COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 39 (2015) 227–267

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12140

Learning a Generative Probabilistic Grammar of Experience: A Process-Level Model of Language Acquisition

Oren Kolodny[a], Arnon Lotem[a], Shimon Edelman[b]

[a]*Department of Zoology, Tel-Aviv University*
[b]*Department of Psychology, Cornell University*

Received 15 November 2012; received in revised form 8 October 2013; accepted 1 November 2013

Abstract

We introduce a set of biologically and computationally motivated design choices for modeling the learning of language, or of other types of sequential, hierarchically structured experience and behavior, and describe an implemented system that conforms to these choices and is capable of unsupervised learning from raw natural-language corpora. Given a stream of linguistic input, our model incrementally learns a grammar that captures its statistical patterns, which can then be used to parse or generate new data. The grammar constructed in this manner takes the form of a directed weighted graph, whose nodes are recursively (hierarchically) defined patterns over the elements of the input stream. We evaluated the model in seventeen experiments, grouped into five studies, which examined, respectively, (a) the generative ability of grammar learned from a corpus of natural language, (b) the characteristics of the learned representation, (c) sequence segmentation and chunking, (d) artificial grammar learning, and (e) certain types of structure dependence. The model's performance largely vindicates our design choices, suggesting that progress in modeling language acquisition can be made on a broad front—ranging from issues of generativity to the replication of human experimental findings—by bringing biological and computational considerations, as well as lessons from prior efforts, to bear on the modeling approach.

Keywords: Generative grammar; Learning; Graph-based representation; Incremental learning; Linguistic experience; Statistical learning; Grammar of behavior; Language learning

1. Introduction

Research into language acquisition and the computational mechanisms behind it has been under way for some time now in cognitive science (e.g., Adriaans & van Zaanen,

Correspondence should be sent to Oren Kolodny, Department of Zoology, Tel-Aviv University, Tel-Aviv 69978, Israel. E-mail: orenkolo@post.tau.ac.il

2004; Bod, 2009; DeMarcken, 1996; Dennis, 2005; Solan, Horn, Ruppin, & Edelman, 2005; Wolff, 1988; see Clark 2001 for additional references). Here, we describe the design and implementation of a computational model of language acquisition, inspired by some recent theoretical thinking in the field (Edelman, 2011; Goldstein et al., 2010; Lotem & Halpern, 2008, 2012). Unlike our own earlier efforts (Solan et al., 2005; Waterfall, Sandbank, Onnis, & Edelman, 2010), this model, U-MILA,[1] is explicitly intended to replicate certain features of natural language acquisition (as reflected in the diverse set of tasks on which it has been tested), while meeting certain performance requirements and adhering to some basic functional-architectural constraints.

1.1. Requirements and constraints in modeling language acquisition

Much useful work within this field focuses on specific developmental phenomena (such as temporary over-generalization in verb past tense formation; McClelland & Patterson, 2002) or characteristics of adult performance (such as "structure dependence" in forming polar interrogatives; Reali & Christiansen, 2005). A comprehensive approach to language acquisition requires, however, that the model be, first and foremost, *generative* in the standard linguistic sense of being capable of accepting and producing actual utterances (as opposed to merely predicting the syntactic category of the next word, a task on which "connectionist" models are often tested), including novel ones (Edelman & Waterfall, 2007). Importantly, a generative model can be evaluated with regard to its precision and recall—two customary measures of performance in natural language engineering, which can address the perennial questions of model relevance and scalability.

An additional requirement is that the model approximates the probability distribution over utterances (Goldsmith, 2007), so as to have low perplexity (Goodman, 2001) on a novel corpus. This requirement combines two senses of generativity: the one from linguistics, mentioned above, and the one from machine learning, which has to do with modeling the joint probability distribution over all variables of interest in a manner that would allow to draw new samples from it (here, to generate new utterances with probability close to that implied by the corpus of experience).

Another fundamental expectation of a comprehensive theory of language acquisition is that the mechanistic explanations that it provides be detailed enough to allow understanding of the reasons behind various performance traits of models derived from it. This requirement can be realized only if the model's functional architecture and operation, including the learning process, are readily interpretable and transparent. (By functional architecture, we mean mechanisms that are defined on the level of their computational function rather than implementational details [neural or other]. A typical example is the phonological loop, used in our model: What matters about it is that it acts like a queue, not how it operates on the inside.)

A viable model of language acquisition must scale up to sizeable corpora of natural language. The traditional connectionist focus on miniature artificial language test environments (e.g., in exploring recursion in a language defined over a handful of symbols; Christiansen & Chater, 1999, 2001) was useful at the time. However, to be able to argue

convincingly that cognitive science is making progress in understanding the computational underpinnings of the human language faculty, modelers can no longer limit their consideration to "toy" corpora.

Finally, a comprehensive model of the language faculty should simultaneously account for a range of phenomena concerning language that have been identified by linguists, and studied by psycholinguists, over the past decades. One example of such a phenomenon is the structure dependence of auxiliary verb fronting, mentioned above (Chomsky, 1980; Reali & Christiansen, 2005); another example is the so-called syntactic island family of effects (Ross, 1967; Sprouse, Wagers, & Phillips, 2012a; c.f. Section 3.8).

1.2. The motivation behind the present model

The functional architecture and the learning method of the model described here have been inspired by the above considerations. Similarly to ADIOS (Solan et al., 2005), U-MILA is structured as a weighted directed graph over elementary units, which can be words in a natural language, syllables in birdsong, or actions in a foraging task, with paths corresponding to sentences, song phrases, or exploration behaviors. This design feature facilitates its interpretation: Admissible sequences can be simply read off the graph structure at each stage of the modeling process. The graph architecture is, of course, reminiscent of neural systems, which also consist of units connected by weighted directed links that can be modified through learning. It is also connectionist, in the original sense of Feldman and Ballard (1982), rather than, say, Elman (1990)—a distinction to which we shall return in the discussion. Unlike ADIOS or the batch algorithms that are common in natural language engineering, U-MILA learns incrementally (c.f. Cramer, 2007; Kwiatkowski, Goldwater, Zettelmoyer, & Steedman, 2012), updating its parameters and structure as each new series of items passes through its sequential working memory ("phonological loop," c.f. Baddeley, Gathercole, & Papagno, 1998).

Evolutionary considerations suggest that learning mechanisms in multiple species and for different tasks are derived from a common origin, are subject to similar constraints, and require flexibility in order to cope with a constantly changing environment (see Kolodny, Edelman, & Lotem, 2014). Accordingly, both the representational approach and the learning mechanism of U-MILA are general-purpose, open-ended, and parameterized so as to allow tuning to different modalities and contexts. We consider U-MILA to be a model for learning grammars of experience and behavior—a broad category of tasks, which includes, besides language acquisition, also tasks such as learning of regularities for efficient foraging (Kolodny et al., 2014) and of birdsong (Menyhart, Kolodny, Goldstein, DeVoogd, & Edelman, unpublished data). In each case, this model meets the three requirements stated earlier: generativity, sensitivity to the probabilistic structure of the domain, and representational transparency.

These evolutionary considerations, alongside the model's endorsement of computational and memory constraints and the incremental, unsupervised, and open-ended nature of its learning process, place it, we believe, at the head of the line with regard to biological realism among the current language learning models.

The rest of this paper is structured as follows. In Section 2, we state in detail the considerations behind the model's design, its functional components, and the learning algorithm, and explain how the grammar that it acquires is used to process and generate new sentences. Section 3 describes the 17 experiments (grouped into five studies) in which we subjected the model to a variety of tests, both general (precision and recall) and specific (ranging from word segmentation to structure-dependent syntactic generalization). Finally, Section 4 offers a discussion of the lessons that can be drawn from the present project.

2. The model and its implementation

2.1. Design principles

Although language learning is increasingly seen as dependent on social and other interactions with the environment (Goldstein et al., 2010; Pereira, Smith, & Yu, 2008; Smith & Gasser, 2005), in the present project we chose to explore a completely unsupervised approach, since the learner-environment interaction only rarely includes explicit feedback to the learner's actions. The performance of U-MILA can, therefore, be seen as a baseline and should improve with the introduction of social and other interactions, as well as with the integration of other modalities such as prosody, joint attention, etc., with linguistic content or "text" (Goldstein et al., 2010).

In dealing with sequential data, U-MILA adheres to certain general computational principles. One such principle is the reliance on the key operations of alignment and comparison for the identification of significant units in the input sequence—for instance, words in a series of syllables (Edelman, 2008a,b; Goldstein et al., 2010). Because the units in question are not available to the learner ahead of time as such, they can be discovered by comparing the input stream to time-shifted versions of itself; a local alignment then signals the presence of a recurring unit, which can be retained provisionally, until its statistical significance can be ascertained. Given the incremental nature of the input and the likely cost of memory and computation, such comparison should only be carried out within a relatively narrow time window—a design feature which happens also to boost the reliability of unit inference, insofar as a unit that re-appears within a short time is likely to be significant (Goldstein et al., 2010; Lotem & Halpern, 2008). We shall highlight additional computational principles incorporated into U-MILA as we proceed with its detailed description.

2.2. The functioning of the model

In each learning cycle, the current input item (e.g., a word, or morpheme) is added to a short-term memory queue, or the phonological loop (Baddeley, 2003; Burgess & Hitch, 1999)—the time window through which the model "sees" the world (Goldstein et al., 2010). Next, this item is analyzed in the context of the existing graph-based representation of the model's experience to date (initially a "clean slate") and the graph is updated as needed. Operations that use this representation, such as the construction of a (possibly

novel) output sequence or the estimation of the probability of a test sequence (a stand-in for acceptability), can be performed at any time during learning.

The input is read from a text file, in which the tokens are either separated by white-spaces (as when the basic units are words) or not, in which case a whitespace is inserted between every two adjacent tokens. The tokens may represent morphemes or words, but also syllables of birdsong, music notation, actions in physical space, or any other type of discrete sequential data.[2]

Items in the short-term memory queue are subject to temporal decay; a token whose activation drops below a threshold is deleted. In all the experiments described in this paper the decay was exponential; the half-life parameter for each run, $D_{\text{short\_term}}$, is listed in Appendix S5. The resulting effective length of the queue was typically 50–300 tokens.

The model's directed graph-like representation of experience is inspired by the higraph formalism proposed by Harel (1988), which combines the idea of a multi-graph with that of Venn diagrams, and which we refer to in this paper simply as "the graph" (Fig. 1). The graph's nodes are of two types: base nodes, which stand for basic input tokens, and supernodes, which are concatenations of nodes—either base nodes or, recursively, other supernodes, thus accommodating the hierarchical structure of language (Phillips, 2003). Supernodes represent *collocations*: sequences of basic tokens that the learning mechanism deems significant enough to be made into units in their own right. A special type of supernode, referred to as a slot collocation, contains a slot that can be occupied by certain other nodes, as in *the* ____ *boy*, with *big* and *nice* as possible fillers (Fig. 1). In other words, a slot collocation contains a constituent that is variable and can accept a number of nodes in the graph as fillers. Slot collocations enable the model to represent recursion and to capture non-local dependencies, such as between the words "the" and "boy" in the above example. A Boolean parameter, $B_{\text{FillerSetSizeSensitivity}}$, controls whether the learner would be sensitive to the fillers' set sizes: Allowing *the* ____ *boy* to contain as fillers

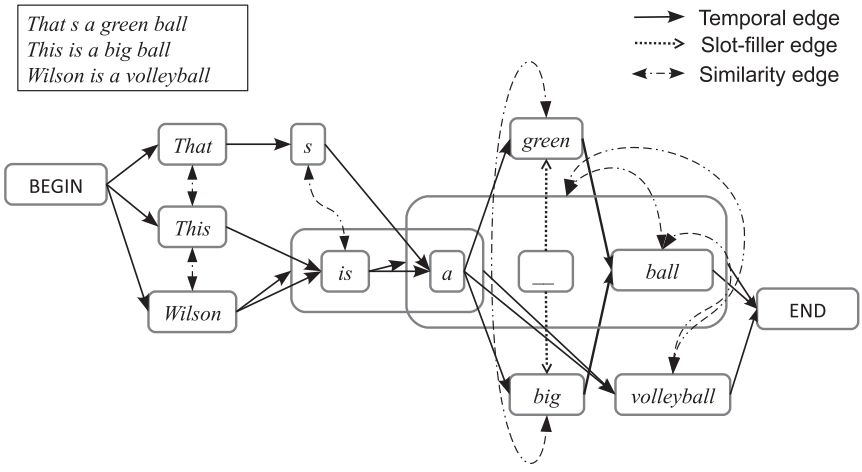

Fig. 1. The graph constructed by U-MILA after training on the three-sentence corpus shown in the inset in the upper left corner. Note that similarity edges, denoted by doubleheaded arrows, are not necessarily symmetric in the model (i.e., the similarity of *s* to *is* may be different from that of *is* to *s*). For clarity, the weights of nodes and edges are not shown.

both *big* and *highly talented*, or requiring the latter to be a filler only in a multi-slot collocation such as *the* ____ ____ *boy*.

Supernodes are implemented as pointers referencing their constituents. The same supernode can have multiple alternate compositions. For example, in a graph that contains the base nodes *I* (1), *want* (2), and *to* (3), and the supernodes *I want* (4) and *I want to* (5), the supernode (5) would contain pointers that signify its composition both as (1) + (2) + (3) and as (4) + (3).

The nodes of the graph may be connected by three types of weighted directed edges: (a) a temporal edge, representing the non-normalized probability[3] of occurrence of one node after another; (b) a slot-constituency edge, representing the non-normalized probability of a certain node acting as a filler in a slot collocation; and (c) a substitutability edge, representing the similarity among two nodes (see Fig. 1 and a detailed explanation below).

Each node, besides representing a token or a sequence of tokens, contains a number of internal fields: a weight, which is a function of its number of occurrences; a counter that denotes the number of cases in which this node could be switched with another within the short-term memory (discussed below, denoted *slot-interchangeability-within-window*); a state of activation, which allows for priming effects (and can be used for "top-down" goal-oriented sequence production, which is outside the scope of this paper); and the criterion that sanctioned the node's creation (discussed below).

