When Suboptimal Behavior is Optimal and Why: Modeling the Acquisition of Noun Classes in Tsez

Annie Gagliardi (acg39@umd.edu)
Naomi H. Feldman (nhf@umd.edu)
Jeffrey Lidz (jlidz@umd.edu)
Department of Linguistics, 1401 Marie Mount Hall, University of Maryland, College Park, MD 20742 USA

Abstract

Children acquiring languages with noun classes (grammatical gender) have ample statistical information available that characterizes the distribution of nouns into these classes, but their use of this information to classify novel nouns differs from the predictions made by an optimal Bayesian classifier. We propose three models that introduce uncertainty into the optimal Bayesian classifier and find that all three provide ways to account for the difference between children's behavior and the optimal classifier. These results suggest that children may be classifying optimally with respect to a distribution that doesn't match the surface distribution of these statistical features.

Keywords: language acquisition; noun classes; Bayesian classification; statistical learning.

Learners are surrounded by statistical information. Considerable evidence suggests that they can make use of statistics to learn about their environment. For example, when acquiring artificial languages, children track distributional cues that allow them to discover phonetic categories (Maye, Werker & Gerken, 2002), word boundaries (Saffran, Newport & Aslin, 1996), grammatical categories (Mintz, 2003; Reeder, Newport & Aslin 2009, 2010), grammatical dependencies (Gomez & Maye, 2005; Saffran, 2001) and phrase structure (Takahashi, 2009). This leads to a commonly held belief in the language acquisition literature that children are perfect statistical learners (e.g. Elman, Bates, Johnson Karmiloff-Smith, Parisi & Plunkett 1996).

The hypothesis that children are perfect statistical learners predicts that when tested on their ability to generalize aspects of their native language in an experimental setting, children's linguistic knowledge should always reflect the distribution of statistical information in the input. However, this is not always the case. Work by Hudson-Kam and Newport (2009), for example, suggests that children are not perfectly veridical learners, in that they sometimes override statistical patterns in the service of amplifying some other facet of the language they are acquiring. As this work has largely focused on artificial language learning, here we examine another type of non-veridical statistical learning involving the acquisition of noun class (grammatical gender) in a natural language, Tsez. We present evidence showing that children exhibit behavior that is inconsistent with the statistical information available in the input when assigning novel nouns to noun classes. This inconsistent behavior suggests that there is more to language acquisition than a simple mapping of external statistical information to an internal representation of this distribution. In particular it suggests that properties of the learner shape the statistical information in the input into the subset of information that is used to guide inferences in language acquisition: the intake. We use a Bayesian model of noun classification to probe what underlies the difference in the measureable *input* and the *intake* that children use to acquire noun classes.

As a general framework, we assume that optimal performance in an experimental task involves the following four components:

(1) Accumulation of knowledge of the statistical distribution of features relating to some phenomenon
(2) Observation of features in a novel experimental item
(3) Knowledge of which features are relevant for the statistical computation
(4) Bayesian computation to determine how to generalize the phenomenon in question to the novel instance

(1) depends on the learner's ability to observe and encode a statistical distribution of features pertaining to some phenomenon. (2) is similar to (1), but refers to encoding these features given a situation where the learner will be performing a computation to classify or otherwise deal with a novel instance. (3) requires the learner to know which features are relevant for a computation and is by no means trivial, as not every feature related to every phenomenon is relevant to the associated computation. (4) is an assumption that we are making about the kind of computations that learners use distributional information for. While step (4) is often assumed to be the culprit when learners show suboptimal performance in experimental tasks, in principle steps (1) through (3) can also contribute to suboptimal performance.

Our case study on Tsez noun classification examines how each of these pieces could result in a reshaping of the statistical information in the input. We begin with an outline of the distributional information that characterizes Tsez noun classes. We then compare children's use of this information in classification with that of a naïve Bayesian classifier. Finally, we explore three models that introduce uncertainty in levels (1)-(3) from above, in an effort to determine what underlies the difference between children's performance and predictions made by the Bayesian model.

