COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 41 (2017) 188–217

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12373

Modeling Statistical Insensitivity: Sources of Suboptimal Behavior[†]

Annie Gagliardi,[a] Naomi H. Feldman,[b,c] Jeffrey Lidz[b]

[a]*School of Informatics, University of Edinburgh*
[b]*Department of Linguistics, University of Maryland*
[c]*Institute for Advanced Computer Studies, University of Maryland*

Received 29 September 2013; received in revised form 23 September 2015; accepted 24 September 2015

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 use rational analysis to investigate the hypothesis that children are classifying nouns optimally with respect to a distribution that does not match the surface distribution of statistical features in their input. We propose three ways in which children's apparent statistical insensitivity might arise, and find that all three provide ways to account for the difference between children's behavior and the optimal classifier. A fourth model combines two of these proposals and finds that children's insensitivity is best modeled as a bias to ignore certain features during classification, rather than an inability to encode those features during learning. These results provide insight into children's developing knowledge of noun classes and highlight the complex ways in which statistical information from the input interacts with children's learning processes.

Keywords: Language acquisition; Bayesian modeling; Statistical learning; Tsez; Noun classes

1. Introduction

Language learners are surrounded by statistical information. Considerable evidence suggests that they can make use of this information to learn about their linguistic environment. For example, when acquiring artificial languages, children track distributional cues that allow them to discover phonetic categories (Maye, Werker, & Gerken, 2002), word

Correspondence should be sent to Naomi H. Feldman, Department of Linguistics and Institute for Advanced Computer Studies, University of Maryland. E-mail: nhf@umd.edu.

[†]Previous versions of this work were presented at the 2012 LSA workshop on Psychocomputational Models of Language Acquisition and the 34th Annual Conference of the Cognitive Science Society.

boundaries (Saffran, Aslin, & Newport, 1996), grammatical categories (Mintz, 2003; Reeder, Newport & Aslin, 2009, 2010), grammatical dependencies (Gomez & Maye, 2005; Saffran, 2001), and phrase structure (Takahashi, 2009). The apparent abundance of statistical information, combined with children's ability to draw inferences from distributional information, leads to a common approach in the language acquisition literature to study language acquisition by examining the way in which a perfect statistical learner would acquire language (e.g., Elman et al., 1996; Frank, Goodman, & Tenenbaum, 2009; Goldwater, Griffiths, & Johnson, 2009; Perfors, Tenenbaum, & Regier, 2011). This parallel depends on children being able to track statistical information reliably and use available and relevant information to draw inferences about the language being acquired. If children cannot track all of the available information, or do not use the right information to solve a given problem, their inferences could cause them to look like less than perfect statistical learners.

When children are tested on their ability to generalize aspects of their native language in experimental settings, their linguistic knowledge does not always reflect the distribution of statistical information in the input. Work by Hudson-Kam and Newport (2009), for example, suggests that children are not perfectly veridical learners, at least in an artificial language context, in that they sometimes override statistical patterns in the service of amplifying some other facet of the language they are acquiring. Singleton and Newport (2004) show a similar pattern wherein a child exposed exclusively to non-native input regularized and hence amplified certain statistical patterns in the input. We also see cases of overgeneralization, for example with regular past tense morphology, which can be seen as a temporary amplification of the pattern of regular past tense found in the input (Brown, 1973). Lidz, Gleitman, and Gleitman (2003) found that Kannada acquiring children rely more on argument number than causative morphology to learn about novel verbs, even though the causative morphology is a more statistically reliable cue. This type of pattern, where the inferences children make do not match the distributional information in the environment, allows us to learn more about how children draw inferences, and what kinds of biases they bring to the task of language acquisition.

Here, we investigate this type of pattern by examining the acquisition of noun class (grammatical gender) in a natural language, Tsez, where children acquiring noun classes do not appear to make optimal use of the statistical information available. We use computational modeling as a probe into the source of this pattern. We review evidence showing that children exhibit behavior that is inconsistent with the statistical information available in the input when assigning novel nouns to noun classes (Gagliardi & Lidz, 2014). 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 (Fodor, 1998; Gagliardi & Lidz, 2014). This work highlights the need to separate the input available in the environment from the intake, or the information learners make use of in language acquisition.

We use a Bayesian model of noun classification to probe what underlies the difference between the measureable input and the intake that children use to acquire noun classes. We

adopt the approach of rational analysis (Anderson, 1990), building a formal model that makes minimal assumptions about computational costs and then revising this model based on detailed comparisons with children's behavior. As a general characterization of the problem, 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. Computation to determine how to generalize the phenomenon in question to the novel instance.

Building a model with all of these components intact allows us to characterize the behavior of a learner with minimal constraints. As the use of a statistical cue for learning depends on the learner's ability to perceive it, and on the ability of their computational system to deem it relevant for learning, we can revise (1)–(3) to probe how learners' behavior is optimized with respect to the environment, given the constraints of the learner (c.f. Pearl, Goldwater, & Steyvers, 2011). (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 subjects show suboptimal performance in experimental tasks (Sternberg & McClelland, 2011; Tversky & Kahneman, 1974; Wason, 1968), in principle steps (1) through (3) can also contribute to suboptimal performance. That is, when people behave suboptimally, we must ask whether they have some problem with parts (1)–(3), yet are generalizing optimally with respect to whatever they have available, or whether (1)–(3) are non-problematic but they are not inferring the optimal solution.[1]

Our case study on Tsez noun classification examines how each of these pieces could result in a behavioral pattern that on the surface appears to be suboptimal use 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 naive Bayesian classifier. Finally, we build four models that explore what classification would look like if uncertainty were introduced into levels (1)–(3) from above, in an effort to determine what underlies the difference between children's performance and predictions made by an optimal Bayesian model.

2. Tsez noun classes

Many languages treat subclasses of nouns differently for the purposes of grammatical agreement and concord processes. The presence and number of these noun classes[2]

(sometimes called genders), as well as the distribution of individual nouns into classes, vary greatly across languages, but several features remain constant. All noun class systems are characterized by distributional information, both internal and external to the noun (Corbett, 1991). 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 6,000 people in the Northeast Caucasus (Bokarev, 1959; Comrie & Polinsky, 1998; Comrie, Polinsky, & Rajabov, 1998; Polinsky, 2000), 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 features on the nouns themselves) (Table 2). The existence of these noun internal features, and in particular, the fact that these features differ in their reliability as cues to noun class, make Tsez an ideal language in which to study differences between the input available in the linguistic environment and the intake that children make use of. This combination of data from an understudied language and computational modeling techniques allows us to probe not only children's acquisition of Tsez but also the role of statistical information in language acquisition.

Gagliardi and Lidz (2014) 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, Polinsky, & Harizanov, 2009). The features shown in Table 2 are only a selection of the most predictive

Table 1
Noun external distributional information

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

Table 2
Most predictive noun internal distributional information

Feature	Value	Class Predicted	Probability of Feature Given Class	Probability of Class Given Feature
Semantic	male	1	1	1
Semantic	female	2	.22	1
Semantic	animate	3	.13	1
First Segment	b-	3	.10	.51
First Segment	r-	4	.09	.61
Last Segment	-i	4	.34	.54

Table 3
Structure of features

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

features of class, with only the most predictive values of these features shown.[3] The full structure of each feature that we assume in our model is given in Table 3. Each feature has specified values that are highly predictive of some class and an unspecified value that ranges over all other possible values that are not predictive.

In this paper, we use a Bayesian model of noun classification to investigate how children use noun internal distributional information. In particular, we look at whether a child can make use of the predictive phonological and semantic information when classifying novel nouns, how they perform when a noun has two features that make conflicting predictions, and what factors could underlie seemingly suboptimal behavior. A Bayesian model requires us to make explicit what information is available for a learner to use, and demonstrates what optimal use of this information would look like. If we see divergences between children's performance and the predictions of the Bayesian model, we can revise our model to ask whether some disturbance in the input (causing a difference between the input and the encoded intake) would predict results in line with children's performance. This allows us to examine what an optimal solution given suboptimal encoding of the input would look like. 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.

We assume for the purposes of our analysis that the computation they make based on this information is Bayesian.

3. 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. Below, we describe the experimental data reported in Gagliardi and Lidz (2014), and then describe our model.

Table 4
Example experimental trial

Speaker	Utterance	Action
Assistant	"kid" girl (Class 2) *girl*	*explains task, points to human character and labels it*
Child	'sis, q'ano, ɬono j-iʃ one, two three, Cl2-eat *One, two three, Eat!*	*instructs character*
Assistant	"buq" sun (Class 3) *sun*	*points to sun, labels it*
Child	"buq b-ac'xosi aanu" sun Cl3-eat-pres.part neg *pro isn't eating the sun*	*instructs character, describes scene*
Assistant	"k'uraj" onion (Class 4) *onion*	*points to onion, labels it*
Child	'k'uraj r-ac'o' onion Cl4-eat *eat the onion*	*instructs character, describes scene*
Assistant	"zamil" nonce (target Class 3) *zamil*	*points to nonce animal, labels it*
Child	"zamil b-ac'xosi aanu" zamil Cl3-eat-pres.part neg *pro isn't eating the zamil*	*instructs character, describes scene*

3.1. 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, Gagliardi and Lidz tested 10 native Tsez-speaking children (mean: 6 years, range: 4–7 years) in a classification task. Here, we give an overview of the experiment; for further details, see Gagliardi and Lidz (2014).

