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

Sepideh Sadeghi
Computer Science Department
Tufts University
Medford, MA 02155
Email: sepideh.sadeghi@tufts.edu

Matthias Scheutz
Computer Science Department
Tufts University
Medford, MA 02155
Email: matthias.scheutz@tufts.edu

***Abstract*—Word learning in ambiguous contexts is a challenging task which is intertwined with the process of understanding the referential intentions of the speaker. It has been suggested that infant word learning occurs across learning situations and is bootstrapped by syntactic regularities such as word order. 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. 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. Comparing learning results in the presence and absence of joint acquisition of word order, word order regularities can quickly converge on real world statistics even using small datasets, relying on imperfect learned interpretations of word meanings, and even given wrong prior biases. Improvement in word learning results in the presence of joint acquisition of word order, however, were limited and only pronounced in the presence of high referential ambiguity and delayed syntactic bootstrapping where word order acquisition was slowed down through the use of priors.**

I. Introduction

Infant word learning often occurs in noisy and ambiguous linguistic and visual contexts, where an utterance with multiple words co-occurs with multiple events or actions in the scene. Learning the meaning of words in such situations involves two intertwined tasks: (a) inferring which events, actions, or objects the speaker is referring to, and (b) mapping each word to its correct referent. [1] bootstrapped the process of learning the meaning of words with the model's belief about the referential intentions of the speaker. Their model exhibited significant outperformance on noisy child-directed corpus over other models including the IBM Machine Translation Model I [2] and the statistical machine translation model [3]. There are similar bootstrapping success stories using syntactic regularities to bootstrap word learning [4]–[8]. [5] ignores the incremental nature of input and assumes full access to all observations. [6] showed that adult-like knowledge of lexical categories learned prior to the onset of word learning improves word learning results. [4], [7] went further and showed that imperfect knowledge of syntactic regularities learned in parallel with words' meaning improves word learning results. [8] went even further and proposed a truly joint learner in which the learned meanings is used to refine the syntactic knowledge, a quality which was missing in the previously proposed joint learners. However, all of these models studied the problem of joint acquisition in the context of ideal learners, ignoring the memory and computational limitations that human learners are subject to.

Ideal learners which assume full access to all data and/or statistical regularities in it are not useful in establishing the true role of joint acquisition in language acquisition. These models often provide theoretical guarantees for converging on real world statistics or the correct lexicon even in the absence of joint acquisition (made possible by making unrealistic assumptions about memory and computational resources). Therefore, it is possible that the true role of joint acquisition is masked by the effect of their unrealistic assumptions. Simulation results from ideal learners suggest that joint acquisition of information can make learning easier, yet it is not suggested that joint acquisition is necessary for converging to the standard spoken language. Furthermore, most ideal learners assume that the result of joint acquisition of different types of information instantly becomes available for constraining the acquisition of other types which assumes that joint acquisition and bootstrapping have the same onsets. However, top-down (information acquired in higher levels of abstraction constraining the acquisition of information in lower levels of abstraction) and bottom-up bootstrapping may have different onsets which has not been studied in experiments with ideal learners. Answering theoretical questions about the role of joint acquisition in language acquisition requires taking into account the memory and computational limitations of human learners, since only memory limited models can truly mimic human performance [9] and reflect the potential advantage of joint acquisition. Moreover, recent findings in [10] challenge the existing statistical accrual-based models of cross-situational word learning. [10] found out that cross-situational word learning is sensitive to input order which is incompatible with the prediction of ideal learners assuming full access to statistical regularities in data. Their findings also suggest that neither alternative hypothesized meanings nor details of past learning situations were retained during cross-situational word learning. These results challenge evidence-accrual models of cross-situational word learning which assume full access to all observations, with no memory limitations.

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. We believe that only memory-limited models qualify as scalable models which remain tractable as the amount of data grows. In this paper,

we augment the model proposed in [1] with word order and event representations to allow for learning verbs (in addition to nouns), as well as accommodating some primitive notion of syntax in the model. We then propose an incremental learning algorithm taking into account the incremental nature of input, as well as memory and computational limitations. Our learning algorithm is a variation of the incremental algorithm proposed in [11] for the word learning model in [1]. The process of learning word referent and word order interleave in our model as imperfect acquired knowledge of word order constrains the acquisition of meaning and vice versa, in each learning trial. The memory of our model is limited to the word-referent mappings stored in the lexicon. Furthermore, the model only sees one situation at a time. Our model departs from ideal learners in that it is not fully Bayesian (only locally in the context of a single learning trial, where only context-appropriate word-referent mappings available in memory are used for hypothesis generation and hypothesis evaluation). The incremental update aggregates the mini-hypotheses inferred in each situation. In doing so, it applies mutual exclusivity constraints between situations to produce a preference for one-to-one mappings in the overall lexicon.

II. Word Learning Assumptions

Our model reduces the problem of learning the meaning of words into the problem of learning the referents of words as oppose to learning a distributed semantic representation for each word. Furthermore, the model is limited to learning the action referent of verbs and words with concrete object referents. In addition to these, the model assumes that infants are capable of correct object and action categorization, prior to the onset of word learning. This is yet another simplifying assumption as the process of object and action categorization probably interleaves with word learning.

A. Input Representation

The input to the model is word learning situations, each of which consists of a scene description paired with an utterance, analogous to the original model [1]. The scene description consists of a list of semantic predicates corresponding to the events happening in the scene (unlike [1]). The utterance consists of an ordered set of words (unlike [1]). Scene and utterance may be empty lists. The events listed in the scene description are not necessarily the ones talked about by the speaker and the model relies on what it already knows about the words and their referents (*lexicon* which is a many to many mappings between words and their referents) to identify the referential intentions of the speaker. We use the term *referential intentions* in the rest of this paper to refer to the events that are listed in the scene description and the one that the speaker is talking about. Inferring the intentions of the speaker in our model refers to the process of identifying which event the speaker is talking about, whereas inferring the intentions of the speaker in [1] signifies the process of identifying which set of objects (from the powerset of objects present in the scene) are referred to by the speaker. The original model in [1] is only capable of learning the meaning of nouns with concrete object referents while our model is also capable of leaning the meaning of verbs with action/event referents that are represented as semantic predicates.

B. Event Representation

Each event happening in the scene, is presented as a semantic predicate with several input arguments corresponding to event participants. The input arguments of the semantic predicates fall into different categories corresponding to their semantic roles. For example, the event of "mom gave Lily a doll" would be represented as "GIVE(MOM,LILY,DOLL)", where "GIVE" is the semantic predicate corresponding to the action of giving, "MOM" (first argument) is the *agent* of the action, "LILY" (second argument) is the *patient 1*, and "DOLL" (third argument) is the *patient 2*. The model has no semantic representation for the semantic roles. However, it assigns the arguments with similar order of appearance in the semantic predicates to the same semantic role category. The model assumes that limited number of semantic roles are known prior to the onset of word learning and used in the event representations.

C. Word Order Representation

The notion of word order in our model, refers to the associations between the words' syntactic position and the semantic role of their referents. This is built on two assumptions. First, The infant assumes that all utterances made by the speaker follow a consistent word order. Second, the infant knows that it is the syntactic position of the words and the semantic role of their referents which are the important pieces of information to be tracked, and that the associations between them allows for learning the structural rules of the language. Note that syntactic position of a word specifies the identity of the word as *subject*, *object*, or *verb*. The learner needs to be able to track the relative order of the headwords in NPs and VPs in each sentence to be able to identify the syntactic position of each word. However we assume that the learner has no knowledge of NPs, VPs, or lexical categories to be used for extracting the headwords. Instead, we assume that learners are only capable of tracking the associations between semantic roles and syntactic positions in short sentences due to their limited cognitive capacity in early infancy. Following this assumption, we limit the input data to our model to short sentences consisting of maximum 3 words which can fall into 2 different lexical categories: noun, verb. These are highly simplifying assumptions, but they allow us to study the problem of joint acquisition of different language capabilities in early stages of acquisition with no linguistic knowledge prior to word learning. We use $\{w_1,w_2,w_3\}$ to represent the syntactic positions and $\{arg_1,arg_2,pred\}$ to represent the semantic roles in our model. The notion of word order Θ consists of three multinomial probability distributions corresponding to three semantic roles. We define $\Theta = \{\theta_{pred}, \theta_{arg_1}, \theta_{arg_2}\}$, where θ_{rol} refers to the multinomial distribution $P(.|rol)$ defined over the three syntactic positions. Each θ_{rol} consists of three $\pi_{pos|rol}$ mass probabilities corresponding to three syntactic positions. The model starts with uniform probability distributions over syntactic positions for each θ_i. Over time, as the model learns more referential words (nouns and verbs), given the word order of English (SVO), it is expected to assign higher probability masses to $\pi_{w_1|arg_1}$, $\pi_{w_2|pred}$, and $\pi_{w_3|arg_2}$ reflecting the expectation that *agent* (arg1), *event* (predicate), and *patient* (arg2) are correspondingly most likely to appear in the position of *subject* (w1), *verb* (w2), and *object* (w3).

III. MODELS

In order to examine the advantage of joint acquisition of word order in our model, we compare two different models: (a) the M-WO model which jointly learns word order and word referent, and (b) the M-B model (baseline model) which only learns word referent. Fig. 1 represents the architecture of the M-WO (with Θ) and M-B models (without Θ), along with their word learning variables and their probabilistic dependencies.

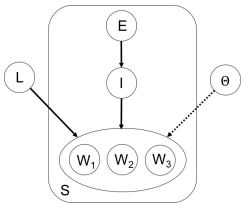

Fig. 1. Graphical model describing the generation of words (w_1,w_2,w_3) given the intention (I), lexicon (L) and word order (Θ). The intention (I) is drawn uniformly from the events (E) present in the situation (S). The building blocks of word order are conditional probabilities ($\pi_{pos|rol}$) for syntactic positions (pos) and semantic role (rol), reflecting the associations between the semantic roles and syntactic positions in the language. The plate indicates multiple copies of the model for different scene-utterance pairs in each situation (S).

In each situation, the model uniformly samples an event from the repertoire of events present in the scene as the referential intention of the speaker (I_s). Each word in the utterance is assumed to be used referentially with probability γ and non-referentially with probability $1-\gamma$. The probability of non-referential use of words P_{NR}, is set to κ for words in the model lexicon (to penalize the non-referential use of such words), and is set to 1 for other words. The referential use of a word in reference to a particular perceptual input P_R is the probability of the word being chosen uniformly from the set of all words linked to the perceptual input in the lexicon.

A. M-WO

In each situation the model infers a context-appropriate mini-lexicon corresponding to the current events and utterance. In doing so, the model tries to maximize the joint posterior probability of mini-lexicon and word order hypotheses according to the Bayes equation and the probability distribution that the model defines over unobserved lexica (L), word order (Θ) and the available context-appropriate evidence (C) including the current situation as well as other context-appropriate situations extracted from the lexicon. The existing word-referent mappings in the lexicon, whose referent or word is observed in the current situation are transformed into situations with utterance=word and scene=referent.

$$P(L,\Theta|C) \propto P(C|L,\Theta)P(L)P(\Theta) \tag{1}$$

Given the probabilistic structure of the model and the fact that speaker's referential intentions are not observable, we marginalize over all possible intentions in each situation and rewrite the likelihood term $P(C|L,\Theta)$ as:

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

Assuming that $P(I_s|E_s) \propto 1$ and that the words of the utterance are generated independently, we can rewrite the term $P(W_s|I_s,L,\Theta)$ as:

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

where $P(pos(w_j)|role(x_i),\Theta)$ is equal to $\pi_{pos(w_j)|role(x_i)}$. We assume that $P(\Theta) \propto 1$ for different word orders, and $P(L) \propto e^{-\beta\cdot|L|}$ serving as a soft mutual exclusivity constraint to produce a preference for one-to-one mappings in the mini-lexicon inferred in each situation.

B. M-B

The goal of the M-B model is to find the the MAP ("maximum a posteriori") lexicon according to $P(L|C) \propto P(C|L)P(L)$, where the likelihood term $P(C|L)$ can be rewritten as:

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

and we can rewrite the term $P(W_s|I_s,L)$ as:

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

IV. INCREMENTAL WORD LEARNING

We need an incremental learning algorithm which operates using limited memory and computational resources available to the learner. In order to fulfill that we specify a set of constraints on incremental learning. First, It only sees each situation once (no iteration over data). Second, the model can only use the knowledge in its current lexicon and current observation for hypothesis generation and evaluation. Third, The model can only maintain a single global hypothesis across different situations motivated by recent findings in [10]. The model can make local revisions to this global hypothesis incrementally, as it receives more data. Our constraints on the data used for hypothesis generation and hypothesis evaluation may be too strict compared to the resources available to human learners, but we believe that they are plausible approximation to the actual constraints, for establishing the utility of joint acquisition in memory-limited and non-ideal learners. Our constraints exclude the use of many proposed incremental algorithms in the literature [12]–[15].

Our proposed incremental learning algorithm has two components: (1) inferring the MAP mini-lexicon in each situation, (2) merging the new mini-lexicon with the current lexicon, while applying mutual exclusivity constraints. The process of inferring the MAP mini-lexicon, subsequently has two distinct components: (1) generating lexicon proposals, and (2) scoring the generated lexica. Scoring is performed by computing the relative posterior probability of the lexicon proposals based on the Bayes equations described earlier for M-WO and M-B. Generating lexicon proposals is guided by stochastic search techniques. The stochastic search in [1] is performed on all the possible links assuming full access to all observations (batch model). Our stochastic search instead is performed only on the

context-appropriate word-referent mappings available in the memory (current lexicon and the current situation). Therefore our stochastic search is focused on small parts of the current and past observations. Focusing on smaller domains is in line with the "less-is-more" hypothesis [16] and, furthermore, more cognitively plausible.

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

The incremental model, at each point in time, extracts the unique words and perceptual referents (actions or objects) observed in the current situation and stores the union of them in *entities*. The model then initializes all possible combinations of word-referent pairs and stores them in *links*. *extract-L* extracts the word-to-referent pairs from the previous lexicon where word or referent can be found in *entities*. The model updates *links* to be the union of *links* and *prevLinks*. *stats* contains some useful statistical measures such as point wise mutual information (PMI) of word-referent pairs. These statistical measures are extracted from all situations observed so far and are incrementally updated as new situations are encountered. We use PMI of links as a *goodness heuristic* for links, employed in *init* and *mutate*. *init* generates *nInit* new lexicon proposals in two steps: (1) sampling the length of the lexicon (we use a uniform distribution over all possible length values going from zero to the size of *links*), and (2) for each proposal, sampling links from *links* according to a distribution created by normalizing exponentiated links' PMIs, where the exponent is the inverse of a temperature parameter. The temperature value can be used to adjust the stochasticity of the outcome of sampling, where higher temperature values make the outcome of sampling more stochastic. *proposals* is a list of *nInit* lexica, and each lexicon is a list of word-referent pairs. *best* computes the posterior probability of its input lexica (hypotheses) given its input situation as data. Then it samples one lexicon as the best one, from a distribution created by normalizing the exponentiated computed lexicon posterior probabilities, where the exponent is the inverse of a temperature parameter. *extract-S* extracts all the mappings of any item in *entities* which exist in the previous lexicon *prevLex* and for each of those mappings it creates a new situation with *link.word* as utterance and *link.referent* as the scene description. The union of the current situation and the situations extracted from the previous lexicon creates *situations* which is used for evaluating the posterior probability of different "mutations" of the best proposed lexicon (*bestLex*). "mutation" of a lexicon simply refers to adding, deleting, or swapping a word-referent pair to/from/in the lexicon. In each mutation step, the model generates 3 new mutated lexica from the base lexicon, using all three mutation moves. It, then uses each of the new generated lexica as the base lexicon for the next "mutation step". Therefore in two mutation steps we will have 3^2 lexica, which are the mutated versions of the base lexicon. *exclude* removes the word-referent pairs whose word or referent are a member of *entities* from the previous lexicon and stores the result in *lex1*. *lex1* shares no item with the current situation which removes the possibility of any inconsistency between *lex1* and the new lexicon to be inferred using *mutate*. *mutate* takes *bestLex* as its base lexicon and tries all three mutation moves described above on it for *nStep*. In each mutation step, it uses mutated lexica in the previous mutation step as the input lexicon for the next mutation step. After *nStep* mutations are completed, it evaluates the mutated lexica as well as the first base lexicon (*bestLex*) using *situations* as data and selects one lexicon as the best one, by sampling from the distribution created by normalizing the exponentiated posterior probability for these lexica, where the exponent is the inverse of the temperature parameter. *mutate* repeats these steps for *nIter* number of times and returns the result. *mutate* can be performed in two modes: *random* or *smart*. We can adjust the probability of performing smart mutations versus random ones in our model. Different from random choices of adding/deleting/swapping links, smart mutation makes proposals based on the links' PMI values. Finally, *merge* merges the non-conflicting part of the previous lexicon *lex1* with the new inferred lexicon *lex2* and returns that as the new lexicon after integration of the new situation. The search for the best lexicon is partly guided by a heuristic where PMI of the links serves as a goodness measure and informs the search (*initLex*), and partly by local optimization (mutating the lexica to maximize the posterior probability in *mutate*). This optimization is local since it maximizes the posterior probability given partial observations. [1] model differs from ours in that it combines heuristic based search with global optimization (maximizing the posterior probability of lexicon given all data). Global optimization is not a choice in the incremental model as past data is no longer available. However, the knowledge in model lexicon serves as model's interpretation of the past observations to evaluate the new proposals and their mutations.

V. Incremental Word Order Learning

Word order learning consists of only one component which updates each $\theta_{rol_i} \in \Theta$ based on the current best lexicon, and the syntax-semantics associations inferred from the current situation. We use a symmetric Dirichlet distribution (with parameter α) as the conjugate prior for each multinomial distribution θ_{rol_i}. Large values of α represent a strong prior bias toward nonsparsity and small values represent a strong bias toward sparsity of θ_{rol_i} (multinomial distributions). The value of each $\pi_{pos|rol}$ at initialization is $\alpha/(3\alpha)$. As the model receives more input incrementally, it updates each $\pi_{pos|rol} \in \Theta$:

$$\pi_{pos|rol} = \frac{Count(rol, pos) + \alpha}{Count(rol) + 3\alpha} \tag{6}$$

VI. Evaluation Data

We evaluated M-WO and M-B on two artificially generated datasets (D1 and D2), each consisting of 60 situations. These datasets were generated based on imaginary interactions in the kitchen at home. For each situation in the dataset, we first generated an utterance. Then, the corresponding semantic predicate representation of the utterance was added to the scene description of the same situation. In some situations, in addition to the semantic representation of the utterance, the semantic representation of up to 3 more co-occurring events were added to the scene description, in order to increase the referential ambiguity. The main difference between D1 and D2 is in their degree of referential ambiguity. D1 situations consist of one event (semantic predicate) paired with one utterance, e.g., the utterance "john drink water" and the scene "DRINK(JOHN,WATER)". D2 situations consist of up to 4 events paired with one utterance, e.g., the scene "DRINK(SUE,WATER)" and "OPEN(MOM,DOOR)" in addition items in the D1 example. Each utterance and event representation accordingly include 2-3 words and 2-3 referents (action/objects).

VII. Results

All results demonstrated in this section are averaged over 10 runs, due to which the learning curves are a bit spiky. The choice of best parameter values to maximize the word learning results depends on the input dataset. We ran M-B on D1, using different parameter values to find the best set of parameters which are used in all of our simulations with both M-B and M-WO: $\gamma = 0.9$, $\beta = 5$, $\kappa = 0.1$, $nIter = 5$, $nStep = 3$.

A. Word Order Learning Results

Table I demonstrates the acquisition of word order in M-WO, over time as more data becomes available. We could measure the accuracy of the learned word order in two ways: (a) computing the Kullback-Leibler divergence (*KL*) between the probability distributions θ_{rol_i} inferred by the model and the target probability distributions, (b) assessing the target conditional probabilities $\pi_{w1|arg1}$, $\pi_{w3|arg2}$, and $\pi_{w2|pred}$ and how they change over time. We chose the second method as the target probability distributions θ_{rol_i} are sparse distributions which assign all the probability mass to one syntactic position. Therefore, computing KL values would not be a good way of representing the direction of change in the inferred conditional probabilities ($\pi_{pos|rol}$) for the correct syntactic positions given semantic roles. We used different values for α (the parameter of the symmetric Dirichlet distribution prior) to enforce different levels of prior bias toward sparsity of θ_{rol_i}. As can be seen, using stronger sparsity prior biases (smaller values) allows the model to form better notions of word order as the conditional probability of the correct syntactic positions given each semantic role increase over time and become significantly higher than the probability of the wrong syntactic positions. This is not surprising as the target θ_{rol_i} are in fact sparse probability distributions. However, the results show that even if the model starts with a wrong sparsity prior (large values), it still can correct its wrong prior over time and move toward sparse θ_{rol_i} by incrementally assigning larger probability masses to the correct syntactic position given each semantic role. Note that word order is better learned on D1 compared to D2. This is probably due to higher context ambiguity in D2, which makes the word-to-referent mapping problem harder which in turn makes the problem of learning the association between the syntactic position and semantic role of the words and their referents, harder.

B. Word Learning Results

Table II demonstrates the average (over 10 runs) f-score scores of the lexica found by M-WO and M-B models when ran on D1 and D2. We ran M-WO with different values for α (the parameter of the symmetric Dirichlet distribution prior for $\theta_{rol_i} \in \Theta$) to enforce different levels of prior bias toward sparsity of θ_{rol_i} (small α values enforce strong sparsity biases and large α values enforce strong non-sparsity biases). There are a couple of things to note. First, referential ambiguity has an inhibitory effect on the word learning results as the fscore scores on D2 (more ambiguous data) are less than those on D1. Second, M-WO exhibits slight outperformance over M-B on D2 (higher referential ambiguity), but not on D1. This may suggest that joint acquisition of word order is more advantageous in improving the word learning results in the presence of high referential ambiguity (D2). Third, M-WO exhibits its worse performance when using stronger sparsity bias (small α values). This may suggest that top-down bootstrapping where the imperfect knowledge of Θ constrains the acquisition of words' meaning, should be limited or delayed until perfection, to avoid possible misguidance in early stages of acquisition. Similar effects were observed in the learning curves capturing the incremental acquisition of words' meaning.

VIII. Discussion and Conclusion

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. Our model's memory of past observations is limited to the word-referent mappings stored in the lexicon and its incremental learning algorithm only sees one situation at a time (no iteration over data). Bayesian inference is thus only applied locally in the context of single learning trials based on context-appropriate word-referent mappings available in the memory (current lexicon and current situation). Different hypotheses (groups of word-referent mappings) are generated during hypothesis generation and evaluated during hypothesis evaluation. The mappings in the best hypothesis (mini-lexicon) are added to the lexicon, while existing alternative mappings are removed, applying a strict mutual exclusivity constraint between situations. We also apply a soft mutual exclusivity constraint, in each situation, through the use of our lexicon prior probability function which is exponential in the size of lexicon and produces a preference for smaller mini-lexica.

The model allows for the acquired word order information to constrain the acquisition of word' meanings and vice versa. Therefore, joint acquisition and bootstrapping are assumed to have similar onsets and delaying the acquisition of word order serves to delay the syntactic bootstrapping. Our results showed that the benefit of joint acquisition of word order in improving word learning results was only pronounced when using weak sparsity biases (delayed syntactic bootstrapping), while learning word order regularities using imperfect interpretations

TABLE I. THE VALUES INCREMENTALLY INFERRED FOR $\pi_{w1|arg1} \in \Theta$ (FIRST COLUMN), $\pi_{w2|pred} \in \Theta$ (SECOND COLUMN), AND $\pi_{w3|arg2} \in \Theta$ (THIRD COLUMN) BY M-WO.

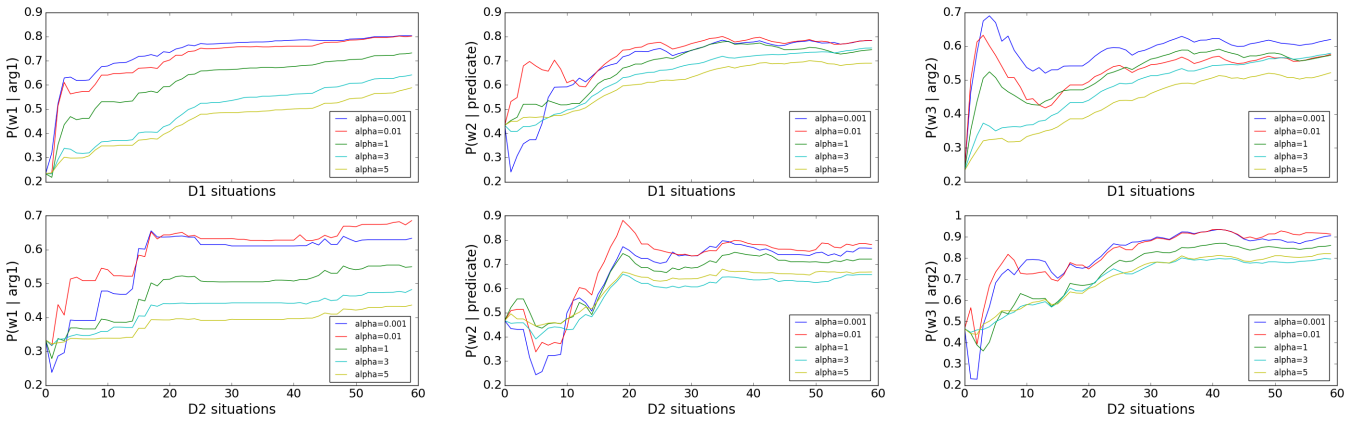

TABLE II. F-SCORE OF THE BEST LEXICON FOUND BY MODELS ON D1 AND D2, AVERAGED OVER 10 RUNS.

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

of the words' meaning was possible regardless of the used sparsity biases. This asymmetry regarding the benefit of joint acquisition suggests that top-down and bottom-up bootstrapping should have different onsets during language acquisition in order to facilitate better learning results. However, our strict mutual exclusivity constraints and memory limitations may have influenced the results. In future research, we plan to study the utility of joint acquisition as we modulate the mutual exclusivity constraints and memory limitations.

The notion of joint acquisition refers to the co-evolution of different language (cognitive) capabilities which is of direct relevance to cognitive information communication [17]. In our model, the notion of co-evolution goes beyond simple transfer of information between different cognitive capabilities as the acquisition of one capability constrains the acquisition of the other and vice versa. The utility of joint acquisition in improving word learning demonstrates the utility of co-evolution of different cognitive capabilities regardless of how the information transfer occurs (in one human, between two humans, or between a human and artificially cognitive system).

IX. ACKNOWLEDGMENTS

This work was in part funded by Vienna Science and Technology Fund project ICT15-045 and by ONR grant N00014-14-0149.

REFERENCES

[1] M. C. Frank, N. D. Goodman, and J. B. Tenenbaum, "Using speakers' referential intentions to model early cross-situational word learning," *Psychological Science*, vol. 20, pp. 578–585, 2009.

[2] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[3] C. Yu and D. H. Ballard, "A unified model of early word learning: Integrating statistical and social cues," *Neurocomputing*, vol. 70, pp. 2149–2165, 2007.

[4] C. Yu, "Learning syntax–semantics mappings to bootstrap word learning," in *Proceedings of the 28th annual conference of the cognitive science society*, vol. 36, 2006.

[5] L. Maurits, A. F. Perfors, and D. J. Navarro, "Joint acquisition of word order and word reference." Cognitive Science Society, 2009.

[6] A. Alishahi and A. Fazly, "Integrating syntactic knowledge into a model of cross-situational word learning," in *Proc. of CogSci*, vol. 10, 2010.

[7] A. Alishahi and G. Chrupała, "Concurrent acquisition of word meaning and lexical categories," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012, pp. 643–654.

[8] O. Abend, T. Kwiatkowski, N. J. Smith, S. Goldwater, and M. Steedman, "Bootstrapping language acquisition," *Cognition*, vol. 164, pp. 116–143, 2017.

[9] M. C. Frank, S. Goldwater, T. L. Griffiths, and J. B. Tenenbaum, "Modeling human performance in statistical word segmentation," *Cognition*, vol. 117, no. 2, pp. 107–125, 2010.

[10] T. N. Medina, J. Snedeker, J. C. Trueswell, and L. R. Gleitman, "How words can and cannot be learned by observation," *Proceedings of the National Academy of Sciences*, vol. 108, no. 22, pp. 9014–9019, 2011.

[11] S. Sadeghi, M. Scheutz, and E. Krause, "An embodied incremental bayesian model of cross-situational word learning." in *Proceedings of the 2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (icdl-epirob)*, 2017.

[12] P. Liang, M. I. Jordan, and D. Klein, "Probabilistic grammars and hierarchical dirichlet processes," *The handbook of applied Bayesian analysis*, 2009.

[13] L. Pearl, S. Goldwater, and M. Steyvers, "Online learning mechanisms for bayesian models of word segmentation," *Research on Language & Computation*, vol. 8, no. 2, pp. 107–132, 2010.

[14] B. Börschinger and M. Johnson, "A particle filter algorithm for bayesian word segmentation," in *Proceedings of the Australasian Language Technology Association Workshop*. D. Mollá & D. Martinez, 2011, pp. 10–18.

[15] ——, "Using rejuvenation to improve particle filtering for bayesian word segmentation," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, 2012, pp. 85–89.

[16] E. L. Newport, "Maturational constraints on language learning," *Cognitive science*, vol. 14, no. 1, pp. 11–28, 1990.

[17] P. Baranyi and A. Csapo, "Definition and synergies of cognitive infocommunications," *Acta Polytechnica Hungarica*, vol. 9, no. 1, pp. 67–83, 2012.