All the node and edge weights in the graph decay exponentially (normally with a very long half-life, controlled by D_{graph}); the number of input tokens received so far acts as a clock that paces the decay. This feature makes it possible for errors, such as perceptual miscategorizations, to decay and eventually become negligible if they are not reinforced.

2.3. The learning process

As already noted, learning in U-MILA is incremental. For every incoming token, the following steps are carried out (see Appendix S1 in the supplementary material for explanations within a more detailed pseudo-code listing of the process[4]):

1. Add to graph: If not encountered previously, add the token to the graph as a base node.
2. Update short-term memory and search for alignments (top-down segmentation):
   (a) Insert the new token into the short-term memory queue.
   (b) Search for newly completed alignments (recurring elements) within the queue.
   (c) Add to the graph each new element, at a probability inversely proportional to the memory decay factor and to the distance between the element's recurrences.
3. Update temporal relations and construct collocations (bottom-up chunking):
   (a) Create a list of all nodes in the graph that terminate the short-term memory sequence.

(b) Create a secondary list of sequences that fills the slot of slot collocations in the primary list.
(c) Update or add temporal edges between each node in the current list (X) and the nodes in a previously found list that contains the nodes preceding X.
(d) Update slot-candidacy edges of all nodes that are within slot collocations in the primary list.
(e) For each pair of nodes A, B between which a temporal link has been updated, create a new supernode, A + B, if sanctioned by Barlow's (1990) principle of suspicious coincidence, subject to a prior.

2.4. Addition of nodes to the graph

As stated above, nodes are added to the graph in three cases: (a) when a new token is encountered, (b) when a recurring sequence, composed of more than one base token, is found within the short-term memory by alignment of the sequence to a shifted version of itself, and (c) when two existing nodes in the graph occur in the same order often enough so that their combination is deemed a significant unit in itself and the two are then "co-located" into a collocation and added to the graph as a supernode. The two latter cases are effectively two modes of chunking: top-down and bottom-up, respectively. Previous work supports the use of both modes in language learning (van Zaanen & van Noord, 2012; Wolff, 1988).

U-MILA supports four run-time learning modes, corresponding to different ways of creating new nodes: a "flat Markov" mode, in which only base tokens are added (thus not allowing for hierarchical structures, hence "flat"); a "phonoloop collocation" mode, which adds base tokens and recurring sequences from the short-term memory (top-down segmentation); a "bottom-up collocation" mode, which adds only new tokens and significant runs of adjacent units; and a "normal" mode, which combines all of the above.

When a new node is added to the graph, a search is conducted for existing nodes with related content; pointers are updated if the new node is found to be a supernode or a subnode of any of them.

2.5. Calculating similarity among nodes

Estimating similarity among nodes is important for the grammar's open-ended generativity. For example, the model would produce the novel sentence *John ran to school* based on previous encounters with the two sentences *John went to school* and *Dan ran to school* only if it recognizes *went* and *ran* as sufficiently similar. There are multiple cues that hint at such similarity, whose weights should depend on the use to be made of the similarity estimate. The present implementation focuses exclusively on substitutability: the degree to which a unit may be acceptably replaced by another (c.f. Section 3.1).

U-MILA calculates the similarity of two nodes by combining their edge profiles (vectors of weights on both temporal and slot-candidacy edges leading to other nodes) with

their interchangeability in slot collocations (see detailed explanation below).[5] The rationale—both biological and computational—of this approach is that in a network of neurons, a unit is best individuated by its connection profile to the other units: A unit, in other words, is known by the company that it keeps (c.f. Hudson's (2007) Word Grammar). Moreover, the decision about where to proceed from the current node is also based on its edge profile.

A Boolean parameter controls the choice between symmetrical and asymmetrical similarity (to see that substitutability need not be symmetrical, consider that *him* can replace *John* successfully in many utterances, but not the other way around). All the runs reported in this paper allowed for asymmetric similarity, as defined below (the symmetric calculation is simpler). The estimation of similarity can be improved by considering non-adjacent contextual data, which the present model does not retain.

The similarity of A to B is calculated as the weighted average of the following three measures:

1. Analogous temporal edges (ATE):

$$\mathrm{ATE} = \frac{\sum_{\mathrm{x}\in\mathrm{X}} (\mathrm{Weight}(\mathrm{x} \Rightarrow \mathrm{A}) \cdot \theta(\mathrm{x} \Rightarrow \mathrm{B}))}{\sum_{\mathrm{x}\in\mathrm{X}} \mathrm{Weight}(\mathrm{x} \Rightarrow \mathrm{A})} + \frac{\sum_{\mathrm{x}\in\mathrm{X}} (\mathrm{Weight}(\mathrm{A} \Rightarrow \mathrm{x}) \cdot \theta(\mathrm{B} \Rightarrow \mathrm{x}))}{\sum_{\mathrm{x}\in\mathrm{X}} \mathrm{Weight}(\mathrm{A} \Rightarrow \mathrm{x})}$$

where X denotes all vertices in the graph, $\Rightarrow$ denotes a temporal edge, and θ is the Heaviside step function. A non-existent temporal edge is treated as having a weight of zero.

2. Common occurrence in slot (COS):

$$\mathrm{COS} = \sum_{x\in X} (\mathrm{Weight}(\mathrm{FE}(A, x)) \cdot \theta(\mathrm{FE}(B, x))) / \sum_{x\in X} (\mathrm{Weight}(\mathrm{FE}(A, x)))$$

where X denotes all nodes that are slot collocations, FE(A, x) denotes a candidacy link of the node A as filler in the slot in x (FE stands for *Filler Edge*), and all non-existent edges are treated as being of weight zero.

3. Within-slot interchangeability within a short time window (WSI):

$$\mathrm{WSI} = \mathrm{SIWW}(A, B) / \sum_{x\in X} (\mathrm{SIWW}(A, X))$$

where X denotes all vertices in the graph, and SIWW(A, X) denotes the weight of the *slot-interchangeability-within-window* variable, which is a count of the number of times in which some (any) slot collocation was found twice within the phonological loop, once with A as a filler and once with B.

The implementation allows similarity to be re-calculated with every update of a potentially relevant edge in the graph, or following some subset of the updates; time-wise, it

may be calculated periodically, or at the end of a learning phase, or only ad-hoc before fulfilling a production request. Such "offline" updating of similarity representations may be thought of as analogous to memory consolidation.

2.6. Production: Generation of sentences

The sentence generation process consists of traversing the graph, starting with the special *BEGIN* node, and ending upon reaching the *END*.[6] At each node, the next item to be appended to the sentence is chosen as follows:

1. (a) With (very low) probability P_{rand}, choose a node from the graph at random with probability proportional to its weight (*this effectively smoothens the model's estimate of the probability distribution over all possible sentences*).
   else:
   (b) Choose a node from among those that the outgoing temporal edges go to, drawing among them randomly with proportion to $W_{\text{edge}} \cdot L$, where W_{edge} is the weight of the directed edge and L is the length of the node's base token sequence. *(i.e., drawing with a higher probability nodes that contain longer sequences).*
2. With probability $P_{\text{generalize}}$, replace the node by another node, chosen with proportion to its similarity (substitutability) index to the node chosen in (1).
3. If the chosen node contains a slot, choose with (a very low) probability P_{rand} a filler from among all the nodes in the graph with proportion to their weight; with probability $1-P_{\text{rand}}$ choose a filler from among the slot filler candidates in the slot, with proportion to weights of the slot-candidacy edges. If the chosen slot filler is itself a slot collocation, step 3 is re-iterated, in order to find a filler for the internal slot collocation, and so on until a filler which is not a slot collocation is reached.

2.7. Assigning probability to input sentences

The same statistical principles used for producing sentences can also be used for evaluating the probability that a given sentence could have been produced by the model—a capability that is essential for turning the learner into a *language model* (in the usual sense of computational linguistics; c.f. Goodman, 2001), which allows the estimation of perplexity and assessment of grammaticality, as explained below. In addition to the smoothing implied by a non-zero value of P_{rand} as described earlier, the model can also assign a small non-zero probability to a completely novel word (when this is set to 0, any sentence with a novel word would have zero probability).

To estimate the probability of a sentence, the model must find all possible covers of it in terms of paths through the graph; the probability of the sentence is equal to the sum of production probabilities of these covers.[7] To do so, U-MILA conducts a search, in each

stage of which it attempts to cover the sentence using a certain number of nodes, ranging from 1 to the number of base tokens in the sentence. The recursive search routine finds all the possible single-node covers of the beginning of the sentence, then for each of these calls itself on the remainder of the sentence, until it finds a complete cover or determines that such a cover does not exist (note a parallel to left-corner parsing: Resnik, 1992). Once all full covers of a sentence are found, the probability of production of each of these is calculated, using a process analogous to the one described in the production section. The probability assigned to the sentences is the sum of production probabilities of all covers.

In cases where the probability of a sequence that is not a full sentence must be estimated (as in some of the experiments described in the results section), the calculation starts with the actual initial node instead of the standard *Begin* node, and the overall probability is weighted by that node's relative weight in the graph.

3. Testing the model: Results

While the present computational approach applies to a variety of sequential-structural learning situations, in this paper we focus on its performance in language-related tasks. To the best of our knowledge, U-MILA is the first model that can deal with as wide a range of language tasks as reported here, while preserving a modicum of biological realism.

The tests reported below include both (a) the replication of a dozen or so published results in sequence segmentation, artificial grammar learning, and structure dependence, and (b) the estimation of the model's ability to learn a generative grammar—a structured representation that selectively licenses natural-language utterances and is capable of generating new ones (Chomsky, 1957)—from a corpus of natural language. Because a model's explanatory power with regard to language acquisition remains in doubt unless it can learn a generative representation (Edelman & Waterfall, 2007; Waterfall et al., 2010), we begin with an account of the model's generative performance, then proceed to describe its replication of various specific phenomena of interest. The experiments we have conducted were grouped into five studies:

1. Study 1: Measures of generative ability of a grammar learned from a corpus of natural language: recall, perplexity, and precision (defined and stated in Section 3.1).
2. Study 2: Characteristics of the learned representation: equivalence (substitutability) of phrases and the similarity structure of the phrase space (Section 3.2).
3. Study 3: Replication of a variety of results in sequence segmentation and chunking (Section 3.3).
4. Study 4: Replication of results in artificial grammar learning (Sections 3.4–3.6).
5. Study 5: Replication of results regarding certain types of structure dependence (Sections 3.7–3.8).

All studies and results are discussed in additional detail in Appendix S2.

3.1. Study 1: Generative performance

A key purpose of learning a grammar is the ability to generate acceptable utterances that transcend the learner's past experience. This ability is typically tested by evaluating the model's *precision*, defined as the proportion of sentences generated by it that are found acceptable by human judges, and *recall*, defined as the proportion of sentences in a corpus withheld for testing that the model can generate (see Solan et al., 2005; for an earlier use of these measures and for a discussion of their roots in information retrieval). Given that sentence acceptability is better captured by a graded than by an all-or-none measure (Schütze, 1996), we employed graded measures in estimating both recall and precision.

A commonly reported graded counterpart for recall is *perplexity*: the (negative logarithm of the) mean probability assigned by the model to sentences from the test corpus (see, e.g., Goodman, 2001; for a definition). Because in practice, perplexity depends on the size and the composition of the test set, its absolute value has less meaning than a comparison of per-word perplexity values achieved by different models; the model with the lower value captures better the language's true empirical probability distribution over sentences (c.f. Goldsmith, 2007). In the experiment described below, we compared the perplexity of U-MILA to that of a smoothed trigram model implemented with publicly available code (Stolcke, 2002).

For precision, a graded measure can be obtained by asking subjects to report, on a scale of 1–7, how likely they think each model-generated sentence is to appear in the context in question (Waterfall et al., 2010). Because our model was trained on a corpus of child-directed speech, we phrased the instructions for subjects accordingly. The test set consisted of equal numbers of sentences generated by the two models and taken from the original corpus.

Perplexity and the precision of a model must always be considered together. A model that assigns the same non-zero probability to all word sequences will have good perplexity, but very poor precision; a model that generates only those sentences that it has encountered in the training corpus will have perfect precision, but very poor recall and perplexity. The goal of language modeling is to achieve an optimal trade-off between these two aspects of performance—a computational task that is related to the bias-variance dilemma (Geman, Bienenstock, & Doursat, 1992). Striving to optimize U-MILA in this sense would have been computationally prohibitive; instead, we coarsely tuned its parameters on the basis of informal tests conducted during its development. We used those parameter settings throughout, except where noted otherwise (see Appendix S5).

For estimating perplexity and precision, we trained an instance of the model on the first 15,000 utterances (81,370 word tokens) of the Suppes corpus of transcribed child-directed speech, which is part of the CHILDES collection (MacWhinney, 2000; Suppes, 1974). Adult-produced utterances only were used. The relatively small size of the training corpus was dictated by considerations of model design and implementation (as stated in Section 2, our primary consideration in designing the model was functional realism rather

than the speed of its simulation on a serial computer). For testing, we used the next 100 utterances that did not contain novel words.

3.1.1. Perplexity over withheld utterances from the corpus

We used a trained version of the model to calculate the production probability of each of the 100 utterances in the test set, and the perplexity over it, using a standard formula (Jelinek, 1990; Stolcke, 2010):

$$\text{Perplexity} = 10^{-\frac{\sum_{s} \log(P(s))}{n}}$$

where *P(s)* is the probability of a sentence *s*, the sum is over all the sentences in the test set, and *n* is the number of words in the test set.

The resulting perplexity was 40.07, for the similarity-based generalization and smoothing parameters used throughout the experiments (see Appendix S5). This figure is not as good as the perplexity achieved over this test set, after the same training, by a trigram model (SRILM; see: Stolcke, 2002) using the Good-Turing and Kneser-Ney smoothing: respectively, 24.36 and 22.43. As already noted, there is, however, a tradeoff between low perplexity and high precision, and, indeed, the precision of the tri-gram model fell short of that of U-MILA (see below). By modifying our model's similarity-based generalization and smoothing parameters, perplexity could be reduced to as low as 34 (with $P_{\text{generalize}} = 0.2$, $P_{\text{rand}} = 0.01$) and perhaps lower, at a cost to the precision performance. At the other extreme, precision results are expected to rise as the similarity-based generalization parameter is lowered; when it is set to zero, the perplexity rises to 60.04.

