The Acquisition of Noun Classes in Tsez: Computational and Experimental Results

Annie Gagliardi[1], Jeff Lidz[1] and Maria Polinsky[2]

[1]University of Maryland, [2]Harvard University

Background

Category Learning Problem in Natural Language

How do kids discover that they have multiple categories (of words, nouns)?

How do they learn what belongs in what class, and learn to assign new items to a class?

Correlated Cues in Artificial Language Learning

(Braine 1989, Frigo & McDonald 1998, Gerken et al 2005)

Lexicon: 2 classes of **words** (1 and 2), and 2 classes of **dependents** (1 and 2)

Dependency: words from Class *i* can only cooccur with Class *i* dependents

Unlearnable Language		Learnable Language	
Class 1	Class 2	Class 1	Class 2
mul-ja, mul-du	sif-no, sif-bi	mula-ja, mula-du	sifo-no, sifo-bi
don-ja, don-du	jav-no, jav-bi	dona-ja, dona-du	javo-no, javo-bi
kap-ja, kap-du	bip-no, bip-bi	kap-ja, kap-du	bip-no, bip-bi
gav-ja, gav-du	dit-no, dit-bi	gav-ja, gav-du	dit-no, dit-bi

To learn **regular dependencies** between items, learners need **partially correlating information** on some members of each class. The dependency and the correlating information make up the **correlated cue.**

Investigating the correlated cue in natural language acquisition:

(1) do they exist in natural language?
(2) are they in children's input?
(3) are children sensitive to the correlating information?
(4) is category learning dependent on the correlated cue?

Future work will determine how the correlated cue works (computational models), and whether the artificial language results really parallel what appears to go on in natural language

Tsez (Dido)

Nakh-Dagestanian language with 4 noun classes spoken by about 6,000 people in Dagestan

Dependency

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

Regular noun class agreement overt on most vowel initial verbs, adjectives and adverbs

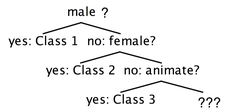
Is there Partially Correlating Information for each class?

Class 1	Class 2	Class 3	Class 4
all male humans	all female humans	all other animates	many other things
only male humans	many other things	many other things	
~13% of words	~12% of words	~41% of words	~34% of words

(1) Do Correlated Cues Exist in Tsez?

Is there information on a subset the nouns in each class that correlates with class? This, in conjunction with agreement, would constitute a correlated cue.

Decision Trees are built by a **supervised learning algorithm** that takes words with specified features and determines which features are most predictive of class



Plaster & Harizanov (2009): Built decision trees classifying Tsez nouns from a dictionary (Xalilov, 1999), but we want to use input that reflects actual language use.

Need a corpus, Build a corpus
~3,000 lines (10 hours) of child directed Tsez speech, transcribed with the help of native speakers

Use nouns from the corpus instead of the dictionary to build a decision tree reflecting words actually in use => set of highly predictive features

(2) Do Correlated Cues Exist in the Children's Input?

- From Corpus: predictive feature on nouns triggering overt agreement
  Class 1: 100%, Class 2: 52%, Class 3: 51%, Class 4: 45%
- 84% of verb tokens and 77% of adjective tokens show overt agreement
- Are there enough examples to be useful?
  - future work with artificial languages will help determine what is enough

(3) Are Speakers Sensitive to the Correlating Information?

Classification Experiment to elicit classification of Real and Nonce words containing different predictive features by native Tsez speakers

Class	Semantic (SC)	Weak Semantic (WCP, WCC)	Phonological (PC, PCR, PCI)	2 agreeing (AC)	2 conflicting (CCG, CCB, CCR, CCI)
1	male human (3/3)	-----	-----	-----	male human & ɣ- initial (0/3) male human & b-initial (0/3)
2	female human (3/3)	paper (3/3) clothing (3/3)	ɣ- initial (3/3)	female human & ɣ-initial (0/3)	female human & r-initial (0/3)
3	animate (3/3)	-----	b- initial (3/3)	animate & b-initial (3/3)	animate & r-initial (2/3) animate & i final (0/3)
4	-----	-----	r-initial (3/3) i final (3/3)	r-initial & i final (2/3)	b-initial Cl4 real words (3/0)

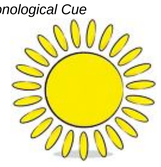
Subjects: 32 Native Tsez speakers in Shamkhal and Kizilyurt, Dagestan
(10 young children (~ 6yr), 12 older children (~ 9yr), 10 adults)

Eat/Don't Eat Task: -iš (*eat*, intransitive), and -ac'o (*eat*, transitive) show overt agreement with the subject and object respectively

Assistant introduces each character and item, participant tells the character to eat, then what to eat/not eat. Class assignment is evident in the agreement.

kid (*girl)*
Class 2, Semantic Cue

buq (*sun)*
Class 3, Phonological Cue



k'uraj (*onion)*
Class 4, no Cue

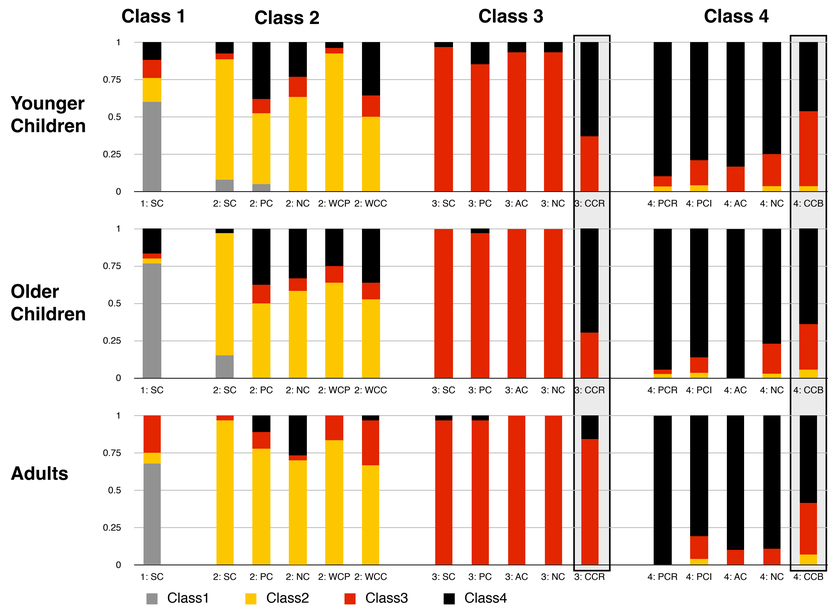
zamil (*nonce)*
Class 3, Semantic Cue

Critical Findings

- Correlated Cues exist in Tsez
- **Not all statistically predictive information is used equally**
  - not just an effect of differences in predictiveness
  - for children phonological cues on real words more powerful than semantic cues
  - children ignore weak semantic cues entirely
- Differences between children and adults suggest a bias to use certain kinds of cues when discovering classes

Experimental Results

Real Words:



Nonce Words: