Psych 229: Language Acquisition

Lecture 7

Categories & Models

Gómez & Lakusta 2004: Categorization

Category abstraction task

Table 1 *A paradigm for investigating category abstraction. Learners are exposed to the pairings shown below except for those denoted by empty cells. Learners are then tested to see if they will generalize correctly to the withheld strings (denoted by empty cells)*

	X_1	X_2	X_3	X_4	X_5
a_1 = the	boy	girl	ball	dog	cat
a_2 = a	boy	girl	ball	dog	
	Y_1	Y_2	Y_3	Y_4	Y_5
b_1 = will	jump	run	play	sleep	eat
b_2 = can	jump	run	play	sleep	

Previous work (aX, bY paradigm)

Interestingly, although learners readily acquire the legal positions of words with respect to which occur first versus second (Smith, 1969), categories and their relationships (i.e. that words belong to particular a, b, X, and Y classes, and that a-words go with Xs and not Ys) are virtually impossible to acquire unless some subset of the X- and Y-category members are marked with salient conceptual or perceptual cues

Gómez & Lakusta 2004: Categorization

What's going on (aX, bY paradigm) - a 2 step process

According to Braine (1987, see also Frigo & McDonald, 1998), there are two essential steps in aX bY category abstraction. Learners must first associate a- and b-elements with cues differentiating X and Y categories. They can then categorize a- and b-elements based on their co-occurrence (or distributional) relationships to the X- and Y-cues. In the second step, learners group (or categorize) a- and b-elements by merit of their joint association with particular distinguishing cues. Once a- and b-categories are formed, learners can rely on memory for a pair they have heard (e.g. 'the cat' in Table 1) and the fact that a_1 ('the') and a_2 ('a') belong in the same category to make inferences about a pair they have not heard (e.g. 'a cat'). By this view, Step 1 learning is evidenced by the ability to discriminate correct from incorrect pairings of functional and lexical test items with distinguishing cues present. Step 2 learning is evidenced by discrimination of test items in the absence of distinguishing cues. Frigo and McDonald (1998) found that there were no learners who had mastered Step 2, but not Step 1. Thus these steps appear to be serial in nature.

Initial Association
- a --> X (with cue), b --> Y (with cue)

Category Membership (abstract)
- D1 is type a since it goes with X
  - how is D1 like other as?
- D2 is type b since it goes with Y
  - how is D2 like other bs?

Gómez & Lakusta 2004: Categorization

17-month olds can do both steps

What about younger learners? Recent research with 17-month-old infants from an English-speaking environment found that they too were able to induce Russian gender categories (Gerken, Wilson & Lewis, 2003; see also Gerken, 2002).

What about younger children?

The ability to abstract categories from distributional information in speech is a significant milestone in cognitive and language development, not only for what it implies about early abstraction abilities but also because of its potentially important contribution to syntactic development. Thus it is necessary to ask about the developmental trajectory of such learning. Do infants younger than 17 months of age show similar abilities?

7- 12 month abilities: previous work

We know that by 7 and 12 months of age, infants are able to abstract patterns from artificial grammars as evidenced by their ability to discriminate grammatical from ungrammatical strings in new vocabulary (Gómez & Gerken, 1999; Marcus, Vijayan, Bandi Rao & Vishton, 1999). However, the kind of abstraction demonstrated by these studies involved noting patterns of repeating or alternating elements (Gómez, Gerken & Schvaneveldt, 2001), a very different kind of abstraction than might be involved in learning form-based categories

We also know from Gerken *et al.* (2003) that 12-month-olds do not show Step 2 learning. However, these younger infants might be able to engage in a more preliminary form of category-based abstraction, therefore providing clues to the developmental trajectory of such learning.

So let's look at 12-month olds…

Gómez & Lakusta 2004: Categorization

General procedure

Infants in all the experiments were exposed to one of two training languages. In Experiment 1, one language consisted of aX and bY pairings. The other training language consisted of aY and bX pairs. Xs were instantiated as disyllabic words and Ys were monosyllabic. During testing, infants were exposed to *new* phrases from their training language versus phrases from the other language. Thus some phrases conformed to the training language whereas others did not. Importantly, all X- and Y-words were novel at test.[2]

Novel "vocabulary"

Table 2 *Training vocabulary*

Category			
a	b	X[a]	Y[b]
alt	*ong*	*coomo*	*deech*
ush	*erd*	*fengle*	*ghope*
		kicey	*jic*
		loga	*skige*
		paylig	*vabe*
		wazil	*tam*

[a] 'X' elements are disyllabic words.
[b] 'Y' elements are monosyllabic words.
Note: L1 phrases are formed by aX and bY pairings. L2 phrases take the form aY and bX.

