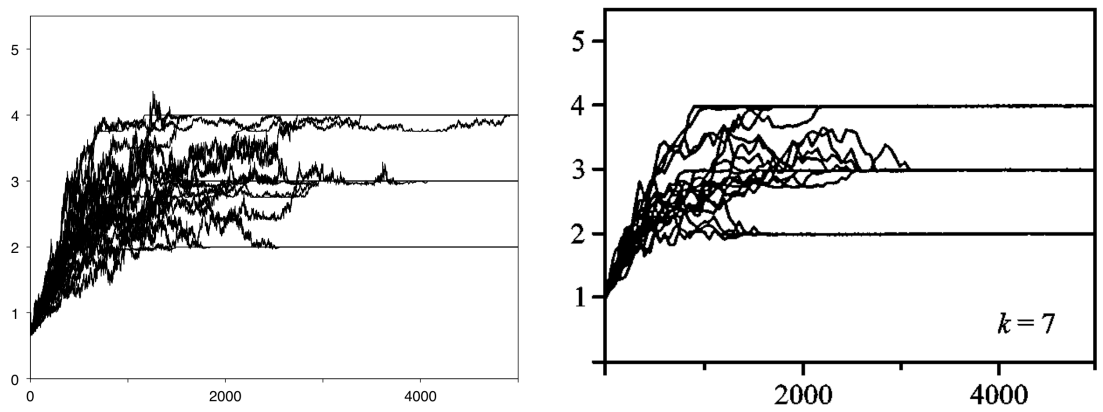
(a) Simulated (b) Published (NOWAK et al., 1999)

Figure 42: Evaluation Results - Random Learning, $K = 1$ - Generations on horizontal axis, Normalized Payoff on vertical axis



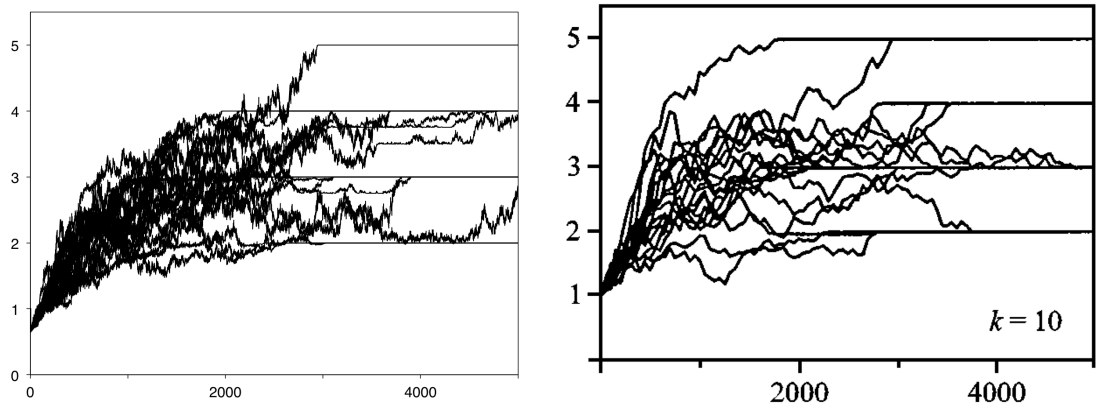
(a) Simulated (b) Published (NOWAK et al., 1999)

Figure 43: Evaluation Results - Random Learning, $K = 4$ - Generations on horizontal axis, Normalized Payoff on vertical axis



(a) Simulated (b) Published (NOWAK et al., 1999)

Figure 44: Evaluation Results - Random Learning, $K = 7$ - Generations on horizontal axis, Normalized Payoff on vertical axis



(a) Simulated (b) Published (NOWAK et al., 1999)

Figure 45: Evaluation Results - Random Learning, $K = 10$ - Generations on horizontal axis, Normalized Payoff on vertical axis

ANNEX A – TG1

Bayesian language change models and a general framework

Matheus Proença[1], **Aline Villavicencio**[1], **Marco Idiart**[2], **Rodrigo Wilkens**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

[2]Instituto de Física – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.051 – 91.501-970 – Porto Alegre – RS – Brazil

