Modeling Uncertainty in Novel Noun Classification in Tsez

Annie Gagliardi, Naomi Feldman and Jeff Lidz (University of Maryland)

A growing literature shows that children are highly sensitive to statistical features of their linguistic input (Saffran et al 1996; Gomez & Gerken 2000). This literature assumes that children can reliably encode all of the information available in this input, contrasting with the observation that children are sometimes sensitive to features of their language out of proportion with their statistical reliability. This contrast highlights the difference between the input, the linguistic information available in the environment and the intake, the proportion of the input the child actually uses. The current paper explains this difference as it is apparent in noun classification by introducing uncertainty in the detection of certain features.

Many languages classify nouns according to grammatical gender. Cross-linguistically, these noun classes correlate with both semantic and phonological features of nouns. For example, in Tsez, a Nakh-Dagestanian language with 4 noun classes spoken in the Northeast Caucasus, semantic features such as natural gender (of humans) and animacy (of non humans) are very reliable predictors of noun class. Phonological features, such as the first segment of the noun, can also predict noun class but does so less reliably. Behavioral experiments have shown that adults and children are sensitive to these semantic and phonological regularities and can use them classify novel nouns. Surprisingly, while adults behave in line with the statistical reliability of the cues in question, 4-7 year olds prefer to use phonological features, rather than the more predictive semantic features, when the two types make conflicting predictions (Gagliardi & Lidz, under review). Here we present a Bayesian model of noun classification to show how this behavior might arise from simple misperception of semantic features.

We propose that children's classification patterns for novel nouns are not random, but instead reflect children's beliefs about the features on the nouns in their lexicon. One possibility is that if semantic features are more difficult to perceive than phonological features, children may be misperceiving semantic cues. Thus when they try to estimate the predictiveness of a semantic cue, they have sparse or distorted data to from which to make this estimation. We test this hypothesis by making a formal link between the feature counts in a child's lexicon and classification behavior through a Bayesian model.

Our model assumes that children use a naïve Bayesian classifier to compute the posterior probability of noun class membership:

$$p\left(c \mid f_1, f_2 \ldots f_n\right) = \frac{p\left(f_1 \mid c\right)p\left(f_2 \mid c\right)\ldots p\left(f_n \mid c\right)p\left(c\right)}{\sum_c p\left(f_1 \mid c_i\right)p\left(f_2 \mid c_i\right)\ldots p\left(f_n \mid c_i\right)p\left(c_i\right)}$$

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. Using a multinomial model with a uniform Dirichlet prior distribution to estimate the proportion of items in class c that contain a particular value for feature f, each likelihood $p(f|c)$ is equal to:

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

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 misperception of semantic features into this model by manipulating the number of observations of a noun with a certain feature in each class. Since our hypothesis is that children misperceive semantic features some proportion of the time, we reduce the count of nouns in each class that contain the relevant semantic features. 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.

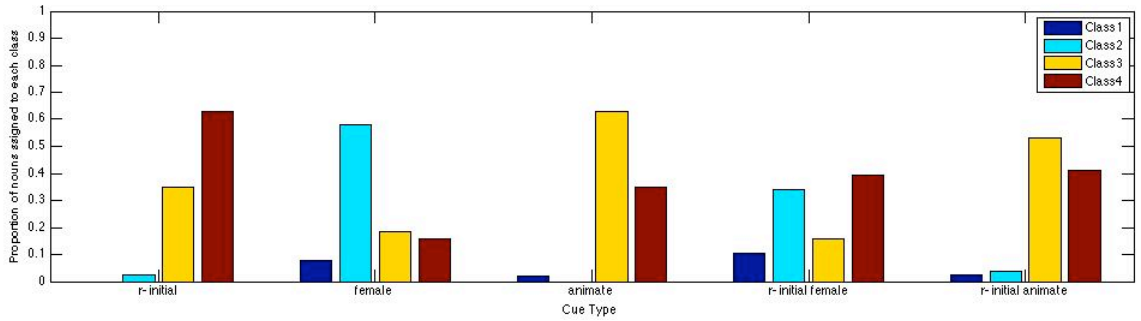
We evaluated the model by comparing its behavior to children's behavior from Gagliardi & Lidz (in prep). G&L elicited classification based on predictive phonological and semantic cues, as well as combinations of these cues that made conflicting predictions. The misperception model produced a close fit to the data in each condition (Figure 1). Furthermore, the estimated degree of misperception 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 misperception was held constant across simulations (0.95) demonstrates that our misperception model significantly outperforms the optimal naïve Bayesian classifier ($p<0.0001$).