Basic (Example) Task:

Data
- alt, ush --> 2-syllable words
- ong, erd --> 1-syllable words

1) Notice correlation.
2) Realize *n*-syllable abstraction.
3) Form new rules:
   e.g. alt/ush --> 2-syllable word
4) Apply new rule to new items.

Table 3 *Test vocabulary*

Category			
a	b	X[a]	Y[b]
alt	*ong*	*bevit*	*vot*
ush	*erd*	*meeper*	*pel*
		gackle	*tood*
		roosa	*rud*
		nawlup	*biff*
		binnow	*foge*

[a] 'X' elements are *novel* words, yet they remain disyllabic.
[b] 'Y' elements are *novel* words, yet they remain monosyllabic.
Note: L1 phrases are formed by aX and bY pairings. L2 phrases take the form aY and bX.

Gómez & Lakusta 2004: Categorization

Fair amount of data, short exposure

Thus, each language contained 24 unique phrases, which were then combined to form grammatical strings. L1 strings took the form, S → [aX] [bY] or S → [bY] [aX]. L2 strings took the form S → [aY] [bX] or S → [bX] [aY]. Altogether there were 288 possible strings in each of the training languages, 72 of which occurred once during training. A subset of 72 strings was used so that the training phase lasted only 3 minutes. Strings were selected such that each of the 24 phrases in the training language was presented six times during training and so that the order of phrases within strings (e.g. [aX] [bY] versus [bY] [aX]) was counterbalanced.

Table 4 *Test materials*

Test sets			
L1-test1	L1-test2	L2-test1	L2-test2
[*alt nawlup*] [*ong pel*]	[*alt roosa*] [*ong tood*]	[*alt biff*] [*ong meeper*]	[*alt rud*] [*ong gackle*]
[*alt gackle*] [*ong foge*]	[*alt meeper*] [*ong biff*]	[*alt vot*] [*erd gackle*]	[*alt pel*] [*ong nawlup*]
[*alt bevit*] [*erd tood*]	[*alt binnow*] [*erd rud*]	[*alt tood*] [*ong binnow*]	[*alt foge*] [*erd roosa*]
[*ush binnow*] [*erd vot*]	[*ush gackle*] [*erd foge*]	[*ush rud*] [*ong roosa*]	[*ush vot*] [*erd meeper*]
[*ush roosa*] [*ong rud*]	[*ush nawlup*] [*ong vot*]	[*ush pel*] [*erd nawlup*]	[*ush tood*] [*erd binnow*]
[*ush meeper*] [*erd biff*]	[*ush bevit*] [*erd pel*]	[*ush foge*] [*erd bevit*]	[*ush biff*] [*ong bevit*]
[*ong vot*] [*alt roosa*]	[*ong rud*] [*ush roosa*]	[*ong gackle*] [*ush vot*]	[*ong roosa*] [*ush rud*]
[*ong tood*] [*ush bevit*]	[*ong pel*] [*alt bevit*]	[*ong bevit*] [*alt rud*]	[*ong meeper*] [*alt vot*]
[*ong biff*] [*ush gackle*]	[*ong foge*] [*ush meeper*]	[*ong nawlup*] [*ush tood*]	[*ong binnow*] [*ush pel*]
[*erd foge*] [*alt meeper*]	[*erd biff*] [*ush binnow*]	[*erd binnow*] [*alt pel*]	[*erd nawlup*] [*ush foge*]
[*erd pel*] [*alt binnow*]	[*erd tood*] [*alt gackle*]	[*erd meeper*] [*alt foge*]	[*erd gackle*] [*alt tood*]
[*erd rud*] [*ush nawlup*]	[*erd vot*] [*alt nawlup*]	[*erd roosa*] [*ush biff*]	[*erd bevit*] [*alt biff*]

Note: Test strings were presented in sets (e.g. L1-test1). Notice, X- and Y-categories consist of new one- and two-syllable words. The strings in the test sets were presented in a random order.

