Computational Cognitive Science

Lecture 12: Language Acquisition and Word Learning

Stella Frank
stella.frank@ed.ac.uk

24 October, 2019

Language as a human cognitive ability

Unique: Only humans have such a complex signalling system

Universal: *All* (young) humans have the ability to learn to speak *any* language.

Neccessary? Children will create languages when not given one (e.g. Nicaraguan Sign Language).

Language as implicit knowledge

Native speakers have deep and fluent linguistic abilities —

but describing their implicit linguistic knowledge is difficult.

Linguists & cognitive scientists seek to understand these abilities; modelling can “reverse-engineer” abilities.

We are still far from having a settled theory of language:
Many open questions about what kinds of linguistic representations, processes are involved in language use.

Language as a social tool

Language is used for *communication*.

So, in order to be useful, you have to have the same language as the people around you: language evolves in a social setting.

Language is constrained by learning abilities of young humans.

- All existing languages are by definition learnable.
- Innate learning abilities need to extend to all langauges.

Language Acquisition

Every (typically developing) child will learn to speak (at least) one language

- With very little instruction or feedback
- Regardless of explicit schooling (unlike mathematics or reading/writing)

Language Acquisition

Every (typically developing) child will learn to speak (at least) one language

- With very little instruction or feedback
- Regardless of explicit schooling (unlike mathematics or reading/writing)

Language acquisition involves *unsupervised* learning
— even if some children receive some supervision on some sub-tasks, there's no single supervision signal that is either necessary or sufficient.

Bayesian models are good for unsupervised learning!

Language Acquisition

What are these childen learning?





Language Acquisition

Language acquisition involves tracking the language in the environment and inferring hypotheses about:

Phonetics/phonology: The inventory of sounds in the language

Lexicon: What are words? What are the separable segments of the speech signal? What meaning do they carry?

Syntax: How are words sequenced into larger utterances?

Semantics: How is the meaning of a sequence derived from its component parts?

Pragmantics: How does meaning depend on context?

Computational cognitive modelling of language: The next six lectures

1. Word Learning:
    - How do we learn to associate words with things in the world?
    - A Bayesian model (M. Frank et al., 2009)

Computational cognitive modelling of language: The next six lectures

1. Word Learning:
    - How do we learn to associate words with things in the world?
    - A Bayesian model (M. Frank et al., 2009)
2. Higher-order learning and structure learning:
    - In category learning, how do we know which features to attend to?
    - How could we learn that sentence structure is hierarchical?
    - Hierarchical Bayesian models

Computational cognitive modelling of language: The next six lectures

1. Word Learning:
    - How do we learn to associate words with things in the world?
    - A Bayesian model (M. Frank et al., 2009)
2. Higher-order learning and structure learning:
    - In category learning, how do we know which features to attend to?
    - How could we learn that sentence structure is hierarchical?
    - Hierarchical Bayesian models
3. How do we represent words in our mental lexicon?
    - Correlating semantic vector embeddings and word processing

Readings

Michael C. Frank, N. Goodman, and J. Tenenbaum (2009). Using Speakers Referential Intentions to Model Early Cross-Situational Word Learning. Psychological Science

The assignment involves this paper, so read it carefully!

Amy Perfors, J. Tenenbaum, T. Griffiths, F. Xu (2011). A tutorial introduction to Bayesian models of cognitive development. Cognition.

A good introduction to the Bayesian modelling framework applied to langauge acquisition and other aspects of cognitive development.

Today: Word Learning, Part 1

Task: Match a word with its "meaning"

Meaning simplified to concrete, *grounded*, objects only:
"cat", "spoon", not "also" or "democracy"

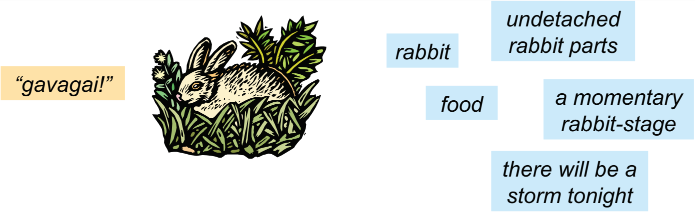
Take as given:

- Word segmentation: "The purring cat", not "thepurr ingcat"
- Phonology: Words are recognised correctly
- Concepts: Words refer to whole, basic-level, objects

Referential Ambiguity: "Gavangai!" (Quine, 1960)

Philosophically, discovering a word's meaning is hard/impossible, especially if you never explicitly define it:

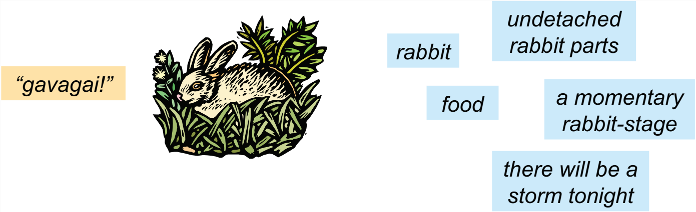
"Look, a rabbit!" vs. `http://en.wikipedia.org/wiki/Rabbit`



Referential Ambiguity: "Gavangai!" (Quine, 1960)

Philosophically, discovering a word's meaning is hard/impossible, especially if you never explicitly define it:

"Look, a rabbit!" vs. `http://en.wikipedia.org/wiki/Rabbit`



However, children figure it out anyway:

- Names, concrete nouns are learned first (one-word stage)
- Then verbs, adjectives, abstract nouns
- "Grammatical"/function words (prepositions, determiners) are learned together with syntax (two-word stage)

Target Cognitive Abilities

What can *children* do
that a *model of word learning* should also be able to do?

Target Cognitive Abilities

What can *children* do
that a *model of word learning* should also be able to do?

Correctness: Learn a plausible lexicon, from plausible data, within a plausible timeline

Target Cognitive Abilities

What can *children* do
that a *model of word learning* should also be able to do?

Correctness: Learn a plausible lexicon, from plausible data, within a plausible timeline

One-shot learning: If context is familiar, children can learn a word's meaning from a single exposure

Target Cognitive Abilities

What can *children* do
that a *model of word learning* should also be able to do?

Correctness: Learn a plausible lexicon, from plausible data, within a plausible timeline

One-shot learning: If context is familiar, children can learn a word's meaning from a single exposure

Mutual exclusivity: Preferentially link new word to new object, rather than linking a second word to a known object

Target Cognitive Abilities

What can *children* do
that a *model of word learning* should also be able to do?

Correctness: Learn a plausible lexicon, from plausible data, within a plausible timeline

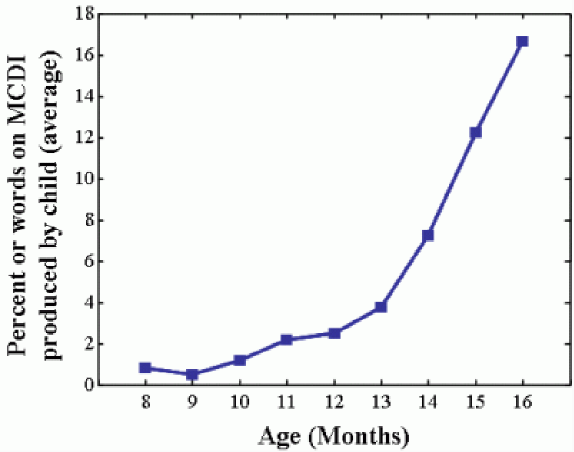
One-shot learning: If context is familiar, children can learn a word's meaning from a single exposure

Mutual exclusivity: Preferentially link new word to new object, rather than linking a second word to a known object

More about all this in the next lecture!

Word Learning: Timeline

Growth curve from the MacArthur-Bates Communicative Development Inventory (MCDI):



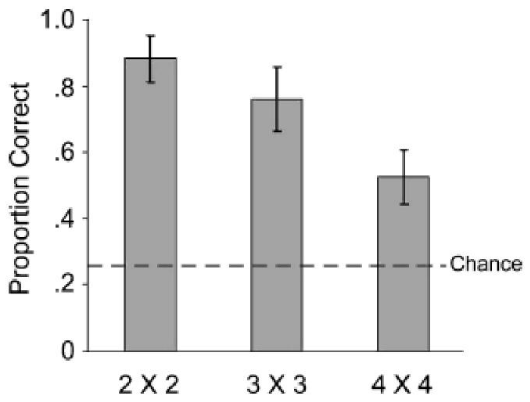
http://www2.psychology.uiowa.edu/faculty/mcmurray/recent/science/

Cross-situational Learning

Track word occurrence statistics over time to resolve referential ambiguity over multiple contexts.