{mcproenca,avillavicencio,rswilkens}@inf.ufrgs.br, idiart@if.ufrgs.br

Abstract. *Language change consists on the variation of linguistic features through time. Previous work was able to account the influences of language acquisition on those changes, establishing the dynamical system grounds behind language change. This work is the first of two parts composing a Bachelor Thesis that has the objective of finding similarities and an overall structure within the models proposed by the literature, providing a proposal of theoretical unifying of them.*

Resumo. *A mudança de linguagem consiste na variação das características linguísticas através do tempo. Trabalhos anteriores puderam quantificar as influências da aquisição de linguagem nessa mudança, estabelecendo o sistema dinâmico fundamentando a mudança de linguagem. Esse trabalho é a primeira de duas partes compondo um Trabalho de Graduação que tem como objetivo encontrar similaridades e uma estrutura geral entre os modelos propostos pela literatura, mostrando uma proposta de unificação destes.*

1. Introduction

The study of language change is a core subject on cognitive sciences. The understanding of its fundamental principles can enlighten the comprehension of how human cognition works. To provide support to that kind of study, computational evaluations have been used for a long time to assess mathematical models of different linguistic features.

In that matter, a well accepted hypothesis is that language acquisition plays a major role in language change. To account those influences, several mathematical models have been developed, proposing different approaches to the subject. Some of those models present common modelling features, suggesting that a general framework can nest several of those models under a single abstraction.

Those facts will be explored in the rest of the document. Section 2 presents the overall objectives of this Bachelor Thesis and of this paper specifically, while section 3 displays the method employed to reach those objectives. Sections 4, 5, 6 and 7 present a short description of some models selected from the literature. Section 8 presents a proposed framework to allow a single general abstraction of the presented work. Section 9 provides the final statements of the paper, presenting the points to be developed in following work.

2. Objectives

As the fist of two parts composing the Bachelor Thesis on Computer Engineering, this paper provides an overall aspect of the phenomena being studied, covering a first theoretical analysis of the literature and a general specification to the second part of this work.

The main objective of this Bachelor Thesis is to provide means to compare some of the models suggested by the literature, verifying its common features and enabling the construction of hybrid approaches, composed by parts of different models.

To allow those assessments to be made, a literature review will be provided in this paper, covering a theoretical basis to the following steps and a initial general framework will be provided to unify the study of the different models. In the second part of this Bachelor Thesis, a computer application will be developed as a tool to perform empirical assessments of the models. That tool will allow the selection of the multiple parts composing a language change mathematical modelling approach, enabling the assessment of the scenarios created by the different models and the comparison of those scenarios with acquired empirical data.

3. Method

The preparation of this work, in order to achieve its objective of providing theoretical basis in language change, and basic specification of a tool to perform empirical comparisons of different language change models, was performed through the following steps, in that order:

1. Gathering of different language change models;
2. Studying the working basis of those models;
3. Studying the results generated by those models;
4. Studying the relationship between the models;
5. Suggest a framework unifying the study of the different models;
6. Building a specification to the second part of this work, based on the models above.

The following sections contain the results of the work performed following those steps.

4. Population versus individual behaviours

The work described in [**?**] describes how the language acquisition mechanism has influences on the distribution of different grammars on the population, inferring that language acquisition is a major driver of the language change phenomena.

4.1. Modelling method

To model those dynamics, the method proposed in [**?**] and [**?**] is used. That model describes a chain of learning individuals, where at each generation, one single individual teaches other individual of the following generation. To formally describe that model, let be defined:

- A class of grammars $\mathcal{G}$;
- A set of expressions $L_g \in \Sigma^*$ generated by $g \in \mathcal{G}$;

- A probability distribution P_g over L_g, given that a speaker of g produces an expression s with probability $P_g(s)$;
- A propability distribution $P^{(t)}$ over $\mathcal{G}$, given that in the generation t, a speaker of $g \in \mathcal{G}$ can be found with the probability $P^{(t)}(g)$;
- A learning algorithm, given that an individual exposed to a n-tuple of expressions $S_n = (s_1, \ldots, s_n) \in (\Sigma^*)^n$ will acquire the grammar $g = \mathcal{A}(S_n)$, defined by the map:

$$\mathcal{A} : \bigcup_{i=1}^{\infty} (\Sigma^*)^i \to \mathcal{G} \tag{1}$$

4.2. Case Study

The work then applies the method to the case of clitic placement change in portuguese, described in [**?**]. During 200 years, starting in 1800, the Portuguese language went to a change in its usual clitic placement. Their work explains that affirmative proclitic constructions ceased to occur on European Portuguese.