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. In Tsez, prefixal verbal agreement with noun class is visible only on vowel initial stems. As both the intransitive (eat) and transitive (eat it) forms for eat are vowel initial in Tsez (–i*s* 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 obligatorily agree with the patient (the thing eaten). An example trial is schematized in Table 4.

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 the 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.

The proportion of nouns that children assigned to each class is shown in Fig. 1. When nouns had no conflicting features, children assigned more nouns to the class most strongly predicted by the feature than to any other class. For example, a novel noun with the [animate] feature, whose label had the [other] values for both phonological features (e.g., a depiction of an invented animal with octopus legs and a duck head called a *zamil*) was likely to be put in Class 3. Similarly, a novel noun with the semantic feature [other] that was labeled with an r-initial label (e.g., a novel inanimate object called a *rega*), was most likely to be put classified as Class 4. However, when nouns had more than one

Table 5
Features used in experiment and simulations

Feature Label	Feature	Value	Class Predicted
f1	Semantic	male	1
f1	Semantic	female	2
f1	Semantic	animate	3
f2	First Segment	b	3
f2	First Segment	r	4
f3	Last Segment	i	4
f1 & f2	Semantic & First Segment	male & b	1 and 3
f1 & f2	Semantic & First Segment	female & r	2 and 4
f1 & f2	Semantic & First Segment	animate & r	3 and 4
f1 & f3	Semantic & Last Segment	animate & i	3 and 4

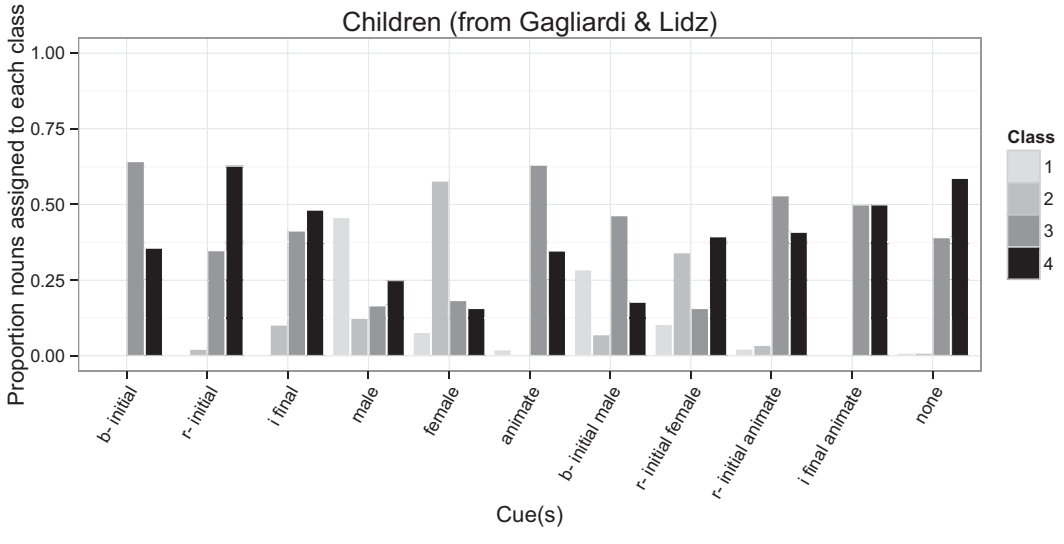

Fig. 1. Proportion of novel nouns assigned to each class (by cue type) in the experimental task by children.

feature that made conflicting predictions (e.g., a novel animal that was labeled *resu*), children relied more heavily on the phonological features ([b-], [r-], and [-i]) 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 [female] and [animate] values of the semantic features never occur in Class 4, and those with the [male] feature do not occur in Class 3.

Gagliardi and Lidz also tested adult Tsez speakers in the same paradigm. Their results are shown in Fig. 2. While the data are again somewhat noisy, we can see that adults, unlike children, appear to rely on the statistically strongest cue. That is, when phonological and semantic cues make conflicting predictions, adults reliably classify the novel noun according to the semantic (and statistically stronger) cue. The only exception to this is when the phonological cue [b-] conflicts with the semantic one [male], where it looks as though speakers rely on the phonology rather than the semantics. However, it looks as though speakers do this with all nonce words with the feature [male], not only those with conflicting phonological information, suggesting that there is more factoring into their classification than the mere combination of probabilities associated with features. This behavior differs from children, who do assign nonce words with the feature [male] to Class 1 when there is no conflicting feature. The reasons for not assigning items that have both a semantic cue [male] and a phonological cue [b-] to Class 1 may therefore differ between children and adults. Due to its heavily constrained semantic space, Class 1 may behave almost like a closed class to adults, and new members would be very unexpected. Children, however, may not have formed this generalization yet. This difference is important to keep in mind throughout the paper. While we are investigating the effect that the probabilities of a noun's features have on classification, we do not pretend that these are the only factors that influence noun classification. In the discussion below, we will return to both the fact that there was some overall noise in the adult classification behavior, as

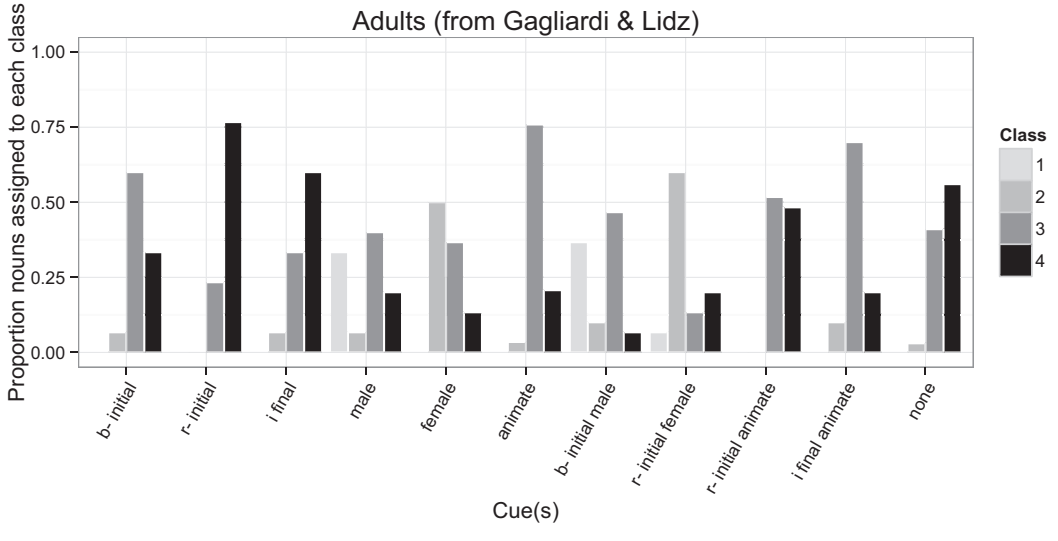

Fig. 2. Proportion of novel nouns assigned to each class (by cue type) in the experimental task by adults.

well as their reluctance to put new nouns in Class 1. For now, we simply state that the process is undoubtedly more complex, and our goal in the paper is to understand how one component of it works.

3.2. Classification by an optimal Bayesian classifier

Given these experimental data, we can evaluate our intuition that children are not 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 do when asked to classify novel words, if exposed to input reflecting the distribution of these features in the corpus of child-directed Tsez speech?

Our model is shown in Eq. 1. We use this model to calculate the posterior probability of each class for a noun, given the features on that noun ($\mathbb{P}(c|f)$). This way we can see how the model would classify a noun with certain features, based on the probability of finding a noun in any class (the prior) and the distribution of features across classes (the likelihood). The prior probability of a class $\mathbb{P}(c)$ corresponds to its frequency of occurrence (by types), and the likelihood terms $\mathbb{P}(f|c)$ for each of n independent features f can be computed from smoothed feature counts in the lexicon.[4] For example, the posterior probability of a novel noun being in Class 3 given the feature values (f_1 = [animate], f_2 = [r-initial] and f_3 = [other], ($\mathbb{P}(c|f_1,f_2,f_3)$) is proportional to the size of Class 3 ($\mathbb{P}(c)$) and the number of nouns in Class 3 with each of these feature values ($\mathbb{P}(f_1|c)$, $\mathbb{P}(f_2|c)$, $\mathbb{P}(f_3|c)$). This formulation of the model relies on the assumption that each of the features used by the model is independent. To investigate the independence of these features, we calculated the marginal entropy of each feature on nouns in each class, and compared the sum of this to the joint entropy of all features on nouns in a given class. We found the numbers to be comparable, supporting our assumption that the features are independent.[5]

$$\mathbb{P}(c_i|f_1 \ldots f_n) = \frac{\mathbb{P}(f_1|c_i) \ldots \mathbb{P}(f_n|c_i) \cdot \mathbb{P}(c_i)}{\sum_{c_j \in \{\text{all classes}\}} \mathbb{P}(f_1|c_j) \ldots \mathbb{P}(f_n|c_j) \cdot \mathbb{P}(c_j)} \tag{1}$$

