Psych 215L: Language Acquisition

Lecture 11

Morphology

Computational Problem

Identify parts of words that indicate functional information

"Jack hugged Lily."

"Jack is hugging Lily."

-ed: past, completed action
-ing: continuing action

"Jack hugged her."

"He hugged Lily."

he = masculine gender
her = feminine gender

Gagliardi, Feldman, & Lidz 2012

Acquisition problem: Acquire noun classes that differ by grammatical gender

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

Gagliardi, Feldman, & Lidz 2012

Acquisition problem: Acquire noun classes that differ by grammatical gender

Language: Tsez
"These classes can be characterized based on noun external distributional information (e.g. prefixal agreement on vowel initial verbs and adjectives) (Table 1), and noun internal distributional information (semantic and morphophonological features on the nouns themselves)"

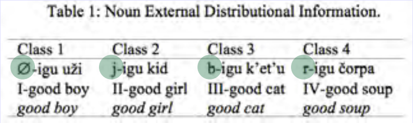

Table 1: Noun External Distributional Information.

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

Gagliardi, Feldman, & Lidz 2012

Acquisition problem: Acquire noun classes that differ by grammatical gender

Language: Tsez
"These classes can be characterized based on noun external distributional information (e.g. prefixal agreement on vowel initial verbs and adjectives) (Table 1), and noun internal distributional information (semantic and morphophonological features on the nouns themselves)"

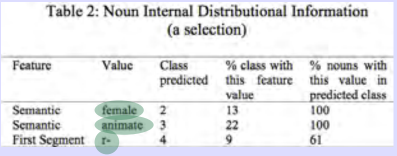

Table 2: Noun Internal Distributional Information (a selection)

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

Recall low Precision high

Gagliardi, Feldman, & Lidz 2012

Acquisition problem: Acquire noun classes that differ by grammatical gender

Question: How are children using the statistical information available?

"…what underlies the difference in the measureable *input* and the *intake* that children use to acquire noun classes."

Components in the acquisition process that can help with assigning a novel word in the experimental task to one of the classes:
(1) Accumulate knowledge of statistical distribution of features
(2) Observe features on novel item
(3) Know which features are relevant
(4) Generalize statistical knowledge to new item (using strategies such as Bayesian inference)

Gagliardi, Feldman, & Lidz 2012

Potentially useful features, which could be relevant (step 3)
"Each feature has specified values that were highly predictive of some class and an unspecified value that ranges over all other possible values that were not predictive."

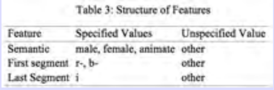

Table 3: Structure of Features

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

Modeling focus
"…how children use noun internal distributional information. In particular we will look at whether a child can make use of the predictive phonological and semantic information when classifying novel nouns, and how they perform when a noun has two features that make conflicting predictions."

Gagliardi, Feldman, & Lidz 2012

Are children behaving in an optimal fashion (i.e., similar to what an ideal Bayesian learning model would predict)?
"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])."

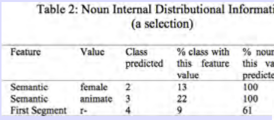

Table 2: Noun Internal Distributional Information (a selection)

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

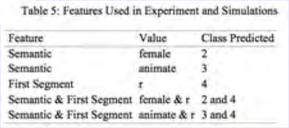

Table 5: Features Used in Experiment and Simulations

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

Gagliardi, Feldman, & Lidz 2012

Are children behaving in an optimal fashion (i.e., similar to what an ideal Bayesian learning model would predict)?

“When nouns had no conflicting features, children assigned more nouns to the class most strongly predicted by the feature than to any other class.”

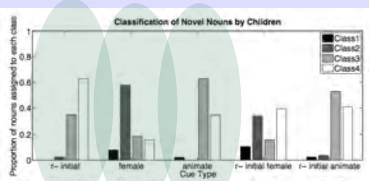

Figure 1: Proportion of novel nouns assigned to each class (by cue type) in the experimental task

Gagliardi, Feldman, & Lidz 2012

Are children behaving in an optimal fashion (i.e., similar to what an ideal Bayesian learning model would predict)?

“However, when nouns had more than one feature that made conflicting predictions, children relied more heavily on the phonological feature [r-] than on the semantic feature.”

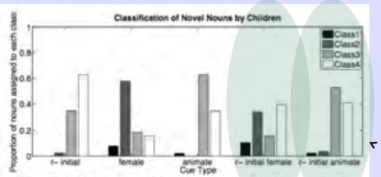

Figure 1: Proportion of novel nouns assigned to each class (by cue type) in the experimental task

Animate seems more important here…

Gagliardi, Feldman, & Lidz 2012

Are children behaving in an optimal fashion (i.e., similar to what an ideal Bayesian learning model would predict)?