To apply the model to this specific case, the paper describes three stress contours, each referring to a type of production:

- c_1, to affirmative proclitic productions;
- c_2, to affirmative enclitic productions; and
- c_3, to proclitic productions with quantified or wh-subjects.

Also, two grammars are defined:

- Classical Portuguese, G_{CP}, where all three stress contours happen; and
- European Portuguese, G_{EP}, c_1 does not occur.

Afterwards, the work further defines the production distributions:

- Classical Portuguese:
  - c_1 is produced with probability p;
  - c_2 is produced with probability $1 - 2p$;
  - c_3 is produced with probability p.
- European Portuguese:
  - c_1 is not produced;
  - c_2 is produced with probability $1 - q$;
  - c_3 is produced with probability q.

To study the evolution of the dynamical system, the population distribution of the grammars is as follows:

- The proportion of G_{CP} speakers in a generation i is given by α_i;
- The proportion of G_{EP} speakers in a generation i is given by $1 - \alpha_i$.

Finally, the studies defines the learning algorithm as the Maximum Likelihood Method: One chooses between G_{CP} and G_{EP} by selecting the grammar that maximizes the probability of generating the given data.

Therefore, all the modelling pattern is defined:

- The class of grammars $\mathcal{G} = \{G_{CP}, G_{EP}\}$;
- The set of expressions, represented by the stress contours, $L_g = \{c_1, c_2, c_3\}$;

- The probability distribution over $L_{G_{CP}}$, $P_{G_{CP}} = [p, 1-2p, p]$;
- The probability distribution over $L_{G_{EP}}$, $P_{G_{EP}} = [0, 1-q, q]$;
- The probability distribution over $\mathcal{G}$, $P^{(i)} = [\alpha_i, 1-\alpha_i]$;
- The learning algorithm defined as the Maximum Likelihood Method.

The following step is to analyse the individual learning algorithm to infer population dynamics. The first step to do so is to, given a set of linguistic data $S_n = \{s_1, \ldots, s_n\}$, calculate the likelihoods $P(S_n|G_{CP})$ and $P(S_n|G_{EP})$. Assuming that the linguistic data set was drawn in i.i.d. fashion, one can assume that:

$$P(S_n|G_k) = \prod_{i=1}^{n} P(s_i|G_k) \tag{2}$$

The likelihoods are, therefore, defined by the equations 3 and 4, given that the linguistic data set has a draws of c_1, b draws of c_3 and $n-a-b$ draws of c_2:

$$P(S_n|G_{CP}) = \prod_{i=1}^{n} P(s_i|G_{CP}) = p^a(1-2p)^{(n-a-b)}p^b \tag{3}$$

$$P(S_n|G_{EP}) = \prod_{i=1}^{n} P(s_i|G_{EP}) = 0^a(1-q)^{(n-a-b)}q^b \tag{4}$$

The learning algorithms define that a learner chooses G_{EP} if, and only if there is no occurrence of c_1, and equation 5 verifies:

$$P(S_n|G_{EP}) > P(S_n|G_{CP}) \Leftrightarrow (1-q)^{(n-b)}q^b > (1-2p)^{(n-b)}p^b \tag{5}$$

The work then proceeds analysing the outcomes of the conditions above, verifying the evolutionary dynamics of the system. The conclusions are that, using the Maximum Likelihood Method, the system tends to a population of European Portuguese speakers. Furthermore it is shown that language change behaviours in the individual level need not to be the same in population level. Also, it is shown that language learning has some accountability in language change overtime.

5. On cultural exposure of learners

The phenomenon modelled by the work of [**?**] describes a different approach to the way agents relate, using the same abstraction presented in the section 4.1, where language change is described as an effect of iterating language learning through several generations.

This work discusses the hypothesis that at each generation each learner acquires linguistic data from one sole individual, suggesting that the learner would obtain linguistic data from its whole community, rather than only one person.

