Psych 215L: Language Acquisition

Lecture 8

Word-Meaning Mapping

Computational Problem

"Look! There's a goblin!"

Smith & Yu (2008)

Learning in cases of referential ambiguity:

Why? "…not all opportunities for word learning are as uncluttered as the experimental settings in which fast-mapping has been demonstrated. In everyday contexts, there are typically many words, many potential referents, limited cues as to which words go with which referents, and rapid attentional shifts among the many entities in the scene."

Also, "…the evidence indicates that 9-, 10-, and certainly 12-month-old infants are accumulating considerable receptive lexical knowledge …Yet many studies find that children even as old as 18 months have difficulty in making the right inferences about the intended referents of novel words…infants as young as 13 or 14 months…can link a name to an object given repeated unambiguous pairings in a single session. Overall, however, these effects are fragile with small experimental variations often leading to no learning."

Smith & Yu (2008)

New approach: infants accrue statistical evidence across multiple trials that are individually ambiguous but can be disambiguated when the information from the trials is aggregated.

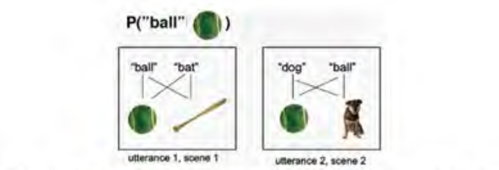


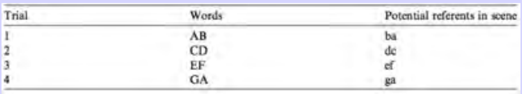
Fig. 1. Associations among words and referents across two individually ambiguous scenes. If a young learner calculates co-occurrences frequencies *across* these two trials, s/he can find the proper mapping of "Ball" to BALL.

Smith & Yu (2008)

A more complicated example:

Trial 1: A = a (.5), b (.5)?   B = a (.5), b (.5)?
Trial 2: C = c (.5), d (.5)?   D = c (.5), d (.5)?
Trial 3: E = e (.5), f (.5)?   F = e (.5), f (.5)?
Trial 4:
A = g (.3), a (.3), b (.3)?   G = g (.5), a(.5)?
(but wait! b isn't present, so A = b has prob 0)
A = g (.5), a (.5)?
(but wait! G wasn't present in trial 1, A = g has prob 0)
A = a   G = g

Requirements:
(1) Learner notices absence of b in Trial 4
(2) Learner remembers absence of g in Trial 1
(3) Learner registers occurrences & non-occurrences
(4) Learner calculates correct statistics based off this information



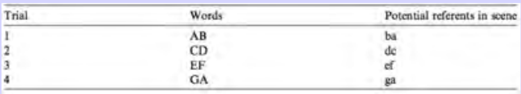
Trial	Words	Potential referents in scene
1	AB	ba
2	CD	dc
3	EF	ef
4	GA	ga

Smith & Yu (2008)

Yu & Smith (2007): Adults seem able to accomplish this.

Smith & Yu ask: Can 12- and 14-month-old infants do this? (Relevant age for beginning word-learning.)

Requirements:
(1) Learner notices absence of b in Trial 4
(2) Learner remembers absence of g in Trial 1
(3) Learner registers occurrences & non-occurrences
(4) Learner calculates correct statistics based off this information

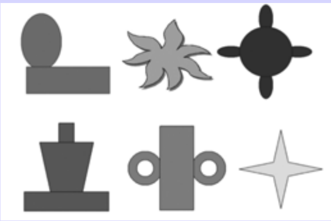


Trial	Words	Potential referents in scene
1	AB	ba
2	CD	dc
3	EF	ef
4	GA	ga

Smith & Yu (2008): Experiment

Six novel words obeying phonotactic probabilities of English:
bosa, gasser, manu, colat, kaki, regli

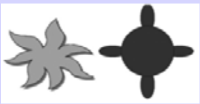
Six brightly colored shapes (sadly greyscale in the paper)



Smith & Yu (2008): Experiment

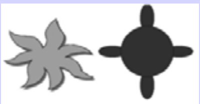
Training: 30 slides with 2 objects named with two words (total time: 4 min)

manu
colat



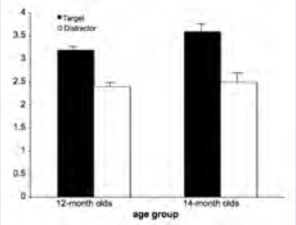
Testing: 12 trials with one word repeated 4 times and 2 objects (correct one and distracter) present

manu
manu
manu
manu



Smith & Yu (2008): Experiment

Results: Infants preferentially look at target over distracter, and 14-month-olds looked longer than 12-month-olds.