The results of classification with this model are shown in Fig. 3. 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. With only one predictive feature (e.g., f_1 = [animate], f_2 = [*other*] and f_3 = [*other*], or f_1 = [*other*], f_2 = [$r-initial$] and f_3 = [*other*]), the model assigns most of the probability to the class predicted by the predictive feature value (Class 3 with the value animate, Class 4 with the value r-initial). Of course, depending on the strength of the feature (Table 2), the amount of probability that the model assigns to each class varies: it assigns more probability for Class 3, given f_1 = [animate], than to Class 4 given f_2 = [r-initial]. 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, which is

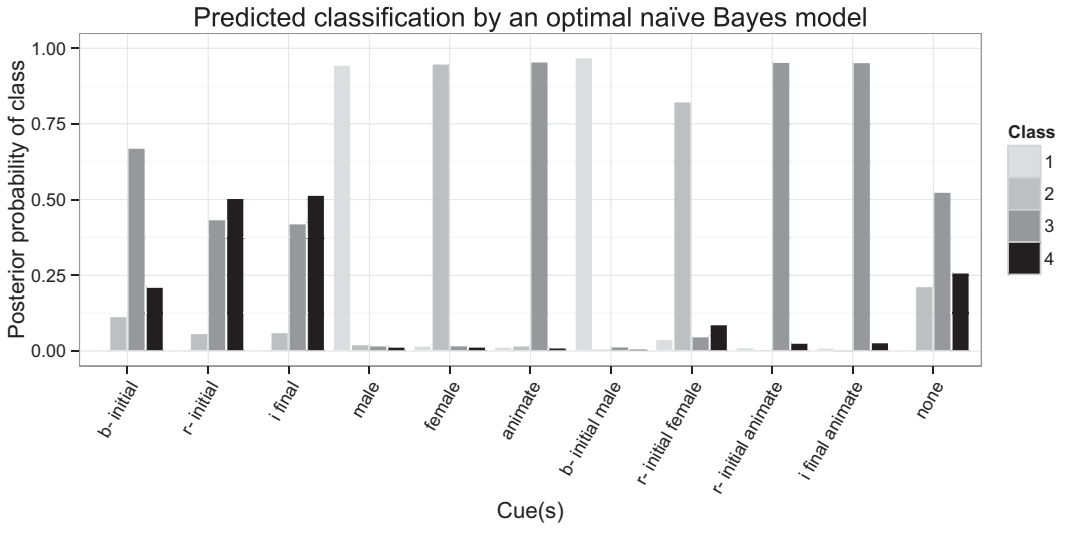

Fig. 3. Predicted classification of novel nouns by an optimal naive Bayesian classifier.

stronger. For example, when $f_1 = [\text{animate}]$, $f_2 = [\text{r-initial}]$ and $f_3 = [\text{other}]$, most of the probability is assigned to Class 3, as predicted by the feature value animate.

Crucially, the model's classification differs from that of the children, in that when features made conflicting predictions, the model relies on the statistically strongest cue (the semantic feature), while the children did not rely so heavily on this. This can be seen especially well in the following cases: the model puts the vast majority of probability in Class 1 for b-initial males while children put more nouns in Class 3; the model puts most of the probability in Class 2 for r-initial females while children split their classification between Classes 2 and 4; and the model puts the majority of probability in Class 3 for both r-initial and i-final animates, but children split the classification between Classes 3 and 4. Additionally, when the model had "no cue" (that is, when all values of all features were set to "other"), the model put the most probability mass in Class 3, while children tended to assign nouns to Class 4 (but did not do so exclusively). Returning to the theme of rational analysis, we can see that the model, when given no limitations, does not seem to match children's behavior. From this point forward we will iterate, building models that incorporate one or more limitations that children might bring to the task of noun classification, to see if we can find a closer fit between model predictions and children's behavior.

At this point, it is necessary to return briefly to adult behavior as well. Comparing their data from the experimental task to the ideal learner, it looks as though they are performing suboptimally as well. While this is in some sense true, in that there is a great deal of noise in the classifications elicited from adult speakers for novel nouns, there is one major difference between adults and children—namely their treatment of nouns with conflicting cues. In the adult case, we see that adults generally use the semantic cue (with the exception of novel males which we attribute to other factors), while children seem to prefer the phonological one. It is this qualitative shift, which is mirrored when comparing

children to both adults and the ideal learner that we seek to capture in our models. There are many more factors at work in both noun classification and participation in an experimental task than we include in our model, meaning we do not expect to find a perfect fit to adults in the ideal learner, nor to children in any one of our models.

4. Predicting seemingly 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 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 what kinds of differences between semantic and phonological features 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 (cf. Bloom, 2000; Carey & Bartlett, 1978; Gillette, Gleitman, & Gleitman, 1999; Gleitman, 1990). 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.

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 investigate several hypotheses concerning where the uncertainty in noun classification could lie. By building uncertainty into different components of noun classification, we are able to see how well each hypothesis predicts children's behavior, as well as what kind of assumptions we need to make about noun classification in order for the model to work well.

4.1. Hypothesis 1: Misrepresentation

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. That is, if children represent phonological features of nouns more reliably than semantic ones, and their classification reflects the intake, what is represented in the child's lexicon (as opposed to what is measureable in the input), their

classification may look like it relies more heavily on phonological features than semantic ones.

In our first manipulation (the Misrepresentation hypothesis), we wanted to see how classification by the model would be affected if the learner was misrepresenting some proportion of the features that they should have encoded on nouns in their lexicon. We predict that a model with a lexicon where phonological features are represented faithfully, but semantic features are represented fairly unreliably, will best fit the children's data. We assume that learners classify 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), but in doing so, rely on a lexicon that does not accurately represent the features on the nouns in the input.

We assume that learners' beliefs about which features are predictive of which class are 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 $\mathbb{P}(f|c)$ from Eq. 1. Recall that we assumed 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:

$$\mathbb{P}(f = k|c) = \frac{N_{c,f=k} + 1}{N_c + K} \tag{2}$$

where N_c denotes the number of 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. The 1 added to each feature count is a smoothing parameter denoting a uniform prior distribution over the probabilities of different feature values.

We introduce misrepresentation of features into this model by manipulating the number of observations of a noun with a certain feature value in each class. Since the Misrepresentation hypothesis posits that children misrepresent feature values some proportion of the time, we use a discounting parameter to reduce the count of nouns in each class that contain the relevant features, changing them to the underspecified "other" feature value.[6] For each feature type (semantic and phonological), we have a parameter that indexes how likely that feature type is to be misrepresented. We then compute the posterior probability of noun class membership using the adjusted feature counts. By finding the best fitting values of the two discounting parameters, we can see how the feature counts would have to shift in order for children's behavior to be optimal with respect to their beliefs. As mentioned above, we predict that more inaccurate representation of semantic features (as opposed to phonological ones) will best fit the children's data, as children appear to put less weight on the semantic features.[7]

We found the best fitting value of these two parameters using a built-in Matlab optimization procedure (fminsearch). The best fitting value of the semantic parameter was quite high (0.95), while the best fitting value of the phonological parameter was 0. This means that the best fitting model heavily discounted semantic information and did not

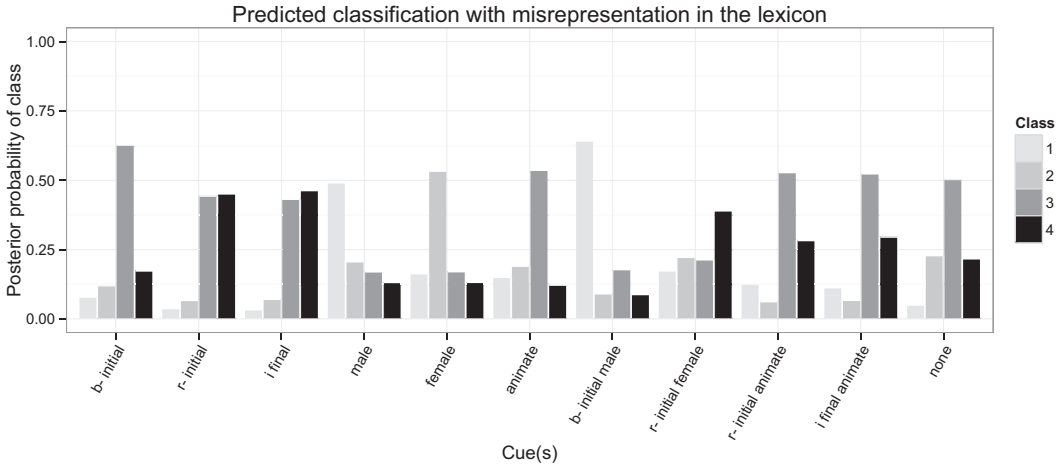

Fig. 4. Classification of novel nouns as predicted by a Naive Bayes Classifier with 95% of predictive semantic features misrepresented as [other].

discount phonological information, in line with our hypothesis that children are underusing semantic information. We evaluated the model with these parameter values by comparing its behavior to children's behavior from the classification task. This model produced a closer fit to the data in each condition than the model with no misrepresentation (Fig. 4). This is in line with our prediction that children's behavior could stem from semantic features being less reliably encoded than phonological ones. The log likelihood of the model with its best fitting parameter given the experimental data was -538, compared to a log likelihood of -746 from the classifier that was constructed from the true empirical probabilities of the feature values. A generalized likelihood ratio test demonstrates that our semantic incompetence model significantly outperforms the optimal naive Bayesian classifier ($\chi^2(2) = 416$, $p < .0001$).