5.1. Iterated Learning Modelling

According to the modelling described in the section 4.1, the learner is exposed to linguistic data coming from a single source. The learner then applies the method and chooses its

grammar, generating linguistic data to a next learner. This approach yields a sequence of grammars chosen by the learners and later used as linguistic data sources. This chain corresponds to a Markov chain with $\mathcal{G}$ as state space. The transition operator of this chain can be defined: For any $g \in \mathcal{G}$ and $h \in \mathcal{G}$, the probability of mapping from g to h is given by the equation 6,

$$T[g, h] = prob[g \to h] = prob[\mathcal{A}(D) = h | D \text{ generated by } P_g] \tag{6}$$

with $T[g, h]$ being the probability the learner would acquire h through a learning algorithm $\mathcal{A}$ using the linguistic data D generated according to P_g.

Considering a population whose initial state is given by $P^{(0)}$, the distribution will evolve according the Markov chain dynamics above:

$$P^{(t+1)}(h) = \sum_{g \in \mathcal{G}} P^{(t)} T[g, h] \tag{7}$$

The equation 7 states that, disregarding the learning algorithm employed, the points below can be drawn:

1. The probability distribution over the grammars spoken in the population must evolve according to a linear rule;
2. This linear dynamics converges to a single stable equilibrium, given Markov chain theory characterization [**?**];
3. This linear dynamics makes it impossible to any bifurcations to happen. Bifurcations are accepted as a empirical need of language change, making Iterated Learning a dynamically insufficient model;
4. Also, the Iterated Learning approach cannot model a frequent language change situation, language stability.

5.2. Social Learning Modelling

This approach suggest an improvement of the model described in the section 5.1, including the fact that learners are exposed to linguistic data from multiple sources. This improved approach is called Social Learning, or SL, and its dynamics are described below.

Unlike Iterated Learning, a learner as described in SL is exposed to a set of linguistic data providing from more than one individual. To formalize this amelioration, one can state that a learner is exposed from data drawn from a distribution μ_t given by the equation[1] 8.

$$\mu_t = \sum_{g \in \mathcal{G}} P^{(t)}(g) P_g \tag{8}$$

A learner exposed to linguistic data drawn from that distribution will learn a grammar h with a probability given by the equation 9.

[1] This approach assumes that the population is perfectly distributed, with no network effects, i.e., differente levels of exposure to speakers.

$$prob[\mathcal{A}(D) = h|D \text{ drawn according to } \mu_t] \tag{9}$$

The proportion of speakers of h in the next generation is also given by that probability. The update rule to the generation is, therefore:

$$P^{(t+1)}(h) = prob[\mathcal{A}(D) = h|D \text{ drawn according to } \mu_t] \tag{10}$$

which yields the map

$$f_{\mathcal{A}} : \mathcal{S} \to \mathcal{S}, \tag{11}$$

being $\mathcal{S}$ the state of possible distributions linguistic populations[2].

The SL model, as described above, infers some substantial differences from IL:

1. Social Learning's iterated map $s_{t+1} = f(s_t)$ is generically nonlinear;
2. As an outcome of that condition, and as parameters change continuously, bifurcations are made possible;
3. For the same reasons, multiple stable states are possible, allowing a more wide range of behaviours concerning multiple grammar systems;
4. Also, every learning algorithm $\mathcal{A}$ yields a corresponding evolutionary dynamics map $f_{\mathcal{A}}$. Hence, different learning algorithm can yield potentially different evolutionary outcomes.

6. Bayesian learning and cultural exposure hypothesis checking

The work of [**?**] investigates some of the hypothesis commonly used in language change modelling. Namely it consider the influences generated by the idealization of language change as a chain of single learners (as described in the Iterated Learning, section 5.1); and also verifies some aspects of Bayesian learning, as the differences between Bayesian sampling and a posteriori maximising.

To perform that task, the author uses the Bayesian learning as described in [**?**] to check the differences between the sampling learner (SAM) and the maximum a posteriori (MAP), and investigates the outcomes of three cultural scenarios: A chain of single agents, a chain of pair of agents and complex[3] populations.

To better explain the checking method, firstly the Bayesian learning modelling pattern will be described, including its conclusions about the behaviours of the learning algorithms under Iterated Learning. Afterwards, the case proposed by [**?**] is described, exploring its outcomes.

[2]Each $s \in \mathcal{S}$ refers to a $P^(t)$.

[3]Complex refers to populations with learners beign exposed to multiple linguistic data sources, geographical influences and generation overlapping.

6.1. Bayesian learning modelling

The approach to model language change through a Bayesian learner fits in the same kind of approach as the Iterated Learning (as described in the section 5.1). The theory relies in describing the agents[4] using the Bayes' theorem, as described below.