Smith & Yu (2008)

Interesting point: More ambiguity within trials may lead to better learning overall

"Yu and Smith (2007; Yu et al., 2007), using a task much like the infant task used here, showed that adults actually learned more word-referent pairs when the set contained 18 words and referents than when it contained only 9. This is because more words and referents mean better evidence against spurious correlations. Although much remains to be discovered about the relevant mechanisms, they clearly should help children learn from the regularities that accrue across the many ambiguous word-scene pairings that occur in everyday communication."

Smith & Yu (2008)

This kind of statistical learning vs. transitional probability learning

"The statistical regularities to which infants must attend to learn word-referent pairings are different from those underlying the segmentation of a sequential stream in that word-referent pairings require computing co-occurrence frequencies across two streams of events (words and referents) simultaneously for many words and referents. Nonetheless, the present findings, like the earlier ones showing statistical learning of sequential probabilities, suggest that solutions to fundamental problems in learning language may be found by studying the statistical patterns in the learning environment and the statistical learning mechanisms in the learner (Newport & Aslin, 2004; Saffran et al., 1996)"

Also, Ramscar et al. (2011)

Kids vs. adults: word-meaning mapping in cases of ambiguity

"These findings…are consistent with other cross-situational approaches to word learning (Yu & Smith, 2007; Smith & Yu, 2008), which have established that in word learning tasks, both children and adults can "rapidly learn multiple word-referent pairs by accruing statistical evidence across multiple and individually ambiguous word-scene pairings"…. However, in this experiment, we explicitly tested for children's sensitivity to the *information* provided by cues, rather than their co-occurrence rates…pattern of children's responses indicates that they can and do use informativity in learning to use words…what a child learns about any given word is dependent on the information it provides about the environment, in relation to other words…it is quite clear that the adults we tested did not place the same value on informativity in their learning that the children did…"

However…

See Medina, Snedecker, Trueswell, & Gleitman (2011) for evidence against learners having multiple meaning hypotheses and cross-tabulating them via statistical procedures. (One issue - the sheer number of items in real world situations, and the different perceptual instances of the items in question.)

Instead, learners "appear to use a one-trial 'fast-mapping' procedure, even under conditions of referential uncertainty."

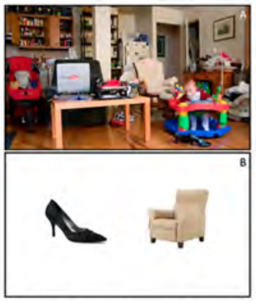


Fig. 1. (A) A plausible word learning environment for the word shoe. (B) The simulated word-learning environment for shoe found in most cross-situational word-learning experiments.

Frank, Goodman, & Tenenbaum (2009)

Redefining the problem: (It's harder)

Not just about learning stable lexicon of word-meaning mappings, but also about the intention of the speaker at the moment.

"Social theories suggest that learners rely on a rich understanding of the goals and intentions of speakers…once the child understands what is being talked about, the mappings between words and referents are relatively easy to learn (St. Augustine, 397/1963; Baldwin, 1993; Bloom, 2002; Tomasello, 2003). These theories must assume some mechanism for making mappings, but this mechanism is often taken to be deterministic, and its details are rarely specified. In contrast, cross-situational accounts of word learning take advantage of the fact that words often refer to the immediate environment of the speaker, which allows learners to build a lexicon based on consistent associations between words and their referents (Locke, 1690/1964; Siskind, 1996; Smith, 2000; Yu & Smith, 2007)."

[How different are these accounts, really?]

Frank, Goodman, & Tenenbaum (2009)

Problems for learning based on cross-situational idea that referents are present:

"…speakers often talk about objects that are not visible and about actions that are not in progress at the moment of speech (Gleitman, 1990), adding noise to the correlations between words and objects."