Tsez Noun Classes

Many languages make use of subclasses of nouns, called noun classes or grammatical gender. The presence and number of noun classes, as well as the distribution of individual nouns into classes varies greatly across languages, but several features remain constant. All noun class systems exhibit some degree of distributional information both internal and external to the noun. Noun internal distributional information consists of commonalities among the nouns in a class, such as semantic or phonological features. Noun external distributional information is made up of class defining information that is separate from the noun, such as agreement morphology that is contingent on noun class. We will look at noun class acquisition in Tsez as a case study.

Tsez, a Nakh-Dagestanian language spoken by about 6000 people in the Northeast Caucasus, has four noun classes. These classes can be characterized based on noun external distributional information (e.g. prefixal agreement on vowel initial verbs and adjectives) (Table 1), and noun internal distributional information (semantic and morphophonological features on the nouns themselves) (Table 2).

Table 1: Noun External Distributional Information.

Class 1	Class 2	Class 3	Class 4
∅-igu uži	j-igu kid	b-igu k'et'u	r-igu čorpa
I-good boy	II-good girl	III-good cat	IV-good soup
good boy	*good girl*	*good cat*	*good soup*

Table 2: Noun Internal Distributional Information (a selection)

Feature	Value	Class predicted	% class with this feature value	% nouns with this value in predicted class
Semantic	female	2	13	100
Semantic	animate	3	22	100
First Segment	r-	4	9	61

Gagliardi and Lidz (under review) measured noun internal distributional information by taking all nouns from a corpus of Tsez child directed speech, tagging them for potentially relevant semantic and morphophonological cues and using decision tree modeling to determine which features were most predictive of class (cf. Plaster, Harizanov & Polinsky, in press). The features shown in Table 2 are only a selection of the most predictive features of class, with only the most predictive values of these features shown.[1] The full structure of each feature that we assume in our model is given below in Table 3. Each feature has specified values that were highly predictive of some class and an unspecified value that ranges over all other possible values that were not predictive.

Table 3: Structure of Features

Feature	Specified Values	Unspecified Value
Semantic	male, female, animate	other
First segment	r-, b-	other
Last Segment	i	other

In this paper we will focus on how children use noun internal distributional information. In particular we will look at whether a child can make use of the predictive phonological and semantic information when classifying novel nouns, and how they perform when a noun has two features that make conflicting predictions. Returning to the four components of statistical learning outlined above, we will be looking at

(1) Whether Tsez children have knowledge of the noun internal distributional information
(2) Whether they can observe these features on novel nouns
(3) Whether they assume all features are relevant for classification
(4) We assume for the purposes of our analysis that the computation they make based on this information is Bayesian.

Classifying Novel Nouns in Tsez

To assess whether children can use the statistics of noun internal information available in their input, we compare classification of novel nouns by Tsez acquiring children to the classification behavior that is predicted by a Bayesian model trained on the input data from our corpus. We describe the experimental data and the model in turn.

Classification by Tsez Children

To determine whether or not children classified novel nouns consistently with the predictions made by the probabilities associated with their noun internal features, 10 native Tsez speaking children (mean: 6yrs, range: 4-7yrs) participated in a classification task. Here we give an overview of the experiment; for further details, including adult data, see Gagliardi and Lidz (under review).

Method Children were presented with unfamiliar items labeled with novel nouns by a native Tsez speaker. They were instructed to first tell a character to begin eating and then tell the character whether or not to eat the other labeled items. As the both the intransitive (eat) and transitive (eat it) forms for *eat* are vowel initial in Tsez (*–iš* and *–ac'o* respectively), classification of the novel word could be seen on the agreement prefix. Furthermore, intransitive verbs in Tsez agree with the agent (the eater) and transitive verbs agree with the patient (the thing eaten). An example trial is schematized in Table 4.