Gómez & Lakusta 2004: Categorization

Expt 1: infants make the association (familiarity preference)

A Wilcoxon Signed Ranks Test showed that infants listened significantly longer to strings from their training language than to strings from the other language ($T = 251$, $p = .004$). Eighteen out of 24 infants showed this pattern, suggesting that they had acquired some sensitivity to the category-based structure of their training grammar.

Infants can make abstract rule and apply it new items. But what's the relationship of this to "real" categorization?

The ability to discriminate legal from illegal marker-feature pairings, despite the fact that X- and Y-elements were novel at test, reflects sensitivity to the co-occurrence relations between markers and X- and Y-categories based on their distinguishing features. Such learning is striking given its complexity – infants had to track four similar sounding markers, associate them with particular distinguishing features and generalize to pairings containing novel words. The fact that infants were able to generalize to novel X- and Y-elements suggests that learning was to some degree abstract (involving grouping of the X- and Y-elements according to syllable number).

Does such grouping count as *categorization*? We would argue 'yes' to the extent that categorization involves distinguishing elements according to some feature or set of features. When individuals treat elements similarly, they are responding to these elements *as if* they were members of the same category.

Gómez & Lakusta 2004: Categorization

Expt 2: Noisy data (more like real life)

Thus, in Experiment 2 we investigated whether young learners are able to separate more probable from less probable structure by exposing them to artificial languages with varying degrees of probabilistic structure. In Condition 83/17, approximately 83% of the training strings were from the infants' 'predominant' training language, whereas 17% of the strings were from the other language. In Condition 67/33, the split between the predominant and non-predominant training languages was 67% and 33%, respectively. Infants who are able to distinguish more probable from less probable structure should show the same pattern of discrimination observed in Experiment 1 (from now on the 100/0 condition).

83/17: yes!   67/33: no

Table 5 *Median differences and 95.1% confidence intervals for listening times (in seconds) to strings from the predominant versus the non-predominant training language. Infants showed significant discrimination even when 17% of the strings encountered during training were from the non-predominant training language*

Probability ratio	Median difference	95.1% confidence interval
Exp. 1: 100/0	1.618*	(0.665–2.605)
Exp. 2: 83/17	1.248*	(0.130–2.850)
Exp. 2: 67/33	−0.125	(−0.925–0.835)

Note: * Listening time differences in these conditions were statistically significant, $p \leq .05$.

Question: Familiarization vs. novelty preference?

Question: How would adults do in these tasks?
(ref: Hudson Kam & Newport 2005)

Gómez & Lakusta 2004: Categorization

So 12-month olds can make & apply this kind of rule (evidenced by novel test items)

We asked whether 12-month-olds would learn the marker-feature relationships and generalize these to *new* X and Y vocabulary at test (thus showing evidence for Step 1 learning). The answer was affirmative. Infants were able to discriminate legal from illegal marker-feature pairings (exhibited by longer looking times to the former than the latter). These findings are noteworthy given the brief training period (three minutes) and the complexity of the artificial language. Infants had to track four similar sounding markers and associate them with particular distinguishing features. The fact that infants were able to generalize to novel X- and Y-elements suggests grouping of these elements according to syllable number and hence categorization by means of this feature. It is important to stress that unlike previous studies, the X- and Y-elements were all novel at test, whereas the features used to group X- and Y-elements in previous studies were physically identical.

Abstraction = # of syllables (how realistic is this?)

Importantly, the 12-month-old infants in this study were not simply learning associations between a- and b-elements and physically identical features. Rather, they were generalizing based on the abstract feature of syllable number, demonstrating that they are capable of categorizing at a level at least one step removed from physical identity. Such generalization is an important precursor to that shown by the older infants in the Gerken *et al.* study, who, by 17 months old, can form a- and b-categories comprised of elements with no common features other than their co-occurrence patterns with X- and Y-categories.

Threshold for generalization

The results are important for determining whether infant learners are equipped to tolerate some degree of inconsistency in their linguistic input. Infants were indeed able to focus on the predominant pattern in their training language and generalized to new strings on this basis. There appear to be limits on such learning, however. Infants exposed to greater noise in Condition 67/33 failed to show learning.