Although introducing misrepresented features into the model yields a closer fit to the children's data, it is not a perfect fit. In particular, the Misrepresentation model still predicts that children would put the vast majority of novel nouns with the feature [male] into Class 1, even when these have the word initial feature [b-]. Additionally, while we see some shift toward Class 4 in the nouns with [animate] and [r- initial] or [-i final], the pattern is not as strong as what we see in the children's data. Similarly, this model does not capture the pattern that we see in children's classification when all feature values are [other]. Importantly, while these patterns are not well-fit by the simple model we have proposed, this model captures what we take as the basic pattern of interest in the data: the reliance on phonological over semantic cues when the two conflict. We return below to a discussion of children's overuse of Class 4 relative to the model predictions.

Additionally, although this model produces a fairly close fit to the empirical data, it predicts an extremely high degree of misrepresentation. To understand why this is the case, consider that using likelihood terms for each class that are proportional to the true

empirical counts would yield optimal noun classification performance, regardless of the exact proportion of time children are misrepresenting features. That is, substituting $(1 - \beta) * \mathbb{P}(f1|c)$ for each term $\mathbb{P}(f1|c)$ in Eq. 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 the smoothing parameter from the Dirichlet prior distribution. When enough nouns are represented correctly in the lexicon so that the actual counts are much larger than the smoothing parameter, lexical misrepresentations of this sort are not predicted to have a substantial effect on children's behavior.

Due to the high level of misrepresentation necessary to get a close fit to the children's data, misrepresentation is unlikely to be the driving factor behind children's suboptimal performance. While it is no doubt possible that children have a higher proportion of semantic features misrepresented than phonological ones, and that semantic incompetence makes some contribution to children's performance, it does not seem likely that at this age, these basic semantic features would be so regularly misrepresented. That is, it seems unreasonable that 4- to 7-year-old children would not know that 95% of animates are animate, 95% of females are female, and 95% of males are male. There are several reasons that these numbers seem high. First, children have been shown to use animacy as a cue in verb learning at as early as 2–3 years (Becker, 2007; Bunger & Lidz, 2006). Second, while reliable knowledge of natural gender appears to come online later than animacy, children do appear to make some gender based distinctions by 2 years (Martin, Ruble, & Szkrybalo, 2002). Even if children's somewhat delayed knowledge of natural gender were behind the effect we see, we might expect it to be limited to natural gender, and that the usefulness of non-human animacy as a cue would not be affected. We do not see such a difference, however. To summarize, while it is entirely possible that children misrepresent some proportion of semantic features, it does not seem likely that they misrepresent them to the degree predicted by Model 1. Because of this unreasonable level of semantic misperception, we move on from this hypothesis to look at other ways in which uncertainty could affect noun classification.

4.2. Hypothesis 2: Misperception

A second possibility for explaining why children appear to use phonological information out of proportion with its statistical reliability is that children have little trouble perceiving, encoding, and representing features on the words in their lexicon, but that the features (importantly, the semantic features) on the experimental items (which are presented as flat pictures in a book) are unreliably perceived and encoded. We call this the Misperception 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 did not reliably encode features on experimental items. To do this, we use a mixture model

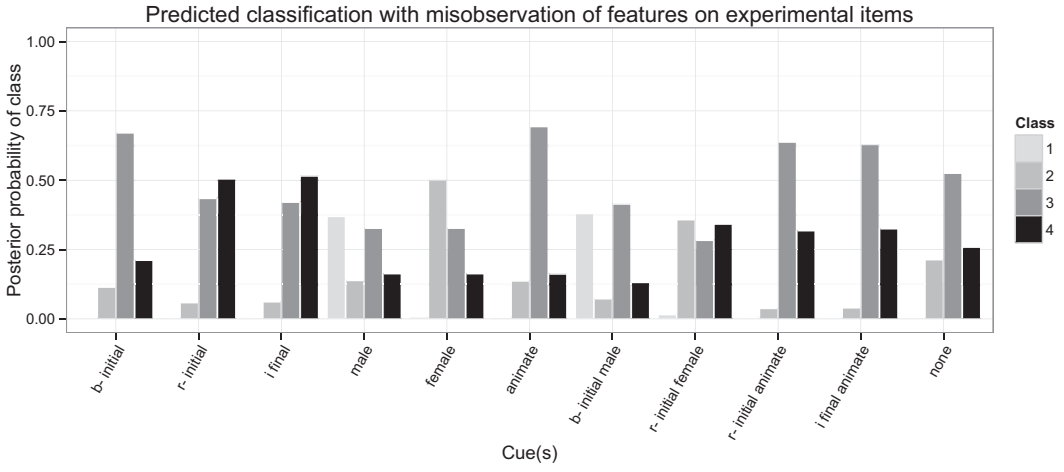

Fig. 5. Classification of novel nouns as predicted by a model that misobserves semantic features on experimental items 61% of the time, and correctly observes phonological ones.

with two parameters β_s and β_p, which specify the proportion of semantic and phonological features, respectively, that are misperceived. The quantities $(1 - \beta_s)$ and $(1 - \beta_p)$ specify the proportion of classification decisions made with semantic and phonological features having their correct values, consisting of either specified values ([male], [female], [animate], [r initial], [b initial], [i final]), or the [other] value. This correct value is denoted as [correct] in the equation below. Conversely, β_s and β_p specify the proportion of decisions made with the semantic or phonological feature correctly or incorrectly holding the [other] value. This yields the following mixture model:

$$\begin{aligned}\mathbb{P}(c_i|f_1,f_2,f_3) = &(1-\beta_s)(1-\beta_p)\mathbb{P}(c_i|f_1=[correct],f_2=[correct],f_3=[correct])\\ &+(\beta_s)(1-\beta_p)\mathbb{P}(c_i|f_1=[other],f_2=[correct],f_3=[correct])\\ &+(1-\beta_s)(\beta_p)\mathbb{P}(c_i|f_1=[correct],f_2=[other],f_3=[other])\\ &+(\beta_s)(\beta_p)\mathbb{P}(c_i|f_1=[other],f_2=[other],f_3=[other])\end{aligned} \tag{3}$$

As with the Misrepresentation model, we found the best-fitting value for each of the two β parameters and evaluated the model by comparing it to children's behavior. This model again produced a close fit for all feature values (Fig. 5). The model showed a consistent degree of misperception across all semantic features and feature combinations. The best fitting level value of β_s (semantic parameter) was 0.61, meaning that children would be misperceiving semantic features on 61% of the experimental items. The best fitting value of β_p (phonological parameter) was 0. The log likelihood of the model given the experimental data was −507. A generalized likelihood ratio test indicates that the Misperception model also significantly outperforms the optimal naive Bayesian Classifier ($\chi^2(2) = 478$, $p < .0001$).

As is reflected in the log likelihood favoring the Misperception model, overall, in some conditions, we can see how it appears to fit the children's data slightly better than the Misrepresentation model. This is most apparent in the predicted classification for words that are both [male] and [b- initial]. The Misperception model accurately captures the strong shift toward using the phonological cue for Class 3 when it conflicted with the semantic cue for Class 1. Like the Misrepresentation model, it does not fully capture the extent to which children shifted to Class 4 when phonological cues for Class 4 ([r- initial], [-i final]) conflicted with semantic cue for Class 3 ([animate]), nor the preference for classifying nouns with all values [other] as Class 4.

Again, we can consider the reasonableness of the best fitting parameters. Anecdotally, it seems as though misencoding 61% of experimental items is quite high, given participants' reactions to the items (children made comments about the novel animals, for example, and found it funny when it was suggested that these, or the novel humans, might be edible). Of course, further experimentation could show how reliably children can infer animacy of such items. Additionally, children appear to do reasonably well classifying items based on semantic features when no phonological feature conflicts with this prediction. This model predicts that children's performance with these items would be somewhat worse than it actually is. For example, on experimental items that were male or female humans, with no phonological cues, the model predicts a higher proportion of Class 3 responses than actually occur in the children's data. A final piece of evidence that points toward a deeper factor than misperception of semantic features during the experiment is children's performance with classifying real words. Gagliardi and Lidz (2014) present data where children make more errors when classifying real words with conflicting cues than those without. That is, a word like "recenoj" (ant), is r- initial but animate, and is a Class 3 word. Children mix up classification of words like this just the way they do with experimental items, putting a higher proportion than expected in Class 4, consistent with the phonological feature, but inconsistent with animacy and the actual class of the word. Thus, it looks as though children's preferential use of phonological information extends more deeply than an encoding problem during the experiment.