Smoothing and generalization enable the model to assign a certain probability even to previously unseen sequences of units within utterances and thus prevent the perplexity from rising to infinity in such cases. It is interesting to note that when the generalization parameter is set to its default value (0.05), smoothing has only a negligible quantitative effect on the perplexity, and setting it to zero leads to perplexity of 40.76, as opposed to 40.07 when it is set to 0.01.

3.1.2. Precision: Acceptability of sentences produced by the learner

To estimate the precision of the grammar learned by U-MILA and compare it to a trigram model, we conducted two experiments in which participants were asked to rate the acceptability of 50 sentences generated by each of the two models, which had been mixed with 50 sentences from the original corpus (150 sentences altogether, ordered randomly). Sentences were scored for their acceptability on a scale of 1 (not acceptable) to 7 (completely acceptable; Waterfall et al., 2010). As the 50 sentences chosen from the original corpus ranged in length between three and eleven words, in the analysis we excluded shorter and longer sentences generated by U-MILA and by the trigram model (SRILM).

In the first precision experiment, the smoothing parameters in the SRILM were set to achieve perplexity of ppl = 40.07, the same value achieved by U-MILA with the "standard" parameter settings used elsewhere in this paper. Six subjects participated in this experiment. The results (see Fig. 2A) indicated an advantage of U-MILA over SRILM ($t = 3.5$, $p < .0005$, R procedure *lme*: D. Bates, 2005). Sentences from the original corpus received a mean score of 6.59; sentences generated by U-MILA, 5.87; sentences generated by SRILM, 5.41. Further, mixed-model analysis (R procedure *lmer*: Bates, 2005) of results broken down by sentence length (see Fig. 2B) yielded a significant interaction between sentence source and length for both models (U-MILA: $t = -3.2$; SRILM, $t = -3.8$). A comparison of the interaction slopes, for which we used a 10,000-iteration Markov Chain Monte Carlo run to estimate the confidence limits on the slope parameters (R procedures *mcmc* and *HPDinterval*), did not yield a significant difference.

In the second precision experiment, the smoothing parameters in SRILM were set to achieve its lowest perplexity and its precision was compared to that of U-MILA with the "standard" settings. See Appendix S2, 1.1.2.

3.2. Equivalence-class inference

To illustrate U-MILA's ability to learn similarities over words and phrases, we offer two characterizations of such relations, for the same version of the model, trained on a corpus of child-directed speech, as in Section 3.1. First, in Table 1, we list the five nodes that are most similar to each of the 20 most common nodes in the graph, as well as to

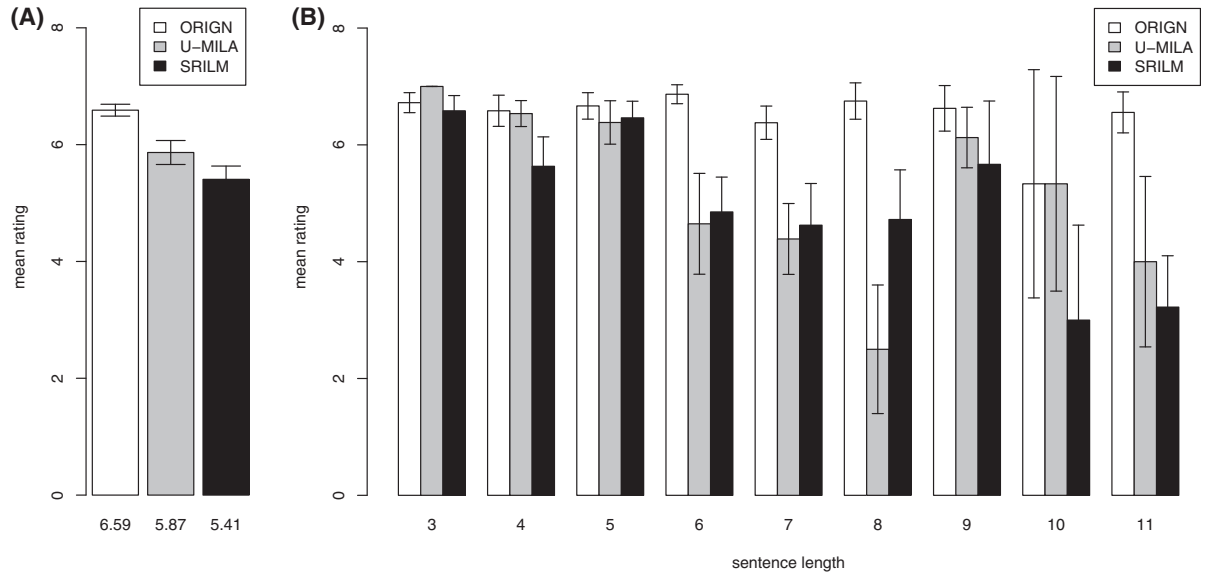

Fig. 2. The mean precision scores assigned by human judges to sentences from the original corpus and to those produced by U-MILA and SRILM (a standard tri-gram model, see text) following training on the first 15,000 utterances in a corpus of child-directed speech (Suppes, 1974). (A) Results for sentences of all lengths pooled together. (B) Results pooled into bins according to sentence length in words. Error bars denote 95% confidence limits. Both models in this experiment were tuned to achieve perplexity of ppl = 40.07.

each of 11 other chosen nodes. Not surprisingly, the most common nodes are function words or slot collocations built around function words; their similarity neighborhoods generally make sense. Thus, in example 1, the neighbors of *the* are all determiners, and the neighbors of *you* are pronouns. Likewise, verbs are listed as similar to verbs or verb phrases (sometimes partial) and nouns—to other nouns or noun phrases (examples 24 and 27). Occasionally, the similarity grouping creates openings for potential production errors, as in example 31, where the list of nodes similar to *which* contains words from both its main senses (interrogative and relative).

The second glimpse into the similarity space learned by U-MILA is a plot produced from similarity data by multidimensional scaling (Shepard, 1980). To keep the plots

Table 1
The five most similar nodes to each of 31 nodes from the repertoire

	Node	Five Most Similar Nodes[18]
	The 20 most frequent nodes in repertoire	
1	The	a; this; your; that; the ___ ?
2	You	we; they; Nina; he; I
3	S	is; was; does; s on; s not
4	what	who; where; it; there; here
5	the ___ ?	the; your; your ___ ?; the ___; it
6	A	the; this; that; your; a ___ ?
7	To	on; in; to ___ to; into; to play with
8	Is	s; was; does; did; what is
9	That	it; this; there; here; he
10	It	that; he; there; this; she
11	you ___ ?	you; you ___ the; we; you ___ to; the
12	what ___ ?	what; it; the; Nina; where
13	On	in; did; to; for; under
14	s ___	s; s ___ the; s ___ ?; ?; s ___ s
15	you ___ to	you; to; we; you want to; going to
16	Are	did; were; do; color are; what are
17	Do	did; are; have; see; eat
18	I	you; we; they; fix; she
19	He	she; it; Nina; that; there
20	In	on; to; inside; at; on top of
	Additional examples from the repertoire	
21	where	what; who; there; here; it
22	is it	is that; are they; were they; is he; was it
23	Go	have; went; do; get; going
24	know	want; remember; see; want to; see it
25	By	in; on; at; where; up
26	bunny	rabbit; boy; elephant; dolly; doll
27	the horse	Nina; it; the boy; the fish; he
28	white	purple; red; big; doll; present
29	pretty	soft; cute; good; called; wet
30	Me	you; her; Linda; Mommy; it
31	which	this; the; that; what ___ that; where

legible, we sorted the words by frequency and focused on two percentile ranges: 95–100 and 75–80 (Fig. 3A and 3B, respectively). As before, the first plot, showing the more frequent items, contains mostly function words and auxiliary verbs, while the second contains open-class words. In both plots, proximity in the map generally corresponds to intuitive similarity.

3.3. Comparison to the TRACX model (French, Addyman, & Mareschal, 2011)

Our next set of studies has been inspired by a recent paper by French et al. (2011) that described a connectionist model of unsupervised sequence segmentation and chunk extraction, TRACX, and compared its performance on a battery of tests, most of them reproductions of published empirical experiments, to that of several competing models, including PARSER (Perruchet & Vinter, 1998) and a generic simple recurrent network (SRN; Elman, 1990). Each of the Sections 3.3.1 through 3.3.10 states a particular earlier result considered by French et al. (2011) and describes briefly its replication by U-MILA (for details, see Appendix S2).

3.3.1. Words versus non-words, infants (Saffran, Aslin, & Newport, 1996, experiment 1)

Following Saffran et al. (1996), French et al. (2011) created a language of four tri-syllabic words and trained their model on a sequence of 180 words with no immediate word repetitions. The model was then tested for its ability to discriminate between words and non-words, and did so successfully.

We used the stimuli of French et al. (2011, supporting online material) as the training set for U-MILA and tested it on the same 4 words and 4 non-words. All test words were assigned higher probability scores (Section 2.7) than non-words, achieving perfect discrimination, with the difference approaching significance despite the small number of items (Wilcoxon signed rank test, one-sided; $V = 10$, $p < .0625$). Running the model in the flat Markov mode (by disabling the acquisition of hierarchical representations) led to perfect discrimination. This is not surprising, as the distinction between words and non-words here is based by definition solely on forward transition probabilities, which is the (only) feature represented by such a Markov model.

3.3.2. Words versus non-words, infants (Aslin, Saffran, & Newport, 1998, experiment 1)

The words in the Saffran et al. (1996) experiment were heard three times as often as their counterpart non-words. To explore the effects of frequency, Aslin et al. (1998) constructed a training sequence composed of four tri-syllabic words, two of which occurred at a high frequency and two half as often. Thus, the non-words spanning the boundary between the two high-frequency words had the same number of occurrences as the low-frequency words; the within-word transition probabilities remained higher than those in the non-words. French et al. (2011) replicated the results of Aslin et al. (1998), with a 270-word training sequence. Both the TRACX and the SRN models successfully discriminated between the words and non-words in the analogous test. Using the same training

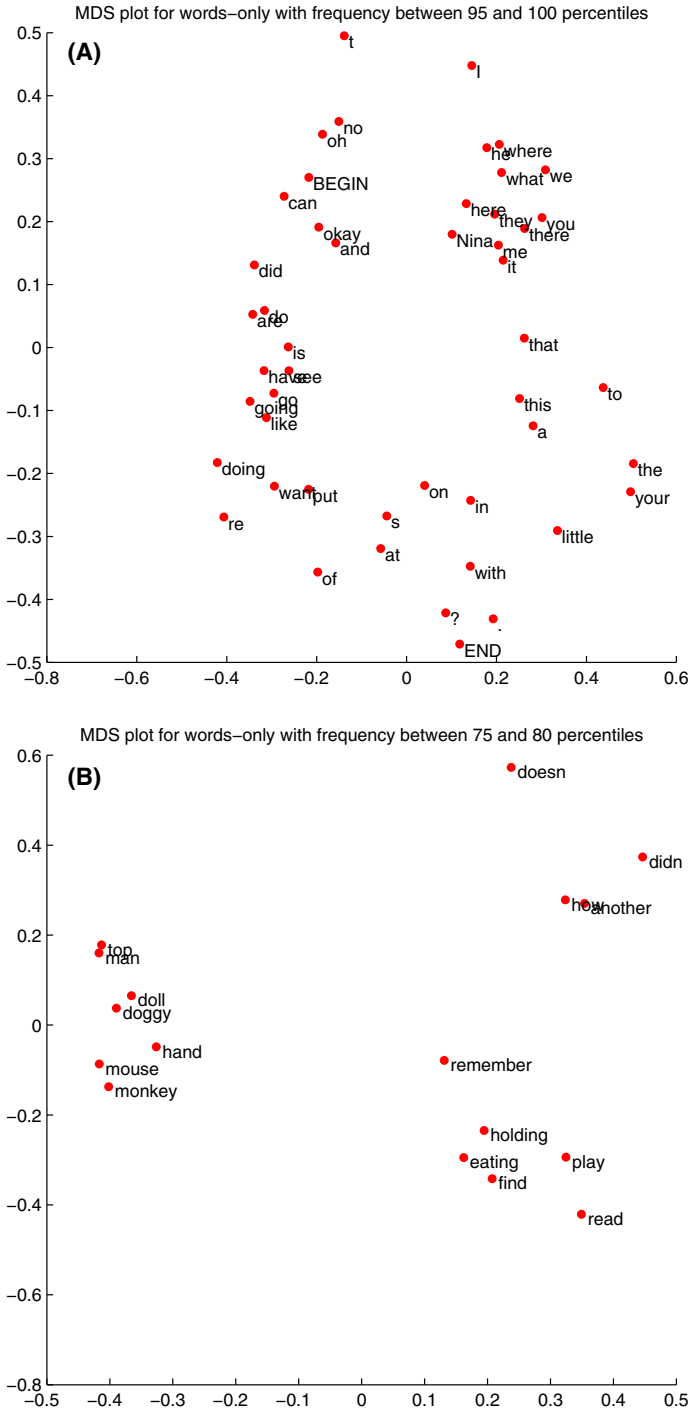

Fig. 3. Similarities among learned words, plotted by applying multidimensional scaling to the tables of similarity scores. (A) The most frequent words in the corpus; percentile range: 95–100. (B) Percentile range: 75–80. Proximity among words in these plots generally corresponds to intuitive similarity among them.

and test sets, U-MILA performed perfectly, always assigning a higher probability to low-frequency words than to non-words (Wilcoxon signed rank test, one-sided; $V = 10$, $p < .0625$; the seemingly low significance value despite the perfect discrimination is due

to the small size of the test set). As in the previous experiment, using our model in the flat Markov mode achieved similar results.