Gómez & Lakusta 2004: Categorization

These findings pose intriguing questions with regard to possible constraints on learning. Were infants in Condition 83/17 learning two forms of structure simultaneously or only the more predominant abstract structure? According to the first possibility, infants were learning specific marker-word phrases from the non-predominant language (there were only two of these) and the more abstract pairings of markers and features (i.e. syllable number) from the predominant language. According to the second possibility, infants were ignoring phrases from the non-predominant language entirely. Because infants were tested on their ability to generalize to new marker-word phrases (rather than old marker-word phrases), we are unable to distinguish these explanations in the present studies.

What about learning in Condition 67/33? Infants in this condition were clearly not generalizing the marker-feature pairing. Nor were they engaged in learning two forms of structure simultaneously (abstract marker-feature pairings and specific marker-word phrases), otherwise they would have shown discrimination on the test as in Condition 83/17. An alternative possibility is that the greater presence of phrases from the non-predominant language disrupted learning entirely. Yet another possibility is that infants learned *only* specific marker-word phrases from the non-predominant language. However, we are unable to distinguish these possibilities with the present data because we did not test infants on marker-word phrases from training

Options for Learning:

1) Learn the rule for the more dominant data type, learn individual items for less dominant. (67/33 case: both "less" dominant)

1) Ignore less dominant data as noise and learn nothing for those items.

2) Learn rules probabilistically? (67/33 case) - forced choice between two options might not reveal this level of probability distinction

Gómez & Lakusta 2004: Categorization

What distribution information is used?

Infants also show some selectivity in terms of their tendency to focus on different types of structure. Given two sources of statistical information, infants will favor the source of greater statistical regularity. Gómez (2002) found that when conditional probabilities between adjacent words are high, learning will reflect a focus on this source of structure. However, when conditional probabilities are low, learners will focus on some other, more reliable, source of statistical information, such as non-adjacent dependencies (see also Gómez, Welch & Lany, 2004). As such, it is reasonable to hypothesize that learners will only focus on a particular source of information to the extent that it yields some degree of statistical regularity. Beyond this point, learners will seek out alternative sources of information (Gómez, 2002).

How long will they wait before "backing off" to another source of information?

Still, some level of noise tolerance (= good)

Instead, we see that learning is both powerful and constrained. Learners can tolerate some degree of inconsistency in their linguistic input, but learning diminishes as inconsistency increases.

Decay under noise

Another issue raised by this research is the question of whether learning degrades gradually or catastrophically with increases in noise. The present findings suggest that learning degrades gradually in that there were no significant decreases in learning from the 100/0 to the 83/17 conditions, and then a marginal decrease between Conditions 83/17 and 67/33.

Marginal decrease from 83/17 to 67/33?

Probability ratio	Median difference	95.1% confidence interval
Exp. 1: 100/0	1.618*	(0.665–2.605)
Exp. 2: 83/17	1.248*	(0.130–2.850)
Exp. 2: 67/33	−0.125	(−0.925–0.835)

Gómez & Lakusta 2004: Categorization

Main conclusion about categorization

Rather, development appears to reflect a progression in the ability to form abstractions from features that must be present at the time of generalization (e.g. syllable number) to those that may be inferred based on memory of a previously encountered instance.

Will distributional learning get us here?

Gómez & Lakusta 2004: Categorization

Discussion questions

Relation to bilingual learning: if children are equipped to learn the predominant structure (assuming it's more than 83% of the data), what does this mean for bilingual children where the data distributions are far messier? (Related question: what if the two languages have different structures?)

Artificial language vs. real language situations: How valid are artificial language results for explaining real language learning, especially since artificial languages are missing so much information available in real languages?

Related: Is distinguishing between one and two-syllable words a realistic analogy for categorization?

Mintz 2003: Frequent Frames

Just to remind us of the problem again in more detail…

Grammatical categories (e.g. noun verb, etc.) are fundamental building blocks of grammar, yet it is not fully known how child language learners initially categorize words. There has been recent interest in the idea that distributional information carried by the co-occurrence patterns of words in sentences could provide a great deal of information relevant to grammatical categories. For example, words in position X in sentences containing the English fragment in (1) are likely to belong to the same grammatical category, verb.

