Learning a generative probabilistic grammar of experience: a process-level model of language acquisition

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

[a]Department of Zoology, Tel-Aviv University, Tel-Aviv, 69978, Israel.
[b]Department of Psychology, Cornell University, Ithaca, NY 14853, USA.

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, 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; Lotem & Halpern, 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

[1] U-MILA stands for Unsupervised Memory-based Incremental Language Acquisition.

"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 approximate 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; Christiansen & Chater, 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; cf. 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 (cf. 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", cf. 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, in preparation). 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., in preparation) and of birdsong (Menyhart, Kolodny, Goldstein, DeVoogd, & Edelman, submitted). 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 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, 2008b; 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 reappears 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 whitespaces (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_{short_term}, is listed in SM5. 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 which 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

[2] Some applications of our approach to other modalities are described elsewhere (Menyhart et al., submitted; Kolodny et al., in preparation); the extension of the model to multimodal inputs is left for future work.

Boolean parameter, $B_{FillerSetSizeSensitivity}$, controls whether the learner would be sensitive to the fillers' set sizes: allowing *the* ____ *boy* to contain as fillers 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: (i) a temporal edge, representing the non-normalized probability[3] of occurrence of one node after another, (ii) a slot-constituency edge, representing the non-normalized probability of a certain node acting as a filler in a slot collocation, and (iii) a substitutability edge, representing the similarity among two nodes (see Fig. 1 and detailed explanation below).

[Fig. 1 should be here]

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

[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.

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 SM1 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):
   — Insert the new token into the short term memory queue.
   — Search for newly-completed alignments (recurring elements) within the queue.
   — 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):
   — Create a list of all nodes in the graph that terminate the short-term memory sequence.
   — Create a secondary list of sequences that fill the slot of slot-collocations in the primary list.
   — 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.
   — Update slot-candidacy edges of all nodes that are within slot-collocations in the primary list.

[4] We remark that the model was implemented (in Java) as a proof of concept, without any attempt at algorithmic optimization.

— 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: (1) when a new token is encountered, (2) 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 (3) 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 super-node or a sub-node 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 (cf. 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 (cf. 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

[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.

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):

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

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):

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

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*):

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

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 recalculated 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:

[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.

1.

1.1 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:

1.2 Choose a node from among those that the outgoing temporal edges go to, drawing among them randomly with proportion to $W_{edge} \cdot L$, 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 $P_{generalize}$, 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 $1-P_{rand}$ 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; cf. Goodman, 2001), which allows the estimation of perplexity and assessment of grammaticality, as explained below. In addition to the smoothing implied by a nonzero value of P_{rand} as described earlier, the model can also assign a small nonzero 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.

[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, & Sima'an, 1999), section 4).

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 (i) the replication of a dozen or so published results in sequence segmentation, artificial grammar learning, and structure dependence, and (ii) 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:

- 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).

- Study 2: Characteristics of the learned representation: equivalence (substitutability) of phrases and the similarity structure of the phrase space (section 3.2).
- Study 3: Replication of a variety of results in sequence segmentation and chunking (section 3.3).
- Study 4: Replication of results in artificial grammar learning (sections 3.4 — 3.6).
- 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 SM2.

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 (cf. 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 to 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 nonzero 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 SM5).

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):

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

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 SM5). 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 $P_{generalize}$=0.2, $\mathrm{P_{rand}}$=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 ($t = 3.5$, $p < 0.0005$, 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 (MCMC) run to estimate the confidence limits on the slope parameters (R procedures *mcmc* and *HPDinterval*), did not yield a significant difference.

[Fig. 2A and 2B should be here]

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 SM2, 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 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).

[Table 1 should be here]

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 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.

[Fig. 3A and 3B should be here]

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, Addyman & Mareschal (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 SM2).

3.3.1 Words vs. nonwords, 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; $V = 10$, $p < 0.0625$). 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 vs. nonwords, 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, Saffran & Newport (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 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; $V = 10$, $p < 0.0625$; 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 vs. nonwords, adults (Perruchet and 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; $V = 21$, $p < 0.016$). Unlike in the previous experiments, running the model in its flat-Markov mode did not lead to successful discrimination.[8]

[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.

3.3.4 Words vs. nonwords, adults (Perruchet and Desaulty (2008), experiment 2: backward transition probabilities)

The second experiment of Perruchet & 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 < 0.016). As expected, the run in a flat-Markov mode did not differentiate between words and non-words.