The approach defines a set of linguistic hypothesis $\mathcal{H}$. Each hypothesis $h \in \mathcal{H}$ describes the inner unterstanding of the agent about the properties of the language the hypothesis refers to. Therefore, each hypothesis correspond to a language (or a grammar). Also, it is defined a set $\mathcal{D}$ of linguistic data. Each instance d of this set defines the linguistic data at which the agent has been exposed.

Each agent has a prior distribution over $\mathcal{H}$, $\mathbf{P}(h)$, representing the prior beliefs of the learner, or better, the amount of evidence the agent has to be exposed to in order to adopt a hypothesis. These prior distributions are updated using the linguistic data d at which the agent is exposed to, resulting in a posterior distribution $\mathbf{P}(h|d)$. The Bayes' theorem states that:

$$\mathbf{P}(h|d) = \frac{\mathbf{P}(d|h)\mathbf{P}(h)}{\mathbf{P}(d)} \tag{12}$$

In the equation 12, $\mathbf{P}(h)$ is the prior, as described above, $\mathbf{P}(d) = \sum_h \mathbf{P}(d|h)\mathbf{P}(h)$, and $\mathbf{P}(d|h)$ is the probability of the agent producing the linguistic data d in the linguistic hypothesis h. This last probability can be calculated if the agent has access to the linguistic data production algorithm.

To select a target linguistic hypothesis, the agent has to make a choice using the updated information about the hypothesis, $\mathbf{P}(h|d)$. Three methods are proposed:

The first, named sampling learner (or SAM), consists in sampling the target linguistic hypothesis h_w according to its posterior probability ($\mathbf{P}(h_w|d)$). It is shown in [**?**] that a chain of agents using SAM as the learning algorithm behaves like a Gibbs sampler [**?**], and therefore converges to the prior when the number of agents in the chain increases.

The second, maximum a posteriori learner (or MAP), consists on selecting the linguistic hypothesis that has the maximum posterior probability:

$$h_w = \arg\max_h \mathbf{P}(h|d) \tag{13}$$

The work of [**?**] successfully interprets a chain of agents using MAP as learning algorithm as a expectation-maximisation algorithm [**?**]. The outcomes of that formalisation is a more complex[5] one, but still dependent on the prior.

The third, proposed by [**?**], uses the two extremes above to build a continuous spectrum of learning algorithms. In its scheme the linguistic hypothesis is randomly cho-

[4] i.e. describing the part concerning the learning algorithm and the generation of utterances. Namely, in the formality described in the section 4.1 these components refer to the learning algorithm $\mathcal{A}$ and the probability distribution P_g.

[5] The hypothesis distribution of a chain of single identical agents, using MAP as their learning algorithm, is, when the number of agents increase, centered at the maximum of the prior.[**?**]

sen with probability $(\mathbf{P}(d|h)\mathbf{P}(h))^r$, in a way that with $r = 1$ the learner corresponds to a sampling learner, and with $r = \infty$ the learner is a maximum a posteriori.

6.2. Case Study

The work describes a language with two linguistic features, F_1 and F_2, each with two possible values, denoted by F_i and F_i^*. To ease the notation, this linguistic features are mapped as utterances of the form $u = f_1 f_2$, where f_i corresponds to the linguistic feature F_i. Hence, four possible utterances are possible (where 1 means that * is present on the feature): 00, 01, 10 and 11. This language can be described by a vector of probabilities of the four possible utterances $\mathbf{p} = [p_{00}, p_{01}, p_{10}, p_{11}]$. Also, the agents will be exposed to linguistic data represented by the vector of occurences of the utterances $\mathbf{n} = [n_{00}, n_{01}, n_{10}, n_{11}]$.

The agent's produces linguistic data following a multinomial distribution $\mathbf{n} \sim Multinom(\mathbf{p})$ with the probability mass function given in the equation 14.

$$f(\mathbf{n}; \mathbf{p}) = \frac{(n_{00} + n_{01} + n_{10} + n_{11})!}{n_{00}! \cdot n_{01}! \cdot n_{10}! \cdot n_{11}!} p_{00}^{n_{00}} p_{01}^{n_{01}} p_{10}^{n_{10}} p_{11}^{n_{11}} \tag{14}$$

Also, assuming that the agent has full access to its own learning algorithms and that it assumes that the best model for language production is its own, its prior distribution is defined using the conjugate Dirichlet, $\mathbf{p} \sim Dirichlet(\alpha)$, with the probability density function given in the equation 15.

$$f(\mathbf{p}; \alpha) = \frac{1}{B(\alpha)} p_{00}^{\alpha_{00}-1} p_{01}^{\alpha_{01}-1} p_{10}^{\alpha_{10}-1} p_{11}^{\alpha_{11}-1} \tag{15}$$

where $\alpha = (\alpha_{00}, \alpha_{01}, \alpha_{10}, \alpha_{11}) > 0$ are parameters and $B(\alpha)$ is the beta function.