Solution: appeal to external social/communication cues

"…cross-situational and associative theories often appeal to external social cues, such as eye gaze (Smith, 2000; Yu & Ballard, 2007), but these are used as markers of salience (the ''warm glow'' of attention), rather than as evidence about internal states of the speaker, as in social theories."

Frank, Goodman, & Tenenbaum (2009)

Task: Identify lexicon items for object nouns

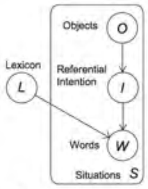
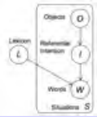


Fig. 1. Illustration of the dependence relations in our model. O, I, and W represent the objects present in the context, the objects that the speaker intends to refer to, and the words that the speaker utters, respectively. These variables are related within each situation s. The words that the speaker utters are additionally determined by the lexicon of the speaker's language, L, which does not change from situation to situation (and hence lies outside the representation of the set of situations).

Frank, Goodman, & Tenenbaum (2009)



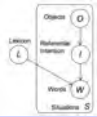
Assumption:

What people intend to say (I) is a function of the world around them (specifically, the objects O present).

Assumption:

The words people say (W) are a function of what people intend to say (I = objects intended) and how those intentions can be translated with the language they speak (using lexicon items L)

Model



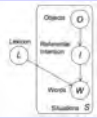
$$P(L|C) \propto P(C|L)P(L)$$

Model learns a probability distribution over unobserved lexicons L (one L = set of lexicon items), given an observed corpus C of situations.

Prior P(L) favors parsimony (fewer lexical items): exponentially penalized for each additional lexical item, using constant α

$P(L) \propto e^{-\alpha|L|}$

Model



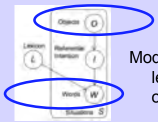
$$P(L|C) \propto P(C|L)P(L)$$

Model learns a probability distribution over unobserved lexicons L (one L = set of lexicon items), given an observed corpus C of situations.

Likelihood P(C|L) is product of the words, objects, and intentions given the lexicon L for all situations in C:

$$P(C|L) = \prod_{s \in C} P(W_s, O_s, I_s|L)$$

Model



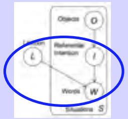
$$P(L|C) \propto P(C|L)P(L)$$

Model learns a probability distribution over unobserved lexicons L (one L = set of lexicon items), given an observed corpus C of situations.

W & O are conditionally independent, so P(W_s, O_s, I_s | L) can be rewritten…

$$P(C|L) = \prod_{s \in C} P(W_s, O_s, I_s|L)$$

Model



$$P(L|C) \propto P(C|L)P(L)$$

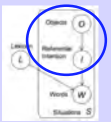
Model learns a probability distribution over unobserved lexicons L (one L = set of lexicon items), given an observed corpus C of situations.

…as the product of the words given the speaker's intended objects and lexicon (P(W_s | I_s, L)…

$$P(C|L) = \prod_{s \in C} P(W_s, O_s, I_s | L)$$

P(W_s | I_s, L) *…

Model



$$P(L|C) \propto P(C|L)P(L)$$

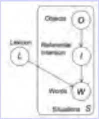
Model learns a probability distribution over unobserved lexicons L (one L = set of lexicon items), given an observed corpus C of situations.

…times the probability of the speaker's intended objects (I) given the objects present (P(I_s | O_s).

$$P(C|L) = \prod_{s \in C} P(W_s, O_s, I_s | L)$$

P(W_s | I_s, L) * P(I_s | O_s)

Model



$$P(L|C) \propto P(C|L)P(L)$$

Model learns a probability distribution over unobserved lexicons L (one L = set of lexicon items), given an observed corpus C of situations.

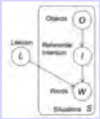
Since we can't observe speaker's intended referent directly, we sum over all possible values of intended referent *I*, assuming the object is present ($I \in O_s$).

$$P(C|L) = \prod_{s \in C} P(W_s, O_s, I_s | L)$$

$\Sigma_{I \subseteq O}$ P(W_s | I_s, L) * P(I_s | O_s)

Note that I_s can be empty if speaker is not referring to an object that is present.