“This is not likely to be predicted by the distribution of these features in the input, where nouns with the [animate] and [female] values of the semantic feature never occur in Class 4.”

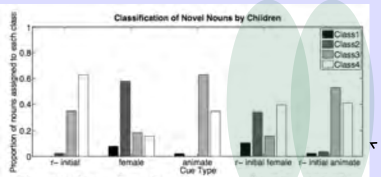

Figure 1: Proportion of novel nouns assigned to each class (by cue type) in the experimental task

Animate seems more important here…

Gagliardi, Feldman, & Lidz 2012

What does optimal behavior look like here, anyway?

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

“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.”

Gagliardi, Feldman, & Lidz 2012

Comparing children’s behavior to optimal behavior

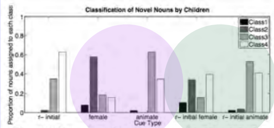

Figure 1: Proportion of novel nouns assigned to each class (by cue type) in the experimental task

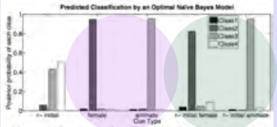

Figure 2: Predicted classification of novel nouns by an optimal naïve Bayesian classifier

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

(Also, children are generally a little noiser, even on the unambiguous features.)

Gagliardi, Feldman, & Lidz 2012

Explaining children’s behavior: Disparity in cue use

“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.”

Three explicit hypotheses:
(1) Semantic incompetence: Misrepresentation of semantic features

(2) Experimental reject: The semantic features in the experimental conditions were hard to pick up

(3) Phonological preference: Salience & reliability of phonological cues leads to preference for those cues

Gagliardi, Feldman, & Lidz 2012

Modeling the Semantic Incompetence Hypothesis
“…how classification by the model would be affected if the learner was misrepresenting some proportion of the semantic features that they should have encoded on nouns in their lexicon.”

“One way of quantifying this is by modeling the learner’s belief about the likelihood terms *p(f|c)* from Equation 1 under the assumption that these beliefs are derived from the counts that a learner accumulates of nouns in each class that contain a given feature.”

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

Gagliardi, Feldman, & Lidz 2012

Modeling the Semantic Incompetence Hypothesis
“We assume learners use a multinomial model with a uniform Dirichlet prior distribution to estimate the proportion of items each class *c* that contain a particular value *k* for feature *f.* Under this assumption, each likelihood term is equal to:

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

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.”

Gagliardi, Feldman, & Lidz 2012

Modeling the Semantic Incompetence Hypothesis

"Since the semantic incompetence hypothesis posits that children misrepresent semantic feature values some proportion of the time, we reduce the count of nouns in each class that contain the relevant semantic features, changing them instead to the unspecified feature value [other]. We then compute the posterior probability of noun class membership using these adjusted feature counts."

"We can use this model to ask how low the counts would have to be in order for children's behavior to be optimal with respect to their beliefs."

Gagliardi, Feldman, & Lidz 2012

Modeling the Semantic Incompetence Hypothesis

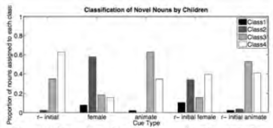

Figure 1: Proportion of novel nouns assigned to each class (by cue type) in the experimental task

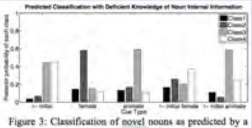

Figure 3: Classification of novel nouns as predicted by a Naïve Bayes Classifier with 95% of predictive semantic features misrepresented as [other].

"The model produced a close fit to the data in each condition…best fitting level of uncertainty ranged from 0.96-0.91, meaning that children would be only using 4-9% of the semantic cues available to them. A generalized likelihood ratio test in which the level of misrepresentation was held constant across simulations (0.95) demonstrates that our semantic incompetence model significantly outperforms the optimal naïve Bayesian classifier ($p < 0.0001$)."

Gagliardi, Feldman, & Lidz 2012

Modeling the Semantic Incompetence Hypothesis

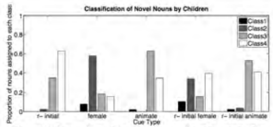

Figure 1: Proportion of novel nouns assigned to each class (by cue type) in the experimental task

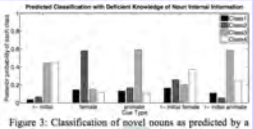

Figure 3: Classification of novel nouns as predicted by a Naïve Bayes Classifier with 95% of predictive semantic features misrepresented as [other].