From equations 12, 14 and 15, it can be derived that the posterior follows a Dirichlet distribution, $\mathbf{p}|\mathbf{n} \sim Dirichlet(\alpha + \mathbf{n})$.

With those definitions, the author proceeds to compare the sampling learner and the maximum a posteriori learning algorithms across three different cultural configurations.

The first configuration consists on the standard model of the literature, featuring discrete, non-overlapping generations. Each of these generations contain only one agent learning from the agent on the directly precedent generation. These assumptions fit in the description of Iterated Learning, as described in the section 5.1. Also, it is the cultural model used in the work of [**?**] to produce the conclusions about MAP and SAM laid on the section 6.1. The outcomes of the author's analysis are similar to other works: chain of agents using SAM converge on the prior, and those using MAP are able to amplify weak biases, but evolve with a more complex dynamics.

The second approach represents a modification of the Iterated Learning, considering that each generation has two agents. Those agents are exposed to a mixed primary linguistic data, using as source both the agents from the previous generation. The results of this approach when two agents of the same type are paired (SAM-SAM and MAP-

MAP) are similar to as if there were only one agent[6] However, chains of SAM-MAP agents behave like the corresponding single SAM chains. This result suggests that, removing the idealisation of having a sole agent per generation, MAP learning algorithm has the same properties of SAM one.

The third approach uses the model of complex populations described in [**?**], where the cultural space is defined as a square grid of 10x10 regions, where each region can hold a different population. The model supports generation overlapping, and it includes several learning patterns. The results obtained by applying the model to Bayesian leaners is that the language finds a asymptotic stability, and MAP and SAM models are indistinguishable.

7. Bayesian Decision Theory based agents

The work described in [**?**] proposes a novel approach to model the language acquisition accounts on the language change patterns described in the section 4.2 (i.e. clitic change on Portuguese). This novel approach has the objective of overcoming some drawbacks of the frequentist approach to the subject[7].

The paper suggests the Bayesian decision theory [**?**] as a candidate to solve the problem. The approach aims in selecting a grammar that will maximize its expected utility to communication, while accounting prior information in the decision. This method has as advantage the fact that it assumes a more subjective perspective to the matter, taking into account issues of communicability, processing and production of language.

7.1. Modelling Method

A Bayesian decision consists in choosing an action $\hat{a}$ from a decision set Θ given an unknown paramater θ for which there is some observed data y. Bayesian Decision Theory defines that one should take the action the maximizes its expected utility given the possible value of the parameters. That can be expressed as:

$$\hat{a} = \arg\max_a \mathbf{E}[U(a,\theta)|y] = \arg\max_a \int_\Theta U(a,\theta)\mathbf{P}(\theta|y)d\theta \tag{16}$$

In the case of Portuguese clitic change, the action a can be regarded as choosing between the grammar G_{EP} or G_{CP}. The parameter θ assumes the form of two distinct parameters to model: α, the proportion of G_{EP} speakers in the population; and p, the rate at which G_{CP} speakers produce enclitic constructions. Finally, the observed data y is correspondent to the linguistic data at which the learner has been exposed, S_n.

The Bayesian decision rule to the portuguese clitic change case is reduced, hence, to the equation below.

$$\hat{a} = \arg\max_a \mathbf{E}[U(a,\alpha,p)|S_n] = \arg\max_a \int_{[0,1]^2} U(a,\alpha,p)d\mathbf{P}(\alpha,p|S_n) \tag{17}$$

[6]MAP-MAP chains behave identically to a chain of single MAP agents, while SAM-SAM have the same final outcomes, reaching the asymptotic tendencies faster.

[7]Here, *frequentist* is used to described the practice of some methods of using as a learning algorithm a method that selects the grammar seeking to match observed frequencies to grammar intrinsic probabilities, e.g. the Maximum Likelihood Estimation.