3.3.5 Hierarchical chunking (Giroux and Rey, 2009)

Giroux & 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 sub-chunks (*kl, lm, mn, klm, lmn, klmn*) as independent units. We note that for a language-oriented model, eliminating sub-chunks 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]

[9] In contrast, the version of the model that was applied to birdsong (Menyhart, et al., submitted) does implement this step, and thus eliminates from the grammar units that are

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

In their first experiment, Frank, Goldwater, Griffiths & Tenenbaum (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.

[Fig. 4 should be here]

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.

3.3.7 Word segmentation: effects of vocabulary size (Frank, Goldwater, Griffiths & Tenenbaum (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) 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 Frank 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 ($t = 3.0$, $p < 0.0039$) and vocabulary size ($t = -5.46$, $p < 0.0000015$) and a significant interaction ($t = 2.1$, $p < 0.04$). 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.

[Fig. 5A and 5B should be here]

3.3.8 Word segmentation, phonetic encoding (French, Addyman & Mareschal (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 ($t = 2.1$, $p < 0.035$) and number of syllables ($t = -7.08$, $p < 2.9X10^{-12}$) and no interaction.

[Fig. 6 should be here]

3.3.9 Word clustering by category (French, Addyman & Mareschal (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).

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*)+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.

[Fig. 7 should be here]

[10] This is equivalent to using Levenshtein distance over strings (e.g., Ristad & Yianilos, 1998).

3.3.10 Word segmentation: effects of frequency and transitional probability (French, Addyman & Mareschal (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 ($t = 2.213$, $p < 0.04$; Fig. 8).

[Fig. 8 should be here]

3.4 Non-adjacent dependencies (Gomez (2002), experiment 1)

Gomez (2002) reported that both adults and infants can learn nonadjacent 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 Gomez (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.

[Table 2 should be here]

The results are summarized in Fig. 9. Nonadjacent 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 ($t=3.0$, $p < 0.0045$).[11] There was, however, no abrupt change in performance between pool sizes 12 and 24, contrary to the effect reported by Gomez (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.

[Fig. 9 should be here]

[11] This interaction amounted to a small (in absolute terms) difference in the slopes of the grammaticality effect, rather than in a change of the sign of the effect. As such, it does not reflect on the rest of the discussion of this experiment.

3.5 Syntactic categories (Gomez and Lakusta (2004), experiment 1)

Gomez 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 Gomez and Lakusta (2004), with spaces inserted between each two consecutive syllables and a random ordering of the sentences. The learner then assigned a probability score to each of the test sentences in Gomez and Lakusta (2004). The model's parameter that controls its sensitivity to slot filler length, $B_{FillerSetSizeSensitivity}$ (see section 2.2), was set so the learner would be sensitive to the filler set size, measured in syllables.

[Table 3 should be here]

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.

[Fig. 10 should be here]

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

Onnis, Waterfall & Edelman (2008) examined the effects of variation sets[12] on artificial grammar learning in adults. As in that study, we trained multiple instances of

[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).

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 ($t = 13.7$, $p < 0.0001$; all p values estimated by Markov Chain Monte Carlo sampling with 10,000 runs, procedure *pvals*, R package *languageR*) and condition ($t = -69.8$, $p < 0.0001$). Crucially, the word-hood × condition interaction was significant ($t = 57.8$, $p < 0.0001$).

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

[Fig. 11 should be here]

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

Reali & 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?*

Their training corpus is composed of sentences uttered by nine mothers addressing their children, recorded over a period of 4 to 5 months, while the children were of ages 1:1 to 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 & 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 & 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 ($F = 26.35$, $p < 7.08 \times 10^{-07}$).

[Fig. 12 should be here]

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 & Kluender (2011) conducted a quantitative study of the interaction between grammatical island constraints and short- and long-term dependencies in determining sentence acceptability. They used a factorial design, with four types of sentences: (i) short-term dependency + no island, (ii) long-term dependency + no island, (iii) short-term dependency + island, (iv) 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,

[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.

Casasanto, & Sag, 2012a, 2012b; Sprouse, et al., 2012a; Sprouse, Wagers, & Phillips, 2012b).

In an attempt to account for this finding by a statistical learning model, Pearl & 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 dominate 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 & 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 & 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

[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; cf. Lashley, 1951).

(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 (Elman, 1990; Christiansen & Chater, 1999) 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 SM3 and SM4).

[Fig. 13 should be here]

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.

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

[15] We thank the anonymous referees for pointing out the importance of explicitly demonstrating the model's ability to learn such grammars.

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 (section S2, 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 SM2).

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 Gomez (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;

[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 SM5). 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.

Lotem & Halpern, 2008; Lotem & Halpern, 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 (cf. 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 Gomez (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 Gomez (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 Gomez & 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 & 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., (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 the present 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 & 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 is, 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 the 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.