These results support the idea that the noun classification data from G&L can be explained by an account in which children behave optimally with respect to the features of nouns in their lexicons, but nevertheless show what looks like suboptimal behavior due to misperception of semantic features. Children can perceive and discriminate sounds well before attaching meaning to these sounds; to perceive and track phonological features of a word, a child does not need to know or be sure of the meaning of the word. This is not so in with semantic features. This means that for phonological features the intake may match the input but for semantic features it does not. Our results show that while children appear to behave suboptimally with respect to the data available in the input, they may actually be behaving optimally with respect to their intake. Additionally, our work suggests that this asymmetry between phonological and semantic features lasts well into preschool age.

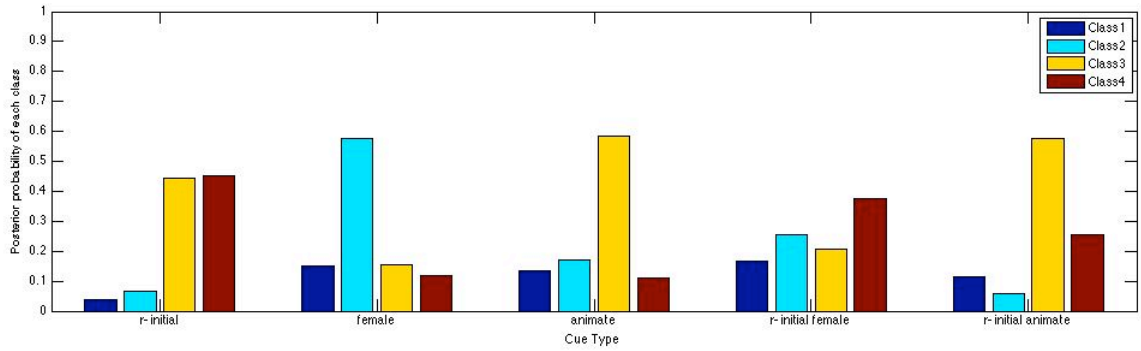
While it is reasonable to assume that children have some degree of difficulty perceiving and subsequently encoding semantic features in the lexicon, the suggestion that children use on 4-9% of the available features seems remarkably low for preschool children. We consider two alternative hypotheses for children's decreased reliance on semantic cues. First, it could be that children are reasonably good at encoding semantic features of nouns in their lexicon, but that they have difficulty perceiving these features on experimental items. Like the account outlined above, this would suggest an asymmetry in the input and intake, where children appear to behave suboptimally with respect to their input, but are actually behaving optimally with respect to their intake. A second possibility is that children come to the noun classification problem with a bias toward using phonological information rather than semantic information to classify nouns. This bias might require extensive experience to overcome, and this experience would be gated by whatever difficulties children have perceiving and encoding semantic features.

Figure 1: Proportion of nouns assigned to each class by cue type (a) Child behavioral data (b) Model with 95% uncertainty (c) Optimal Naïve Bayes classifier

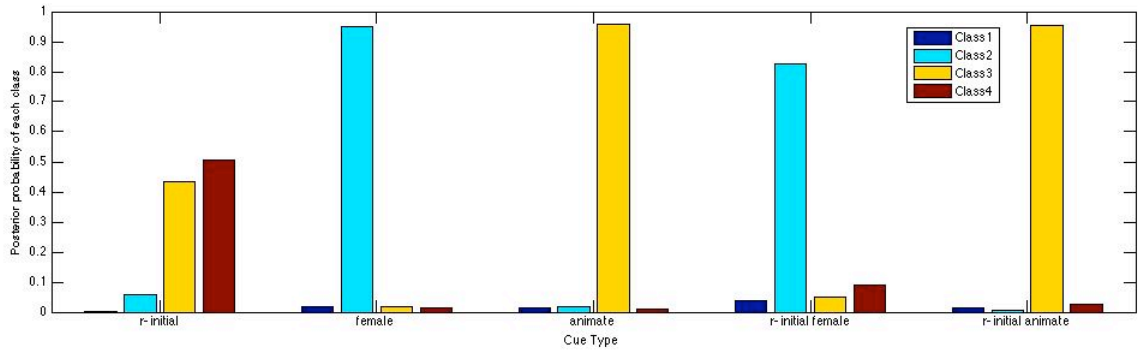
(a)



(b)



(c)



References

GAGLIARDI, A. & LIDZ, J. Under review at *Language.* Separating input from intake: Acquiring noun classes in Tsez.

GOMEZ, R. & GERKEN, L. 2000. Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, 4, 178-186.

SAFFRAN, J. R., NEWPORT, E. L., & ASLIN, R. N. 1996. Word segmentation: The role of distributional cues. *Journal of Memory and Language* 35(4), 606–621.