To solve the equation 17, the utility function $U(a, \alpha, p)$ needs to be defined. The author then proceeds to define a function that captures the facts that the learner wants to acquire the same grammar as the rest of its community, and also that she wants to be able to play both the roles of speaker and hearer. The basis to construct this function is that, since G_{EP} is regarded as a subset of G_{CP}, a speaker of Classic Portuguese will be able to understand both grammars. However, a speaker of European Portuguese will experience some difficulty understanding G_{CP} speakers. Conversely, EP speakers can talk with no penalty with CP speakers, but not vice-versa. An utility function satisfying those premisses is given on the equation 18.

$$U(a, \alpha, p) = \begin{cases} -\frac{1}{2}\alpha & \text{if } a = G_{CP} \\ -\frac{1}{2}(1-\alpha) & \text{if } a = G_{EP} \end{cases} \tag{18}$$

Also in equation 17, to define $d\mathbf{P}(\alpha, p|S_n)$, the author proceeds to calculate $\mathbf{P}(\alpha, p|S_n)d(\alpha, p)$. To do so, she uses the Bayes' theorem:

$$\mathbf{P}(\alpha, p|S_n) = \frac{\mathbf{P}(\alpha, p)\mathbf{P}(S_n|\alpha, p)}{\mathbf{P}(S_n)} \tag{19}$$

To calculate the elements on the equation 12, one must define the probability density of the prior function, $f(\alpha, p)$, and the likelihood function, $\mathbf{P}(S_n|\alpha, p)$.

The prior function should reproduce prior beliefs of the learner concerning the possible combinations of α and p. In the case of Portuguese clitic change, two points have to be addressed. First, if $\alpha = 1$, all the population is composed of G_{EP} speakers. In that case, the value of p is irrelevant[8]. The hypothesis assumed is that $p = 1$ when $\alpha = 1$[9]. Secondly, if $\alpha = 0$ the population is entirely composed by G_{CP} speakers. The rate of enclisis should, therefore, reflect stable rates observed for Classical Portuguese. The exact figures for that scenario are $p = 0.05$, according to [**?**]. According to those definitions, the author assumes a probability density to the prior with the distribution of a Gaussian with mean $0.95\alpha + 0.05$:

$$f(\alpha, p) = \frac{1}{c} e^{-[p-(0.95\alpha+0.05)]^2} \tag{20}$$

being c a normalizing constant, as defined below.

$$c = \int_{[0,1]^2} e^{-[p-(0.95\alpha+0.05)]^2} d(\alpha, p) \tag{21}$$

The likelihood function, in the other hand, is a direct consequence of the parameters. It evaluates the probability of having S_n given the parameters α and p[10]. The function

[8] As it only applies to G_{CP} speakers.

[9] This hypothesis assumes that, before disappearing, the speakers of G_{CP} would use a higher proportion of enclitic constructions, in order to fit the rest of the population.

[10] i.e. the probability of being exposed to the linguistic data set S_n, given that the population is composed of a proportion of G_{EP} speakers defined by α and that the rate at which G_{CP} produces enclitic constructions is defined by p

can be defined as n Bernoulli trials. Let k be the number of c_2 constructions observed, the likelihood is defined as below:

$$\begin{aligned}\mathbf{P}(S_n|\alpha,p) &= \binom{N}{k}\mathbf{P}(c_2)^k\mathbf{P}(c_1)^{N-k} \\ &= \binom{N}{k}[\mathbf{P}(G_{CP})\mathbf{P}(c_2|G_{CP}) + \mathbf{P}(G_{EP})]^k \times [\mathbf{P}(G_{CP})\mathbf{P}(c_1|G_{CP})]^{N-k} \\ &= \binom{N}{k}[(1-\alpha)p + (1-\alpha)]^k \times [(1-\alpha)(1-p)]^{N-k} \end{aligned} \tag{22}$$

With the equations 18, 12, 20 and 22, one can solve the decision criterion on the equation 17. Actually, since the action parameter space is discrete and contain only two possibilities (i.e. G_{CP} and G_{EP}), the criterion can be reduced to:

$$\mathbf{E}[U(G_{CP},\alpha,p)|S_n] > \mathbf{E}[U(G_{EP},\alpha)|S_n] \tag{23}$$