4.3. Hypothesis 3: Featural preference

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, either because phonological information has been reliably available for a longer period of time or because children have a bias that privileges information when tracking morphophonological dependencies.

Our third model, embodying the Featural Preference hypothesis, therefore looked at what would happen if we had a learner that was biased to only use one type of feature in classifying 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

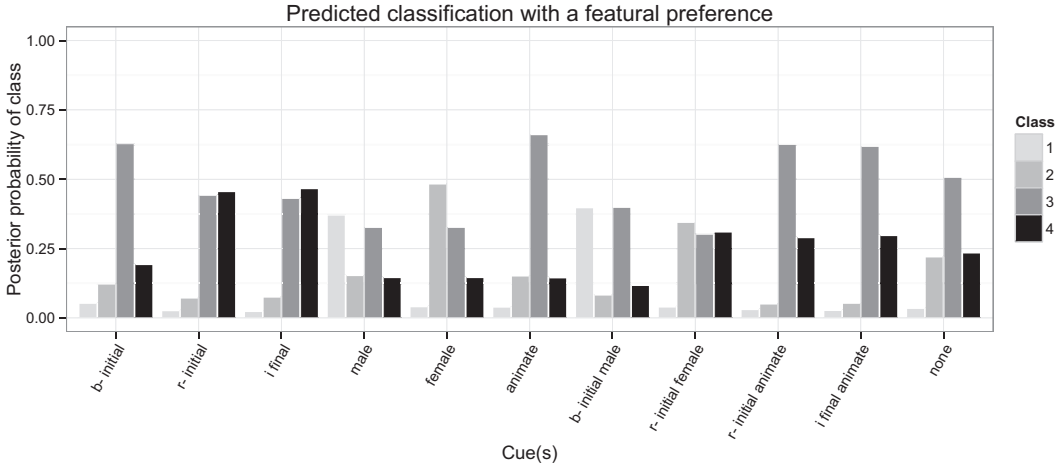

Fig. 6. Classification as predicted by a model biased not to use semantic information 64% of the time and not to use phonological information 7% of the time.

both semantic and phonological features, one that only used phonological features, one that only used semantic features, and one that used neither kind of feature. The crucial difference between this model and the Misperception model is that in the Misperception model, all features are always used, but are encoded as the wrong value some proportion of the time, whereas in the Featural Preference model, some features do not factor into the calculation at all some proportion of the time (again, where β_s is the semantic parameter and β_p is the phonological one). The model can be seen in Eq. 4.

$$\begin{aligned}\mathbb{P}(c_i|f_1,f_2,f_3) = (1-\beta_s)(1-\beta_p)\mathbb{P}(c_i|f_1,f_2,f_3) + (\beta_s)(1-\beta_p)\mathbb{P}(c_i|f_2,f_3) \\ + (1-\beta_s)(\beta_p)\mathbb{P}(c_i|f_1) + (\beta_s)(\beta_p)\mathbb{P}(c_i)\end{aligned} \tag{4}$$

Again, we evaluated the model against the children's classification data and found a closer fit than the ideal learner (Fig. 6). The best fitting value of β_s was 0.64, meaning that children would be choosing not to use semantic features on 64% of classification decisions. The best fitting value of β_p was 0.07 meaning phonological information was not used on 7% of classification decisions. The log likelihood of the model given the data was -512. A generalized likelihood ratio test showed that this model also significantly outperformed the optimal naive Bayesian classifier ($\chi^2(2) = 468$, $p < .0001$). Although the best fitting value of the phonological parameter was non-zero, adding the phonological parameter resulted in only 0.733 increase in the log likelihood compared to a model with only the semantic parameter. This difference in log likelihood did not reach significance ($\chi^2(1) = 1.47$, $p = .23$). This indicates that the model of the children's data that achieves the best trade-off between high likelihood and few free parameters would be one that discounts only semantic information.

Like the Misperception model, this model captures children's behavior quite well for most noun types, but underpredicts the extent to which children prefer Class 4 when a phonological cue predicts it and contrasts with the semantic cue for Class 3, or when all features are [other]. This continuously prevalent pattern will be discussed below. While this model did not use phonological information as faithfully as it is present in the input, it still followed the general pattern that we saw in the other models, where the match to children's data was driven primarily by less use of semantic information in the model. As with the Misrepresentation and Misperception models, we can consider the "reasonableness" of the best-fitting parameter estimates (not using semantic information in 64% of classification decisions and not using phonological information on 7%). Whether or not these parameters are reasonable greatly depends on what underlies this preference. The preference could stem from the fact that phonological information has been more reliably available throughout development, and children only slowly move away from this source of information to incorporate semantic information. Future modeling work looking at how children move away from a phonologically based system once semantic information is reliably available could shed light on how quickly we might expect this to happen. Alternatively, children could prefer to use (morpho)phonological information when determining the distribution of a (morpho)phonological dependency due to an expectation that only information in one domain will matter within that domain. Given that agreement appears as morphophonological information on the verb, this could bias children to attend to phonological cues for noun classification. Again, depending on the process by which the learner revises this hypothesis and begins to track and rely on information from another domain (semantics), we might be able to see whether ignoring semantic information 64% of the time seems reasonable or not.

4.4. Comparing the models

It is not immediately obvious how to best evaluate the alternative models with respect to one another. Model comparison methods such as AIC, BIC, and so on trade off the log likelihood of the models against the number of free parameters. Because the models proposed above all have the same number of free parameters, this amounts to a comparison of their log likelihoods. These log likelihoods were −538 for the Misrepresentation model, −507 for the Misperception model, and −512 for the Featural preference model. While a comparison of the log likelihoods of the models favors Model 2, the Misperception model, it's not clear that this metric alone is enough to say that this is the best model of children's performance. It is important to consider what assumptions go into these models. For example, each model yielded a different set of best-fitting parameters, corresponding to different degrees 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 is not immediately clear how to measure reasonableness, or how to compare it across models. Despite these limitations, some discussion of the merits of each model, in addition to the "reasonableness" of the best-fitting parameters and the premises of the models in general, can be enlightening.

The best-fitting parameters for Model 1, the Misrepresentation model, suggest that 95% of semantic features are misrepresented. As discussed above, for children 4 years old and older, this number seems quite high. Moreover, as mentioned above, with the overall best-fitting semantic parameter, Model 1 fails to capture children's shift away from classifying males as Class 1 when they are b-initial. To best capture children's performance in this condition, the model needs 99% misrepresentation of the semantic feature male. What seems like an extremely high level of misrepresentation necessary to counteract such a strongly predictive feature makes the model even less plausible.

Model 2, the Misperception model, predicts that children have difficulty encoding semantic features of experimental items 61% of the time, with the range of parameters that give high likelihood being fairly tightly clustered around that best-fitting value (Fig. 5). Again, we have some doubts with respect to the reasonableness of such a percentage, given that children show similar patterns when classifying real words, can use both semantic and phonological features in isolation and are generally good at perceiving features like animacy. One possibility that might merit further consideration is that children are sensitive to the degree of animacy of different items (e.g., something fluffy with eyes might be "more animate" than something like a bug), and that systematically manipulating this degree of animacy during the experiment could shed light on whether all versions of animacy are discounted by the learners, or if just the less canonically animate items are affected. Unfortunately, there is not enough variability in the experimental items used by Gagliardi and Lidz to investigate this hypothesis. Even with this being a possible contributor to children's behavior, it looks as though children's preferential use of phonological information extends more deeply than an encoding problem during the experiment.

Model 3, the Featural Preference model, is perhaps the most difficult to evaluate with respect to the reasonableness of the best-fitting parameters. Overall, it provides what looks like the most reasonable fit to the children's data, though like all models it underpredicts the degree to which children will assign nouns to Class 4 when no cues (and conflicting phonological cues) are present. Evaluating the reasonableness of the 64% semantic parameter is difficult, however, without a better understanding of where this preference could come from. Moreover, we need to consider the fact that the best-fitting parameters of this model did not always use all available phonological information. This may be more similar to children's behavior (if they indeed do not use phonological information 100% of the time), but a more targeted inquiry into their behavior would be necessary to confirm this. Below, we will further explore the Featural Preference Model and discuss how we might model its source.