Model



$$P(L|C) \propto P(C|L)P(L)$$

Model learns a probability distribution over unobserved lexicons L (one L = set of lexicon items), given an observed corpus C of situations.

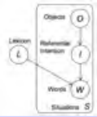
$$P(C|L) = \prod_{s \in C} P(W_s, O_s, I_s | L)$$

$$P(C|L) = \prod_{s \in C} \sum_{I_s \subseteq O_s} P(W_s | I_s, L) \cdot P(I_s | O_s)$$

Simplicity assumption: P(I_s | O_s) ∝ 1
(all intentions equally likely)

Remaining term: P(W_s | I_s,L)

Model

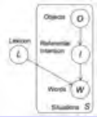


$$P(L|C) \propto P(C|L)P(L)$$

Model learns a probability distribution over unobserved lexicons L (one L = set of lexicon items), given an observed corpus C of situations.

$$P(W_s|I_s,L) = \prod_{w\in W_s}\left[\gamma \cdot \sum_{o\in I_s}\frac{1}{|I_s|}P_R(w|o,L) + (1-\gamma)\cdot P_{NR}(w|L)\right]$$

Model



$$P(L|C) \propto P(C|L)P(L)$$

Model learns a probability distribution over unobserved lexicons L (one L = set of lexicon items), given an observed corpus C of situations.

$$P(W_s|I_s,L) = \prod_{w\in W_s}\left[\gamma \cdot \sum_{o\in I_s}\frac{1}{|I_s|}P_R(w|o,L) + (1-\gamma)\cdot P_{NR}(w|L)\right]$$

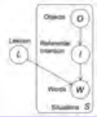
Assumption: words are generated as a bag of words (no order or dependencies, so can multiply them together)

Assumption: words are generated because

(1) they are referential to some item present [P_R]

(2) they are non-referential [P_{NR}]

Model



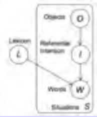
$$P(L|C) \propto P(C|L)P(L)$$

Model learns a probability distribution over unobserved lexicons L (one L = set of lexicon items), given an observed corpus C of situations.

$$P(W_s|I_s,L) = \prod_{w\in W_s}\left[\gamma \cdot \sum_{o\in I_s}\frac{1}{|I_s|}P_R(w|o,L) + (1-\gamma)\cdot P_{NR}(w|L)\right]$$

Υ = probability a word is used referentially, given context
(1 – Υ) = probability word is not used referentially (specifically, not referring to objects: function words, adjectives, verbs)

Model



$$P(L|C) \propto P(C|L)P(L)$$

Model learns a probability distribution over unobserved lexicons L (one L = set of lexicon items), given an observed corpus C of situations.

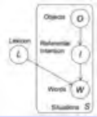
$$P(W_s|I_s,L) = \prod_{w\in W_s}\left[\gamma \cdot \sum_{o\in I_s}\frac{1}{|I_s|}P_R(w|o,L) + (1-\gamma)\cdot P_{NR}(w|L)\right]$$

P_R(w|o, L) = probability of word used referentially for an object = probability of word being chosen, given the object and the lexicon

Uniform over words linked to object in the lexicon. If a word is not linked to an object, its referential probability is 0 for that object.

Averaged over all possible intended referents (I_s).

Model



$$P(L|C) \propto P(C|L)P(L)$$

Model learns a probability distribution over unobserved lexicons L (one L = set of lexicon items), given an observed corpus C of situations.

$$P(W_s|I_s, L) = \prod_{w \in W_s} \left[\gamma \cdot \sum_{o \in I_s} \frac{1}{|I_s|} P_R(w|o, L) + (1-\gamma) \boxed{P_{NR}(w|L)} \right]$$

$P_{NR}(w|L)$ = probability of word used non-referentially w.r.t objects = probability of word being chosen, given lexicon.

If word not in lexicon already, probability of choosing word ∝ 1.
If word in lexicon already, probability of choosing word ∝ κ.

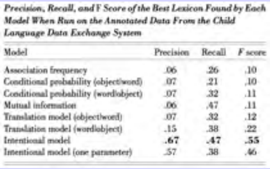
When κ < 1, words in lexicon less likely to be uttered non-referentially than words not in lexicon.