Cross-situational Learning in the Lab (Yu & Smith 2007)

Adults learn successfully, even with high ambiguity.

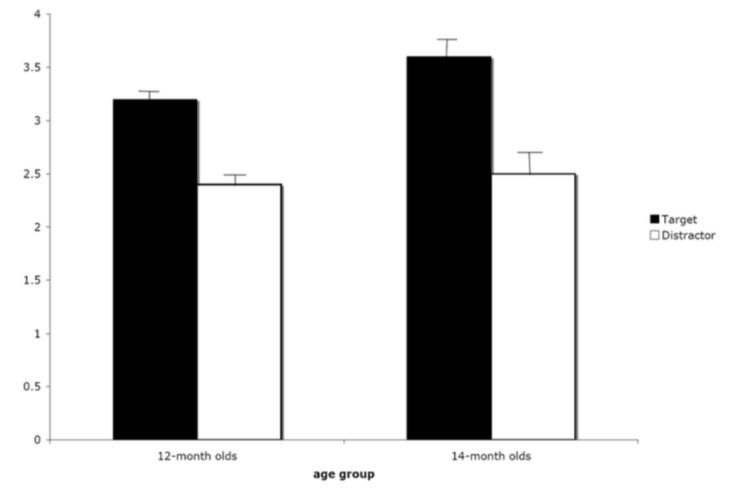


$n \times n$ condition: trial presented n words and n possible referents.

Many participants were "quite sure they had learned nothing from the training and were amazed at their own success."

Cross-situational Learning in the Lab (Smith & Yu 2008)

Infants can learn too: 2 words x 2 objects setting, 6 total pairs.



Baselines: Co-occurrence Statistics

$$P(w|o) = \frac{C(w,o)}{C(o)} \quad \text{or} \quad P(o|w) = \frac{C(w,o)}{C(w)}$$

w: words; o: objects; C: counts

Baselines: Co-occurrence Statistics

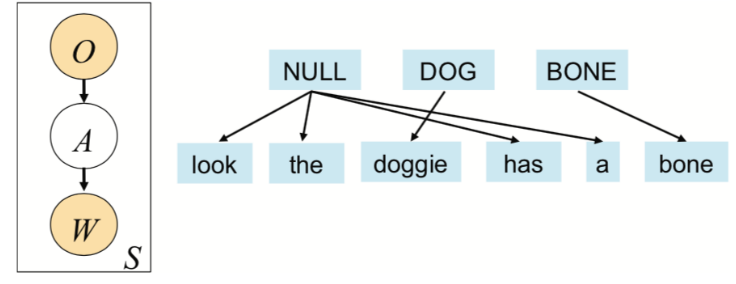
$$P(w|o) = \frac{C(w,o)}{C(o)} \quad \text{or} \quad P(o|w) = \frac{C(w,o)}{C(w)}$$

w: words; o: objects; C: counts

How will these two statistics differ?

Machine Translation Formulation (Yu & Ballard 2007)

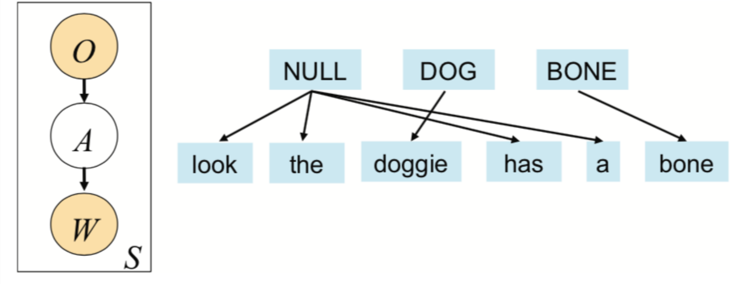
Given a set/sequence of objects ('French'), estimate the probability of a set/sequence of words ('English'):
This is the famous IBM model 1 (Brown et al. 1994).



