Language Acquisition and Probabilistic Models: keeping it simple

Aline Villavicencio[♣], Marco Idiart[♡] Robert Berwick[♢], Igor Malioutov[♠]

[♣]Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[♡]Institute of Physics, Federal University of Rio Grande do Sul (Brazil)
[♢]LIDS, Dept. of EECS, Massachusetts Institute of Technology (USA)
[♠] CSAIL, Dept. of EECS, Massachusetts Institute of Technology (USA)
avillavicencio@inf.ufrgs.br, marco.idiart@if.ufrgs.br
berwick@csail.mit.edu, igorm@mit.edu

Abstract

Hierarchical Bayesian Models (HBMs) have been used with some success to capture empirically observed patterns of under- and overgeneralization in child language acquisition. However, as is well known, HBMs are "ideal" learning systems, assuming access to unlimited computational resources that may not be available to child language learners. Consequently, it remains crucial to carefully assess the use of HBMs along with alternative, possibly simpler, candidate models. This paper presents such an evaluation for a language acquisition domain where explicit HBMs have been proposed: the acquisition of English dative constructions. In particular, we present a detailed, empirically-grounded model-selection comparison of HBMs vs. a simpler alternative based on clustering along with maximum likelihood estimation that we call linear competition learning (LCL). Our results demonstrate that LCL can match HBM model performance without incurring on the high computational costs associated with HBMs.

1 Introduction

In recent years, with advances in probability and estimation theory, there has been much interest in Bayesian models (BMs) (Chater, Tenenbaum, and Yuille, 2006; Jones and Love, 2011) and their application to child language acquisition with its challenging combination of structured information and incomplete knowledge, (Perfors, Tenenbaum, and Wonnacott, 2010; Hsu and Chater, 2010; Parisien, Fazly, and Stevenson, 2008; Parisien and Stevenson, 2010) as they offer several advantages in this domain. They can readily handle the evident noise and ambiguity of acquisition input, while at the same time providing efficiency via priors that mirror known pre-existing language biases. Further, hierarchical Bayesian Models (HBMs) can combine distinct abstraction levels of linguistic knowledge, from variation at the level of individual lexical items, to cross-item variation, using hyper-parameters to capture observed patterns of both under- and over-generalization as in the acquisition of e.g. dative alternations in English (Hsu and Chater, 2010; Perfors, Tenenbaum, and Wonnacott, 2010), and verb frames in a controlled artificial language (Wonnacott, Newport, and Tanenhaus, 2008).

HBMs can thus be viewed as providing a "rational" upper bound on language learnability, yielding optimal models that account for observed data while minimizing any required prior information. In addition, the clustering implicit in HBM modeling introduces additional parameters that can be tuned to specific data patterns. However, this comes at a well-known price: HBMs generally are also ideal learning systems, known to be computationally infeasible (Kwisthout, Wareham, and van Rooij, 2011). Approximations proposed to ensure computational tractability, like reducing the number of classes that need to be learned may also be linguistically and cognitively implausible. For instance, in terms of verb learning, this could

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1321–1330,
Sofia, Bulgaria, August 4-9 2013.

take the form of reducing the number of subcategorization frames to the relevant subset, as in (Perfors, Tenenbaum, and Wonnacott, 2010), where only 2 frames are considered for 'take', when in fact it is listed in 6 frames by Levin (1993). Finally, comparison of various Bayesian models of the same task is rare (Jones and Love, 2011) and Bayesian inference generally can be demonstrated as simply one class of regularization or smoothing techniques among many others; given the problem at hand, there may well be other, equally compelling regularization methods for dealing with the bias-variance dilemma (e.g., SVMs (Shalizi, 2009)). Consequently, the relevance of HBMs for cognitively accurate accounts of human learning remains uncertain and needs to be carefully assessed.

Here we argue that the strengths of HBMs for a given task must be evaluated in light of their computational and cognitive costs, and compared to other viable alternatives. The focus should be on finding the simplest statistical models consistent with a given behavior, particularly one that aligns with known cognitive limitations. In the case of many language acquisition tasks this behavior often takes the form of overgeneralization, but with eventual convergence to some target language given exposure to more data.

In particular, in this paper we consider how children acquire English dative verb constructions, comparing HBMs to a simpler alternative, a linear competition learning (LCL) algorithm that models the behavior of a given verb as the linear competition between the evidence for that verb, and the average behavior of verbs belonging to its same class. The results show that combining simple clustering methods along with ordinary maximum likelihood estimation yields a result comparable to HBM performance, providing an alternative account of the same facts, without the computational costs incurred by HBM models that must rely, for example, on Markov Chain Monte Carlo (MCMC) methods for numerically integrating complex likelihood integrals, or on Chinese Restaurant Process (CRP) for producing partitions.