3.3.3. Words versus non-words, adults (Perruchet & Desaulty, 2008, experiment 2: Forward transition probabilities)

In the study by Perruchet and Desaulty (2008), adult subjects listened to a training sequence in which words and non-words had the same frequency, and differed in that transition probabilities were equal to 1 within words and lower within non-words. In the replication by French et al. (2011), both TRACX and SRN learned successfully to discriminate between words and non-words.

Following training with the same dataset, U-MILA also successfully differentiated between words and non-words (Wilcoxon signed rank test, one-sided; $V = 21$, $p < .016$). Unlike in the previous experiments, running the model in its flat Markov mode did not lead to successful discrimination.[8]

3.3.4. Words versus non-words, adults (Perruchet & Desaulty, 2008, experiment 2: Backward transition probabilities)

The second experiment of Perruchet and Desaulty (2008) was the first to show that adults can segment a continuous auditory stream on the basis of backward transition probabilities. The TRACX model of French et al. (2011) replicated this finding; the SRN model did not.

In our replication, using the same training and test sets, U-MILA successfully assigned significantly higher scores to words than to non-words (Wilcoxon signed rank test, one-sided; $V = 21$, $p < .016$). As expected, the run in a flat Markov mode did not differentiate between words and non-words.

3.3.5. Hierarchical chunking (Giroux & Rey, 2009)

Giroux and Rey (2009) showed that once a lexical unit ("sub-chunk") is assimilated into a larger one ("chunk"), it becomes harder to recognize. French et al. (2011) trained TRACX on a corpus composed of two-, three-, and four-syllable words, including *klmn*. At first, the model recognized *kl*, *lm*, and *mn* as separate chunks, which it then gradually merged into larger units (*klm* and then *klmn*). As learning proceeded, the shorter chunks were forgotten.

When trained on this corpus, U-MILA recognized all chunks and subchunks (*kl, lm, mn, klm, lmn, klmn*) as independent units. We note that for a language-oriented model, eliminating subchunks after they are incorporated into larger units would be counterproductive. For instance, it would cause the word *dead* to be forgotten after learning the word *deadline*.[9]

3.3.6. Word segmentation: Effects of sentence length (Frank, Goldwater, Griffiths, & Tenenbaum, 2010, experiment 1)

In their first experiment, Frank et al. (2010) explored the effect of sentence length on the subjects' ability to extract words from it. To do so, they used a set of 18 syllables to

construct two 2-syllable words, two 3-syllable words, and two 4-syllable words, with no shared syllables among the six words. Participants heard a sound stream consisting of 144 of these words, randomly ordered and divided into "sentences" by short pauses. They tested eight groups of participants, all of whom heard the same sequence, but for each group it was divided into a different number of sentences: 144, 72, 48, 36, 24, 18, 12, corresponding to sentences of lengths 1, 2, 3, 4, 6, 8, 12, 24.

French et al. (2011) trained and tested TRACX on a similar dataset, and found that it discriminated between words and part-words better as the sentences got shorter, achieving a correlation of 0.92 with the human results; the correlation of the SRN model's results with the human data was 0.60.

We ran U-MILA in a variety of modes and parameter values, training and testing it as did French et al. (2011), and found the same qualitative trend: The model exhibits better discrimination between words and non-words as the sentences get shorter (Fig. 4). This result held for a range of parameters, with correlation with the human data ranging from 0.49 to 0.87.

3.3.7. Word segmentation: Effects of vocabulary size (Frank et al., 2010, experiment 3)

The next experiment of Frank et al. (2010) replicated by French et al. (2011) explored the effect of the size of the language's vocabulary on learning word/non-word discrimination. The training set in this experiment consisted of four-word sentences, in which the words were drawn from a cadre of differing size, from three to nine words. Word length varied from two to four syllables, and there was an equal number of two-, three-, and four-syllable words in the training corpora for the various conditions. Frank et al. (2010)

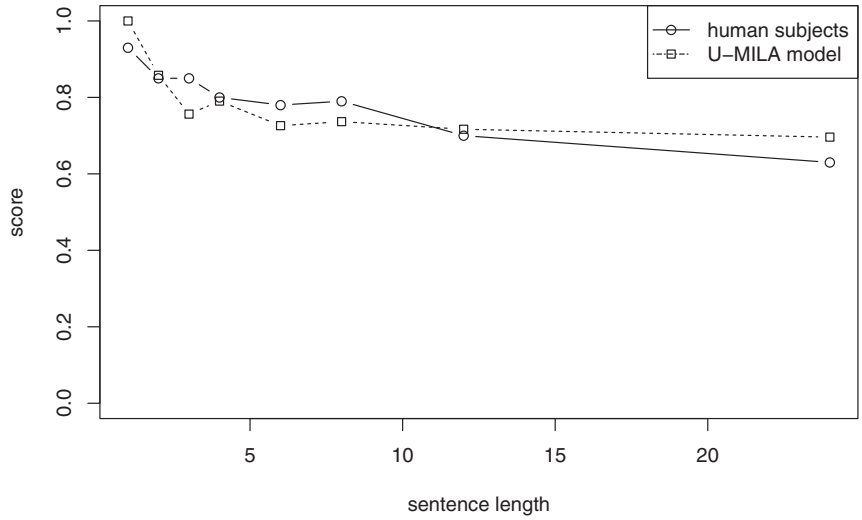

Fig. 4. Discrimination between words and part-words by human participants (Frank et al., 2010, exp. 1) and by U-MILA, after training on constant-size corpora that differed in sentence length. Both humans and the model perform better when trained on shorter sentences. Similar results are achieved for a range of model parameters; discrimination scores presented here are for a simulation in *flat Markov* run mode, using the proportionbetter score described by French et al. (2011).

found that as the word cadre got smaller, the subjects' performance improved. French et al. (2011) replicated this finding with the TRACX model, but not with SRN.

We applied U-MILA to the same dataset used by French et al. (2011) in a range of modes and run parameters. Learning was successful in all cases, but the trend in which a larger word cadre leads to weaker discrimination was found only for some settings: specifically, in the flat Markov mode, or when the prior against creating collocations was strong and the phonological loop decay was very large or the alignment module disabled. An analysis of covariance (R procedure *lm*) applied to a typical case (see Fig. 5A, 5B) yielded significant effects of word-hood ($t = 3.0$, $p < .0039$) and vocabulary size ($t = -5.46$, $p < .0000015$) and a significant interaction ($t = 2.1$, $p < .04$). The absence of the effect of vocabulary size for some parameter settings can be explained by observing that our implementation (unlike humans) has no limitations on simultaneously tracking the statistics of as large a number of syllables as required by the task, and thus finds it as easy to keep tabs on 27 syllables as on 9.

3.3.8. Word segmentation, phonetic encoding (French et al., 2011, simulation 8)

In this experiment, French et al. (2011) applied their model to a phonetically encoded corpus of natural child-directed speech (Bernstein-Ratner, 1987; Brent & Cartwright, 1996), consisting of 9,800 sentences and 95,800 phonemes. French et al. (2011) presented TRACX with each sentence six times in succession, completing five passes through the corpus.

We trained U-MILA with a single run on the same dataset and tested it as in the previous simulations by having it assign probabilities to each word/part-word in the test set. The model assigned significantly higher probability scores to words than to part-words (Fig. 6). An analysis of covariance (R procedure *lm*) yielded significant effects of word-hood ($t = 2.1$, $p < .035$) and number of syllables ($t = -7.08$, $p < 2.9 \times 10^{-12}$) and no interaction.

3.3.9. Word clustering by category (French et al., 2011, simulation 10)

In their experiments 9 and 10, French et al. (2011) explored their model's ability to cluster its internal representations so as to correspond to categories in the training data. We reproduced the second, more complex of these experiments, the stimuli in which came from two microlanguages, each composed of three-letter words. Each word in language A was constructed as follows: The first letter was randomly chosen from $\{a,b,c\}$, the second letter from $\{d,e,f\}$, and the third letter from $\{g,h,i\}$. Similarly, each word in language B consisted of a letter from $\{d,e,f\}$, a letter from $\{a,b,c\}$, and a letter from $\{g,h,i\}$.

A 10,000-word training sequence (approximately 5,000 from each language) contained no markers indicating word or language boundaries. The words in the corpus were drawn from a subset of two-thirds of the possible words in each language. The words were ordered as follows: for each new word, a random draw from among all possible words in one language took place, with a probability of 0.025 of switching to the other language (thus creating within the corpus runs of words from the same language).

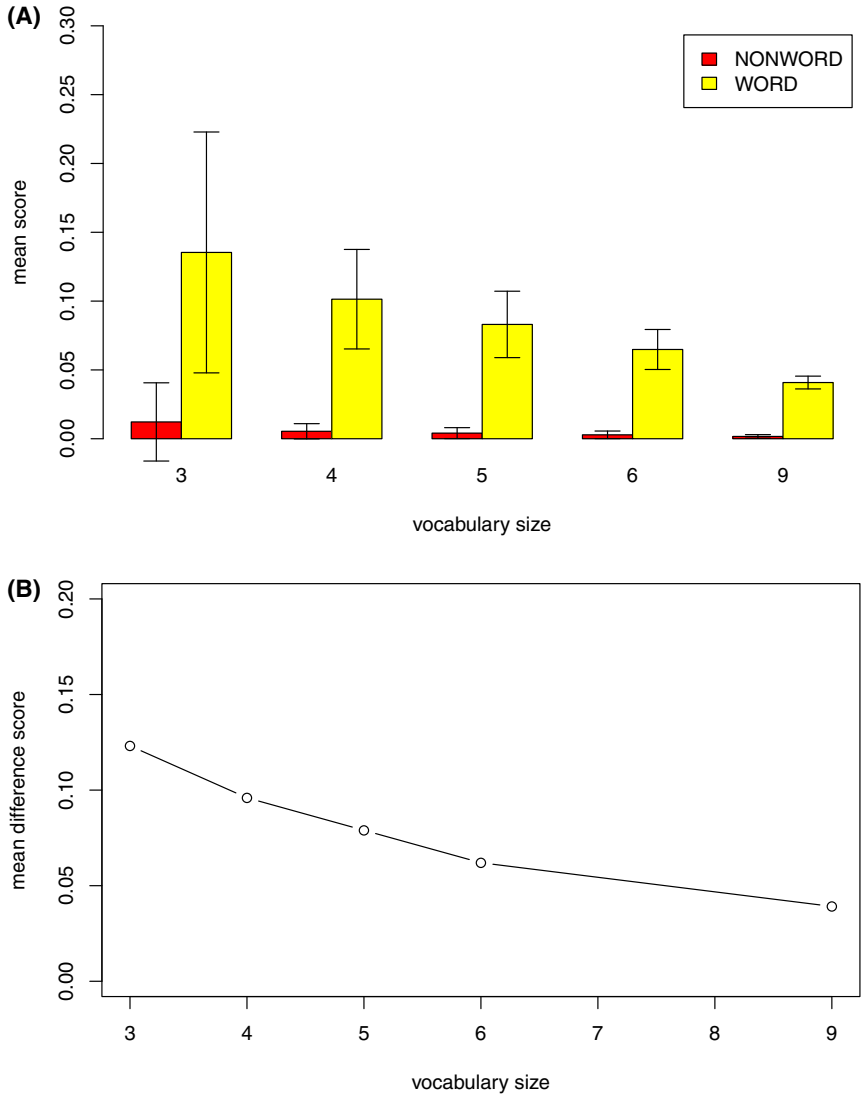

Fig. 5. Discrimination between words and part-words by U-MILA after training on a corpus of constant length, composed of words from a cadre of differing size (cadres of 3, 4, 5, 6, and 9 words). As in humans (Frank et al., 2010, exp. 3), the model achieves better discrimination after being trained on small word cadres. This result holds only for a certain range of parameters (see text). (A) The mean probability scores assigned to words and to part-words for each condition. (B) The difference between the mean probability scores for words and part-words for each condition.

Although U-MILA does not commit to "crisp" categorical distinctions among units (see Section 3.2), the similarity relations that it builds up can be used to cluster words into categories. After training, U-MILA correctly recognized all three-letter words, in both languages, as such, making the similarity scores among them immediately available. Similarity scores between words of which one or both did not appear in the training corpus were defined as an equally weighted sum of the similarity scores between their components; thus, the similarity between *abc* and *def* was defined as (sim(*a,d*) + sim(*b,e*)

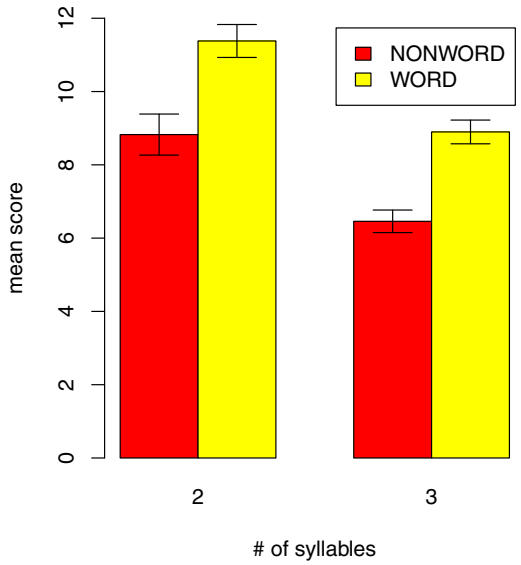

Fig. 6. The mean log-probability scores assigned to bi- and tri-syllabic words and non-words by U-MILA after training on a phonetically encoded corpus of natural child-directed speech (see text). The test set was composed of 496 words and 496 non-words (sequences that straddled word boundaries) that occurred in the training set.

+ sim(c,f))/3.[10] A clustering algorithm (Matlab procedure *linkage* with default values of the parameters) was applied to the resulting similarity matrix among all words in both languages. A dendrogram plot of the cluster structure (Fig. 7) indicated that the model correctly classified all the words, including novel words that did not appear in the training corpus.