Finally, it is possible that a combination of all three of these processes (and perhaps more that we have not considered here) is influencing children's classification decisions. This is what we explore in the next section, by building a model that combines the possibilities of misrepresentation and featural preference (Models 1 and 3). By combining these two models, we can square off a degradation of featural representation in the lexicon with a preferences for one type of information over another in classification.[8]

4.5. Uncertainty in more than one place

The last possibility we will consider in this paper is that there is some combination of the above processes that leads to children's classification behavior. To explore this, we combined Models 1 and 3, looking at whether some level of misrepresentation of features combined with a preference to use some information over other could predict children's behavior. As the best fitting values for all parameters consistently favored discounting semantics, rather than phonology, across all models, we omit the phonological discounting parameter here and only look at values of parameters for misrepresenting and not using semantic information in classification,

$$\mathbb{P}(c_i|f_1,f_2,f_3) = (1 - \beta_{s_3})\mathbb{P}(c_i|f_1,f_2,f_3) + (\beta_{s_3})\mathbb{P}(c_i|f_2,f_3) \quad (5)$$

where $\mathbb{P}(f_1 = k|c)$ is based on the counts of words with feature f_1 reduced at a rate of β_{s_1}

As in all other cases, we evaluated the best fit parameters for the model using a built-in optimization procedure in MATLAB (fminsearch). The process showed the best-fitting parameters to be equivalent to Model 3. The best-fitting parameter for misrepresentation of features in the lexicon was zero, indicating that semantic information was fully represented. The best-fitting parameter for omission of the semantic feature during classification was 0.64, indicating that semantic information was not used on 64% of classification decisions (this is equivalent to the reduction in the use of semantic information found in Model 3). In case this was a case of finding a local maximum, a grid search was also employed to look for the best-fitting parameter, testing all values between 0 and 1, by increments of 0.01 for both parameters. This procedure confirmed that the best-fitting values of these parameters was indeed as reported. Fig. 7 shows how the likelihood of each model changed throughout the parameter space, searched incrementally in the same way as Model 4.

4.6. More exhaustive models

It is of course possible to imagine more exhaustive models. For example, one might differentiate the probability of misperceiving or misrepresenting different semantic cues (perhaps animacy is easier to perceive that natural gender, for example). When exploring such models, we found that the high number of free parameters (e.g., 12 for the semantic feature alone in Model 1) needed to build such a model was large compared to the number of datapoints in the children's data. Moreover, finding the optimal solution to such a model may not be important in answering the question we asked at the outset—we are mainly interested in distinguishing between Models 1, 2, and 3, as these illustrate three different ways in which learners' computations may be constrained, and it is not clear how examining tiny variants of each type of constraint will help us answer that question. While it is of course possible that a model with more parameters would be able to better

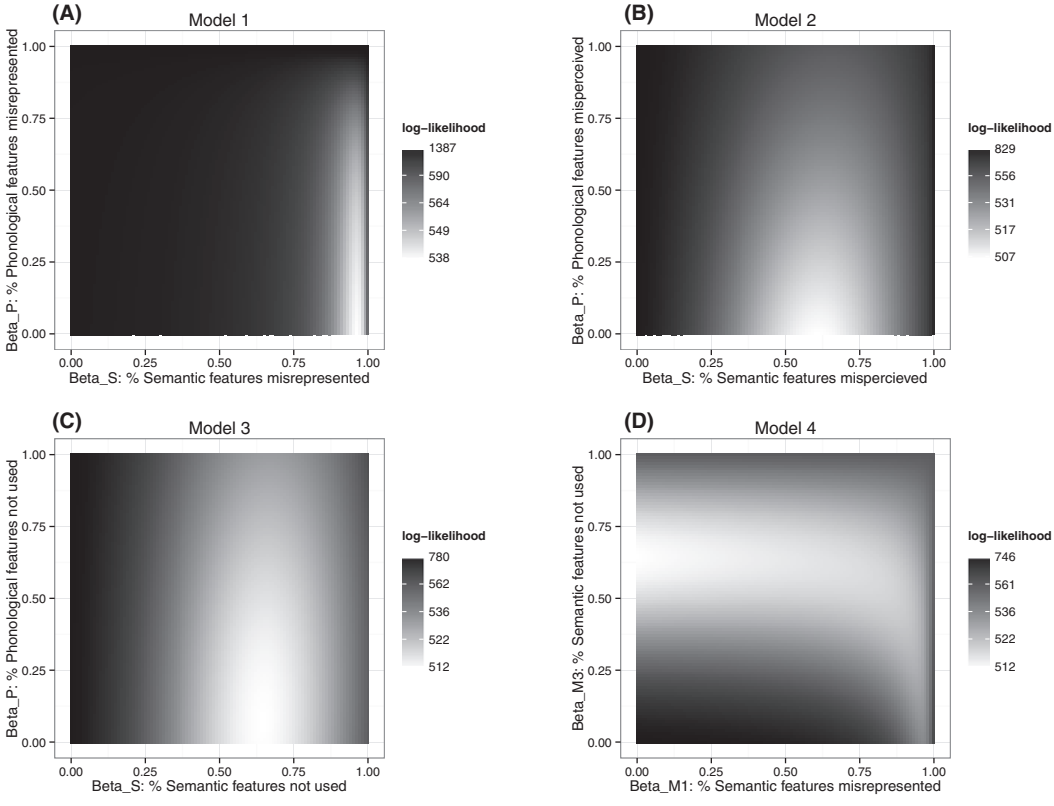

Fig. 7. Log likelihood of each model across the entire parameter space.

fit the data, this is problematic for two reasons. First, we run the risk of overfitting. Second, focusing on minutia can lead to a loss of the bigger picture questions which we set out to investigate.

We did conduct one additional simulation to rule out an alternative account of the data that appeals to children's prior distribution over classes. All the models considered above underpredict the extent to which children assign nouns to Class 4. The fact that we see this shift only when we have nouns with conflicting cues (i.e., not when we have nouns with semantic cues to other classes) suggests that the bias for Class 4 is not due to simple shifts in the prior distribution over classes. Nevertheless, it is important to ensure that the overall pattern of phonological preference seen in the children's data cannot be attributed to a simple preference for particular classes. To examine this quantitatively, we implemented a version of our naive Bayes model that treats the class priors as free parameters (Fig. 8). The priors that best fit the children's data are $p(c_1) = 0.0027$, $p(c_2) = 0.0296$, $p(c_3) = 0.2686$, and $p(c_4) = 0.6991$ (as compared with those based on the corpus estimation: $p(c_1) = 0.053$, $p(c_2) = 0.184$, $p(c_3) = 0.491$, and $p(c_4) = 0.272$). These parameters do indeed favor Class 4. However, despite having three free parameters, this model gives a log likelihood of -572, substantially lower than any of the log likelihoods

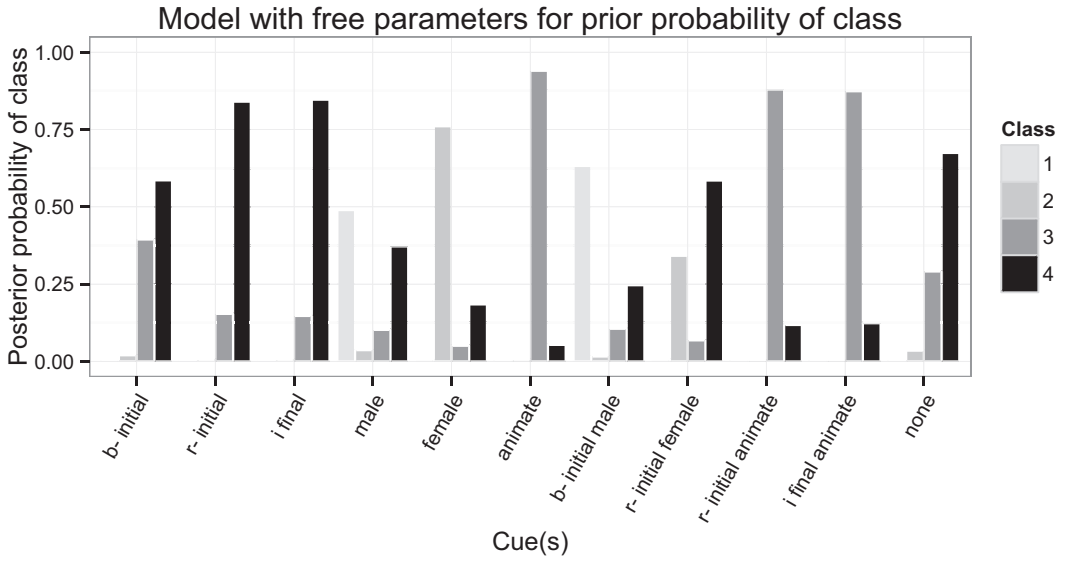

Fig. 8. Classification as predicted by a model with three free parameters, finding the best fit for the prior probability of each class given the children's data.

of the models with two free parameters explored above, and would be disfavored by model evaluation metrics such as the BIC that trade off model likelihood against the number of free parameters. Thus, a simple preference for particular classes does not adequately account for the phonological preference that is the focus of this paper.

5. 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 then presented models exploring three 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 features in the lexicon, (b) misperceived 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. In each of these cases, the children's data were best fit by a model that used degraded semantic information, together with intact phonological information. We also explored what would happen if semantic information could be degraded in both the lexicon and during classification (Model 4), and saw that it is most likely a bias during classification, rather than a severe misrepresentation of features that leads to the behavior we

have seen in children. All four models fit children's data significantly better than the optimal naive Bayesian classifier did, though no model fit the children's data perfectly. The improvement in fit found with some version of filtered input 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, and their beliefs about which features of their input are relevant.