In terms of Marr's hierarchy (Marr, 1982) learning verb alternations is an abstract computational problem (Marr's type I), solvable by many type II methods combining representations (models, viz. HBMs or LCLs) with particular algorithms. The HBM convention of adopting ideal learning amounts to invoking unbounded algorithmic resources, solvability in principle, even though in practice such methods, even approximate ones, are provably NP-hard (cf. (Kwisthout, Wareham, and van Rooij, 2011)). Assuming cognitive plausibility as a desideratum, we therefore examine whether HBMs can also be approximated by another type II method (LCLs) that does not demand such intensive computation. Any algorithm that approximates an HBM can be viewed as implementing a somewhat different underlying model; if it replicates HBM prediction performance but is simpler and less computationally complex then we assume it is preferable.

This paper is organized as follows: we start with a discussion of formalizations of language acquisition tasks, §2. We present our experimental framework for the dative acquisition task, formalizing a range of learning models from simple MLE methods to HBM techniques, §3, and a computational evaluation of each model, §4. We finish with conclusions and possibilities for future work, §5.

2 Evidence in Language Acquisition

A familiar problem for language acquisition is how children learn which verbs participate in so-called dative alternations, exemplified by the child-produced sentences 1 to 3, from the Brown (1973) corpus in CHILDES (MacWhinney, 1995).

1. *you took me three scrambled eggs* (a direct object dative (DOD) from Adam at age 3;6)
2. *Mommy can you fix dis for me ?* (a prepositional dative (PD) from Adam at age 4;7)
3. **Mommy, fix me my tiger* (from Adam at age 5;2)

Examples like these show that children generalize their use of verbs. For example, in sentence (1), the child Adam uses *take* as a DOD before any recorded occurrence of a similar use of *take* in adult speech to Adam. Such verbs *alternate* because they can also occur with a prepositional form, as in sentence (2). However, sometimes a child's use of verbs like

these amounts to an overgeneralization – that is, their productive use of a verb in a pattern that does not occur in the adult grammar, as in sentence (3), above. Faced with these two verb frames the task for the learner is to decide for a particular verb if it is a non-alternating DOD only verb, a PD only verb, or an alternating verb that allows both forms.

This ambiguity raises an important learnability question, conventionally known as Baker's paradox (Baker, 1979). On the assumption that children only receive positive examples of verb forms, then it is not clear how they might recover from the overgeneralization exhibited in sentence (3) above, because they will never receive positive sentences from adults like (3), using *fix* in a DOD form. As has long been noted, if negative examples were systematically available to learners, then this problem would be solved, since the child would be given evidence that the DOD form is not possible in the adult grammar. However, although parental correction could be considered to be a source of negative evidence, it is neither systematic nor generally available to all children (Marcus, 1993). Even when it does occur, all careful studies have indicated that it seems mostly concerned with semantic appropriateness rather than syntax. In the cases where it is related to syntax, it is often difficult to determine what the correction refers to in the utterance and besides children seem to be oblivious to the correction (Brown and Hanlon, 1970; Ingram, 1989).

One alternative solution to Baker's paradox that has been widely discussed at least since Chomsky (1981) is the use of *indirect negative evidence*. On the indirect negative evidence model, if a verb is not found where it would be expected to occur, the learner may conclude it is not part of the adult grammar. Crucially, the indirect evidence model is inherently statistical. Different formalizations of indirect negative evidence have been incorporated in several computational learning models for learning e.g. grammars (Briscoe, 1997; Villavicencio, 2002; Kwiatkowski et al., 2010); dative verbs (Perfors, Tenenbaum, and Wonnacott, 2010; Hsu and Chater, 2010); and multiword verbs (Nematzadeh, Fazly, and Stevenson, 2013). Since a number of closely related models can all implement the indirect negative evidence approach, the decision of which one to choose for a given task may not be entirely clear. In this paper we compare a range of statistical models consistent with a certain behavior: early overgeneralization, with eventual convergence to the correct target on the basis of exposure to more data.

3 Materials and Methods

3.1 Dative Corpora

To emulate a child language acquisition environment we use naturalistic longitudinal child-directed data, from the Brown corpus in CHILDES, for one child (Adam) for a subset of 19 verbs in the DOD and PD verb frames, figure 1. This dataset was originally reported in Perfors, Tenenbaum, and Wonnacott (2010), and longitudinal and incremental aspects to acquisition are approximated by dividing the data available into 5 incremental epochs (E1 to E5 in the figures), where at the final epoch the learner has seen the full corpus.