6 Acknowledgements

We are grateful to Andreas Stolcke, Jon Sprouse, Lisa Pearl, Florencia Reali, and Morten Christiansen for sharing their code and data, and for their advice. 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. O.K. was supported by a Dean's scholarship at Tel Aviv University and a scholarship from the Wolf Foundation.

7 Bibliography

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.

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: Mouton.

Chomsky, N. (1980). *Rules and Representations*. Oxford: 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.* School of Cognitive and Computing Sciences, University of Sussex.

Cramer, B. (2007). Limitations of current grammar induction algorithms. *Proceeding of the 45th Annual Meeting of the ACL: Student Research Workshop*, 43-48.

DeMarcken, C. G. (1996). *Unsupervised Language Acquisition.* 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. *Prediction in the Brain,* pp. 170-189.

Edelman, S., & Solan, Z. (2009). Machine translation using automatically inferred construction-based correspondence and language models, *Proc. 23rd Pacific Asia Conference on Language, Information, and Computation (PACLIC)*. 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*.

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. *Commun. 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*. Readings in Speech Recognition,* pp. 450-506.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*, 1–37.

Joshi, A., & Schabes, Y. (1997). Tree-Adjoining Grammars*. Handbook of Formal Languages,* pp. 69-124.

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. (in preparation). The Evolution of Unsupervised Learning.

Küntay, A., & Slobin, D. (1996). Listening to a Turkish mother: Some puzzles for acquisition*. Social interaction, social context, and language: Essays in honor of Susan Ervin-Tripp,* pp. 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*, 234-244.

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,* pp. 112-146.

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.

Menyhart, O., Kolodny, O., Goldstein, M. H., DeVoogd, T. J., & Edelman, S. (2013, in review). Like father , like son: zebra finches learn structural regularities in their tutors' song.

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*: 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*. Encyclopedia of Cognitive Science,* pp. 319-329.

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.* MIT.

Scha, R., Bod, R., & Sima'an, K. (1999). A memory-based model of syntactic analysis: data-oriented parsing. *J. 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. *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 Proc. Intl. Conf. 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.* 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*. Categories and Processes in Language Acquisition,* pp. 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.

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

Table 2: The stimuli used by Gomez (2002), experiment 1.

Table 3: The word categories used by Gomez & Lakusta (2004), experiment 1.

Figure 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 double-headed 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.

Figure 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.

Figure 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 to 100. (B) percentile range 75 to 80. Proximity among words in these plots generally corresponds to intuitive similarity among them.

Figure 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 proportion-better score described by French et al. (2011).

Figure 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.

Figure 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.

Figure 7: Clustering of tri-syllabic words, according to the similarity values assigned by U-MILA, 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.

Figure 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.

Figure 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 Gomez (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).

Figure 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 Gomez & Lakusta (2004), U-MILA successfully differentiates between sentences with different lengths of dependency, even when these contain novel intervening syllables.

Figure 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*).

Figure 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.

Figure 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.

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). (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. The respective perplexity scores of the two models are $\text{ppl}_{\text{U-MILA}}$=40.07, $\text{ppl}_{\text{SRILM}}$=22.43.

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.