3.3.10. Word segmentation: Effects of frequency and transitional probability (French et al., 2011, simulation 11)

To explore learning based on backward transition probabilities, French et al. (2011) constructed a dataset similar to those previously discussed, composed of a random sequence of two-syllable words, all of which had the same frequency of occurrence and were included in the test. The training sequence was constructed so that words and non-words had the same forward transition probabilities; the within-word backward transition probabilities were higher than for non-words (1 as opposed to 0.25). The TRACX model was trained on this corpus and learned words significantly better than non-words. French et al. (2011) also reported a behavioral experiment with 8-month-old infants, using a similarly structured dataset, in which the subjects successfully differentiated between words and non-words.

We trained U-MILA on the same corpus and had it assign probabilities to each of the words and non-words in it. The model differentiated between the two groups successfully, assigning words a mean probability of 0.0094, compared to 0.0035 for non-words. An analysis of variance (R procedure *lm*) indicated that this difference is significant (t = 2.213, p < .04; Fig. 8).

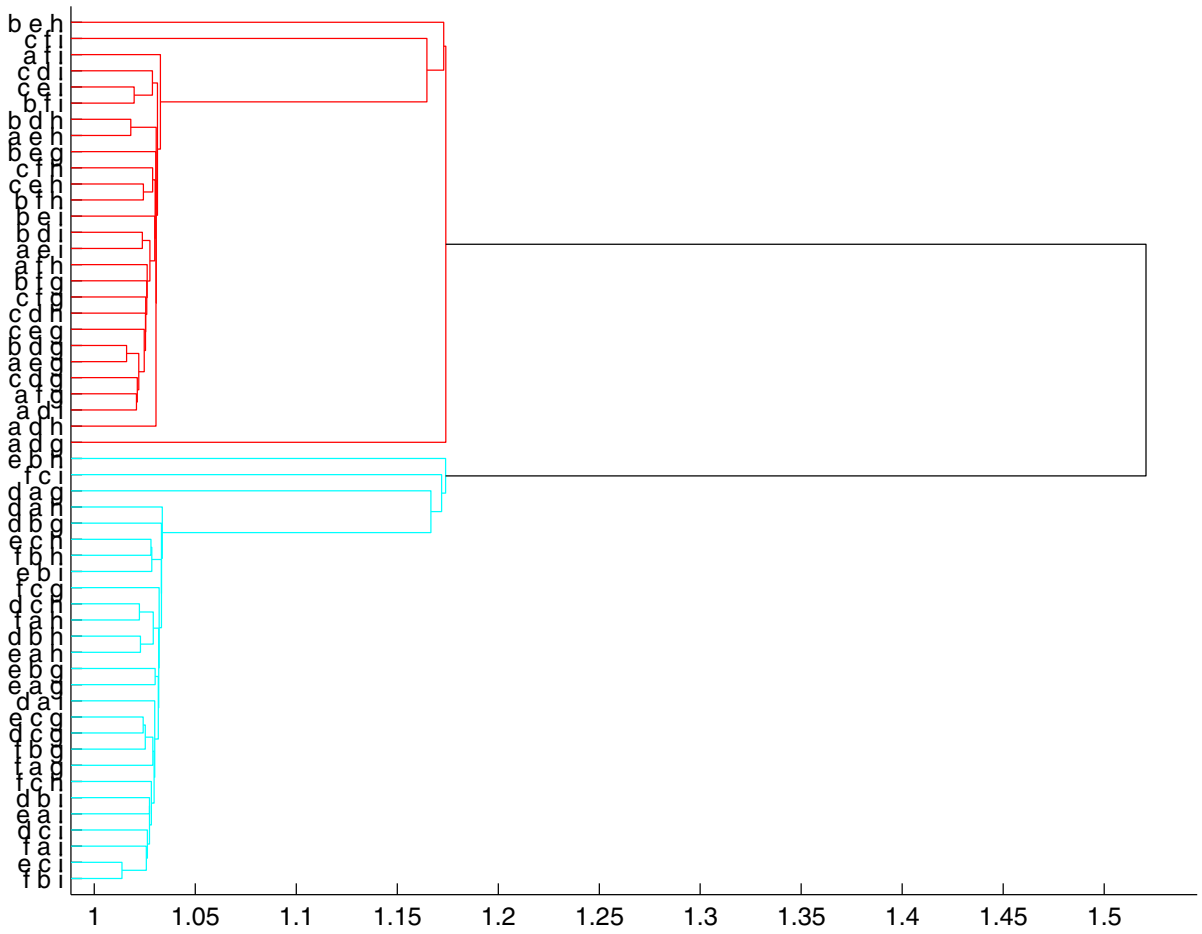

Fig. 7. Clustering of tri-syllabic words, according to the similarity values assigned by UMILA, illustrating implicit categorization. The items, containing both words that appeared in the training set and novel words, belong to two artificial micro-languages, of which the training set was composed. The clustering reveals two clades that discriminate between the two languages with no errors.

3.4. Non-adjacent dependencies (Gómez, 2002, experiment 1)

Gómez (2002) reported that both adults and infants can learn non-adjacent dependencies in an artificial language, solely from statistical cues, and that they do so most successfully in a setting in which the adjacent dependencies are the least reliable. We trained one instance of the U-MILA model on each of the datasets used by Gómez (2002); see Table 2. Each learner was evaluated by the probability scores it assigned to each of 12 sentences, six of which were taken from the language it had been exposed to, and six from the other language.

The results are summarized in Fig. 9. Non-adjacent dependency structure was successfully learned by all learners in all conditions. An analysis of covariance (R procedure *lm*) yielded significant effects of grammaticality (i.e., whether or not the sentences followed the rules of the training set's language) and pool size ($t = -22.7$, $p < 2 \times 10^{-16}$; $t = -14.4$, $p < 2 \times 10^{-16}$) and a significant interaction ($t = 3.0$, $p < .0045$).[11] There

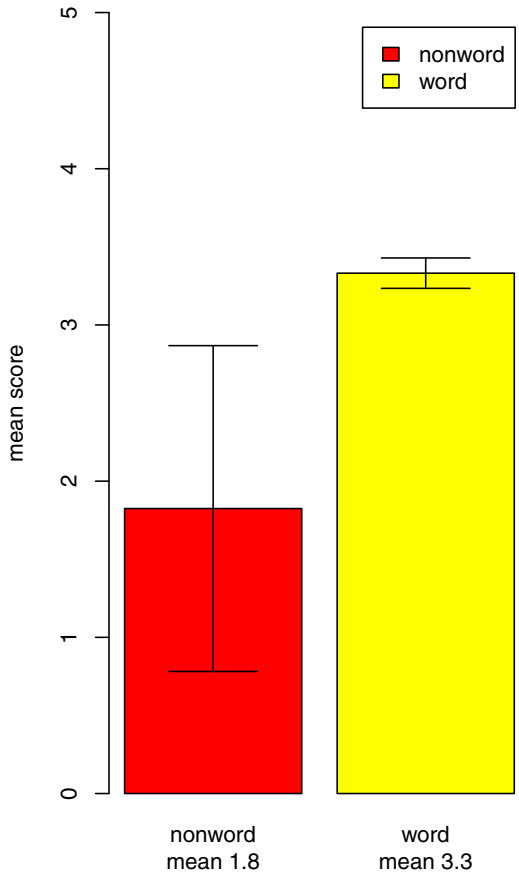

Fig. 8. Mean log-probability scores assigned to words and non-words following a training set in which words differed from non-words only in their backward transition probabilities.

was, however, no abrupt change in performance between pool sizes 12 and 24, contrary to the effect reported by Gómez (2002). This finding supports Gomez's proposed explanation of that effect, according to which the difference between her subjects' performance for pool sizes 12 and 24 is an outcome of human learners' switching between different learning mechanisms in response to a change in the nature of statistical cues in the data —a switch that is not implemented in our model, which by default always applies both adjacent and non-adjacent learning mechanisms (see Section 2.4). In further support of this explanation, the model failed to differentiate between grammatical and ungrammatical sentences in all four set sizes when running in "bottom-up collocation" mode, in which it learns using only adjacent transition probabilities.

3.5. Syntactic categories (Gómez & Lakusta, 2004, experiment 1)

Gómez and Lakusta (2004) showed that infants are capable of unsupervised learning of syntactic categories and rules in an artificial language (see Table 3). We trained a U-MILA instance on a training set patterned after that of Gómez and Lakusta (2004), with spaces inserted between each two consecutive syllables and a random ordering of

Table 2
The stimuli used by Gómez (2002), experiment 1

		Test Strings	
Language 1	Language 2	Language 1	Language 2
S→{*aXd*	S→{*aXe*	*pel wadim rud*	*pel wadim jic*
bXe	*bXf*	*vot wadim jic*	*vot wadim tood*
cXf }	*cXd*}	*dak wadim tood*	*dak wadim rud*
X → x1, x2, . . ., xn; *n* = 2, 6, 12 or 24		*pel kicey rud*	*pel kicey jic*
		vot kicey jic	*vot kicey tood*
		dak kicey tood	*dak kicey rud*

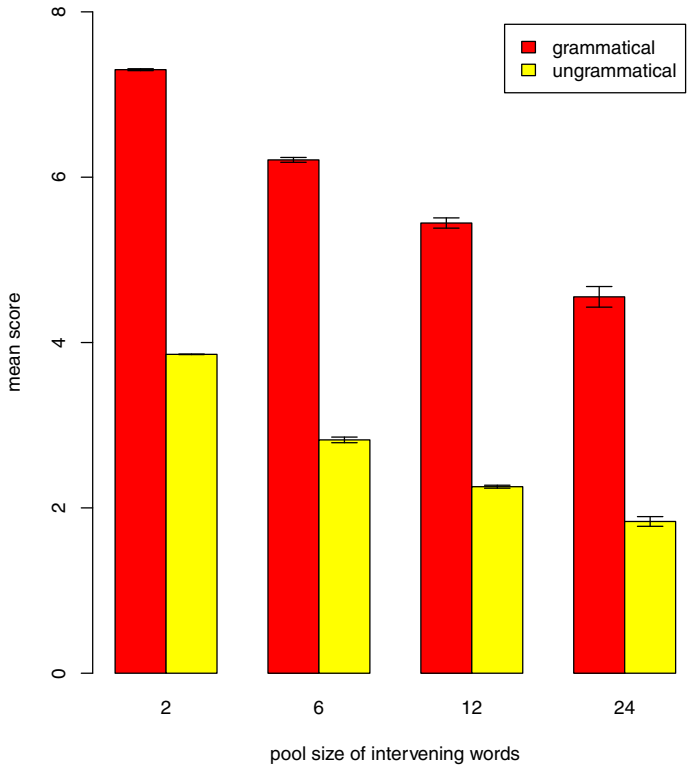

Fig. 9. Mean log-probability scores assigned to grammatical and ungrammatical sentences from an artificial language with long-range dependencies between words, with a single intervening word between them, for different sizes of the word pool from which the intervening words were taken during the training (2, 6, 12, 24). Grammatical sentences are significantly preferred by U-MILA in all conditions, contrary to the finding by Gómez (2002), in which the preference was significant only for the largest word pool size. This is in accord with Gomez's explanation of the finding (see text).

the sentences. The learner then assigned a probability score to each of the test sentences in Gómez and Lakusta (2004). The model's parameter that controls its sensitivity to slot filler length, $B_{\text{FillerSetSizeSensitivity}}$ (see Section 2.2), was set so the learner would be sensitive to the filler set size, measured in syllables.

Table 3
The word categories used by Gómez and Lakusta (2004), experiment 1

a	*b*	*X*	*Y*
alt	*ong*	*coo mo*	*deech*
ush	*erd*	*fen gle*	*ghope*
		ki cey	*jic*
		lo ga	*skige*
		pay lig	*vabe*
		wa zil	*tam*

Sentences from the training language (L1) were assigned higher scores than sentences from L2. An analysis of variance (R procedure *aov*) indicated that this difference was significant ($F = 49.1$, $p < 8.9 \times 10^{-09}$; see Fig. 10). The model's success is due to the alignment mechanism, which creates collocations of the form *alt* ___ ___ *ong*, and *ong* ___ *alt,* that can be thought of as describing rules regarding non-adjacent dependencies. In the test phase, it thus assigns higher scores to sequences that conform to these patterns, even if the slot contains unfamiliar syllables.

3.6. Variation sets (Onnis, Waterfall, and Edelman, 2008, experiment 1)

Onnis et al. (2008) examined the effects of variation sets[12] on artificial grammar learning in adults. As in that study, we trained multiple instances of U-MILA (100 learners), simulating individual subjects, on 105 sentences (short sequences of uni- and disyllabic "words" such as *kosi fama pju*, presented with word boundaries obliterated by introducing spaces between each two syllables: *ko si fa ma pju*). For half of the simulated subjects, 20% of the training sentences formed variation sets in which consecutive sentences shared at least one word (Varset condition); for the other half, the order of the sentences was permuted so that no variation sets were present (Scrambled condition). After training, learners scored disyllabic words and non-words in a simulated lexical decision task.

As with the human subjects, learning occurred in both conditions, with the model demonstrating better word/non-word discrimination (e.g., *fa ma* vs. *si fa*) in the Varset condition, compared to the Scrambled condition (see Fig. 11). A mixed- model analysis of the data, with subjects and items as random effects (R procedure *lmer*), yielded significant main effects of word-hood ($t = 13.7$, $p < .0001$; all p values estimated by Markov Chain Monte Carlo sampling with 10,000 runs, procedure *pvals*, R package *languageR*), and condition ($t = -69.8$, $p < .0001$). Crucially, the word-hood × condition interaction was significant ($t = 57.8$, $p < .0001$).