Model comparison requires a gold standard database for acquisition, reporting which frames have been learned for which verbs at each stage, and how likely a child is of making creative uses of a particular verb in a new frame. An independent gold standard with developmental information (e.g. Gropen et al. (1989)) would clearly be ideal. Absent this, a first step is demonstrating that simpler alternative models can replicate HBM performance on their own terms. Therefore, the gold standard we use for evaluation is the classification predicted by Perfors, Tenenbaum, and Wonnacott (2010). The evaluations reported in our analysis take into account intrinsic characteristics of each model in relation to the likelihoods of the verbs, to determine the extent to which the models go beyond the data they were exposed to, discussed in section 2. Further, since it has been argued that very low frequency verbs may not yet be firmly placed in a child's lexicon (Yang, 2010; Gropen et al., 1989), at each epoch we also impose a low-frequency threshold of 5 occurrences, considering only verbs that the learner has seen at least 5 times. This use of a low-frequency threshold for learning has extensive support in the literature for learning

of all kinds in both human and non-human animals, e.g. (Gallistel, 2002). A cut-off frequency in this range has also commonly been used in NLP tasks like POS tagging (Ratnaparkhi, 1999).

3.2 The learners

We selected a set of representative statistical models that are capable in principle of solving this classification task, ranging from what is perhaps the simplest possible, a simple binomial, all the way to multi-level hierarchical Bayesian approaches.

A Binomial distribution serves as the simplest model for capturing the behavior of a verb occurring in either DOD or PD frame. Representing the probability of DOD as θ, after n occurrences of the verb the probability that y of them are DOD is:

$$p(\ y|\ \theta, n) = \binom{n}{y} \theta^y\ (1-\theta)^{n-y} \tag{1}$$

Considering that $p(y|\ \theta, n)$ is the likelihood in a Bayesian framework, the simplest and the most intuitive estimator of θ, given y in n verb occurrences, is the Maximum Likelihood Estimator (MLE):

$$\theta_{MLE} = \frac{y}{n} \tag{2}$$

θ_{MLE} is viable as a learning model in the sense that its accuracy increases as the amount of evidence for a verb grows ($n \rightarrow \infty$), reflecting the incremental, on-line character of language learning. However, one well known limitation of MLE is that it assigns zero probability mass to unseen events. Ruling out events on the grounds that they did not occur in a finite data set early in learning may be too strong – though it should be noted that this is simply one (overly strong) version of the indirect negative evidence position.

Again as is familiar, to overcome zero count problem, models adopt one or another method of smoothing to assign a small probability mass to unseen events. In a Bayesian formulation, this amounts to assigning non-zero probability mass to some set of priors; smoothing also captures the notion of *generalization*, making predictions about data that has never been seen by the learner. In the context of verb learning smoothing could be based on several principles:

- an (innate) expectation as to how verbs in general should behave;
- an acquired class-based expectation of the behavior of a verb, based on its association to similar but more frequent verbs.

The former can be readily implemented in terms of prior probability estimates. As we discuss below, class-based estimates arise from one or another clustering method, and can produce more accurate estimates for less frequent verbs based on patterns already learned for more frequent verbs in the same class; see (Perfors, Tenenbaum, and Wonnacott, 2010). In this case, smoothing is a side-effect of the behavior of a class as a whole.

When learning begins, the prior probability is the only source of information for a learner and, as such, dominates the value of the posterior probability. However, in the large sample limit, it is the likelihood that dominates the posterior distribution regardless of the prior. In Hierarchical Bayesian Models both effects are naturally incorporated. The prior distribution is structured as a chain of distributions of parameters and hyper-parameters, and the data may be divided into classes that share some of the hyper-parameters, as defined below for the case of a three levels model:

$$\begin{aligned}
\lambda &\sim \text{Exponential}(1) \\
\mu &\sim \text{Exponential}(1) \\
\alpha_k &\sim \text{Exponential}(\lambda) \\
\beta_k &\sim \text{Beta}(\mu, \mu) \\
\theta_{ik} &\sim \text{Beta}(\alpha_k\beta_k, \alpha_k(1-\beta_k)) \\
y_i|n_i &\sim \text{Binomial}(\theta_{ik})
\end{aligned}$$

The indices refer to the possible hierarchies among the hyper-parameters. λ and μ are in the top, and they are shared by all verbs. Then there are classes of different α_k, β_k, and the probabilities for the DOD frame for the different verbs (θ_{ik}) are drawn according to the classes k assigned to them. An estimate for (θ_{ik}) for a given configuration of clusters is given by

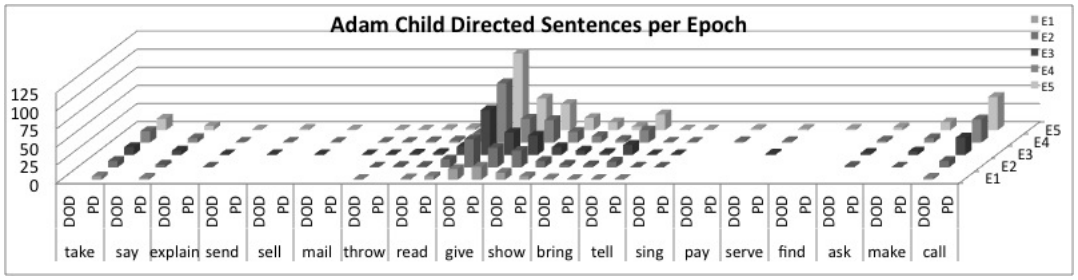

Figure 1: Verb tokens per epoch (E1 to E5)