(1) … to X to …

Pinker (1987) argued that, given sentences in (2a,b), a distributional learner would incorrectly categorize *fish* and *rabbits* together and, hearing (2c), would incorrectly assume that (2d) is also permissible.

(2) a. John ate fish.
b. John ate rabbits.
c. John can fish.
d. *John can rabbits.

The crux of the problem exemplified in (2) is that a given word form (in this case, *fish*) can belong to multiple categories and thus occur in different syntactic contexts (e.g. as a noun in 2a or a verb in 2c), potentially providing misleading category information. Pinker argued that the resulting erroneous generalizations would be common, and would render a distributional approach to categorization untenable.

Mintz 2003: Frequent Frames

Another potential difficulty is that important distributional regularities are not always local, as in (1), but can occur over a variable distance, as in (3) (Chomsky, 1965; Pinker, 1987).

(3) … *to* hurriedly and effortfully *X to* …

Here, the informative verb environment for X in (1) (… *to X to* …) spans many words. The fundamental issue is that lexical adjacency patterns are variable: in any particular utterance, the words in a specific position relative to a target is somewhat accidental, and a learner that categorized only from fixed positions could be led to make erroneous generalizations. Thus, another question is how the learner is to know *which* environments are important and which should be ignored. Distributional analyses that consider all the possible relations among words in a corpus of sentences would be computationally unmanageable at best, and impossible at worst.

What information should children be tracking?

How about local linguistic environments?

Moreover, by showing that local contexts are informative, these findings suggested a solution to the problem of there being too many possible environments to keep track of: focusing on local contexts might be sufficient.

Mintz 2003: Frequent Frames

Frequent frames

Frequent frames are defined as ordered pairs of words that frequently co-occur with exactly one word position intervening (occupied by any word). Any words that occur as the intervening word inside a given frequent frame are categorized together. The motivation for investigating the informativeness of frequent frame contexts as a basis for category learning was twofold: (1) to study a distributional unit for which there is evidence that infants and adults attend to: the logical problem of which distributional context(s) to attend to can be circumvented if one can demonstrate that the distributional contexts that learners *do* attend to can support categorization; and (2) to study a procedure that categorizes only words that occur in contexts that are likely to be informative: frequent frames, as defined here, might provide reliable category information because requiring the joint occurrence of contextual elements eliminates many accidental contexts from the analysis. A brief overview of these

X _Y:
categorize all _ together

18-month olds can use non-adjacent information

In sequence-learning studies, Gómez (2002) has shown that adults and 18-month-old infants can track (slightly) long distance dependency relationships similar to those that define frames here (see also Santelmann & Jusczyk, 1998).

And lexical items are salient

In word-learning research, Childers and Tomasello (2001) have shown that children more easily acquire novel verb meanings when the verbs occur in lexical frames that occur frequently in the children's input.

Mintz 2003: Frequent Frames

In the present approach the word 'W' in the environment '… X W Y …' is stored as "jointly following X and preceding Y", but such would not be the case if W occurred after X and before Y on independent occasions. In contrast, the Mintz et al. (2002) and Redington et al. (1998) studies use "bigram" contexts, which record only independent co-occurrence patterns (e.g. "following X", "preceding Y"). There are potentially important consequences of using frame contexts as opposed to bigram contexts. In particular, the property of joint co-occurrence in the frame contexts involves an additional relationship between the context elements themselves, as well as between context and target word.

Better than bigrams (transitional probability)

…more informative by themselves

Another important difference between frame and bigram contexts is that, as mentioned above, adults will categorize words in an artificial language based on their occurrence within frames (Mintz, 2002), whereas bigram regularity alone has failed to produce categorization in artificial grammar experiments, without additional cues (Braine, 1987; Gerken, Gomez, & Nurmsoo, 1999; Smith, 1966; Wilson, Gerken, & Nicol, 2000).

The plan: Are frequent frames a useful strategy on real data? Do they yield the right information?