The equation 23 can be written, disregarding constants, as:

$$\int_{[0,1]^2} \alpha\mathbf{P}(S_n|\alpha,p)f(\alpha,p)\mathbf{P}(\alpha,p)d(\alpha,p) > \int_{[0,1]^2} (1-\alpha)\mathbf{P}(S_n|\alpha,p)f(\alpha,p)\mathbf{P}(\alpha,p)d(\alpha,p) \tag{24}$$

That is reduced to the following criterion, that if true, will tell the learner to choose G_{CP} over G_{EP}:

$$\int_{[0,1]^2} (2\alpha-1)\mathbf{P}(S_n|\alpha,p)f(\alpha,p)\mathbf{P}(\alpha,p)d(\alpha,p) < 0 \tag{25}$$

Furthermore, the paper describes the rules to update the production rates at each new generation. The rules are constructed in way that they are dependent on the frequencies observed during learning:

$$\mathbf{P}(c_1|G_{CP}) = \frac{N-k}{N} \tag{26}$$

$$\mathbf{P}(c_2|G_{CP}) = \frac{k}{N} \tag{27}$$

$$\mathbf{P}(c_2|G_{EP}) = 1 \tag{28}$$

8. A framework

The work presented in the previous sections, as well as other related work not presented in this selection, can be fashioned as two distinct but intrinsically dependant phenomena, defining two problems to be solved in order to understand acquisitionist language change. Namely, one corresponds to the behaviours intrinsic to the agent while the other represents the environment defining the cultural relationship between agents.

8.1. The agent

An agent can be regarded as a subject who speaks and listens to linguistic data. Regarding language acquisition theories, it is expected that the agent produces linguistic data in a language somewhat similar to the one it has been exposed to.

In that context, a Bayesian agent learns its language, and extract linguistic data from it, using Bayesian constructions. To model that behaviour, the entities below can be used:

Production Method The agent uses a probability distribution over the grammar to extract samples that will form its output linguistic data;

Linguistic Hypothesis The agent's language is represented by the parameters of the probability distribution used to produce linguistic data, defining its internal representation of the language;

Learning Method The agents employs the method to, given a set of sample sentences[11], estimate its linguistic hypothesis.

8.2. The environment

The environment defines the relationships between agents, i.e. to whom they produce linguistic data, and from whom they receive it. To allow more complex scenarios, weights to those relationships are also interesting.

To correctly model those setups, a graph is proposed. Its vertices represent agents, and its edges represent *speak to* relationships, with a parameter representing how many sentences are produced per iteration. For instance, the graph on the figure 1 represents the following relationship: *Agent A speaks n sentences to agent B.*

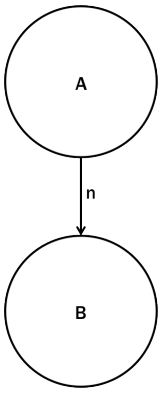


Figure 1. Example environment graph

9. Conclusion and future work

This work selected some of the existant literature about models of language change. While this selection is not exhaustive, it can provide general concepts in the subject. The overall review of these selected works yields a strong basis of reasoning about objectives and paths to be accomplished in the second part of this work.

[11] The agent's input linguistic data

With that structure presented in the section 8 defined, the several atomic concepts presented in the literature can compose several different hybrid models, with various assumptions, or even several levels of idealization. To allow assessments to be performed with those hybrid models, a computational tool has to be developed.

Finally, the language change scenarios generated by those models have to be checked against documented facts, verifying or disproving its effectiveness to model the different phenomena of language change.

To accomplish those tasks, a schedule is proposed in the table 1. To *specify and develop a tool to empirical assessment of the models* consists on the construction of a tool that allow empirical tests to be performed over the unified structure. To *gather well documented language change data* is to review the literature searching for expected behaviours on language change. *Performing empirical assessments of the hybrid models* consists in using the tool developed to test the unified models. Finally, *produce Bachelor Thesis document* represents the task of writing the work's final text.

Table 1. Schedule to the second part of the work

Activity	Dec	Jan	Feb	Mar	Apr	May	Jun
Specify a tool to empirical assessment of the models	x						
Develop the tool according to the specification	x	x	x	x			
Gather well documented language change data	x	x	x	x			
Perform empirical assessments of the hybrid models					x	x	
Produce Bachelor Thesis final document				x	x	x	x