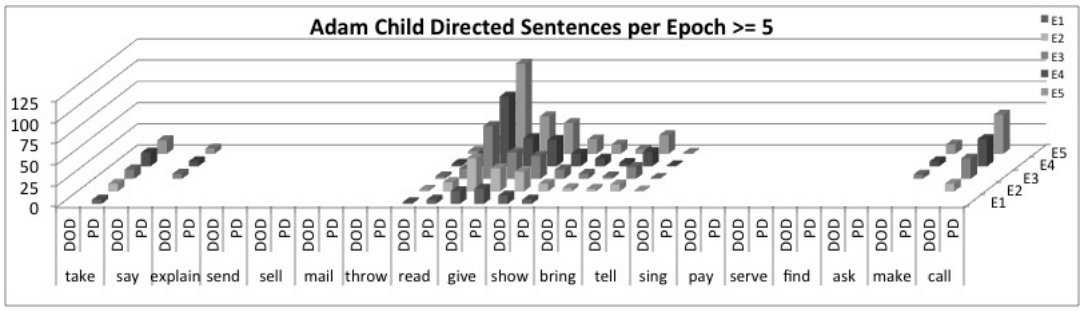

Figure 2: Verb tokens ≥ 5 per epoch (E1 to E5)

$$\theta^{ik}_{HBM} = \frac{1}{P(\mathbf{Y})} \int \frac{y_i + \alpha_k \beta_k}{n_i + \alpha_k} \; P^*_{L3}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\mu} \,|\mathbf{Y})\, d\boldsymbol{\alpha}\, d\boldsymbol{\beta}\, d\boldsymbol{\lambda}\, d\boldsymbol{\mu}$$

where $P(\mathbf{Y})$ is the evidence of the data, the unnormalized posterior for the hyper-parameters is

$$P^*_{L3}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\mu} \,|\mathbf{Y}) = \left(\prod_k P(\{y_i, n_i\}_{i \in k} | \alpha_k, \beta_k)\, \mathrm{Exp}(\alpha_k|\lambda) \mathrm{Beta}(\beta_k|\mu, \mu) \right) * \mathrm{Exp}(\lambda|1) \mathrm{Exp}(\mu|1)$$

and the likelihood for α and β is

$$P(\{y_i, n_i\}_{i \in k} | \alpha_k, \beta_k) \;=\; \prod_{i \in k} \binom{n_i}{y_i} \frac{\mathrm{B}(\alpha_k \beta_k + y_i, \alpha_k(1-\beta_k) + n_i - y_i)}{\mathrm{B}(\alpha_k \beta_k, \alpha_k(1-\beta_k))}$$

The Hierarchical Bayesian Model prediction for θ_i is the average of the estimate ${\theta^{ik}}_{HBM}$ over all possible partitions of the verbs in the task. To simplify the notation we can write

$$\theta_{HBM} = E\left[\frac{y + \alpha\beta}{n + \alpha}\right] \tag{3}$$

where in the expression $E[\dots]$ are included the integrals described above and the average of all possible class partitions. Due to this complexity, in practice even small data sets require the use of MCMC methods, and statistical models for partitions, like CRP (Gelman et al., 2003; Perfors, Tenenbaum, and Wonnacott, 2010). This complexity also calls into question the cognitive fidelity of such approaches.

Eq.3 is particularly interesting because by fixing α and β (instead of averaging over them) it is possible to deduce simpler (and classical) models: MLE corresponds to $\alpha = 0$; the so called "add-one" smoothing (referred in this paper as L1) corresponds to $\alpha = 2$ and $\beta = 1/2$. From Eq.3 it is also clear that if α and β (or their distributions) are unchanged, as the evidence of a verb grows ($n \to \infty$), the HBM estimate approaches MLE's, ($\theta_{HBM} \to \theta_{MLE}$). On the other hand, when $\alpha >> n$, $\theta_{HBM} \sim \beta$, so that β can be interpreted as a prior value for θ in the low frequency limit.

Following this reasoning, we propose an alternative approach, a linear competition learner (LCL), that explicitly models the behavior of a given verb as the linear competition between the evidence for the verb, and the average behavior of verbs of the same class. As clustering is defined independently from parameter estimation, the advantages of the proposed approach are twofold. First, it is computationally much simpler, not requiring approximations by Monte Carlo methods. Second, differently from HBMs where the same attributes are used for clustering and parameter estimation (in this case the DOD and PD counts for each verb), in LCL cluster-

ing may be done using more general contexts that employ a variety of linguistic and environmental attributes.