As expected, the presence of this interaction depended on the value of the phonological loop decay parameter: With slower decay (0.035 compared to 0.075, corresponding to a wider time window in which overlaps are sought), variation sets made no difference on learning the distinction between words and non-words. The length of the phonological loop also influenced the results: The effect of variation sets depended on sentences that

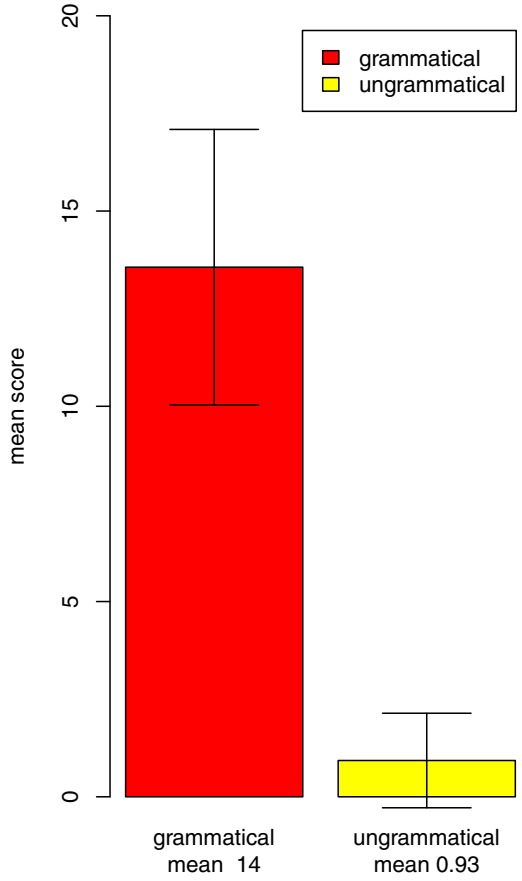

Fig. 10. Mean log-probability scores assigned to grammatical and ungrammatical sentences from an artificial language with long-range dependencies between words. Grammatical and ungrammatical sentences differ in the number of syllables (1 or 2) separating the two dependent elements. Similar to infants in experiment 1 of Gómez and Lakusta (2004), U-MILA successfully differentiates between sentences with different lengths of dependency, even when these contain novel intervening syllables.

form a variation set being simultaneously present within the loop (in addition to not decaying too quickly).

3.7. Structure dependence: Auxiliary fronting (Reali & Christiansen, 2005)

Reali and Christiansen (2005) set out to demonstrate that choosing which instance of the auxiliary verb to front in forming a polar interrogative—as, in the example below, transforming *The man who is hungry is ordering dinner* into form (b) rather than form (a)—is amenable to statistical learning. In their experiment 1, they trained a bigram/trigram model, using Chen-Goodman smoothing, on a corpus of 10,705 sentences from the Bernstein-Ratner (1984) corpus. They then tested its ability to differentiate between correct and incorrect auxiliary fronting options in 100 pairs of sentences such as:

a. Is the man who hungry is ordering dinner?
b. Is the man who is hungry ordering dinner?

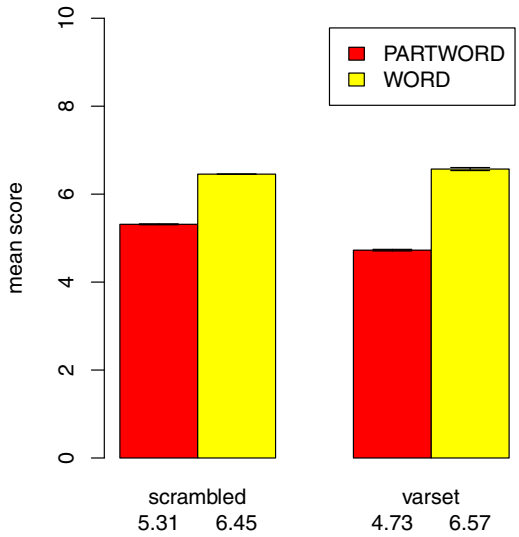

Fig. 11. The mean log-probability scores assigned to words and to non-words after training in one of two conditions, which differed only in the order of sentence presentation. In the *Variation set* condition, a lexical overlap was present in 20% of adjacent sentences; in the *Scrambled* condition, there were no such overlaps. Similar to human participants in Onnis et al. (2008), U-MILA discriminates significantly better between words and part-words in the Variation set condition (*right*).

Their training corpus is composed of sentences uttered by nine mothers addressing their children, recorded over a period of 4–5 months, while the children were of ages 1:1–1:9. The corpus does not contain explicit examples of auxiliary fronting in polar interrogatives. In a forced-choice test, the n-gram model of Reali and Christiansen (2005) chose the correct form 96 of the 100 times, with the mean probability of correct sentences being about twice as high as of incorrect sentences.

We trained U-MILA on all the sentences made available to us by Reali and Christiansen (10,080 sentences for training and 95 pairs of sentences for testing). When forced to choose the more probable sentence in each pair, U-MILA correctly classified all but six sentence pairs, and the mean probability of correct sentences was higher than that of incorrect sentences by nearly two orders of magnitude (see Fig. 12; note that the ordinate scale is logarithmic). An analysis of variance (R procedure *aov*) confirmed that this difference was highly significant ($F = 26.35$, $p < 7.08 \times 10^{-07}$).

3.8. Structure dependence: Island constraints and long-range dependencies (Pearl & Sprouse, 2012)

In the second experiment addressing issues of structure dependence, we examined the ability of U-MILA to learn grammatical islands—structures that, if straddled by a long-distance dependency following a transformation, greatly reduce the acceptability of the resulting sentence (Sprouse et al., 2012a; see footnote for an example). Recently, Sprouse, Fukuda, Ono, and Kluender (2011) conducted a quantitative study of the interaction between grammatical island constraints and short- and long-term dependencies in

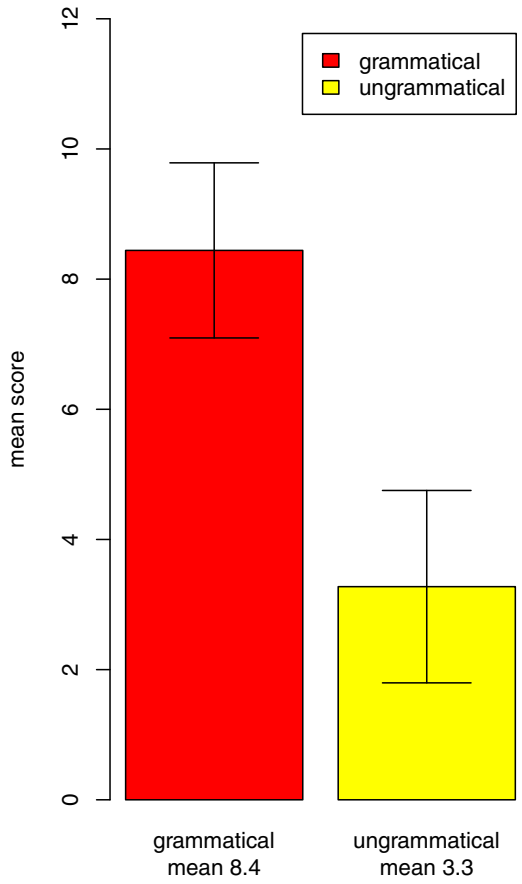

Fig. 12. The mean scores assigned to grammatical and non-grammatical instances of auxiliary verb fronting, following training on a corpus of natural child-directed speech that did not contain explicit examples of auxiliary fronting in polar interrogatives (logarithmic scale; following Reali & Christiansen, 2005). In a forced-choice preference test, 89 of 95 pairs of grammatical and ungrammatical instances of auxiliary verb fronting were classified correctly.

determining sentence acceptability. They used a factorial design, with four types of sentences: (a) short-term dependency + no island, (b) long-term dependency + no island, (c) short-term dependency + island, and (d) long-term dependency + island.[13] The pattern of acceptability judgments exhibited the signature of the island effect: an interaction between island occurrence and dependency distance. In other words, the acceptability of a sentence containing both a long-term dependency and an island was lower than what would have been expected if these two effects were independent. This finding opened an interesting debate regarding its implications for reductionist theories of language (Hofmeister, Casasanto, & Sag, 2012a,b; Sprouse et al., 2012a; Sprouse, Wagers, & Phillips, 2012b).

In an attempt to account for this finding by a statistical learning model, Pearl and Sprouse (2012) trained a parser to recognize shallow phrasal constituents in sentences represented as trees of part of speech (POS) tags, while collecting the statistics of "container node" trigrams covering these parses, with container nodes defined as nodes in a

phrase structure tree that dominates the location of the gap left by extraction. With proper smoothing, such a model can simulate acceptability judgments by assigning probabilities to sentences. The model was trained on 165,000 parses of sentences containing island dependencies, drawn from a distribution mirroring that of various island structures in natural language. When tested on a set of sentences that crossed multiple island types with short and long dependencies, the model qualitatively reproduced the empirical finding described above.

We attempted to replicate this result, hypothesizing that the collocations that U-MILA learns, which are in a sense analogous to trees of POS n-grams, may lead to the emergence of an interaction between islands and dependency length (something of a long shot, to be sure). For this purpose, we tested the instance of U-MILA that had been trained on the first 15,000 sentences of the Suppes (1974) corpus (see Section 3.1) on the same set of sentences as described above (four types of islands types, five factorial blocks in each, four sentences in each block). All sentences were patterned after the test set described in Pearl and Sprouse (2012); words that did not occur in the training corpus were replaced with words of the same part of speech that did. U-MILA assigned probabilities to each of the test sentences, which we then analyzed and plotted as in Pearl and Sprouse (2012). No significant interaction between island presence and dependency length was found for any of the four island types, and there was no consistent trend regarding the direction of a potential interaction. Further probing showed that the results were strongly affected by replacement of certain units in the sentences with grammatically analogous counterparts (e.g., replacing *Nancy* with *she*). We believe that this source of noise in estimating sentence probability, combined with the relatively small training set (much smaller than that used by Pearl & Sprouse, 2012), may explain the failure of our model to replicate the island effect.

4. General discussion

In this section, we discuss the representational power and learnability of U-MILA's graph architecture, suggest some ideas for improving its performance, and outline two key directions for future development.

4.1. Representational power and learnability

4.1.1. On graph-like formalisms in language and other sequential tasks

A potential strength of a graph-like representation is its immediate compatibility with basic associative learning principles, which makes it especially useful for modeling the incremental evolution of complex cognition from simple beginnings (e.g., Lotem & Halpern, 2012). In fact, we are now using the U-MILA platform to pursue such evolutionary research (Kolodny et al., submitted). In language, earlier theoretical approaches that posited a graph-structured grammar, such as the Neurocognitive Grammar of Lamb (1998)[14] and the Word Grammar of Hudson (2007), did not specify how the graph

should be learned from experience. The first such approach that did learn and that scaled up to large corpora of natural language was the ADIOS algorithm (Solan et al., 2005). Learning in U-MILA is more realistic than in ADIOS, because it does not require access to (let alone multiple passes over) the entire training corpus, and because it consists of incremental, experience-driven modifications of "synaptic" weights between graph nodes, rather than all-or-none rewiring of subgraphs as in ADIOS. For this reason, the U-MILA graph can immediately and at all times serve as a probabilistic generative language model rather than requiring a separate training phase as in ADIOS.

While its reliance on a graph makes U-MILA trivially "connectionist," it is different from the connectionism of approaches based on the popular SRN idea (Christiansen & Chater, 1999; Elman, 1990) in several key respects. Unlike the SRN architecture, the graph learned by U-MILA is incrementally built and hierarchically structured; it also consists of several kinds of nodes and links, which fulfill specific functional needs, rather than being the same all over the network. U-MILA is also distinguished by its ability to both accept and generate actual natural language (as opposed to a microlanguage targeting a specific phenomenon being modeled).

4.1.2. Relationships to formal syntax

When attempting to relate U-MILA (or any other heuristically specified model) to formal language theory, both the expressive power and the learnability on each side of the comparison need to be addressed. In particular, a rigorous comparison between two formalisms, such as U-MILA and, for instance, the probabilistic context-free grammar (PCFG), and their associated learning algorithms, must involve a formal reduction of one to the other (Hopcroft & Ullman, 1979). Attempting such a reduction would, however, take us too far away from the main thrust of the present project, which is why we offer instead some informal analogies and observations, in the hope that these would serve as a useful starting point for a more rigorous exploration in the future.

The U-MILA model transcends the power of a finite state automaton by making use of slot collocations (see Section 2.2). Specifically, when a slot in a collocation is encountered during graph traversal, activation shifts to its filler, subsequently returning to the slot collocation's latter part. Because fillers may be slot collocations themselves (including self-referentially), U-MILA can learn and represent an infinite center-embedded recursion, as illustrated in Fig. 13, thus producing a PCFG language.[15] Note that in the current implementation, successful learning of such grammars depends on the model's parameter values and also on the structure of the training set (see details in Appendix S3 and S4).

4.2. Performance highlights and lessons for the model

We now briefly recap and discuss the model's performance in the tasks to which is was applied, grouped into the same five-study order in which it was presented in Section 3.

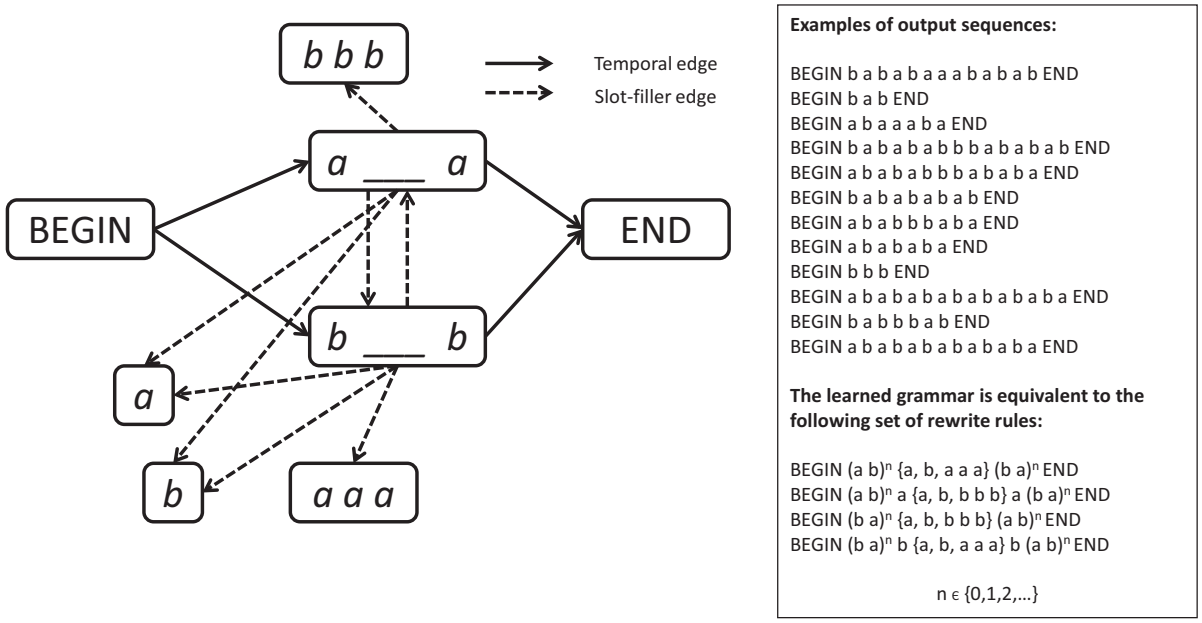

Fig. 13. A grammar learned by U-MILA and the rewrite rules that correspond to it. The graph is a simplified version of the full representation constructed by the model.