Testing the Model: Corpus Evaluation

Input Corpus: Rollins videos of parents interacting with preverbal infants
Annotated with all mid-size objects judged to be visible to the infant.

Other word-learning models evaluated on same data, and all models judged on the accuracy of the lexicons learned and inferences on speaker intentions

Lexicons: Each model produced association probability between word & object. Chose lexicon that maximized F-score (harmonic mean of precision & recall).



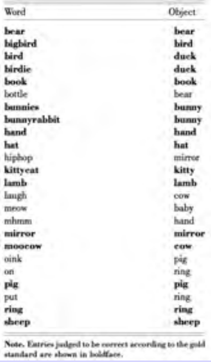
Precision, Recall, and F Score of the Best Lexicon Found by Each Model When Run on the Annotated Data From the Child Language Data Exchange System

Model	Precision	Recall	*F* score
Association frequency	.06	.26	.10
Conditional probability (object|word)	.07	.21	.10
Conditional probability (word|object)	.07	.32	.11
Mutual information	.06	.47	.11
Translation model (object|word)	.07	.32	.12
Translation model (word|object)	.15	.38	.22
Intentional model	**.67**	**.47**	**.55**
Intentional model (one parameter)	.57	.38	.46

Note: Intentional model with “one parameter” is when α is the only free parameter.

Testing the Model: Corpus Evaluation

Best lexicon found by intentional model



Word	Object
bear	**bear**
bigbird	**bird**
bird	**duck**
birdie	**duck**
book	**book**
bottle	bear
bunnies	**bunny**
bunnyrabbit	**bunny**
hand	**hand**
hat	**hat**
hiphop	mirror
kittycat	**kitty**
lamb	**lamb**
laugh	cow
meow	baby
mhmm	hand
mirror	**mirror**
moocow	**cow**
oink	pig
on	ring
pig	**pig**
put	ring
ring	**ring**
sheep	**sheep**

Note. Entries judged to be correct according to the gold standard are shown in boldface.

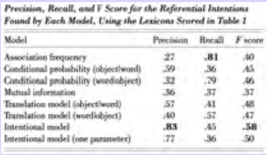
Testing the Model: Corpus Evaluation

Input Corpus: Rollins videos of parents interacting with preverbal infants
Annotated with all mid-size objects judged to be visible to the infant.

Other word-learning models evaluated on same data, and all models judged on the accuracy of the lexicons learned and inferences on speaker intentions

Speaker Intentions:
Intentional model = intention with highest posterior probability given lexicon

Other models = objects for which matching words in best lexicon had been uttered



Precision, Recall, and F Score for the Referential Intentions Found by Each Model, Using the Lexicons Scored in Table 1

Model	Precision	Recall	*F* score
Association frequency	.27	**.81**	.40
Conditional probability (object|word)	.59	.36	.45
Conditional probability (word|object)	.32	.79	.46
Mutual information	.36	.37	.37
Translation model (object|word)	.57	.41	.48
Translation model (word|object)	.40	.57	.47
Intentional model	**.83**	.45	**.58**
Intentional model (one parameter)	.77	.36	.50

Note: Intentional model with “one parameter” is when α is the only free parameter.

Testing the Model: Corpus Evaluation

Why did the intentional model work so well?

“The high precision of the lexicon found by our model was likely due to two factors. First, the distinction between referential and nonreferential words allowed our model to exclude from the lexicon words that were used without a consistent referent. Second, the ability of the model to infer an empty intention allowed it to discount utterances that did not contain references to any object in the immediate context.”

$$P(W_s|I_s, L) = \prod_{w \in W_s} \left[\gamma \cdot \sum_{o \in I_s} \frac{1}{|I_s|} P_R(w|o, L) + (1-\gamma) \cdot P_{NR}(w|L) \right]$$

Using the model to explain experimental results

Cross-situational word-learning (Yu & Smith 2007, Smith & Yu 2008)

All models (even the non-intentional ones) successfully learned the word-meaning mappings, given those experimental stimuli.

Doesn’t help to differentiate – just shows that all these models can use statistical information like this.