For LCL the prior and class-based information are incorporated as:

$$\theta_{LCL} = \frac{y_i + \alpha_C \beta_C}{n_i + \alpha_C} \quad (4)$$

where α_C and β_C are defined via justifiable heuristic expressions dependent solely on the statistics of the class attributed to each verb i.

The strength of the prior (α_C) is a monotonic function of the number of elements (m_C) in the class C, excluding the target verb v_i. To approximate the gold standard behavior of the HBM for this task (Perfors, Tenenbaum, and Wonnacott, 2010) we chose the following function for α_C:

$$\alpha_C = {m_C}^{3/2}(1 - {m_C}^{-1/5}) + 0.1 \quad (5)$$

with the strength of the prior for the LCL model depending on the number of verbs in the class, not on their frequency. Eq.5 was chosen as a good fit to HBMs, without incurring their complexity. The powers are simple fractions, not arbitrary numbers. A best fit was not attempted due to the lack of assessment of how accurate HBMs are on real data.

The prior value (β_C) is a smoothed estimation of the probability of DOD in a given class, combining the evidence for all verbs in that class:

$$\beta_C = \frac{Y_C + 1/2}{N_C + 1} \quad (6)$$

in this case Y_C is the number of DOD occurrences in the class, and N_C the total number of verb occurrences in the class, in both cases excluding the target verb v_i.

The interpretation of these parameters is as follows: β_C is the estimate of θ in the absence of any data for a verb; and α_C controls the crossover between this estimate and MLE, with a large α_C requiring a larger sample (n_i) to overcome the bias given by β_C.

For comparative purposes, in this paper we examine alternative models for (a) probability estimation and (b) clustering. The models are the following:

- two models without clusters: MLE and L1;
- two models where clusters are performed independently: LCL and $\text{MLE}_{\alpha\beta}$; and
- the full HBM described before.

$\text{MLE}_{\alpha\beta}$ corresponds to replacing α, β in eq.3 by their maximal likelihood values calculated from $P(\{y_i, n_i\}_{i \in k} | \alpha, \beta)$ described before.

For models without clustering, estimation is based solely on the observed behavior of verbs. With clustering, same-cluster verbs share some parameters, influencing one another. HBMs place distributions over possible clusters, with estimation derived from averages over distributions. In HBMs, clustering and probability estimation are calculated jointly. In the other models these two estimates are calculated separately, permitting 'plug-and-play' use of external clustering methods, like X-means (Pelleg and Moore, 2000)[1]. However, to further assess the impact of cluster assignment on alternative model performance, we also used the clusters that maximize the evidence of the HBM for the DOD and PD counts of the target verbs, and we refer to these as Maximum Evidence (ME) clusters. In MWE clusters, verbs are separated into 3 classes: one if they have counts for both frames; another for only the DOD frame; and a final for only the PD frame.

4 Evaluation

The learning task consists of estimating the probability that a given verb occurs in a particular frame, using previous occurrences as the basis for this estimation. In this context, overgeneralization can be viewed as the model's predictions that a given verb seen only in one frame (say, a PD) can also occur in the other (say, a DOD) as well, and it decreases as the learner receives more data. In one extreme we have MLE, which does not overgeneralize, and in the other the L1 model, which assigns uniform probability for all unseen cases. The other 3 models fall somewhere in between, overgeneralizing beyond the observed data, using the prior and class-based smoothing to assign some (low) probability mass to an unseen verb-frame pair. The relevant models'

[1]Other clustering algorithms were also used; here we report X-means results as representative of these models. X-means is available from `http://www.cs.waikato.ac.nz/ml/weka/`

predictions for each of the target verbs in the DOD frame, given the full corpus, are in figure 3. In either end of the figure are the verbs that were attested in only one of the frames (PD only at the left-hand end, and DOD only at the right-hand end). For these verbs, LCL and HBM exhibit similar behavior. When the low-frequency threshold is applied, MLE$_{\alpha\beta}$, HBM and LCL work equally well, figure 4.

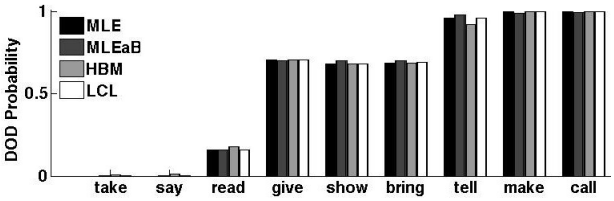

Figure 4: Probability of verbs in DOD frame, Low Frequency Threshold.