[1] Here we talk about 'noun classes' to refer what is often called grammatical gender. One of the cues to noun class is often natural gender, but this is only one of several cues, and many other nouns are in each class that don't have this (or potentially any) cue predicting their class.

The test items had either a single noun internal distributional feature from Table 2, or a combination of these features that made conflicting predictions (e.g. semantic = [animate] and initial = [r]). The exact feature combinations used in this experiment, along with the classes each feature predicts, are shown in Table 5. While these only represent a selection of the most predictive features, we focus on them here as they are a representative set of predictive semantic and phonological features.

Table 4: Example Experimental Trial

Speaker	Utterance	Action/**Conclusion**
Exper-imenter	kid girl(class2) *girl*	Points to girl on page
Child	sis, q'ano, ɬono, **j**-iš one two three CL2-eat *One two three, Eat!*	Tells *kid* to start eating using Class 2 prefix **j** / ***kid*** **is in Class 2**
Exper-imenter	zamil novel[animate]	Points to unfamiliar animal and labels it with the novel noun *zamil*
Child	zamil **b**-ac'xosi aanu zamilCL3-eat-pres.part neg pro *isn't eating the zamil*	Says whether or not the girl is eating the *zamil* using Class 3 prefix **b** / ***zamil*** **is in Class 3**

Table 5: Features Used in Experiment and Simulations

Feature	Value	Class Predicted
Semantic	female	2
Semantic	animate	3
First Segment	r	4
Semantic & First Segment	female & r	2 and 4
Semantic & First Segment	animate & r	3 and 4

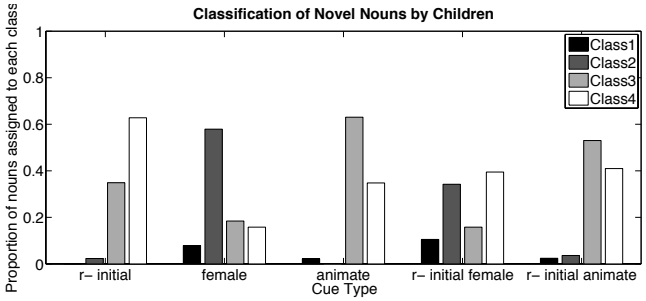

Figure 1: Proportion of novel nouns assigned to each class (by cue type) in the experimental task

Results The proportion of nouns that children assigned to each class are shown in Figure 2. When nouns had no conflicting features, children assigned more nouns to the class most strongly predicted by the feature than to any other class. However, when nouns had more than one feature that made conflicting predictions, children relied more heavily on the phonological feature [r-] than on the semantic feature. This is not likely to be predicted by the distribution of these features in the input, where nouns with the [animate] and [female] values of the semantic feature never occur in Class 4.[2]

Classification by an Optimal Bayesian Classifier

Given these experimental data, we can evaluate whether children are optimally using the statistics in their input by examining how a Bayesian model would classify each novel noun. That is, what would an ideal learner, exposed to input with these features, do when asked to classify novel words?

Our model is shown in Equation 1. The prior probability of a class $p(c)$ corresponds to its frequency of occurrence, and the likelihood terms $p(f|c)$ for each of n independent features f can be computed from feature counts in the lexicon.

$$p(c \mid f_1, f_2 \ldots f_n) = \frac{p(f_1 \mid c)p(f_2 \mid c)\ldots p(f_n \mid c)p(c)}{\sum_i p(f_1 \mid c_i)p(f_2 \mid c_i)\ldots p(f_n \mid c_i)p(c_i)} \quad (1)$$

The results of classification with this model are shown in Figure 2. Just as we did with children, we tested the model on classification with each semantic and phonological feature from Table 2 individually, as well as cases where these features were in conflict with one another. As would be expected based on the relative strength of these features (Table 2), when semantic and phonological features make conflicting predictions the model classifies in line with the predictions made by the semantic feature.