Using the model to explain experimental results

Mutual Exclusivity

“Can you give me the dax?” (“bird” = BIRD already known)

Children give novel object, presumably assuming bird can’t also be called “dax”.

Intentional model has soft preference for one-to-one mappings already, since having multiple words for object reduces consistency of word use with that object.

(Though note that some of the other comparison models can also show this behavior, such as the conditional probability models.)

Using the model to explain experimental results

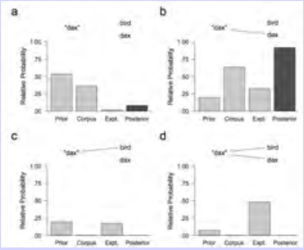
Mutual Exclusivity

“Can you give me the dax?” (“bird” = BIRD already known)

Children give novel object, presumably assuming bird can’t also be called “dax”.

Intentional model scoring for four potential word-referent mappings. Mapping to novel object is the best.

Note also that this is a case of one-trial learning (Carey 1978, Markson & Bloom 1997).



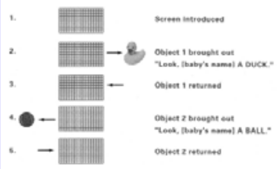
Using the model to explain experimental results

Object Individuation

Xu 2002: Infants use words to individuate objects

Habituation: toys coming out from behind screens

(figure shows two-word habituation, where words are “duck” and “ball” - alternative is one-word habituation, where both objects would be labeled “toy”)



Using the model to explain experimental results

Object Individuation

Xu 2002: Infants use words to individuate objects

Habituation:
“Look, a duck!” “Look, a ball!”

Infant reaction:
Infants didn’t look as long.
(not surprised)

vs.

Habituation:
“Look, a toy!” “Look, a toy!”

Infant reaction:
Infants looked longer.
(surprised to see two objects)

Test: screen removed to reveal…

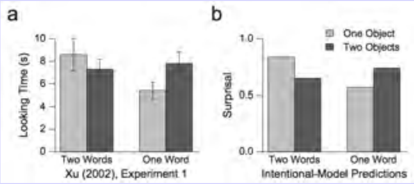
Using the model to explain experimental results

Object Individuation

Xu 2002: Infants use words to individuate objects

Interpretation: Infants expect words to be used referentially. One object = one label, two objects = two labels.

Intentional model: Simulate looking time with surprisal (negative log probability) and get equivalent results.



Using the model to explain experimental results

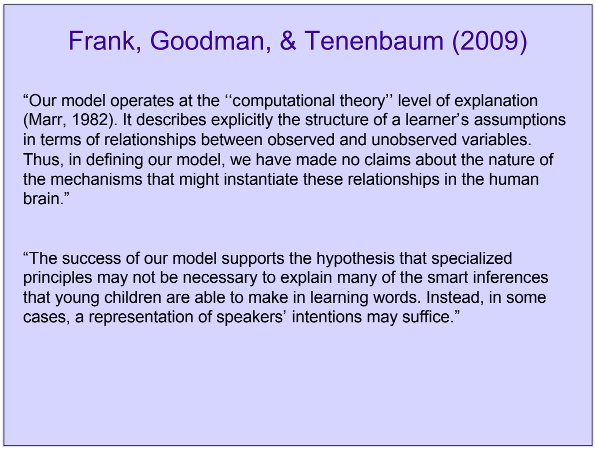
Intention Reading

Baldwin 1993: Children sensitive to intentional labeling, not just timing of labeling.

Children told the name of a toy that was unseen and given a second toy to play with. Children learned to label the first toy with the name.

Easy to simulate in intentional model: Instead of intended objects being unknown, intended objects are known.

Note: Perceptual salience models cannot capture this.



Frank, Goodman, & Tenenbaum (2009)

“Our model operates at the ‘‘computational theory’’ level of explanation (Marr, 1982). It describes explicitly the structure of a learner’s assumptions in terms of relationships between observed and unobserved variables. Thus, in defining our model, we have made no claims about the nature of the mechanisms that might instantiate these relationships in the human brain.”

“The success of our model supports the hypothesis that specialized principles may not be necessary to explain many of the smart inferences that young children are able to make in learning words. Instead, in some cases, a representation of speakers’ intentions may suffice.”