To examine how overgeneralization progresses during the course of learning as the models were exposed to increasing amounts of data, we used the corpus divided by cumulative epochs, as described in §3.1. For each epoch, verbs seen in only one of the frames were divided in 5 frequency bins, and the models were assessed as to how much overgeneralization they displayed for each of these verbs. Following Perfors, Tenenbaum, and Wonnacott (2010) overgeneralization is calculated as the absolute difference between the models predicted θ and θ_{MLE}, for each of the epochs, figure 5, and for comparative purposes their alternating/non-alternating classification is also adopted. For non-alternating verbs, overgeneralization reflects the degree of smoothing of each model. As expected, the more frequent a verb is, the more confident the model is in the indirect negative evidence it has for that verb, and the less it overgeneralizes, shown in the lighter bars in all epochs. In addition, the overall effect of larger amounts of data are indicated by a reduction in overgeneralization epoch by epoch. The effects of class-based smoothing can be assessed comparing L1, a model without clustering which displays a constant degree of overgeneralization regardless of the epoch, while HBM uses a distribution over clusters and the other models X-means. If a low-frequency threshold is applied, the differences between the models decrease significantly and so does the degree of overgeneralization in the models' predictions, as shown in the 3 lighter bars in the figure.

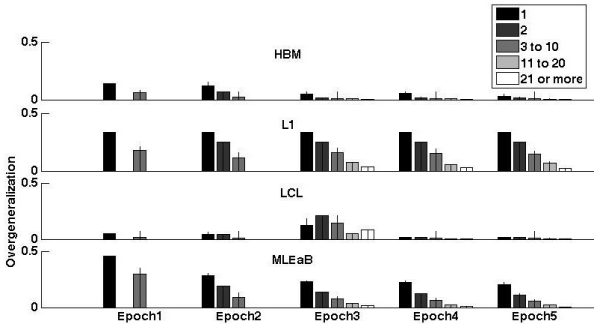

Figure 5: Overgeneralization, per epoch, per frequency bin, where 0.5 corresponds to the maximum overgeneralization.

While the models differ somewhat in their predictions, the quantitative differences need to be assessed more carefully. To compare the models and provide an overall difference measure, we use the predictions of the more complex model, HBM, as a baseline and then calculate the difference between its predictions and those of the other models. We used three different measures for comparing models, one for their standard difference; one that prioritizes agreement for high frequency verbs; and one that focuses more on low frequency verbs.

The first measure, denoted *Difference*, captures a direct comparison between two models, M_1 and M_2 as the average prediction difference among the verbs, and is defined as:

$$D(M_1, M_2) = \frac{1}{N_{verbs}} \sum_{i=1}^{N_{verbs}} abs(\theta^i_{M1} - \theta^i_{M2})$$

This measure treats all differences uniformly, regardless of whether they relate to high or low frequency verbs in the learning sample (e.g. for *bring* with 150 counts and *serve* with only 1 have the same weight). To focus on high frequency verbs, we also define the *Weighted Difference* between two models as:

$$D_n(M_1, M_2) = \frac{1}{\sum_i n_i} \sum_{i=1}^{N_{verbs}} n_i \, abs(\theta^i_{M1} - \theta^i_{M2})$$

Here we expect $D_n < D$ since models tend to

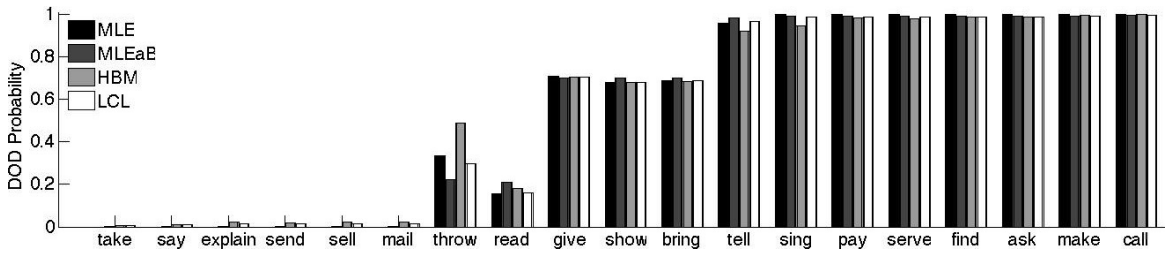

Figure 3: Probability of verbs in DOD frame.

agree as the amount of evidence for each verb increases. Conversely, our third measure, denoted *Inverted*, prioritizes the agreement between two models on low frequency verbs, defined as follows:

$$D_{1/n}(M_1, M_2) = \frac{1}{\sum_i 1/n_i} \sum_{i=1}^{N_{verbs}} \frac{1}{n_i} abs(\theta^i_{M1} - \theta^i_{M2})$$

$D_{1/n}$ captures the degree of similarity in overgeneralization between two models. The results of applying these three difference measures are shown in figure 6 for the relevant models, where grey is for $D(M_1, M_2)$, black for $D_n(M_1, M_2)$ and white for $D_{1/n}(M_1, M_2)$. Given the probabilistic nature of Monte Carlo methods, there is also a variation between different runs of the HBM model (HBM to HBM-2), and this indicates that models that perform within these bounds can be considered to be equivalent (e.g. HBMs and ME-MLE$_{\alpha\beta}$ for Weighted Difference, and the HBMs and X-MLE$_{\alpha\beta}$ for the Inverted Difference).