5.1. Behind the featural preference

As noted above, before we can determine whether the best fitting parameter values for the featural preference model are reasonable ones, we need to know more about where this preference comes from. It could originate from several sources: the asymmetry between phonological and semantic cues across development, the asymmetry between these cues in novel word learning, and the fact that phonological, but not semantic, cues appear to occupy the same linguistic domain as the agreement morphology they (or the noun classes they predict) trigger.

First, the availability of phonological and semantic cues is unbalanced throughout development. Phonological segments are differentiated and encoded earlier than word meanings (Gervain & Mehler, 2010; Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Werker & Tees, 1984) and thus dependencies among phonological segments should be available earlier than those between semantic features or between feature types. In fact, infants have been shown to be able to track (morpho)phonological dependencies at as early as 7 months (Gonzalez & Nazzi, 2012), earlier than they have been shown to have attached detailed meanings to phonological strings.[9] This could mean that infants build up some system of noun classification that relies fully on morphophonological information and that it takes some time, once the semantic features are reliably available, to revise the reliance on only morphophonological cues.

Second (and perhaps relatedly), it could be the case that once infants are learning some nouns and begin dividing them into classes, the phonology of the nouns is going to be more reliably available to the learner for a given token of a word than the semantics. That is, when learning a new word, the learner might be relatively certain early on of what the phonology of this word is, but may take several instances to become certain of the precise meaning and what semantic feature values would go along with this. Again, the learner might then develop a system that relies more heavily on phonology as s/he is still uncertain about word meanings, and consequently less certain about their correlation with noun classes.

Finally, agreement results in overt morphophonological information being expressed on the verb. It could be that the learner has some expectation that information in one domain (say morphophonology) is going to matter for dependencies within that domain (i.e., agreement), and it might take more time to realize that information from another domain (semantics) is also relevant. This type of expectation could be specifically related to linguistic knowledge (e.g., it could be an expectation that certain domains of the grammar

will be more likely to relate to one another than others), or it could be a general cognitive principle (e.g., that for any given phenomenon, features or information from within the domain of that phenomenon are more likely to be relevant to learning that features or information from another domain) (cf. Moreton, 2008; Warker & Dell, 2006).

Additionally, it could be that phonology is a useful cue for making classification distinctions elsewhere in the language and that children have formed an overhypothesis (Goodman, 1955; Kemp, Perfors, & Tenenbaum, 2007) about phonological features being useful. If phonology were supported as a feature that often factored into computations in other domains, this overhypothesis could give children a bias to use phonology in noun classification as well. To determine whether this would be the case, we would have to look at other domains that children would control better or earlier than noun classification and see if phonology can predict classification there (for example, in declension class or lexical category), and if children seem aware of the phonology class correspondence. Unfortunately, at this time, we do not have the relevant data to say whether or not this is true.

Whatever the cause, the phonological preference appears to be a developmental step, and not the characterizing feature of noun classes. This is apparent in the fact that Gagliardi and Lidz show both older children and adults overcoming this preference and relying more on semantic cues at older ages. Additionally, it is apparent in the fact that semantic features such as natural gender and animacy are present in virtually every noun class system found cross linguistically (Corbett, 1991), and remain robust parts of these systems even when phonological regularities are found as well. The fact that these features are so heavily relied on in noun classification systems suggests that there may be a prior bias linking these types of features and classification, but that this bias can only begin to show its effects once the learner begins to reliably track semantic information and overcome the learned preference for phonological information.

5.2. An unaccounted preference for Class 4

While all of our models of uncertainty captured some shift by children to classify novels nouns as Class 4 when a phonological cue for Class 4 ([r- initial], [-i final]) conflicted with a semantic cue for Class 3 ([animate]), none of the models were able to fully capture children's preferences. Additionally, no model accurately predicted children's behavior when classifying nouns with all feature values set to [other]. Children preferred to put these in Class 4 (with some in Classes 2 and 3), while the model predicted most would be in Class 3 (with some in Classes 2 and 4). What both of these observations share is that children appear to have some preference for classifying nouns in Class 4 that is not captured by our models. It appears that some other factor must be influencing children's classification.

This behavior resembles what has traditionally been called a "default" class: a class that is chosen for a word when all else fails. However, this notion of a default is somewhat unsatisfying from a classification standpoint. First, while the majority of nouns with these [other] values are put in Class 4, not all of them are (and not all speakers are putting the same ones there). Second, this sort of "default" label fails to characterize the

process by which speakers are actually classifying nouns, doing little more than providing a concise label for their behavior. In this section, we explore several possibilities as to where this behavior might stem from.

One possibility is that the priors we used for the classes do not accurately reflect class size (these were taken from the sample of child-directed speech from Gagliardi and Lidz (2014)). If Class 4 is significantly larger in the child's lexicon than we have represented it, and because this class consists solely of nouns with the semantic feature [other], some of this effect could be due to a misrepresentation of class size in the model. As mentioned above, we built a model that included free parameters for class size, in effect testing this hypothesis. This model did not do a better job of capturing children's data overall. However, it did appear to predict a shift to Class 4 in items where all feature values are set to [other], and it is possible that this model could capture the shift to Class 4 in relevant conflicting cue conditions if combined with a model that discounts semantic information. However, this still leaves open the question of if and why children assign a higher prior probability to Class 4. As Gagliardi and Lidz point out, while other measures of class size (corpus tokens and dictionary types) show a higher proportion of nouns in Class 4 than we have represented, they do not appear to show enough nouns in Class 4 to predict this effect.

A second possibility is that Class 4 is seen by the learner as having a more diverse makeup, making it more likely that any new noun would be in this class than in any of the other, possibly more well defined, classes (cf. Hare, Elman, & Daugherty, 1995; O'Donnell, 2011). Evaluating this possibility requires specifying on what level we might want look at features and decide if their distribution in a class is heterogeneous or not. In particular, we are required to look into the [other] value of the semantic feature. If it is just treated as [inanimate], then Class 4 looks very homogenous (whereas Classes 2 and 3 have heterogeneity on this level with a mix of [female]-[inanimate] or [animate]-[inanimate] feature values). However, if it is the case that learners are looking for structure beyond this rather coarse level of feature coding (taking into account more semantic features such as shape, function, material, etc), then they may not find it in Class 4. This would mean that Class 4 could look highly heterogeneous, and thus a speaker with a bias to maintain the relative makeup of the classes would tend to assign words denoting novel objects to this class. At this point, it is also useful to remember one aspect of adult classification behavior that our model did not probe. Recall that adult speakers were somewhat unwilling to put any novel noun into Class 1, whether it had conflicting phonological information or not. We suspect that this may also be related to speakers' expectations about a classes makeup and productivity. As Class 1 is unique in that all nouns in the class are males, it may be seen as the least heterogeneous, and end up being the class that is most closed to new assignment. Further experimentation and modeling would be necessary to determine if this is indeed what lies behind speakers' preferences for Class 4, and their relative reluctance to put any novel nouns into Class 1.

A third possibility, following Baayen (1992), is that the frequency distribution of words differs across classes. That is, while Class 3 is larger (both in terms of tokens and types), perhaps if Class 4 had a higher proportion of single token types (hapax legomena),

then a novel word would be more likely to fall in Class 4 (as, by this metric, it appears to be made up of mostly novel words). However, an analysis of the distribution of word frequency across classes shows that Class 3 in fact has the highest proportion of hapax legomena, meaning that if this were the sort of information children were using, then they should be more likely to classify "no cue" words as being in Class 3.

A final possibility is that the use of Class 4 agreement morphology is indeed a "default," but of a particular kind. In many languages with classifier systems (e.g., Mandarin), different classifiers are used with different semantic- or shape-based classes of nouns, but a default classifier exists as well. That is, there is a classifier that can be used with any noun, regardless of its semantic or shape properties. It is possible that Tsez speakers have some view of Class 4 as this sort of default, thinking that many nouns, while they might have a best class based on their featural content, can also be used as if they were in Class 4. When we consider the fact that plural morphology for all nouns (with the exception of plurals of nouns from Class 1, or groups containing nouns from Class 1) is the same as that for Class 4 (the prefix [r]), it looks like this kind of default-like system could be learnable from a mistaken encoding of the input. That is, if a learner heard a noun in any given class, and also heard it in the plural (but did not realize it was plural), he/she might infer that while the noun could be used with agreement from Class 2 (or 3), it could also be used with morphology from Class 4. Seeing enough nouns in the plural and also in Class 2 or 3 (and consistently missing the singular/plural distinction) might lead children to believe that Class 4 was a catch all, even when nouns could also be assigned to other classes. This means that if a novel word has no cue, the child/speaker might not have a good guess about what class it belongs to, but would think that putting it in Class 4 was fine no matter what. To test this, we looked at the distribution across classes of nouns that were found in the plural. If we were to find that many nouns from Classes 2 and 3 were seen in the plural (where the agreement is ambiguous between Class 4 and plural), this could be a source of confusion about the classes of these nouns. However, we found very few instances of plural nouns altogether (only 15 noun types out of 114 total appear in the plural). Out of these 15 types, one was from Class 1, two were from Class 2, four were from Class 3, and eight were from Class 4 (when looking at noun tokens, the numbers vary but the general pattern remains the same). While this is just a small amount of data, it does not appear to support this hypothesis, as the majority of nouns seen in the plural were in Class 4, meaning it is unlikely that what is driving the preferences for putting nouns into Class 4 stems from misanalysis of plural agreement.