Machine Translation Formulation (Yu & Ballard 2007)

Generative story: Given objects O in situation S,

- Choose number of words K
- For each k, choose an alignment $a_1 \ldots a_K$ between objects (including NULL object) and words
- For each k, choose a word given object aligned to it

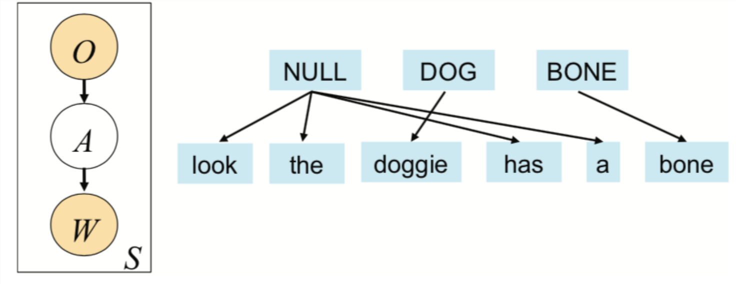


Machine Translation Formulation (Yu & Ballard 2007)

$$P(words|objects) = \prod_{w} \sum_{a} P(w, a|o)$$

This is reversible: could translate "words" into "objects".

What changes if you do this?



Bayesian Formulation

Goal: Infer a lexicon, given corpus data.

Data: Video corpus, annotated with salient objects and transcription of child-directed speech.

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

Lexicon is a set of word-object pairs.

Lexicon only contains the words used referentially: no “NULL” object to map to.

M. Frank et al. (2009): Using Speakers Referential Intentions to Model Early Cross-Situational Word Learning

Lexicon given data:

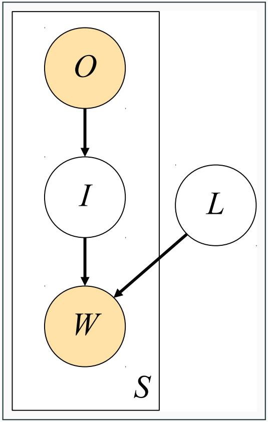
$$P(L|D) \propto P(D|L)P(L)$$

Prior favours smaller lexicons:

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

Likelihood $P(D|L)$ defined using a *generative model*: explains how data is generated from the lexicon.

Generative Model



For each situation (utterance) $s \in D$:

- Objects O are present and observable.
- The speaker chooses a set of intended referents $I \subseteq O$, not visible to the learner.
- The speaker chooses a set of words $W \in L \cup C$
  - Some of these words are used referentially, to refer to referents in I
  - Others are not: these can be words in L or outside (e.g., function words)

Graphical model notation:

- empty/white-background circle: hidden random variable
- shaded/colored circle: observed random variable
- arrow: conditional dependence
- plate: replicated S times.

Generative Model

$$
\begin{aligned}
P(D|L) &= \prod_{s \in D} P(O_s, W_s|L) \\
&= \prod_{s \in D} \sum_{I_s \subseteq O_s} P(O_s, I_s, W_s|L) \\
&= \prod_{s \in D} \sum_{I_s \subseteq O_s} P(O_s)P(I_s|O_s)P(W_s|I_s, L) \\
&\propto \prod_{s \in D} \sum_{I_s \subseteq O_s} P(I_s|O_s)P(W_s|I_s, L)
\end{aligned}
$$

Generative Model

Generate intentions from objects: uniform distribution:

$$P(I_s|O_s) \propto 1$$

Generate words from intentions and lexicon: words are independent. For each word w in W_s:

- choose referential ($p = \gamma$) or non-referential ($p = 1 - \gamma$):

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

- $P_R(w|o, L)$: choose uniformly from lexical items that refer to correct object;
- $P_{NR}(w|L)$: choose uniformly from all words in corpus.

Next Time

- (You: Read the paper!)
- Recap Frank et al. (2009) model description
- Test the models on corpus data
- How well can they capture acquisition phenomena like Mutual Exclusivity?