Joint Acquisition of Word Order and Word Referent in a Memory-Limited and Incremental Learner

Sepideh Sadeghi, Matthias Scheutz

Department of Computer Science, Tufts University

Word Learning In Ambiguous Contexts

Utterance: “Jack is biting the apple.”

Word Learning In Ambiguous Contexts

Utterance: “Jack is biting the apple.”

Word Learning In Ambiguous Contexts

Utterance: “Jack is biting the apple.”

Word Learning In Ambiguous Contexts

Utterance: “Jack is biting the apple.”

Mapping Words to Referents

ball

BAT BALL

ball

BAT BALL

DOG

Quine, W. V. O., *Word and object*, 1960

Mapping Words to Referents

ball

BAT BALL

ball

BAT BALL

DOG

Cross-Situational Word Learning

Quine, W. V. O., *Word and object*, 1960

Mapping Words to Referents

ball

BAT BALL

Incremental

ball

BAT BALL

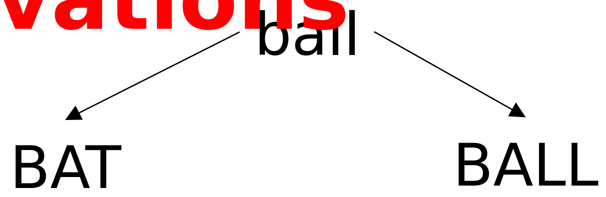
DOG

Cross-Situational Word Learning

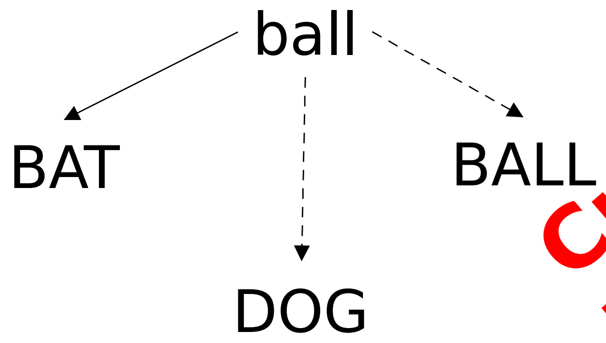
Quine, W. V. O., *Word and object*, 1960

Mapping Words to Referents

With limited memory of past observations



Incremental



Cross-Situational Word Learning

Quine, W. V. O., *Word and object*, 1960

Syntactic bootstrapping

The girl is gorping the boy
vs.
The boy is gorping the girl

Gleitman, L., *The structural sources of verb meanings*, 1990; Fisher et al., *Syntactic bootstrapping*, 2010

Syntactic bootstrapping

The girl is gorping the boy

vs.

The boy is gorping the girl

Gleitman, L., *The structural sources of verb meanings*, 1990; Fisher et al., *Syntactic bootstrapping*, 2010

Objectives

(1) Simulation results from ideal learners suggest that it is possible to jointly acquire word order and meanings and that learning is improved as each language capability bootstraps the other.

(2) A good theory of word learning needs to give clear accounts for hypothesis generation as well as hypothesis evaluation and the information used for these computations, while staying tractable as input size grows.

(3) We study the utility of joint acquisition of simple versions of word order and word meaning in early stages of acquisition in a memory-limited incremental model. We believe that only memory-limited models qualify as scalable models which remain tractable as the amount of data grows.

(4) We allow for the acquired word order information to constrain the acquisition of word’ meanings and vice versa.

Input Representation

Utterance: “Jack is biting the apple”

scene: 