Overall, it remains somewhat unclear where the preference to classify "no cue" words as Class 4 comes from. Here, we have outlined several possible geneses for this preference, and leave it to further research to determine which of these best fits both the data available to the learner and the system that appears to be learned.

6. Conclusions

This work has several important implications for research on statistical learning and language acquisition. First, we identified an area where children's behavior does not

appear to reflect the ideal inferences licensed by the statistical patterns in the input. This sort of pattern is important in the study of language acquisition, as it bears on the question of what constrains children's generalization of the linguistic input. By building a model that insufficiently captured the empirical data and iterating through possible ways to constrain this model, we were able to investigate the source of the asymmetry in children's behavior. While each model differed in where the asymmetry came from, all appeared to employ 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), but these patterns could be related. That is, if children find that phonological information is somewhat useful in the input (perhaps before they know enough word meanings to determine that semantic information is also useful), they may make some generalization that amplifies the import of phonological behavior. This finding demonstrates that we need to be careful when determining whether a statistical learner could or could not draw some kind of inference. It shows that in investigating language acquisition, we need to consider more than just the kind of computation being performed, as what data the computation is performed over is also critical.

Next, our models showed that it is plausible that these children are indeed behaving optimally with respect to some statistical distribution, just not one that includes all of the information available in 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 is not whether or not children are using statistics to acquire language, but what statistics they are using. Even though children were not behaving optimally with respect to the data in principle available, our models allowed us to see that they may be behaving optimally with respect to either a filtered version of the input (Pearl & Lidz, 2009), or the input combined with their own biases (Lidz et al., 2003; Viau & Lidz, 2011). This observation allows us to make hypotheses about what kinds of biases children might entertain, and these hypotheses can be tested in future work.

Finally, 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 different types of features 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 has been found consistently across languages.

Acknowledgments

This research was supported by NSF IGERT DGE-0801465 and an NSF GRF to Gagliardi. We thank Masha Polinsky, the UMD Cognitive Neuroscience of Language Lab, the UMD Project on Childrens Language Learning, and the UMD Computational Psycholinguistics group for helpful discussion and assistance.

Notes

1. Moreton (2008) proposed similar factors to account for phonologization, with his *channel bias* mapping most closely onto (1) and (2) above and his *analytic bias* being most closely related to (3).
2. Here, we talk about "noun classes" to refer to 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 normally in each class that do not have this (or potentially any) cue predicting their class.
3. In our subsequent models, we use this set of features as well, as we can only compare model predictions with behavior on the features used in the behavioral experiment, and that those were chosen selectively, as severe limitations on subject availability meant that only a small number of features could be tested.
4. Smoothed feature counts were computed by adding one to each raw feature count.
5. The joint entropy and sum of marginal entropies for each class were as follows: Class 1 Joint = 0.87, Sum = 0.90; Class 2 Joint = 1.04, Sum = 1.21; Class 3 Joint = 1.25, Sum = 1.25; Class 4 Joint = 0.98, Sum = 0.99.
6. Another way to introduce misrepresentation would be to simply omit the features on some proportion of the nouns that should have them represented. We tried this as well, and we do not include the results here as they are effectively identical to those generated by the method we outline above.
7. The predictions of this model (as well as the next one) could change if children had some way of estimating their own rates of misrepresentation and misperception. However, it is unclear that children would have a way to estimate this information.
8. We selected Model 3 over Model 2 because we believe the considerations noted above provide evidence against the types of misperceptions assumed by Model 2. However, combining Model 1 and Model 2 yields a comparable result
9. Bergelson and Swingley (2012) showed that children as young as 6 months associate some meaning with phonological strings. However, it is not clear how detailed these representations are, or what kinds of semantic features they might rely on, thus we believe that it is still possible that children have some knowledge of phonological strings and dependencies prior to developing detailed semantic representations.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Baayen, H. (1992). Quantitative aspects of morphological productivity. In G. E. Booij & J. van Marie (Eds.), *Yearbook of morphology 1991* (pp. 109–149). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Becker, M. (2007). Animacy, expletives, and the learning of the raising-control distinction. In A. Belikova, L. Meroni, & M. Umeda (Eds.), *Proceedings of the 2nd conference on generative approaches to language acquisition North America (GALANA)* (pp. 12–20). Somerville, MA: Cascadilla.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*, 3253–3258.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Bokarev, E. A. (1959). *Cezskie (didojskie) jazyki dagestana*. Moscow-Leningrad: Nauka.

Brown, R. (1973). *A first language: The early stages*. London: George Allen & Unwin.

Bunger, A., & Lidz, J. (2006). Constrained flexibility in the acquisition of causative verbs. In D. Bamman, T. Magnitskaia & C. Zaller (Eds.), *Proceedings of the 30th annual Boston University conference for language development* (pp. 413–424). Somerville, MA: Cascadilla Press.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, *15*, 17—29.

Comrie, B., & Polinsky, M. (1998). The great Daghestanian case hoax. In A. Siewierska & J. J. Song (Eds.), *Case, typology and grammar: In honor of Barry J. Blake* (pp. 95–114). Amsterdam: John Benjamins.

Comrie, B., Polinsky, M., & Rajabov, R. (1998). Tsezian languages. Unpublished manuscript, Max Planck Institute for Evolutionary Anthropology.

Corbett, G. (1991). *Gender*. Cambridge, UK: Cambridge University Press.

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.

Fodor, J. (1998). Parsing to learn. *Journal of Psycholinguistic Research*, *27*(3), 339–374.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585.

Gagliardi, A., & Lidz, J. (2014). Statistical insensitivity in the acquisition of Tsez noun classes. *Language*, *90*(1), 58–89.

Gervain, J., & Mehler, J. (2010). Speech perception and language acquisition in the rst year of life. *Annual Review of Psychology*, *61*, 191–218.

Gillette, J., Gleitman, H., & Gleitman, L. (1999). Human simulations of vocabulary learning. *Cognition*, *73*, 135–176.

Gleitman, L. (1990). The structural sources of word meaning. *Language Acquisition*, *1*, 3–55.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54.

Gomez, R. L., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy*, *7*, 183–206.

Gonzalez, N., & Nazzi, T. (2012). Acquisition of nonadjacent phonological dependencies in the native language during the first year of life. *Infancy*, *17*, 498–524.

Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Hare, M., Elman, J. L., & Daugherty, K. G. (1995). Default generalization in connectionist networks. *Language and Cognitive Processes*, *10*, 601–630.

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

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*(5044), 606–608.

Lidz, J., Gleitman, H., & Gleitman, L. (2003). Understanding how input matters: The footprint of universal grammar on verb learning. *Cognition*, *87*, 151–178.

Martin, C. L., Ruble, D. N., & Szkrybalo, J. (2002). Cognitive theories of early gender development. *Psychological Bulletin*, *128*, 903–910.

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

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

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology*, *25*, 83–127.

O'Donnell, T. (2011). Productivity and reuse in language. Unpublished doctoral dissertation, Harvard University.

Pearl, L., & Lidz, J. (2009). When domain general learning fails and when it succeeds: Identifying the contribution of domain specificity. *Language Learning and Development*, *5*(4), 235–265.

Pearl, L., Goldwater, S., & Steyvers, M. (2011). Online learning mechanisms for Bayesian models of word segmentation. *Research on Language and Computation*, *8*(2), 107–132.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338.

Plaster, K., Polinsky, M., & Harizanov, B. (2009). Noun classes grow on trees: Noun classification in the North-East Caucasus. *Language and Representations* (*Tentative*). John Benjamins, In Press.

Polinsky, M. (2000). Tsez beginnings. In J. Good & A. C. L. Yu (Ed.), *Papers from the 25th Annual Meeting of the Berkeley Linguistics Society* (pp. 14–29). Berkeley, CA: Berkeley Linguistics Society.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2009). The role of distributional information in linguistic category formation. In N. A. Taatgen & H. v. Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2564–2569). 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* (pp. 2063–2068). Austin, TX: Cognitive Science Society.

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

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: The acquisition of american sign language from inconsistent input. *Cognitive Psychology*, *49*, 370–407.

Sternberg, D. A., & McClelland, J. L. (2011). Two mechanisms of human contingency learning. *Psychological Science*, *23*, 59–68. doi: 10.1177/0956797611429577

Takahashi, E. (2009). Beyond statistical learning in the acquisition of phrase structure. Unpublished doctoral dissertation, University of Maryland.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Viau, J., & Lidz, J. (2011). Selective learning in the acquisition of kannada ditransitives. *Language*, *87*, 679–714.

Warker, J. A., & Dell, G. S. (2006). Speech errors reflect newly learned phonotactic constraints. *Experimental Psychology: Learning, Memory, and Cognition*, *32*, 387–398.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. doi: 10.1080/14640746808400161

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49–63.