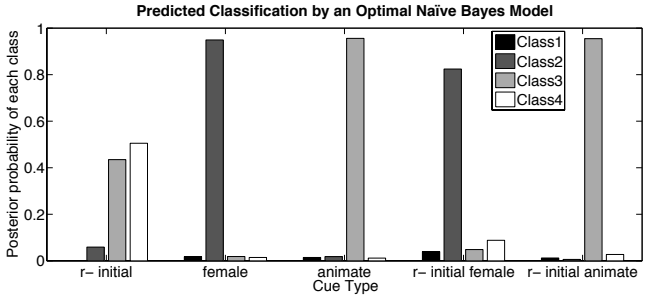

Figure 2: Predicted classification of novel nouns by an optimal naïve Bayesian classifier

The model's classification differs from that of the children in that when features made conflicting predictions the model relied on the statistically strongest cue (the semantic feature), while the children did not rely so heavily on this.

Predicting Suboptimal Performance

While children roughly align with the model when classifying based on one highly predictive feature, they diverge when features make conflicting predictions. Children appear to use phonological features out of proportion with their statistical reliability. That is, children

[2] For a more detailed description of the results of the experiment, see Gagliardi & Lidz (under review).

appear to prefer the weaker predictions made by the phonological feature to the stronger ones made by the semantic feature. In order to determine the source of this asymmetry it is useful to first consider the fundamental differences between semantic and phonological features that could lead to this kind of behavior, and then to determine where and how these factors could affect our model.

There are several differences between semantic and phonological features that could affect their use in noun classification, but here we will focus on a fundamental difference in how reliably perceived and encoded each feature type may be during early acquisition. Every time a word is uttered (or most of the time, allowing for noisy conditions and fast speech) phonological features are present. However, especially during the early stages of lexical acquisition, the meaning of a word, and thus the associated semantic features, is much less likely to be available or apparent. Below we will consider how this sort of asymmetry could lead to a disparity in the way children end up using them in novel noun classification.

Three Models of Uncertainty

The difference between semantic and phonological features could affect each of the three components from the schema of noun classification in different ways. In this section we will model each of these to see how building the asymmetry into each level changes the classification by the model.

Knowledge of Noun Internal Distributional Information

An asymmetry in the reliability with which semantic and phonological features of nouns are perceived and encoded during word learning could lead to a disparity in the way phonological and semantic features are represented as compared with how they are distributed in the input.

In our first manipulation (the Semantic Incompetence Hypothesis) we examined how classification by the model would be affected if the learner was misrepresenting some proportion of the semantic features that they should have encoded on nouns in their lexicon. We assume that learners represented the remaining proportion of nouns as predicted (accurately observing features during the experiment and assuming that both semantic and phonological features were relevant in classification). In doing this, we assume that learners' beliefs about which features are predictive of which class is built up as they observe different feature values on words belonging to different classes. One way of quantifying this is by modeling the learner's belief about the likelihood terms $p(f|c)$ from Equation 1 under the assumption that these beliefs are derived from the counts that a learner accumulates of nouns in each class that contain a given feature. We assume learners use a multinomial model with a uniform Dirichlet prior distribution to estimate the proportion of items each class c that contain a particular value k for feature f. Under this assumption, each likelihood term is equal to:

$$p(f = k \mid c) = \frac{N_{c,f=k} + 1}{N_c + K} \quad (2)$$

where N_c denotes the number if nouns in the class, $N_{c,f=k}$ denotes the number of nouns in the class for which the feature has value k, and K is the number of possible values for the feature.

We introduce misrepresentation of semantic features into this model by manipulating the number of observations of a noun with a certain feature value in each class. Since the semantic incompetence hypothesis posits that children misrepresent semantic feature values some proportion of the time, we reduce the count of nouns in each class that contain the relevant semantic features, changing them instead to the unspecified feature value [other]. We then compute the posterior probability of noun class membership using these adjusted feature counts. We can use this model to ask how low the counts would have to be in order for children's behavior to be optimal with respect to their beliefs.