Comparing the prediction agreement, the strong influence of clustering is clear: the models that have compatible clusters have similar performances. For instance, all the models that adopt the ME clusters for the data perform closest to HBMs. Moreover, the weighted differences tend to be smaller than 0.01 and around 0.02 for the inverted differences. The results for these measures become even closer in most cases when the low frequency threshold is adopted, figure 7, as the

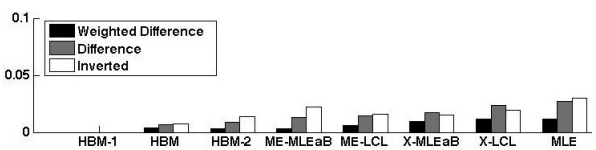

Figure 6: Model Comparisons.

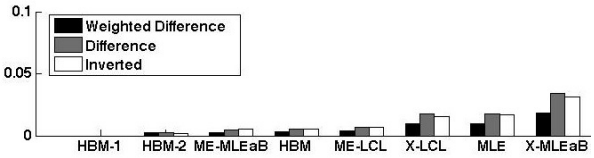

Figure 7: Model Comparison - Low Frequency Threshold.

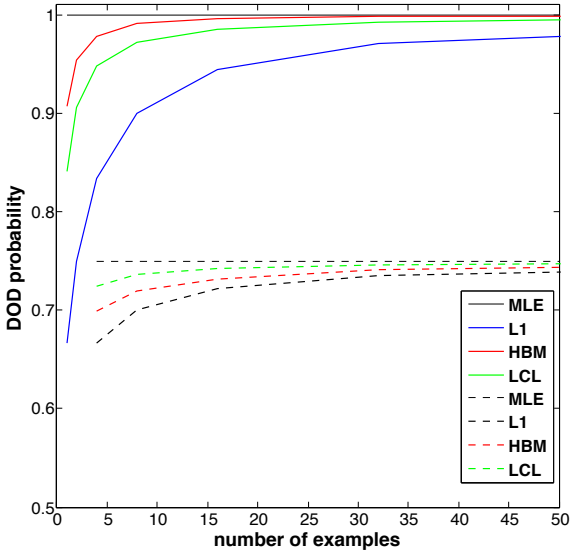

Figure 8: DOD probability evolution for models with increase in evidence

evidence reduces the influence of the prior.

To examine the decay of overgeneralization with the increase in evidence for these models, two simulated scenarios are defined for a single generic verb: one where the evidence for DOD amounts to 75% of the data (dashed lines) and in the other to 100% (solid lines), figures 9 and 8. Unsurprisingly, the performance of the models is dependent on the amount of evidence available. This is a consequence of the decrease in the influence of the priors as the sample size increases in a rate of $1/N$, as shown in figure 9 for the decrease in overgeneralization. Ultimately it is the ev-

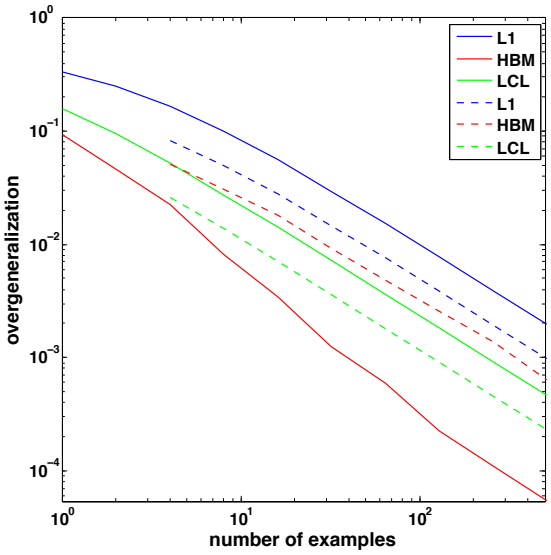

Figure 9: Overgeneralization reduction with increase in evidence

idence that dominates the posterior probability. Although the Bayesian model exhibits fast convergence, after 10 examples, the simpler model L1 is only approximately 3% below the Bayesian model in performance for scenario 1 and is still 90% accurate in scenario 2, figure 8.

These results suggest that while these models all differ slightly in the degree of overgeneralization for low frequency data and noise, these differences are small, and as evidence reaches approximately 10 examples per verb, the overall performance for all models approaches that of MLE.

5 Conclusions and Future Work

HBMs have been successfully used for a number of language acquisition tasks capturing both patterns of under- and overgeneralization found in child language acquisition. Their (hyper)parameters provide robustness for dealing with low frequency events, noise, and uncertainty and a good fit to the data, but this fidelity comes at the cost of complex computation. Here we have examined HBMs against computationally simpler approaches to dative alternation acquisition, which implement the indirect negative approach. We also advanced several measures for model comparison in order to quantify their agreement to assist in the task of model selection. The results show that the proposed LCL model, in particular, that combines class-based smoothing with maximum likelihood estimation, obtains results comparable to those of HBMs, in a much simpler framework. Moreover, when a cognitively-viable frequency threshold is adopted, differences in the performance of all models decrease, and quite rapidly approach the performance of MLE.