situation=<utterance,scene>

Utterance = W_s= {jack, is, biting, the, apple}

Scene = E_s ={SIT<JACK, CHAIR>,
SIT<SARAH,CHAIR>,
SIT<JACK>,
SIT<SARAH>,
BITE<JACK,APPLE>
PICK<SARAH,APPLE>}

I_s = BITE<JACK,APPLE>

Word Order Representation

Syntactic positions = {w_1,w_2 ,w_3}

Roles = {arg1, arg2, pard}
{agent, patient, action}

$\boldsymbol{\Theta} = \{\theta_{arg1},\theta_{arg2},\theta_{pred}\}$

$\theta_{arg1} = P(.|arg1) = <\pi_{w1|arg1},\pi_{w2|arg1},\pi_{w3|arg1}>$

$\theta_{arg2} = P(.|arg2) = <\pi_{w1|arg2},\pi_{w2|arg2},\pi_{w3|arg2}>$

$\theta_{pred} = P(.|pred) = <\pi_{w1|pred},\pi_{w2|pred},\pi_{w3|pred}>$

English word order used for artificial data generation

$\theta_{arg1} = <1, \; 0, \; 0>$

$\theta_{arg2} = <0, \; 0, \; 1>$

$\theta_{pred} = <0, \; 1, \; 0>$

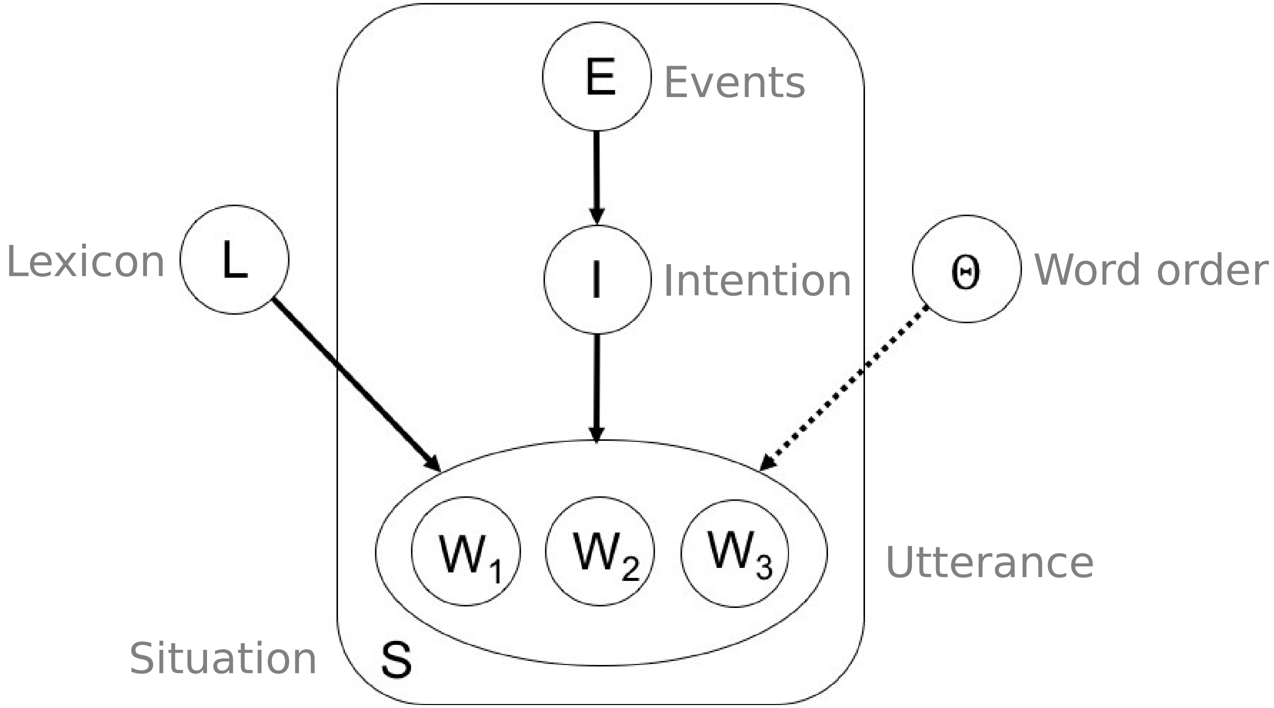
Model Design and Generative Process

E_S = {SIT<JACK, CHAIR>,
SIT<SARAH,CHAIR>,
SIT<JACK>,
SIT<SARAH>,
BITE<JACK,APPLE>
PICK<SARAH,APPLE>
}

I_S = BITE<JACK,APPLE>

Utterance: “Jack is biting the apple”

scene: 



M-WO: The model with Θ

M-B: Baseline model without Θ

Model Design and Generative Process

L = {bite: BITE,
Jack: JACK,
apple: APPLE}

I_S = BITE<JACK,APPLE>
Pred $<arg_1,arg_2>$

$\mathbf{\Theta} = \{\theta_{arg1},\theta_{arg2},\theta_{pred}\}$

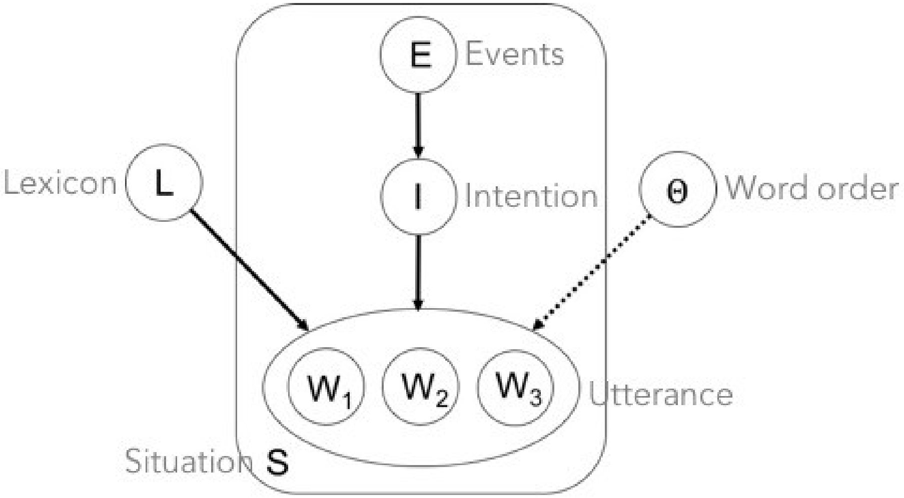
English is SVO :

$\theta_{arg1} = <1, \quad 0, \quad 0>$

$\theta_{arg2} = <0, \quad 0, \quad 1>$

$\theta_{pred} = <0, \quad 1, \quad 0>$

$<w_1, w_2, w_3>$



Utterance: “Jack is biting the apple”

scene:

Model Design and Generative Process

L = {bite: BITE,
Jack: JACK,
apple: APPLE}

$I_S = \text{BITE}<\text{JACK,APPLE}>$

$\text{Pred} <arg_1, arg_2>$

$\Theta = \{\theta_{arg1}, \theta_{arg2}, \theta_{pred}\}$

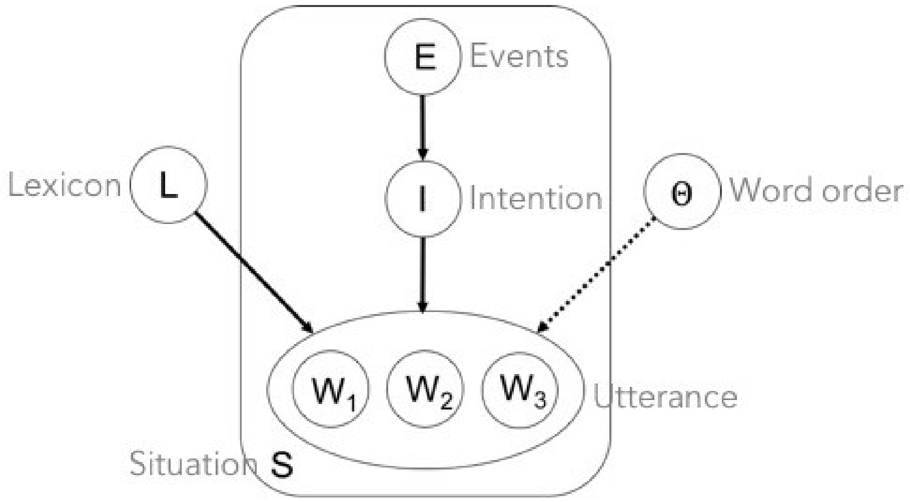
English is SVO :

$\theta_{arg1} = <1, \quad 0, \quad 0>$

$\theta_{arg2} = <0, \quad 0, \quad 1>$

$\theta_{pred} = <0, \quad 1, \quad 0>$

$<w_1, w_2, w_3>$



Utterance: “Jack is biting the apple”

scene: 