We evaluated the model by comparing its behavior to children's behavior from the classification task. The model produced a close fit to the data in each condition (Figure 3). Furthermore, the estimated degree of misrepresentation was highly consistent across all semantic features and conflicting feature combinations. The best fitting level of uncertainty ranged from 0.96-0.91, meaning that children would be only using 4-9% of the semantic cues available to them. A generalized likelihood ratio test in which the level of misrepresentation was held constant across simulations (0.95) demonstrates that our semantic incompetence model significantly outperforms the optimal naïve Bayesian classifier ($p < 0.0001$).

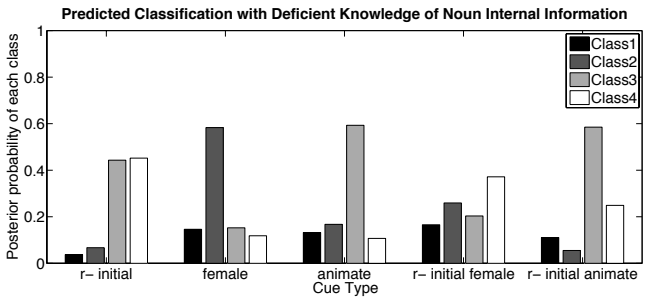

Figure 3: Classification of novel nouns as predicted by a Naïve Bayes Classifier with 95% of predictive semantic features misrepresented as [other].

Although this model produces a close fit to the empirical data, it predicts an extremely high degree of misperception. To understand why this is the case, consider that using likelihood terms for each class that are proportional to the true empirical counts $\frac{N_{c,f=k}}{N_c}$ would yield optimal noun classification performance, regardless of the exact proportion of time children are misrepresenting features. That is, substituting $\beta * p(f_1|c)$ for each term $p(f_1|c)$ in Equation 1, where β is a constant denoting the degree of misperception, does not result in any change in the posterior probability distribution. This analysis suggests that changes in model predictions under this account of feature

misrepresentation occur primarily for low empirical feature counts, when the model relies heavily on pseudocounts from the Dirichlet prior distribution.

Observation of semantic and phonological features on novel nouns A second possibility is that children have little trouble perceiving, encoding and representing features on the words in their lexicon, but that the semantic features on the experimental items (as they are presented as flat pictures in a book) are unreliably perceived and encoded. We call this the Experimental Reject Hypothesis.

In this manipulation we investigate what would happen if a learner had a lexicon that faithfully represented the predictive features as they were distributed in the input and assumed both semantic and phonological features were relevant to classification, but didn't reliably encode semantic features on experimental items. To do this we use a mixture model, where some proportion of the time (1- β) an item that was supposed to have the specified semantic feature value [animate] or [female] (denoted as [spe]) it would be classified as with that value, the rest of the time (β) it would be classified as if it had the unspecified value [other]. This yields the following model:

$$p(c \mid f_1, f_2) = (1-\beta)\frac{p(f_1=[spe]\mid c)p(f_2\mid c)p(c)}{\sum_i p(f_1=[spe]\mid c_i)p(f_2\mid c_i)p(c_i)} + \beta\frac{p(f_1=[other]\mid c)p(f_2\mid c)p(c)}{\sum_i p(f_1=[other]\mid c_i)p(f_2\mid c_i)p(c_i)} \quad (3)$$

As with the semantic incompetence model, we found the best-fitting value of β and evaluated the model by comparing it to children's behavior. This model again produced a close fit for all feature values (Figure 4). The model showed a consistent degree of misperception across all semantic features and feature combinations. The best fitting level value of β ranged from .49 to .83, where 58% was the best fit overall. This means that children would be misperceiving semantic features on 58% of the experimental items. A generalized likelihood ratio test indicates that the experimental reject model also significantly outperforms the optimal naïve Bayesian classifier ($p < 0.05$).