The goal of the work described here was not to provide a model of grammatical categorization by learners (cf. Cartwright & Brent, 1997), but to examine, based on evidence from human cognition, what assumptions would be reasonable to build into such a model. Specifically, the goal was to formulate a unit to which there is some evidence that children and adults attend, and with which adults have been shown to categorize, and examine how predictive it is of grammatical category membership.

Mintz 2003: Frequent Frames

6 corpora *Do we believe these are a realistic representation of data children hear outside the laboratory?

Table 1
Experiment 1 session ranges for analyzed corpora, number of utterances, number of tokens and types categorized, percentage of corpus (tokens) accounted for by categorized types, and percentage of corpus (tokens) analyzed

Child	CHILDES sessions	# of utterances	Tokens categorized	Types categorized	Percentage of corpus accounted for	Percentage of corpus analyzed
Peter	peter01-peter12	19846	5690	446	48%	6%
Eve	eve01-eve20	14922	3513	400	46%	5%
Nina	nina01-nina23	14417	6265	469	51%	8%
Naomi	n01-n58	6950	1617	297	38%	5%
Anne	anne01a-anne23b	26199	4389	405	54%	4%
Aran	aran01a-aran20b	20857	5628	620	61%	5%
Mean			4517	439.5	50%	6%

evaluation metrics

"precision"

$$\text{Accuracy} = \frac{\text{hits}}{\text{hits} + \text{false alarms}}$$

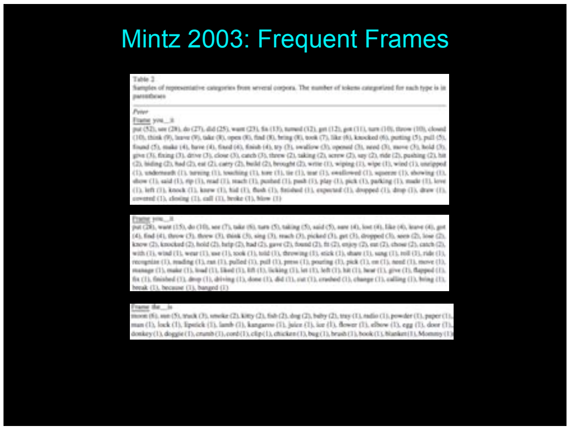
"recall"

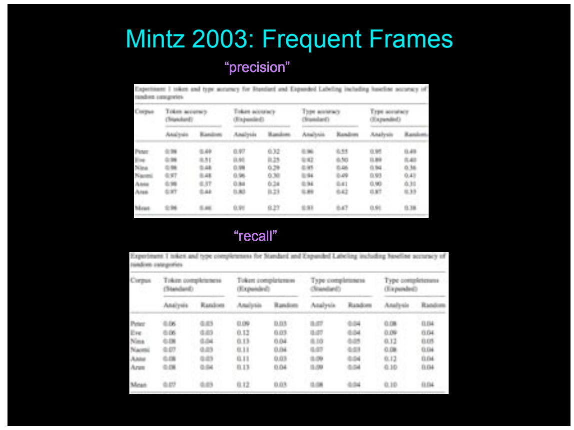
$$\text{Completeness} = \frac{\text{hits}}{\text{hits} + \text{misses}}$$

Using 45 as absolute cut-off for "frequent" frame

Next, a subset of these frames was selected as the set of frequent frames. The principles guiding inclusion in the set of frequent frames were that frames should occur frequently enough to be noticeable, and that they should also occur enough to include a variety of intervening words to be categorized together. While these criteria were not operationalized in the present experiment, a pilot analysis with a randomly chosen corpus, Peter, determined that the 45 most frequent frames satisfied these goals and provided good categorization. Hence, the frames analyzed for each corpus were the 45 most frequent frames for that corpus.

The distributional information provided by frequent frames was robust. The word types that were categorized constituted, on average, 50% of the tokens in a given corpus. This coverage was achieved by analyzing only about 6% of the tokens and their contexts. That is, the tokens of the categorized types making up the 6% contained in frequent frames constituted half the tokens in a given corpus. Frequent frames can thus focus a learner on a relatively small number of contexts that can have broad impact on how words in the input are categorized. The efficiency and accuracy provided by frequent frames could thus be very useful to young language learners, who have limited memory and processing resources.


Mintz 2003: Frequent Frames


Mintz 2003: Frequent Frames
“precision”
“recall”