4.2.1. Generative ability (Study 1)

The generative abilities of the model—perplexity (recall) and precision—were tested after training it on an unannotated corpus of 15,000 natural child-directed utterances (Suppes, 1974), available as part of CHILDES (MacWhinney, 2000). Although U-MILA's perplexity score fell short of that of a standard tri-gram model (Stolcke, 2002) trained on the same corpus, its precision was significantly higher. This was true both when the smoothing parameters of the tri-gram model were set such that both models achieve the same perplexity (Section 3.1) and when the tri-gram model was allowed to achieve its lowest perplexity and its precision compared to that of U-MILA with the "standard" settings (Data S1, 1.1).

4.2.2. Characteristics of learned representations (Study 2)

While it avoids separating words and larger nodes into crisp categories (see Section 3.2), U-MILA supports intuitively satisfying clustering (Table 1; Fig. 2). Insofar as the model does not distinguish between syntax and semantics, the clusters it produces should be seen as ad hoc categories, based on a particular choice of similarity parameter settings, rather than syntactic categories (parts of speech) or semantic (e.g., thematic) ones; a different choice of similarity parameters, stemming, for instance, from different presets for "conceptual" nodes (as in the translation model of Edelman & Solan, 2009), would lead to a different, task-specific pattern of similarities. Retention of wider contextual information about each node is expected to significantly improve the performance of future versions of U-MILA in this field (see Section 1.2 in Appendix S2).

4.2.3. Sequence segmentation and chunking (Study 3)

We believe that U-MILA's success in replicating the entire range of empirical findings in segmentation and chunking tasks is due to its reliance on both mechanisms of collocation construction available to it: the "bottom-up" mechanism based on statistical significance of sequential co-occurrence and the "top-down" mechanism based on recurrence within a short time window. The latter, we observe, is particularly useful for finding sparsely occurring patterns, while the former reflects traditional associative learning and is more effective for more common patterns. In practice, we found that disabling the bottom-up mechanism prevents the model from replicating some of the results of French et al. (2011; see Section 3.3); at the same time, the difference in learning between scrambled and variation-set conditions in the Onnis et al. experiment (2008; Section 3.6) cannot be reproduced without the "top-down" mechanism. We note that the two mechanisms may have different costs in terms of cognitive resources, and the balance between them could be governed by a combination of internal factors and the characteristics of the learning task at hand.[16] U-MILA's default use of both mechanisms may render it too powerful compared to a human learner, preventing it from accounting for some empirical findings such as that of Gómez (2002; see also below).[17]

Our model's successful replication of the variation set effect (Onnis et al., 2008) depended on another parameter whose adjustment—in real time or over the evolutionary process—is important. This parameter, the length of the short-term memory queue, or phonological loop, should fit the typical distribution of unit recurrence in the data stream to which the learner is exposed (Goldstein et al., 2010; Lotem & Halpern, 2008, 2012). In natural child-directed speech, for example, the effective time window in which recurrences are sought should correspond to the typical length of a bout of interaction whose subject remains constant, as in *What is it? It is a doll. Do you like it?*—two to four utterances or so. A longer window might lead to spurious alignments, while a shorter one would not allow the extraction of the recurring element (in this case, *it*). The range of settings of this parameter can conceivably be selected by evolution (c.f. Christiansen & Chater, 2008; and see Lotem & Halpern, 2012), while the precise setting for a given learner could also be a function of its recent experiential history and context.

4.2.4. Artificial language learning (Study 4)

The two mechanisms that allowed U-MILA to replicate the segmentation and chunking results as mentioned above seem to correspond to the two types of learning strategies posited by Gómez (2002) in her discussion of learning of artificial language rules by human subjects. Balancing these two mechanisms dynamically (as suggested above) is what may underlie the subjects' apparent switching between learning based on adjacent transition probabilities to learning based on long-distance dependencies, as proposed by Gomez. At present, U-MILA relies equally on both the bottom-up and the top-down mechanisms, which allows it to learn successfully in all the conditions of the Gómez (2002) experiment, as opposed to humans, who seem to use the former mechanism as a default and switch to the other only when it is obvious that the first one is not working. This switching need not be all-or-none: It may be implemented as a gradual change in

prior probabilities of collocation creation (in particular P_{col}) while adjusting the decay parameter of the phonological loop.

Our model's reproduction of the results of Gómez and Lakusta (2004) is made possible by just such a switch: U-MILA succeeds in this task when it is set to be sensitive to slot fillers' set size (see Section 2.2; this is the only instance in which a major change in the model's parameter values away from the "standard" setting was necessary). Whether the learner should be sensitive to this value or not may depend on the statistics of the data set in question (for instance, it may make sense for one natural language but not for another). A data- and performance-driven mechanism that would adjust this parameter seems realistic and can be implemented in a straightforward and biologically feasible way.

4.2.5. Learning structure dependence (Study 5)

Although U-MILA replicated the result of Reali and Christiansen (2005) in learning auxiliary fronting in polar interrogatives, the conceptual significance of that finding has been disputed (Kam, Stoyneshka, Tornyova, Fodor, & Sakas, 2007). We agree that learning-based approaches will be effective in countering the "Poverty of the Stimulus" argument for the innateness of language (Chomsky, 1980; Legate & Yang, 2002; Pullum & Scholz, 2002) only if they succeed in replicating a wide range of structure-related syntactic phenomena (see, e.g., Phillips, 2003, 2010). A set of syntactic island (Ross, 1967) phenomena, which manifest psycholinguistically as an interaction between two graded acceptability effects (that of dependency length and that of the presence of an intervening island), could not be replicated by U-MILA in this study. We ascribe U-MILA's failure to exhibit this interaction to the relatively short training that it underwent (see Section 3.8). We are encouraged, however, by the success in this task of a model of Pearl and Sprouse (2012), which is based on the statistics of a massive amount of phrase structure tree data. Insofar as this representation is functionally similar to U-MILA's collocations, our model too should fare better when trained on a larger corpus.

4.3. Future directions

4.3.1. Incremental improvements to the present model

There are at least two ways in which better use can be made of U-MILA's short-term memory queue, or the phonological loop. The first, and most straightforward, could undertake segmentation "on the fly" of transient sequences of items passing through the queue, using existing units for guidance. For example, while reading the sequence *a l l y o u n e e d i s l o v e* and given previous familiarity with the units *you* and *is,* the model would be able to infer that *need* is likely also a meaningful unit. This feature could be especially useful in modalities where the probability of noise is relatively low, as in phonetics, where most phonemes within an utterance are non-random; it might be less so in visual tasks such as a fish scanning the sea bottom for food.

The second way in which the short-term memory queue can be made a better use of has to do with exploiting more fully the idea that events that recur within a short time

window are likely to be significant and worth paying special attention to. U-MILA now assigns a special weight to the recurrence of a unit or a sequence; it could also mark the recurrence of a certain temporal relation or to the interchange of one unit with another in a certain context. Thus, encountering *the big ball* and *a blue ball* within a short time period suggests that the similarity index between *big* and *blue* should be increased more than if these two events were widely separated in time. The present implementation does this only with regard to interchange events that take place among nodes as they take on the role of fillers within a slot collocation (as governed by the calculation of *WSI*; see Section 2.5).

4.3.2. The next key step: Learning in parallel from multiple modalities

While U-MILA is architecturally and functionally realistic in many important respects, its ability to model learning in natural biological systems and situations is limited by its exclusive reliance on a single modality. Thus, when applied to language, it can process a stream of symbolically encoded phonemes (or, of course, a stream of text characters), but not, say, parallel streams of phonemes, prosody, and visual cues—a rich, multimodal, synchronized flow of information that is available to human learners of language (Goldstein et al., 2010; Smith & Gasser, 2005).

Integrating multiple cues to boost performance in tasks such as categorization or word learning is, of course, a deservedly popular idea in cognitive science (e.g. Frank, Goodman, & Tenenbaum, 2009; Yu & Ballard, 2007; Yu & Smith, 2007). Our focus on learning a grammar of dynamic experience (which in the project reported here was limited to language) does, however, introduce a number of conceptual complications, compared to "static" tasks such as categorization. Some of these challenges, such as the need to represent parallel sequences of items or events, we have already begun to address (see the discussion of the possible use of higraphs for this purpose in Edelman, 2011). A full treatment of those ideas is, however, beyond the scope of the present paper.

4.3.3. Interactive and socially assisted learning

The human language learner's experience is not only decidedly multimodal but also thoroughly interactive and social (see Goldstein et al., 2010; for a review). Babies learning language do so not from a disembodied stream of symbols: They simultaneously process multiple sensory modalities, all the while interacting with the world, including, crucially, with other language users. The key role of the interactive and social cues in language acquisition (which are also important in birdsong learning, for instance) is now increasingly well documented and understood. Our model at present incorporates such cues only in a limited and indirect fashion. In particular, variation set cues, which U-MILA makes use of, are presumably there in the input stream because of the prevalence of variation sets in child-directed language (Waterfall, 2006). Other aspects of the model that may be adjusted by social interaction are the parameters of weight and memory decay. These are likely to be tuned according to emotional or physiological states that may indicate how important the incoming data are, and therefore how much weight it should receive and for how long it should be remembered (Lotem & Halpern, 2012). We

expect the next version of the model, which will be capable of dealing in parallel with multiple streams of information, to do a much better job of replicating human performance in language acquisition.

5. Conclusion

In cognitive modeling (as in computer science in general), it is widely understood that the abilities of a computational model depend on its choice of architecture. The focus on architecture may, however, hinder comparisons of performance across models that happen to differ in fundamental ways. The question of modeling architecture would be sidelined if a decisive, computationally explicit resolution of the problem of language acquisition (say) became available, no matter in what architecture. In the absence of such a resolution, the way ahead, we believe, is to adopt a systematic set of design choices—inspired by the best general understanding, on one hand, of the computational problems arising in language and other sequentially structured behaviors, and, on the other hand, of the characteristics of brain-like solutions to these problems—and to see how far this approach would get us. This is what the present project has aimed to do.

In this paper, we laid out a set of design choices for a model of learning grammars of experience, described an implemented system conforming to those choices, and reported a series of experiments in which this system was subjected to a variety of tests. Our model's performance largely vindicates our self-imposed constraints, suggesting both that these constraints should be more widely considered by the cognitive science community and that further research building on the present efforts is worthwhile. The ultimate goal of this research program should be, we believe, the development of a general-purpose model of learning a generative grammar of multimodal experience, which, for the special case of language, would scale up to life-size corpora and realistic situations and would replicate the full range of developmental and steady-state linguistic phenomena in an evolutionarily interpretable and neurally plausible architecture.

Acknowledgments

We are grateful to Andreas Stolcke for sharing his code and advice, and to Jon Sprouse, Lisa Pearl, Florencia Reali, and Morten Christiansen for sharing their data. We thank Haim Y. Bar for statistical advice, and Colin Phillips and Roni Katzir for fruitful discussions and helpful suggestions. We thank Doron Goffer, Amir Zayit, and Ofer Fridman for their help and insights with regard to the technical aspects of this project. We also thank Amy Perfors and two anonymous reviewers for their constructive comments and suggestions regarding earlier versions of this manuscript. O.K. was supported by a Dean's scholarship at Tel Aviv University and a scholarship from the Wolf Foundation. The project was supported in part by the Israel Science Foundation grant no. 1312/11 to A.L.

Notes

1. U-MILA stands for Unsupervised Memory-based Incremental Language Acquisition.
2. Some applications of our approach to other modalities are described elsewhere (Menyhart et al., unpublished data; Kolodny et al., 2014); the extension of the model to multimodal inputs is left for future work.
3. Edge weights are only normalized so as to become proper probabilities if the need arises to estimate the probability of a candidate sequence; see Sections 2.6 and 2.7.
4. We remark that the model was implemented (in Java) as a proof of concept, without any attempt at algorithmic optimization.
5. The present implementation of the model allows assigning different relative weights to be assigned to these three data types, but in all runs reported in this paper the weights were equal. Optimizing these with regard to the specific nature of the data may lead to an improvement in the similarity measure, but was set aside for future exploration.
6. The motivational mechanisms that initiate and end the production process are beyond the scope of this paper, but we assume that as in simple foraging tasks, the agent is first motivated to activate a familiar starting point from which it navigates through various potential paths offered by the network until it reaches a familiar goal. Obviously, this implies that a realistic production process also includes steps designed to fit the sentence to the specific goal and context, not only to make it grammatically and logically correct. Note that biological realism requires that nodes in the representation interact with one another only locally. U-MILAU-MILA's production process adheres to this principle.
7. For example, if a grammar contains the node "I" (1), "want" (2), "to" (3), "break" (4), "free" (5), "I want" (6) and "break free" (7), then possible covers of the sentence "I want to break free" are [(1) + (2) + (3) + (4) + (5)], [(6) + (3) + (4) + (5)], [(1) + (2) + (3) + (7)], and [(6) + (3) + (4) + (5)]. For a similar approach, see Scha, Bod, and Sima'an (1999, Section 4).
8. This is due to a frequency difference in the training set between first syllables of words compared to first syllables of non-words: The latter were more frequent. Because the probability estimation procedure (Section 2.7) takes into account the absolute probability of occurrence of the first syllable in the sequence, the frequency difference in favor of non-words balanced the higher internal transition probabilities in words, and the overall effect was that words and non-words were assigned similar probabilities.
9. In contrast, the version of the model that was applied to birdsong (Menyhart et al., unpublished data) does implement this step, and thus eliminates from the grammar units that are wholly contained in others if the weights of the two units (a proxy of their frequency of occurrence) differ by less than a certain threshold (e.g., 10%). In this manner, wholly contained units are eliminated, unless they occur in other contexts as well. This solution seems somewhat artificial and should probably be replaced by a probabilistically motivated weight updating scheme.