Assumption that all features are relevant for classification The asymmetry between the reliability of perceiving and encoding phonological as compared to semantic features could also engender a bias to prefer phonological information for classification decisions, as phonological information has been reliably available for a longer period of time. Our third model, embodying the Phonological Preference Hypothesis, therefore looked at what would happen if we had a learner that was biased not to use semantic features in classification some proportion of the time, even if these features were represented just as distributed in the input and accurately perceived during the experimental task. We used a second mixture model, this time looking at the mixture of a Bayesian classifier that used both semantic and phonological features, and one that only used phonological features.

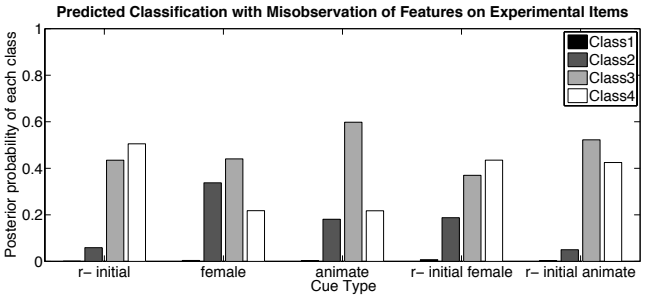

Figure 4: Classification of novel nouns as predicted by a model that misobserves semantic features on experimental items 58% of the time

The crucial difference between this model and the experimental reject model is that in the experimental reject model semantic features are always used, but are encoded as the wrong value (the unspecified [other] value) some proportion of the time, whereas in the phonological preference model, semantic features do not factor into the calculation at all some proportion of the time (β). The model can be seen in Equation 4.

$$p(c \mid f_1, f_2) = (1-\beta)\frac{p(f_1=[sem]\mid c)p(f_2\mid c)p(c)}{\sum_i p(f_1=[sem]\mid c_i)p(f_2\mid c_i)p(c_i)} + \beta\frac{p(f_2\mid c)p(c)}{\sum_i p(f_2\mid c_i)p(c_i)} \quad (4)$$

Again we evaluated the model against the children's classification data and found a close fit (Figure 5). The best fitting value of β ranged from .49 to .83, and was .65 over all, meaning that children would be choosing not to use semantic features on 65% of classification decisions. A generalized log likelihood test showed that this model also significantly outperformed the optimal naïve Bayesian classifier ($p < 0.0001$)

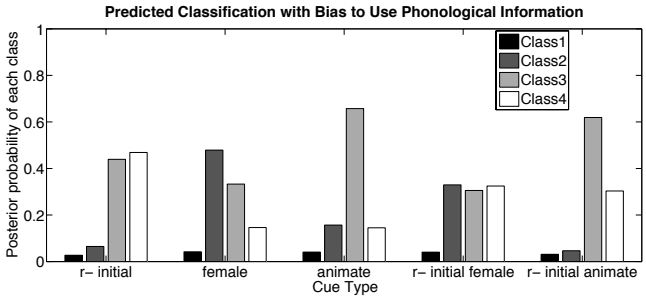

Figure 5: Classification as predicted by a model biased not to use semantic information 65% of the time

Discussion

Tsez noun classes are characterized by both semantic and phonological features. Children have been shown to be able to use these features when classifying novel nouns. Here we showed that their classification patterns differ from those of an optimal Bayesian classifier when nouns have semantic and phonological features that make conflicting predictions.

We also presented three models that take into account ways in which the difference between semantic and phonological features could lead to children's apparent preference to use the less reliable phonological features. These models examined how classification would look if a learner had (a) misrepresented semantic features in the lexicon, (b) misencoded semantic features during the classification experiment, or (c) developed a bias to use phonological information in noun classification due to its higher reliability in the early stages of lexical acquisition. All three models fit children's data significantly better than the optimal naïve Bayesian classifier did. This suggests that although originally children did not look as though they were behaving optimally with respect to the input, they may well be behaving optimally with respect to their intake, that is, the input as they have represented it.