But this only works early on in learning

"This analysis suggests that changes in model predictions under this account of feature misrepresentation occur primarily for low empirical feature counts, when the model relies heavily on pseudocounts from the Dirichlet prior distribution."

Gagliardi, Feldman, & Lidz 2012

Modeling the Experimental Reject Hypothesis

"…what would happen if a learner had a lexicon that faithfully represented the predictive features as they were distributed in the input and assumed both semantic and phonological features were relevant to classification, but didn't reliably encode semantic features on experimental items."

"To do this we use a mixture model, where some proportion of the time (1-*β)* an item that was supposed to have the specified semantic feature value [animate] or [female] (denoted as [spe]) it would be classified as with that value, the rest of the time (β) it would be classified as if it had the unspecified value [other]."

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

Gagliardi, Feldman, & Lidz 2012

Modeling the Experimental Reject Hypothesis

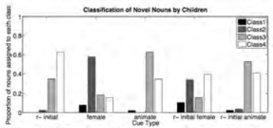

Figure 1: Proportion of novel nouns assigned to each class (by cue type) in the experimental task

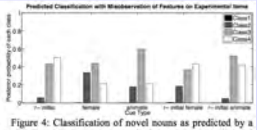

Figure 4: Classification of novel nouns as predicted by a model that misobserves semantic features on experimental items 58% of the time

“The best fitting level value of *β* ranged from .49 to .83, where 58% was the best fit overall. This means that children would be misperceiving semantic features on 58% of the experimental items. A generalized likelihood ratio test indicates that the experimental reject model also significantly outperforms the optimal naïve Bayesian classifier (p < 0.05).”

Gagliardi, Feldman, & Lidz 2012

Modeling the Phonological Preference Hypothesis

“…what would happen if we had a learner that was biased not to use semantic features in classification some proportion of the time, even if these features were represented just as distributed in the input and accurately perceived during the experimental task.”

“We used a second mixture model, this time looking at the mixture of a Bayesian classifier that used both semantic and phonological features, and one that only used phonological features.”

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

Gagliardi, Feldman, & Lidz 2012

Modeling the Phonological Preference Hypothesis

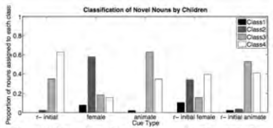

Figure 1: Proportion of novel nouns assigned to each class (by cue type) in the experimental task

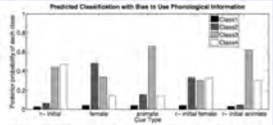

Figure 5: Classification as predicted by a model biased not to use semantic information 65% of the time

“The best fitting value of *β* ranged from .49 to .83, and was .65 over all, meaning that children would be choosing not to use semantic features on 65% of classification decisions. A generalized log likelihood test showed that this model also significantly outperformed the optimal naïve Bayesian classifier (p < 0.0001).”

Gagliardi, Feldman, & Lidz 2012

What the different hypotheses indicate collectively

“This suggests that although originally children did not look as though they were behaving optimally with respect to the input, they may well be behaving optimally with respect to their intake, that is, the input as they have represented it.”

Choosing among the different hypotheses

“It is not obvious how one would best to evaluate the alternative models with respect to one another…it is likely that a combination of all three of these processes (and perhaps more that we haven’t considered here) is influencing children’s classification decisions. This could potentially be explored through a combined model; however, as all of these models fit the data so closely, it would be difficult to determine which and to what extent each type of misrepresentation or bias is involved.”

Gagliardi, Feldman, & Lidz 2012

Implications

“…by combining experimental data from children acquiring an understudied language with computational modeling techniques, we found a better understanding of both children’s acquisition of Tsez, and the role of statistical cues in language acquisition. Tsez was an ideal language to look at, as feature types differed in their reliability as cues to noun class.”

“…we identified an area where children’s behavior does not appear to reflect the ideal inferences licensed by the statistical patterns in the input. Three models allowed us to investigate the source of this asymmetry. While each model differed in where the asymmetry came from, all employed a weakening of the statistical import of semantic features. This is a distinct pattern from the finding that children learning an artificial language amplify an already strong statistical tendency (Hudson-Kam & Newport, 2009).”

Gagliardi, Feldman, & Lidz 2012

Implications

“…while children’s behavior does not align with the predictions made by the optimal Bayesian classifier, it can be predicted by modifying the terms of this classifier in reasonable ways. Thus we were able to model children’s suboptimal behavior using a Bayesian model, rather than adopting some other system of computation.”

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