10. This is equivalent to using Levenshtein distance over strings (e.g., Ristad & Yianilos, 1998).
11. This interaction amounted to a small (in absolute terms) difference in the slopes of the grammaticality effect, rather than in a change in the sign of the effect. As such, it does not reflect on the rest of the discussion of this experiment.
12. A variation set is a series of utterances that follow one another closely and share one or more lexical elements (Küntay & Slobin, 1996; Waterfall, 2006).
13. An example of such a factorial design:
    a. Who __ heard that Lily forgot the necklace? (short-distance dependency, non-island structure)
    b. What did the detective hear that Lily forgot __ ? (long-distance dependency, non-island structure)
    c. Who __ heard the statement that Lily forgot the necklace? (short-distance dependency, island structure)
    d. What did the detective hear the statement that Lily forgot __ ? (long-distance dependency, island structure)
    For a definition and overview of the island phenomena, see Sprouse et al. (2011).
14. Lamb mentions in passing a point that in our opinion is central, namely, that language is a proper subset of a broader category of sequential behaviors: "Those who think it is marvelous that we can produce a new sentence that we have never heard or said before—do they also think it is marvelous that we can go through a cafeteria line and select a meal that we have never eaten before?" (Lamb, 1998, p. 205; c.f. Lashley, 1951).
15. We thank the anonymous referees for pointing out the importance of explicitly demonstrating the model's ability to learn such grammars.
16. The ability of biological learning systems to adjust learning parameters in real time, based on an ongoing estimate of performance, or over a longer time frame, based on contextual data (cognitive state, recent history, etc.), may explain the need that we encountered for the occasional minor adjustments of parameter values between tasks (see Appendix S5). Notably, however, this need arose only rarely and the changes were minor; the only exception is discussed below.
17. U-MILA is also too powerful, compared to a human learner, in that it makes no mistaken alignments. The effects of such mistakes, as well as the possible optimal tuning of the model's various parameters such as the length of the phonological loop and the rate of memory decay of the graph (Section 2.2), were not explored in this paper for reasons of scope.
18. We present the 20 most frequent nodes, because their statistics are the most extensive, and so their categories are likely to be meaningful, and 11 examples of slightly less frequent nodes, which provide some insight into the model's categorization (see main text). The symbol *s* that appears as a node or as part of a node pertains to the sequence *'s*, which is transcribed in the corpus as a stand-alone *s*, as in *that s a bunny*.

References

Adriaans, P., & van Zaanen, M. (2004). Computational grammar Induction for linguists. *Grammars*, *7*, 57–68.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*, 829–839.

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*(1), 158–173.

Barlow, H. (1990). Conditions for versatile learning, Helmholtz's unconscious inference, and the task of perception. *Vision Res*, *30*(11), 1561–1571.

Bates, D. (2005). Fitting linear mixed models in R. *R News*, *5*, 27–30.

Bernstein-Ratner, N. (1984). Patterns of vowel modification in motherese. *Journal of Child Language*, *11*, 557–578.

Bernstein-Ratner, N. (1987). The phonology of parent-child speech. In K. E. Nelson & A. van Kleek (Eds.), *Children's language*. Vol. 6 (pp. 159–174). Hillsdale, NJ: Erlbaum.

Bod, R. (2009). From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science*, *33*, 752–793.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*(1–2), 93–125.

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, *106*, 551–581.

Chomsky, N. (1957). *Syntactic structures*. The Hague, the Netherlands: Mouton.

Chomsky, N. (1980). *Rules and representations*. Oxford, England: Basil Blackwell.

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*(2), 157–205.

Christiansen, M. H., & Chater, N. (2001). Connectionist psycholinguistics: Capturing the empirical data. *Trends in Cognitive Sciences*, *5*, 82–88.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*, 489–509.

Clark, A. (2001). *Unsupervised language acquisition: Theory and practice*. D. Phil.: School of Cognitive and Computing Sciences, University of Sussex.

Cramer, B. (2007). Limitations of current grammar induction algorithms. In *Proceeding of the 45th annual meeting of the ACL: Student research workshop* (pp. 43–48).

DeMarcken, C. G. (1996). *Unsupervised language acquisition*. D. Phil.: MIT.

Dennis, S. (2005). A Memory-based theory of verbal cognition. *Cognitive Science*, *29*, 145–193.

Edelman, S. (2008a). *Computing the mind: How the mind really works*. New York: Oxford University Press.

Edelman, S. (2008b). On the nature of minds, or: Truth and consequences. *Journal of Experimental and Theoretical AI*, *20*, 181–196.

Edelman, S. (2011). On look-ahead in language: Navigating a multitude of familiar paths. In Bar, M. (Ed.) *Prediction in the Brain* (pp. 170–189). Oxford, England: Oxford University Press.

Edelman, S., & Solan, Z. (2009). Machine translation using automatically inferred construction-based correspondence and language models. In *Proceeding 23rd Pacific Asia Conference on Language, Information, and Computation (PACLIC)* (pp. 654–661). Hong Kong.

Edelman, S., & Waterfall, H. R. (2007). Behavioral and computational aspects of language and its acquisition. *Physics of Life Reviews*, *4*, 253–277.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science*, *6*, 205–254.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*, 107–125.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578–585.

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*(4), 614–636.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*, 1–58.

Giroux, I., & Rey, A. (2009). Lexical and sublexical units in speech perception. *Cognitive Science*, *33*(2), 260–272.

Goldsmith, J. A. (2007). Towards a new empiricism. *Recherches Linguistiques de Vincennes*, ed. J.B. de Carvalho, Available online.

Goldstein, M. H., Waterfall, H. R., Lotem, A., Halpern, J., Schwade, J., Onnis, L., et al. (2010). General cognitive principles for learning structure in time and space. *Trends in Cognitive Sciences*, *14*, 249–258.

Goodman, J. T. (2001). A bit of progress in language modeling. *Computer Speech and Language*, *15*, 403–434.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431–436.

Gómez, R. L., & Lakusta, L. (2004). A first step in form-based category abstraction by 12-month-old infants. *Developmental Science*, *7*, 567–580.

Harel, D. (1988). On visual formalisms. *Communications of the ACM*, *31*, 514–530.

Hofmeister, P., Casasanto, L. S., & Sag, I. A. (2012a). How do individual cognitive differences relate to acceptability judgments? A reply to Sprouse, Wagers, and Phillips. *Language*, *88*(2), 390–400.

Hofmeister, P., Casasanto, L. S., & Sag, I. A. (2012b). Misapplying working-memory tests: A reductio ad absurdum. *Language*, *88*(2), 408–409.

Hopcroft, J. E., & Ullman, J. D. (1979). *Introduction to automata theory, languages, and computation*. Reading, MA: Addison-Wesley.

Hudson, R. (2007). *Language networks: The new word grammar*. New York, NY: Oxford University Press.

Jelinek, F. (1990). Self-organized language modeling for speech recognition. In Waibel, A., Lee, K.F., & Kaufmann, M. (Ed.), *Readings in Speech Recognition* (pp. 450–506). San Mateo, California: Morgan Kaufmann Publishers.

Kam, X.-N. C., Stoyneshka, I., Tornyova, L., Fodor, J. D., & Sakas, W. G. (2007). Bigrams and the richness of the stimulus. *Cognitive Science*, *32*(4), 771–787.

Kolodny, O., Edelman, S., & Lotem, A. (2014). Evolution of Continuous Learning of the Structure of the Environment. *Journal of the Royal Society Interface*, *11*(92), 20131091. http://dx.doi.org/10.1098/rsif.2013.1091

Küntay, A., & Slobin, D. (1996). Listening to a Turkish mother: Some puzzles for acquisition. In Slobin, D.I., & Ervin-Tripp S.M. (Ed.). *Social interaction, social context, and language: Essays in honor of Susan Ervin-Tripp*, Mahwah, New Jersey: Lawrence Erlbaum Associates. 265–286.

Kwiatkowski, T., Goldwater, S., Zettelmoyer, L., & Steedman, M. (2012). A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 234–244). Avignon, France.

Lamb, S. M. (1998). *Pathways of the brain: The neurocognitive basis of language*. Amsterdam: John Benjamins.

Lashley, K. S. (1951). The problem of serial order in behavior. Cerebral Mechanisms in Behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–146). New York: Wiley.

Legate, J. A., & Yang, C. D. (2002). Empirical re-assessment of poverty of the stimulus arguments. *Linguistic Review*, *19*, 151–162.

Lotem, A., & Halpern, J. (2008). A Data-Acquisition Model for Learning and Cognitive Development and Its Implications for Autism (Computing and Information Science Technical Reports). Cornell University.

Lotem, A., & Halpern, J. Y. (2012). Coevolution of learning and data-acquisition mechanisms: A model for cognitive evolution. *Philosophical Transactions of the Royal Society B-Biological Sciences*, *367*(1603), 2686–2694.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Erlbaum.

McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences*, *6*, 465–472.

Onnis, L., Waterfall, H. R., & Edelman, S. (2008). Learn locally, act globally: Learning language from variation set cues. *Cognition*, *109*, 423–430.

Pearl, L., & Sprouse, J. (2012). Computational models of acquisition for islands. In J. Sprouse & N. Hornstein (Eds.), *Experimental syntax and island effects* (pp. 109–131). Cambridge, UK: Cambridge University Press.

Pereira, A. F., Smith, L. B., & Yu, C. (2008). Social coordination in toddler's word learning: Interacting systems of perception and action. *Connection Science*, *20*, 73–89.

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & Cognition*, *36*(7), 1299–1305.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*(2), 246–263.

Phillips, C. (2003). Syntax. In Nadel, L. (Ed.) *Encyclopedia of Cognitive Science*, Volume 4: 319–329). London, UK: Macmillan.

Phillips, C. (2010). Syntax at age two: Cross-linguistic differences. *Language Acquisition*, *17*(1–2), 70–120.

Pullum, G. K., & Scholz, B. (2002). Empirical assessment of poverty of the stimulus arguments. *The Linguistic Review*, *19*, 9–50.

Reali, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structural dependence and indirect statistical evidence. *Cognitive Science*, *29*, 1007–1028.

Resnik, P. (1992). Left-corner parsing and psychological plausibility. Paper presented at the International Conference on Computational Linguistics (COLING)

Ristad, E. S., & Yianilos, P. N. (1998). Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, 522–532.

Ross, J. R. (1967). *Constraints on variables in syntax*. D. Phil. Cambridge, MA: MIT.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Scha, R., Bod, R., & Sima'an, K. (1999). A memory-based model of syntactic analysis: Data-oriented parsing. *Journal of Experimental and Theoretical Artificial Intelligence*, *11*, 409–440.

Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago, IL: University of Chicago Press.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*, 390–397.

Smith, L. B., & Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. *Artificial Life*, *11*, 13–30.

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages of the United States of America. *Proceedings of the National Academy of Science*, *102*, 11629–11634.

Sprouse, J., Fukuda, S., Ono, H., & Kluender, R. (2011). Reverse island effects and the backward search for a Licensor in multiple Wh-questions. *Syntax-a Journal of Theoretical Experimental and Interdisciplinary Research*, *14*(2), 179–203.

Sprouse, J., Wagers, M., & Phillips, C. (2012a). A test of the relation between working-memory capacity and syntactic Island effects. *Language*, *88*(1), 82–123.

Sprouse, J., Wagers, M., & Phillips, C. (2012b). Working-memory capacity and island effects: A reminder of the issues and the facts. *Language*, *88*, 401–407.

Stolcke, A. (2002). SRILM—An extensible language modeling toolkit. Paper presented at the Proceeding: International Conference on Spoken Language Processing.

Stolcke, A. (2010). SRILM—The SRI Language Modeling Toolkit.

Suppes, P. (1974). Semantics of childrens language. *American Psychologist*, *29*(2), 103–114.

van Zaanen, M., & van Noord, N. (2012). Model merging versus model splitting context-free grammar induction. *Journal of Machine Learning Research*, *21*, 224–236.

Waterfall, H. R. (2006). *A little change is a good thing: Feature theory, language acquisition and variation sets*. D. Phil.: University of Chicago.

Waterfall, H. R., Sandbank, B., Onnis, L., & Edelman, S. (2010). An empirical generative framework for computational modeling of language acquisition. *Journal of Child Language*, *37*(Special issue 03), 671–703.

Wolff, J. G. (1988). Learning syntax and meanings through optimization and distributional analysis. In Y. Levy, I.M. Schlesinger, M.D.S. Braine (Eds.). *Categories and processes in language acquisition*, Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc., 179–215.

Yu, C., & Ballard, D. (2007). A unified model of word learning: Integrating statistical and social cues. *Neurocomputing*, *70*, 2149–2165.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414–420.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. The mean precision scores assigned by human judges to sentences from the original corpus and to those produced by U-MILA and SRILM (a standard tri-gram model, see text) following training on the first 15,000 utterances in a corpus of child-directed speech (Suppes, 1974).

Figure S2. A grammar learned by U-MILA and the rewrite rules that correspond to it. The graph is a simplified version of the full representation constructed by the model.

Data S1. Supplementary material 2: results.

Data S2. Supplementary material 3: Relationships between U-MILA and formal syntax.

Data S3. Supplementary Material 4: Output sequences from two PCFG grammars learned by U-MILA.

Data S4. Supplementary material 5: Simulation parameters.