Model Design and Generative Process

L = {bite: BITE,
Jack: JACK,
apple: APPLE}

I_S = BITE<JACK,APPLE>
Pred $\langle arg_1, arg_2 \rangle$

$\Theta = \{\theta_{arg1}, \theta_{arg2}, \theta_{pred}\}$

English is SVO :

$\theta_{arg1} = \langle 1, \quad 0, \quad 0 \rangle$

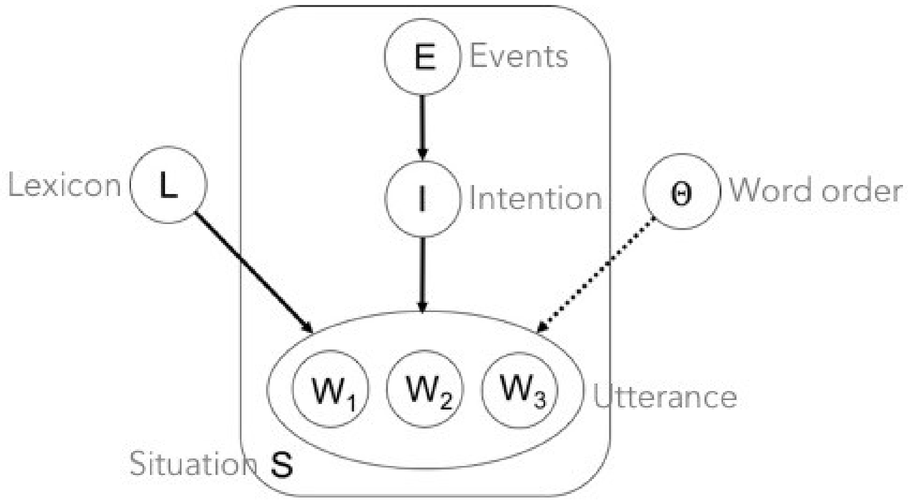
$\theta_{arg2} = \langle 0, \quad 0, \quad 1 \rangle$

$\theta_{pred} = \langle 0, \quad 1, \quad 0 \rangle$

$\langle w_1, w_2, w_3 \rangle$

$P_R(w) = \gamma$

$P_{NR}(w) = 1-\gamma$



Utterance: “Jack is biting the apple”

scene:

Model Design and Generative Process

L = {bite: BITE,
Jack: JACK,
apple: APPLE}

$I_S = \text{BITE}<\text{JACK,APPLE}>$

$\text{Pred} <\text{arg}_1,\text{arg}_2>$

$\Theta = \{\theta_{arg1}, \theta_{arg2}, \theta_{pred}\}$

English is SVO :

$\theta_{arg1} = <1, \; 0, \; 0>$

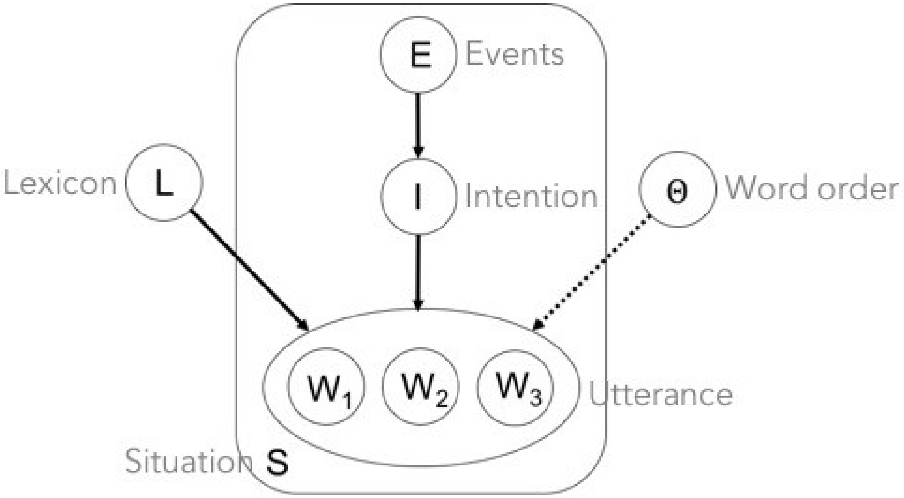
$\theta_{arg2} = <0, \; 0, \; 1>$

$\theta_{pred} = <0, \; 1, \; 0>$

$<w_1, w_2, w_3>$

$\mathbf{P_R(w)} = \gamma$

$\mathbf{P_{NR}(w)} = \mathbf{1}\text{-}\gamma$



Utterance: “Jack is biting the apple”

scene:

Model Design and Generative Process

L = {bite: BITE,
Jack: JACK,
apple: APPLE}

I_S = BITE<JACK,APPLE>
Pred <arg_1,arg_2>

$\Theta = \{\theta_{arg1}, \theta_{arg2}, \theta_{pred}\}$

English is SVO :

$\theta_{arg1} = <1, \ 0, \ 0>$

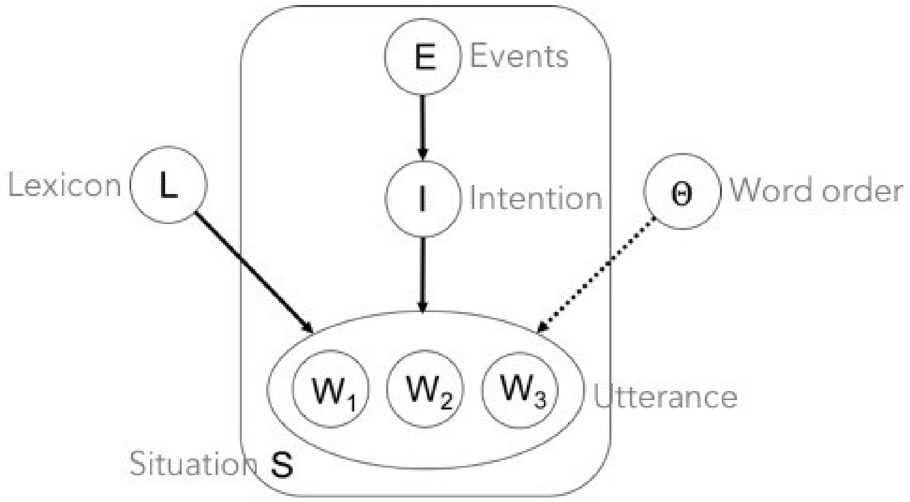
$\theta_{arg2} = <0, \ 0, \ 1>$

$\theta_{pred} = <0, \ 1, \ 0>$

$<w_1, w_2, w_3>$

$P_R(w) = \gamma$

$P_{NR}(w) = 1-\gamma$



Utterance: "Jack is biting the apple"

scene:

Model Design and Generative Process

L = {bite: BITE,
Jack: JACK,
apple: APPLE}

I_S = BITE<JACK,APPLE>
Pred <arg_1,arg_2>

$\Theta = \{\theta_{arg1}, \theta_{arg2}, \theta_{pred}\}$

English is SVO :

$\theta_{arg1} = <1, \quad 0, \quad 0>$

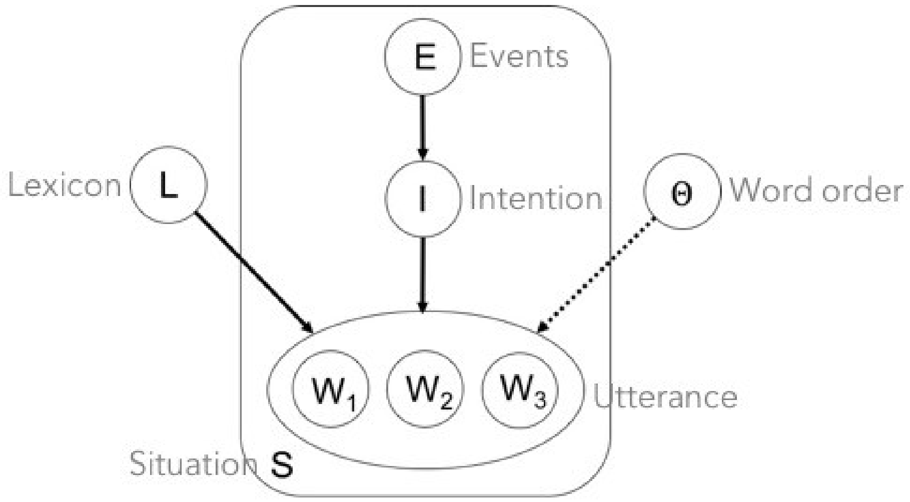
$\theta_{arg2} = <0, \quad 0, \quad 1>$

$\theta_{pred} = <0, \quad 1, \quad 0>$

$<w_1, w_2, w_3>$

$\mathbf{P_R(w) = \gamma}$

$\mathbf{P_{NR}(w) = 1\text{-}\gamma}$



Utterance: “Jack is biting the apple”

scene:

Model Design and Generative Process

L = {bite: BITE,
Jack: JACK,
apple: APPLE}

I_S = BITE<JACK,APPLE>
Pred <arg_1,arg_2>

$\Theta = \{\theta_{arg1}, \theta_{arg2}, \theta_{pred}\}$

English is SVO :

$\theta_{arg1} = <1, \quad 0, \quad 0>$

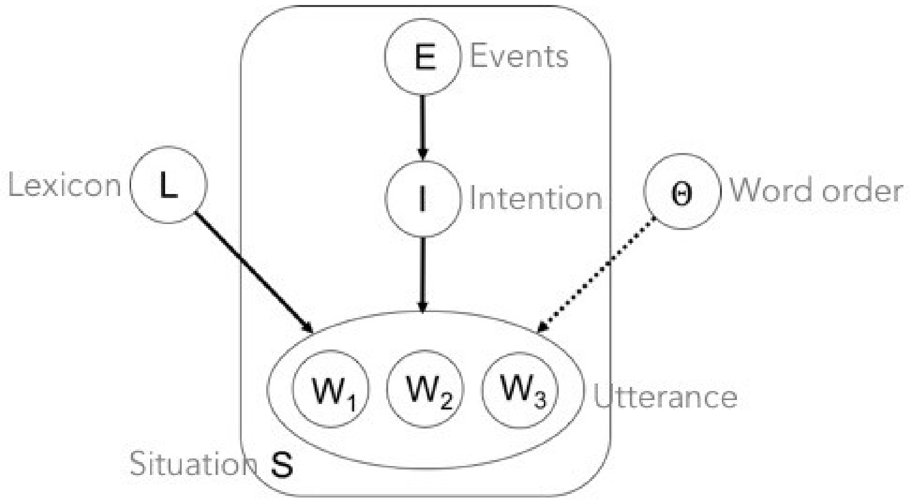
$\theta_{arg2} = <0, \quad 0, \quad 1>$

$\theta_{pred} = <0, \quad 1, \quad 0>$

$<w_1, w_2, w_3>$

$P_R(w) = \gamma$

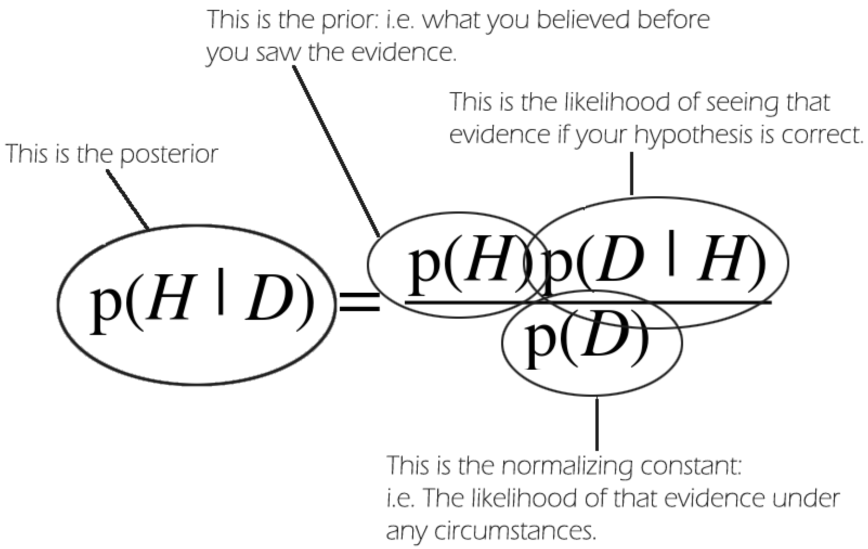
$P_{NR}(w) = 1-\gamma$



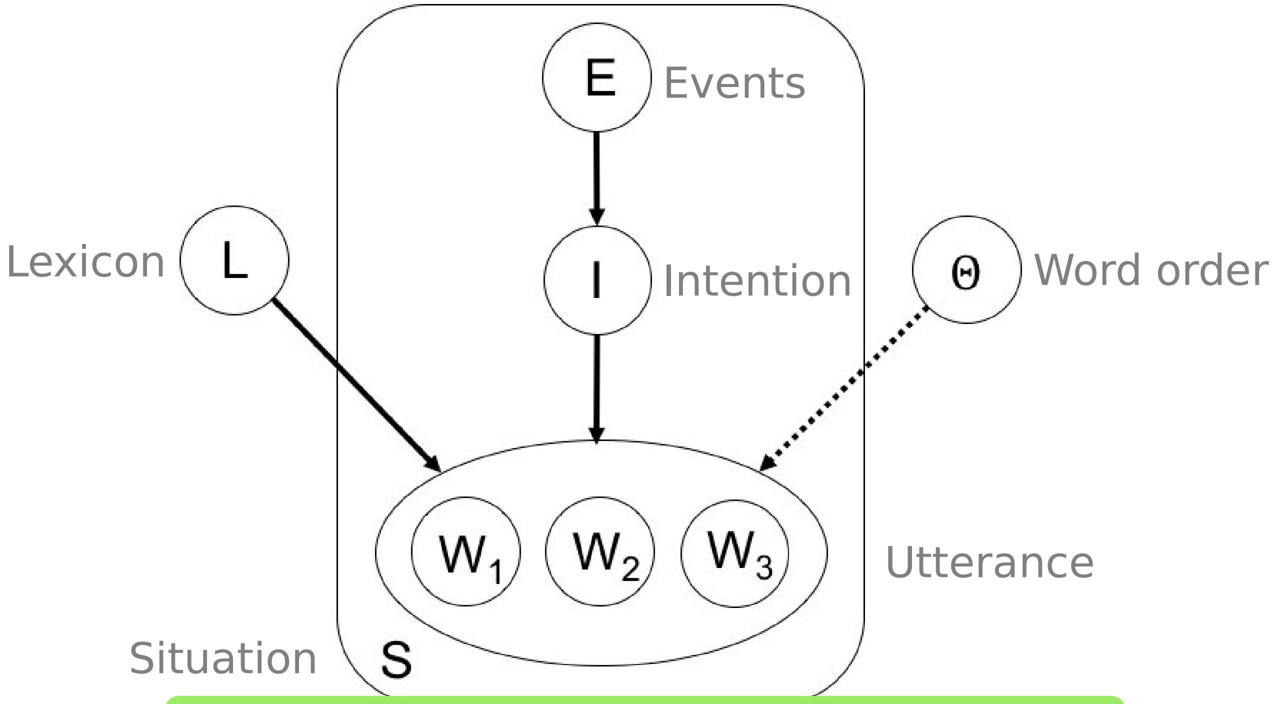
Utterance: "Jack is biting the apple"

scene:

Reversing the Generative Process: Bayesian Inference



Posterior ∝ Likelihood × Prior



M-WO: The model with Θ

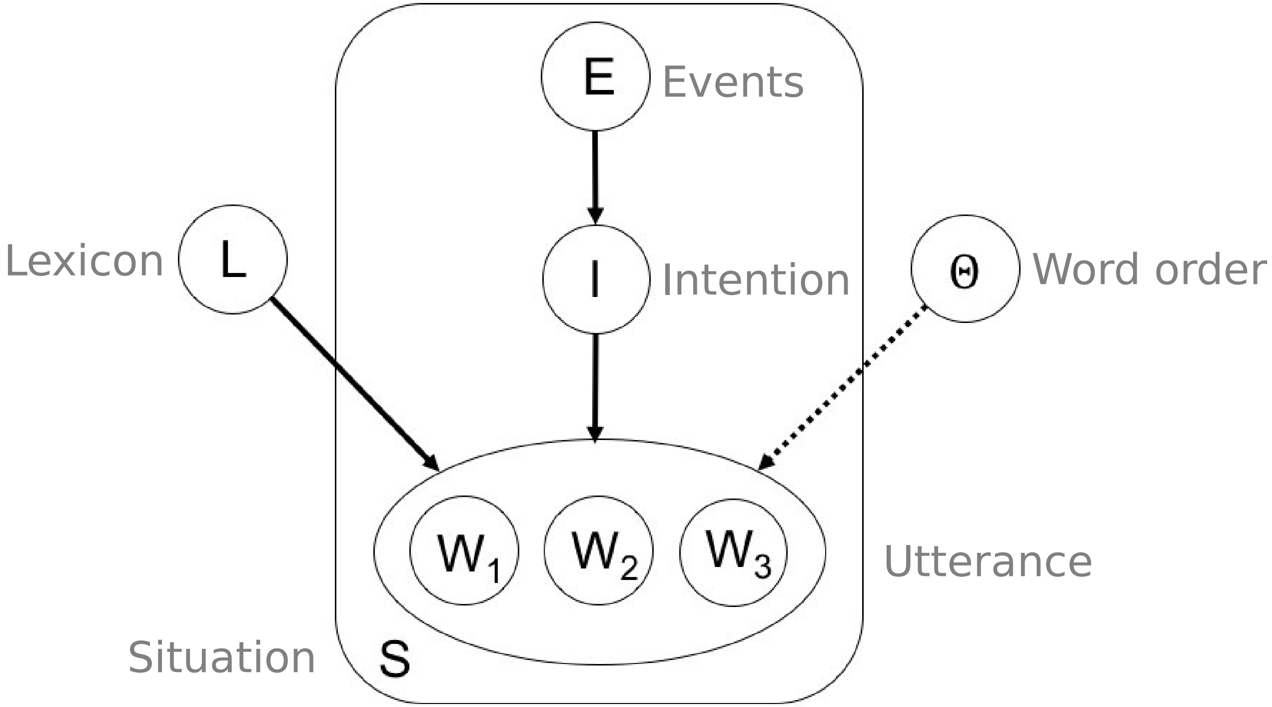
M-B: Baseline model without Θ

Bayesian Inference in M-WO

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

$$P(\Theta) \propto 1$$

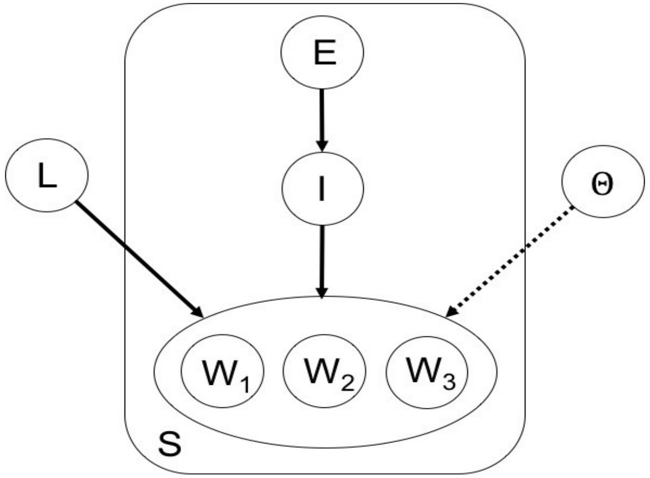
$$P(I_s|E_s) \propto 1$$



$$P(L, \Theta|C) \propto P(C|L, \Theta)P(L)P(\Theta) \quad (1)$$

$$P(C|L, \Theta) = \prod_{s \in C} \sum_{I_s \subseteq E_s} P(W_s|I_s, L, \Theta)P(I_s|E_s) \quad (2)$$

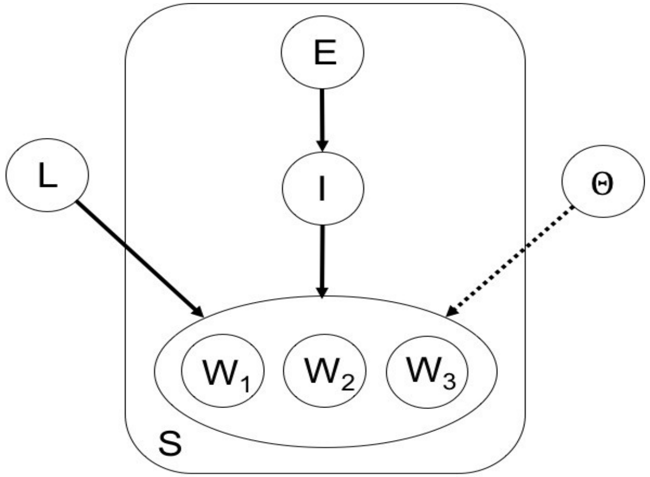
Bayesian Inference in M-WO



$$P(C|L,\Theta) = \prod_{s \in C} \sum_{I_s \subseteq E_s} P(W_s|I_s,L,\Theta)P(I_s|E_s) \qquad (2)$$

$$P(W_s|I_s,L,\Theta) = \prod_{w_j \in W_s} [\gamma \cdot \sum_{x_i \in I_s} \frac{1}{|I_s|} P_R(w_j|x_i,L) \cdot P(pos(w_j)|role(x_i),\Theta) + (1-\gamma)P_{NR}(w_j|L)] \qquad (3)$$

Bayesian Inference in M-B



$$P(C|L) = \prod_{s \in C} \sum_{I_s \subseteq E_s} P(W_s|I_s, L)P(I_s|E_s) \tag{4}$$

$$P(W_s|I_s, L) = \prod_{w \in W_s} [\gamma \cdot \sum_{x \in I_s} \frac{1}{|I_s|} P_R(w|x, L) + (1-\gamma)P_{NR}(w|L)] \tag{5}$$

Incremental and Memory-Limited Learning Algorithm

Model's memory:
The knowledge in its lexicon and current situation.

Incremental and Memory-Limited Learning Algorithm

Incremental Word Learning:
(1) It only sees one situation at a time (no iteration over data).
(2) the model can only use the knowledge in its memory for hypothesis generation and hypothesis evaluation.
(3) The model maintains a single global lexicon (hypothesis) across situations.
(4) The model makes local revisions to the global hypothesis by integrating the inferred mini-lexicon in the global hypothesis.
(5) Bayesian inference is only applied locally in the context of single situations based on context-appropriate word-referent pairs available in the memory (current lexicon and current situation)

Incremental Learning: Updating Lexicon

Inferring the MAP mini-lexicon in each situation:
(1) Generating mini-lexicon proposals (hypothesis generation)
.........Stochastic Search Techniques
(2) Scoring (hypothesis evaluation)
.........Relative posterior probability

Merging the new mini-lexicon with the current lexicon:
(1) Applying mutual exclusivity constraints to produce a preference for one-to-one mappings in the output lexicon.

Incremental Learning: Updating Word order

Using a symmetric Drichlet distribution prior with parameter **α**

$$\pi_{pos|rol} = \frac{Count(rol, pos) + \alpha}{Count(rol) + 3\alpha}$$

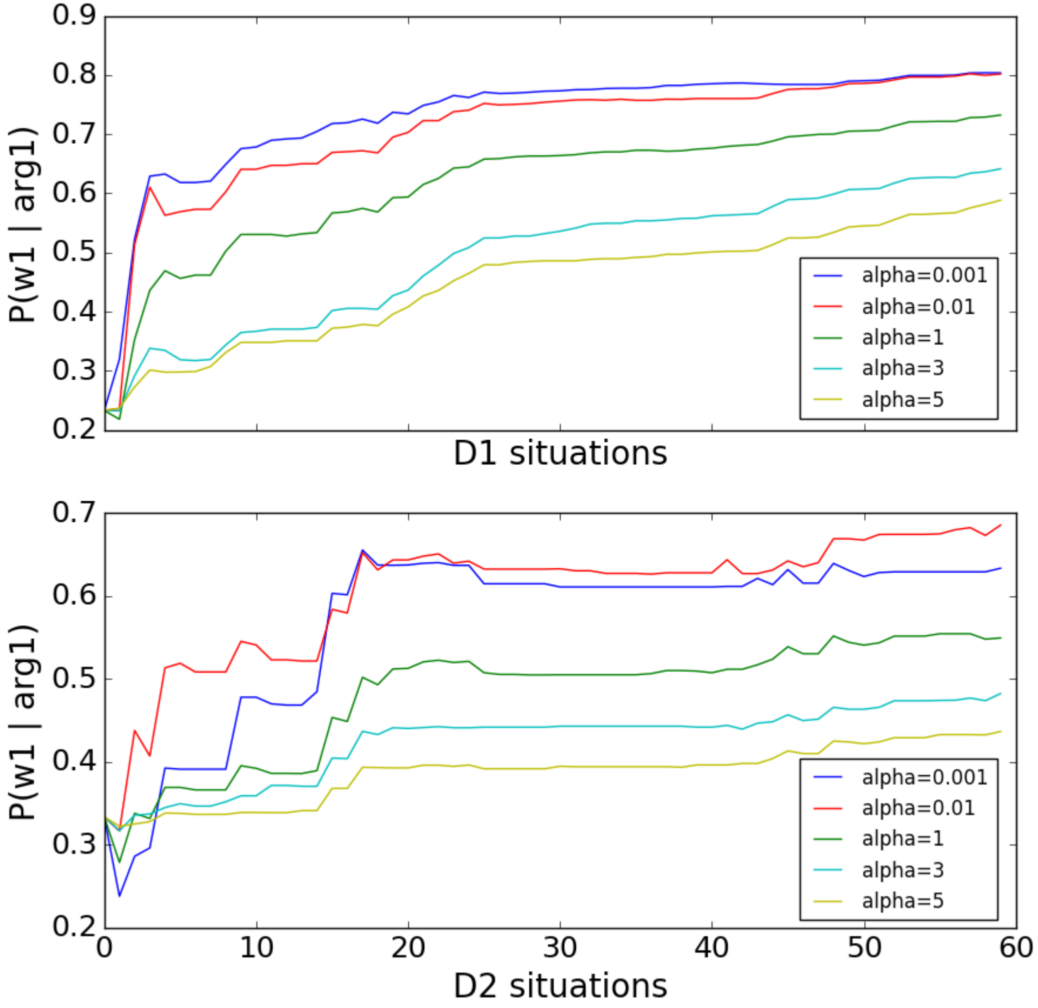
Update Algorithm

Algorithm 1 Algorithm for updating the lexicon incrementally in light of a new situation.

```
procedure UPDATE(prevLex,situation)
    words ← unique(situation.words)
    refs ← unique(situation.refs)
    entities ← union(words, refs)
    links ← initLinks(words, refs)
    prevLinks ← extract-L(prevLex, entities)
    links ← union(links, prevLinks)
    proposals ← init(nInit, links, stats)
    bestLex ← best(proposals, situation)
    prevSits ← extract-S(prevLex, entities)
    situations ← union(situation, prevSits)
    lex1 ← exclude(prevLex, entities)
    lex2 ← mutate(bestLex, links,
        stats, situations)
    lexicon ← merge(lex1, lex2)
end procedure
```

Results: Word Order Learning Curves

Word Order Learning $P(w_1 \mid arg_1)$

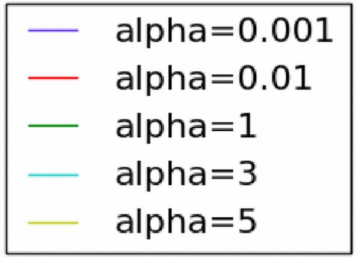


D1

ambiguity

D2

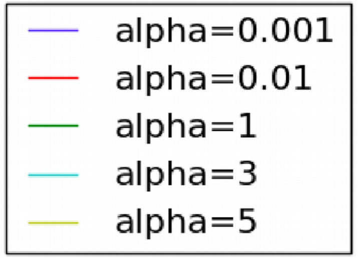
Strong non-sparsity bias for word order distributions θ_i



Strong sparsity bias for word order distributions θ_i

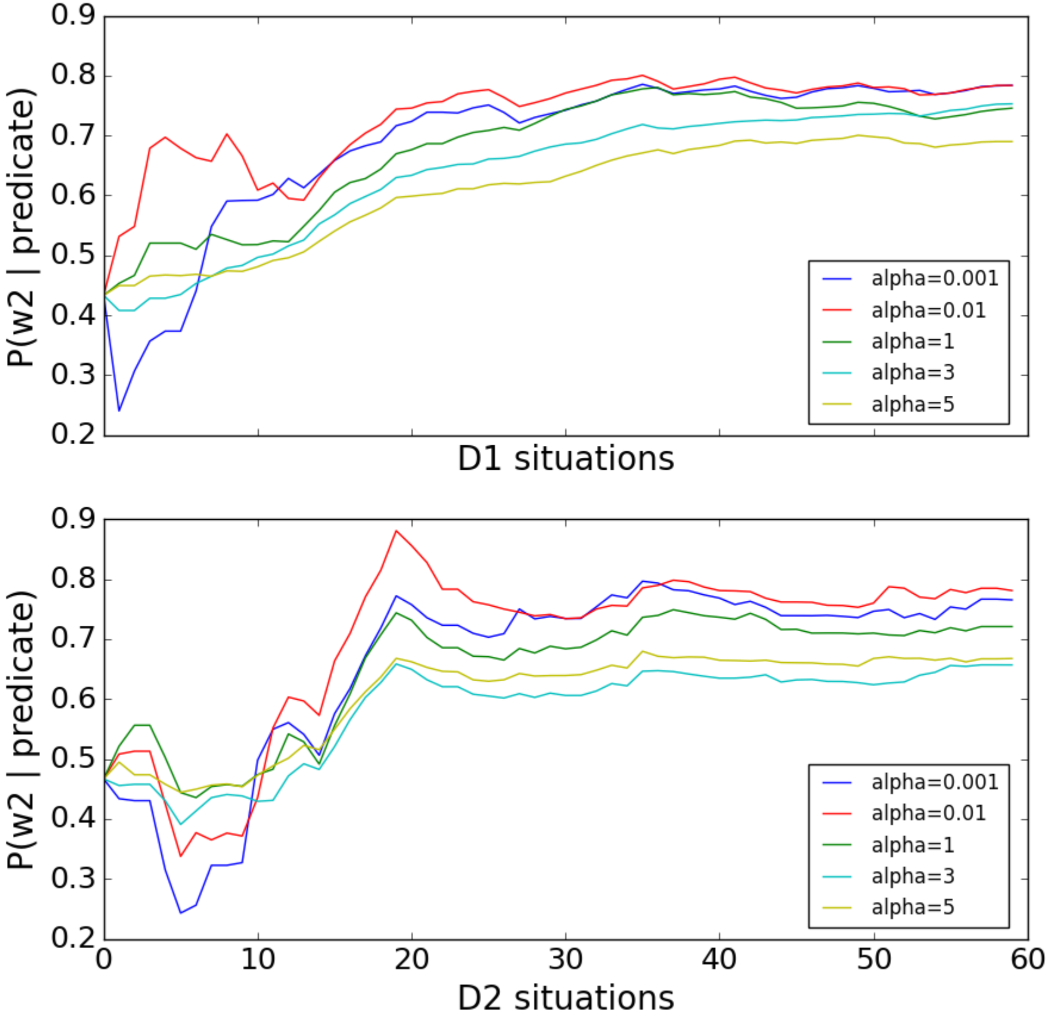
Results: Word Order Learning Curves

Strong non-sparsity bias for word order distributions θ_i



Strong sparsity bias for word order distributions θ_i

Word Order Learning $P(w_2 \mid pred)$



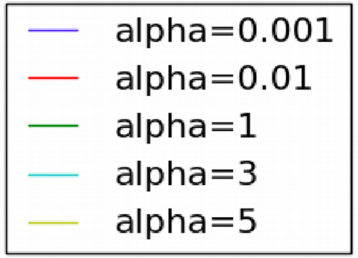
D1

ambiguity

D2

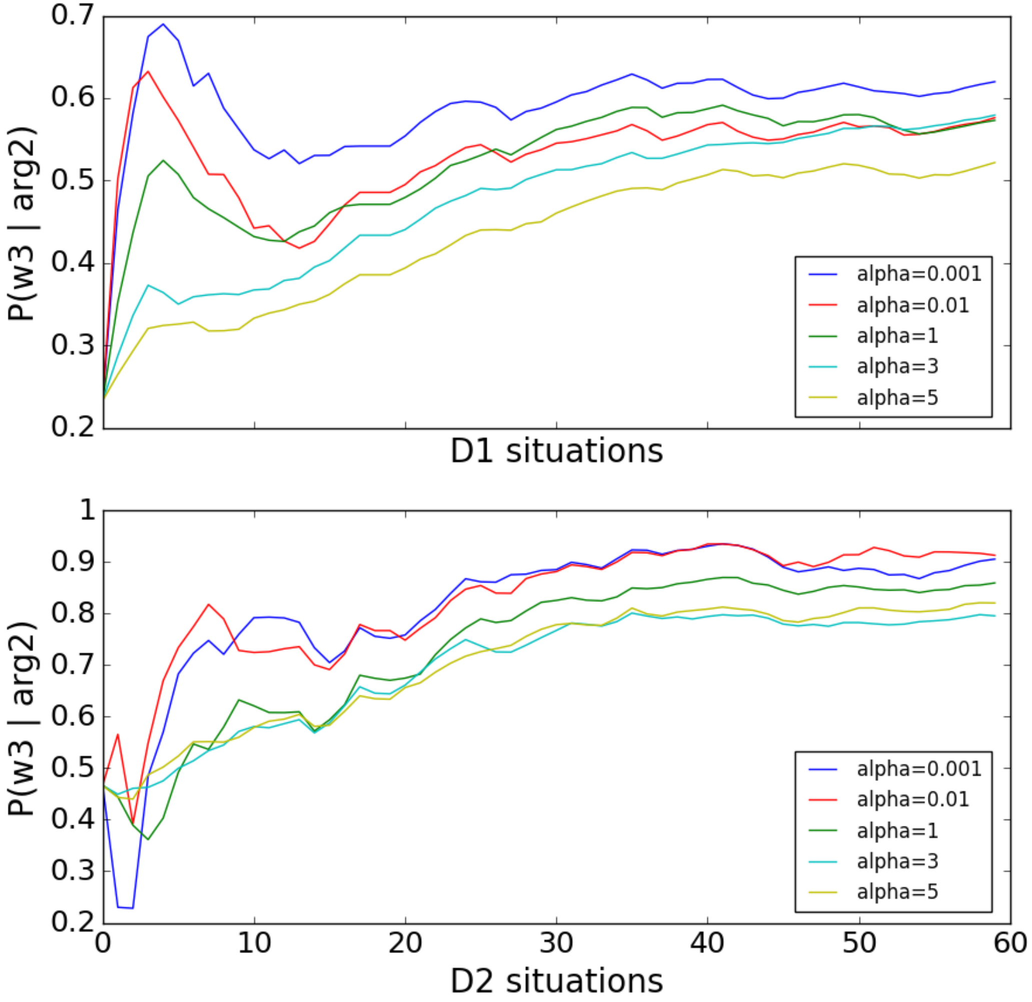
Results: Word Order Learning Curves

Strong non-sparsity bias for word order distributions θ_i



Strong sparsity bias for word order distributions θ_i

Word Order Learning P(w_3 | arg_2)



D1

ambiguity

D2

Results: Word Learning Results

Word Learning

ambiguity
D2
D1

Model	F-Score(D1)	F-Score(D2)
M-WO ($\alpha = 0.001$)	0.718	0.554
M-WO ($\alpha = 0.01$)	0.732	0.548
M-WO ($\alpha = 1$)	0.736	0.568
M-WO ($\alpha = 3$)	0.736	0.543
M-WO ($\alpha = 5$)	0.758	0.576
M-B	0.755	0.522

Conclusion and Discussion

(1) We proposed a memory-limited incremental model of word learning, in order to study the utility of joint acquisition of information in realistic situations under which infant word learning occurs.

(2) Please use the discussion section of the paper to add more elements here

(3) ...

Thank you!

Sepideh Sadeghi
@sepid_s
http://www.eecs.tufts.edu/~ssadeg01/

Matthias Scheutz
@matthiasscheutz
http://hrilab.tufts.edu/people/matthias.php

Tufts UNIVERSITY

hrr lab