In this paper we used standard clustering techniques grounded solely on verb counts to enable comparison with previous work. However, a variety of additional linguistic and distributional features could be used for clustering verbs into more semantically motivated classes, using a larger number of frames and verbs. This will be examined in future work. We also plan to investigate the use of clustering methods more targeted to language tasks (Sun and Korhonen, 2009).

Acknowledgements

We would like to thank the support of projects CAPES/COFECUB 707/11, CNPq 482520/2012-4, 478222/2011-4, 312184/2012-3, 551964/2011-1 and 312077/2012-2. We also want to thank Amy Perfors for kindly sharing the input data.

References

Baker, Carl L. 1979. Syntactic Theory and the Projection Problem. *Linguistic Inquiry*, 10(4):533–581.

Briscoe, Ted. 1997. Co-evolution of language and the language acquisition device. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 418–427. Morgan Kaufmann.

Brown, Roger. 1973. *A first language: Ehe early stages*. Harvard University Press, Cambridge, Massachusetts.

Brown, Roger and Camille Hanlon. 1970. Derivational complexity and the order of acquisition of child's speech. In J. Hays, editor, *Cognition and the Development of Language*. NY: John Wiley.

Chater, Nick, Joshua B. Tenenbaum, and Alan Yuille. 2006. Probabilistic models of cognition: where next? *Trends in Cognitive Sciences*, 10(7):292 – 293.

Chomsky, Noam. 1981. *Lectures on government and binding*. Mouton de Gruyter.

Gallistel, Charles R. 2002. Frequency, contingency, and the information processing theory of conditioning. In P.Sedlmeier and T. Betsch, editors, *Frequency processing and cognition*. Oxford University Press, pages 153–171.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2003. *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 2 edition.

Gropen, Jess, Steve Pinker, Michael Hollander, Richard Goldberg, and Ronald Wilson. 1989. The learnability and acquisition of the dative alternation in English. *Language*, 65(2):203–257.

Hsu, Anne S. and Nick Chater. 2010. The logical problem of language acquisition: A probabilistic perspective. *Cognitive Science*, 34(6):972–1016.

Ingram, David. 1989. *First Language Acquisition: Method, Description and Explanation*. Cambridge University Press.

Jones, Matt and Bradley C. Love. 2011. Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34(04):169–188.

Kwiatkowski, Tom, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2010. Inducing probabilistic CCG grammars from logical form with higher-order unification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1223–1233.

Kwisthout, Johan, Todd Wareham, and Iris van Rooij. 2011. Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science*, 35(5):779–1007.

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL.

MacWhinney, Brian. 1995. *The CHILDES project: tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum Associates, second edition.

Marcus, Gary F. 1993. Negative evidence in language acquisition. *Cognition*, 46:53–85.

Marr, D. 1982. *Vision*. San Francisco, CA: W. H. Freeman.

Nematzadeh, Aida, Afsaneh Fazly, and Suzanne Stevenson. 2013. Child acquisition of multiword verbs: A computational investigation. In A. Villavicencio, T. Poibeau, A. Korhonen, and A. Alishahi, editors, *Cognitive Aspects of Computational Language Acquisition*. Springer, pages 235–256.

Parisien, Christopher, Afsaneh Fazly, and Suzanne Stevenson. 2008. An incremental bayesian model for learning syntactic categories. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, CoNLL '08, pages 89–96, Stroudsburg, PA, USA. Association for Computational Linguistics.

Parisien, Christopher and Suzanne Stevenson. 2010. Learning verb alternations in a usage-based bayesian model. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Pelleg, Dan and Andrew Moore. 2000. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 727–734, San Francisco. Morgan Kaufmann.

Perfors, Amy, Joshua B. Tenenbaum, and Elizabeth Wonnacott. 2010. Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, (37):607–642.

Ratnaparkhi, Adwait. 1999. Learning to parse natural language with maximum entropy models. *Machine Learning*, pages 151–175.

Shalizi, Cosma R. 2009. Dynamics of bayesian updating with dependent data and misspecified models. *ElectroCosmanic Journal of Statistics*, 3:1039–1074.

Sun, Lin and Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. In *EMNLP*, pages 638–647.

Villavicencio, Aline. 2002. *The Acquisition of a Unification-Based Generalised Categorial Grammar*. Ph.D. thesis, Computer Laboratory, University of Cambridge.

Wonnacott, Elizabeth, Elissa L. Newport, and Michael K. Tanenhaus. 2008. Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, 56:165–209.

Yang, Charles. 2010. Three factors in language variation. *Lingua*, 120:1160–1177.