It is not obvious how one would best to evaluate the alternative models with respect to one another. For example, each model yielded a different best-fit parameter, corresponding to a different degree of misrepresentation or bias. While these best fitting parameters may differ in terms of their 'reasonableness' (i.e. misrepresenting 95% of semantic features in the lexicon at age 6 seems quite high), it isn't immediately clear how to measure reasonableness, or how to compare it across models. Furthermore, it is likely that a combination of all three of these processes (and perhaps more that we haven't considered here) is influencing children's classification decisions. This could potentially be explored through a combined model; however, as all of these models fit the data so closely, it would be difficult to determine which and to what extent each type of misrepresentation or bias is involved.

This work has several important implications for research statistical learning and language acquisition. First, and most broadly, by combining experimental data from children acquiring an understudied language with computational modeling techniques, we found a better understanding of both children's acquisition of Tsez, and the role of statistical cues in language acquisition. Tsez was an ideal language to look at, as feature types differed in their reliability as cues to noun class. However, we expect that these results will be generalizable across languages, as the relative difficulty of acquiring semantic, as compared to phonological, features of words will be consistent cross linguistically.

Second, we identified an area where children's behavior does not appear to reflect the ideal inferences licensed by the statistical patterns in the input. Three models allowed us to investigate the source of this asymmetry. While each model differed in where the asymmetry came from, all employed a weakening of the statistical import of semantic features. This is a distinct pattern from the finding that children learning an artificial language amplify an already strong statistical tendency (Hudson-Kam & Newport, 2009). Further research will determine whether or not these patterns could be in some way related.

Next, we showed that it is possible for a learner to be suboptimal and Bayesian at the same time. That is, we demonstrated that while children's behavior does not align with the predictions made by the optimal Bayesian classifier, it can be predicted by modifying the terms of this classifier in reasonable ways. Thus we were able to model children's suboptimal behavior using a Bayesian model, rather than adopting some other system of computation.

Finally, our models showed that it is plausible that these children are indeed behaving optimally with respect to some statistical distribution, just not one directly measureable from the input. This point is crucial as researchers extend accounts of statistical learning to a greater range of problems, highlighting the fact that the critical question isn't whether or not children are using statistics to acquire language, but what statistics they are using.

Acknowledgments This research was supported by NSF IGERT 0801465 and a NSF GRF to Gagliardi. We would like to thank Masha Polinsky, the UMD Cognitive Neuroscience of Language Lab, the UMD Project on Children's Language Learning and the UMD Computational Psycholinguistics group for helpful discussion and assistance.

References

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: MIT Press.

Gagliardi, A., & Lidz, J. (Under review) Separating input from intake: Acquiring noun classes in Tsez.

Gómez, R.L., & Maye, J. (2005). The Developmental Trajectory of Nonadjacent Dependency Learning. *Infancy*, *7*, 183–206.

Hudson Kam, C.L., & Newport, E.L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*, 30–66.

Mintz, T.H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*. *90*, 91–117.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101–B111.

Plaster, K., Polinsky, M., & Harizanov, B. (In Press). Noun Classes Grow on Trees: Noun Classification in the North-East Caucasus. *Language and representations* (tentative). John Benjamins

Reeder, P.A., Newport, E.L., & Aslin, R.N. (2009). The role of distributional information in linguistic category formation. In N. Taatgen and H. van Rijn (eds), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Reeder, P.A., Newport, E.L., & Aslin, R.N. (2010). Novel words in novel contexts: The role of distributional information in form-class category learning. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*, 606–621.

Saffran, J.R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, *44*, 493–515.

Takahashi, E. (2009). *Beyond statistical learning in the acquisition of phrase structure*. College Park, MD: University of Maryland dissertation.