Figure 1

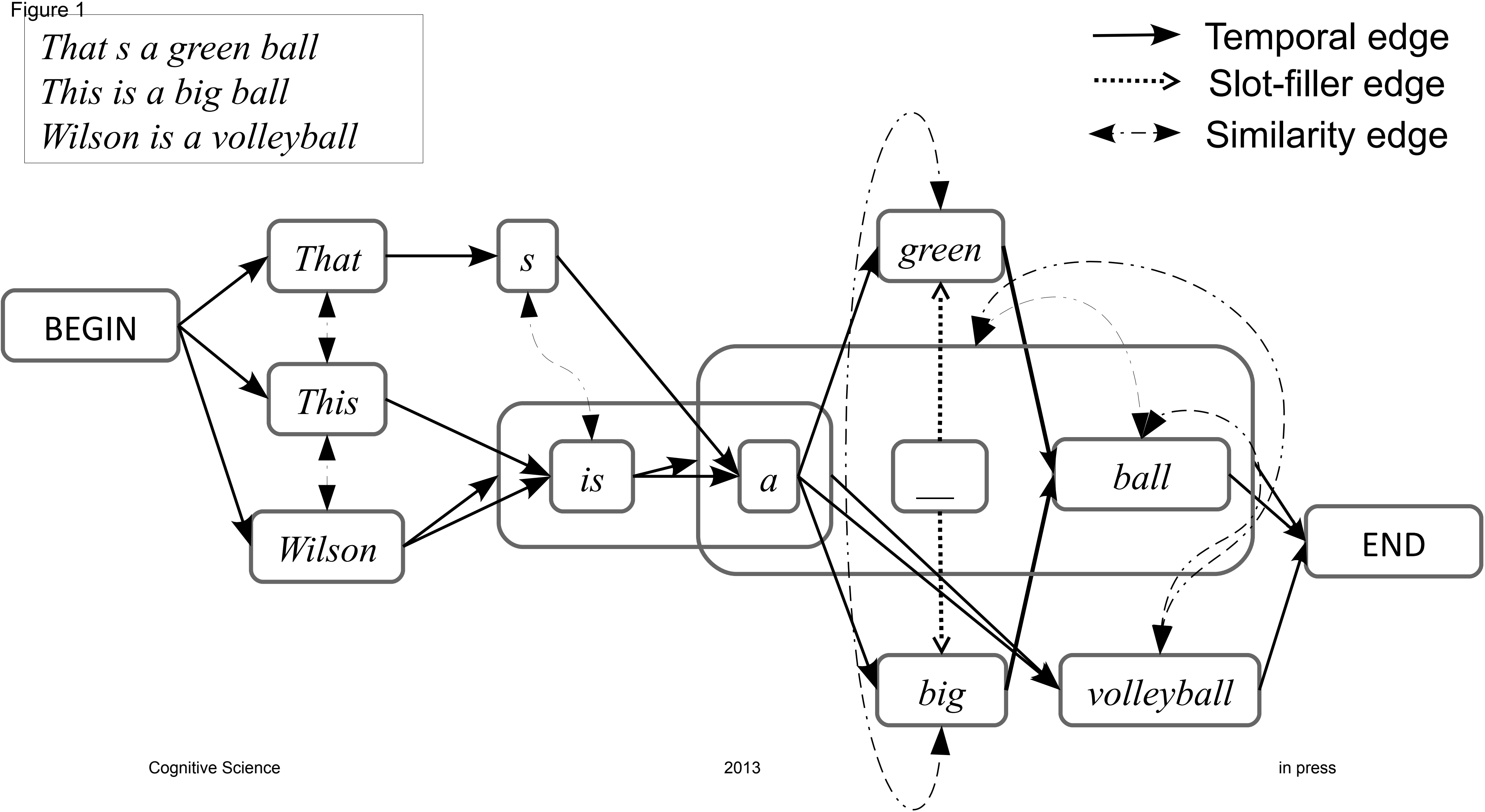

That s a green ball
This is a big ball
Wilson is a volleyball
Temporal edge
Slot-filler edge
Similarity edge
BEGIN
That
This
Wilson
s
is
a
green
__
big
ball
volleyball
END

Figure 2A

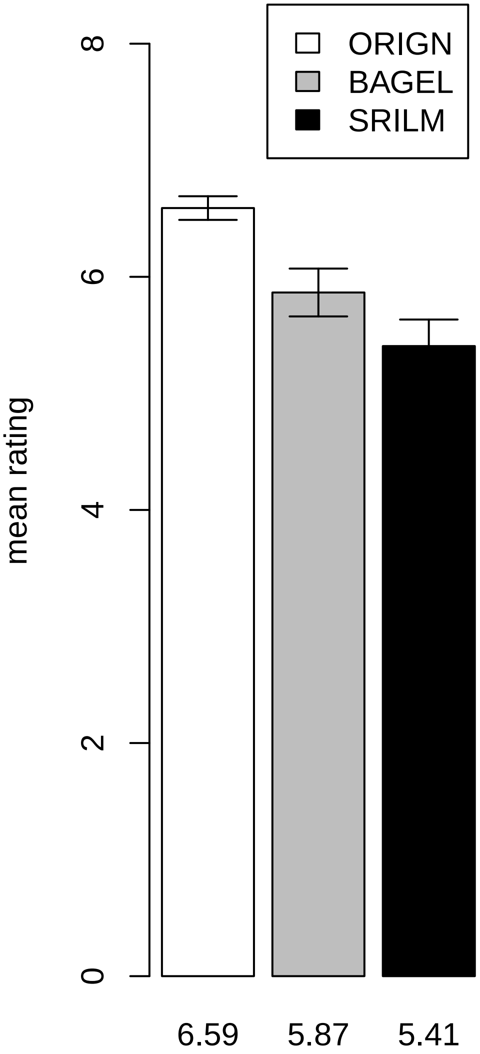
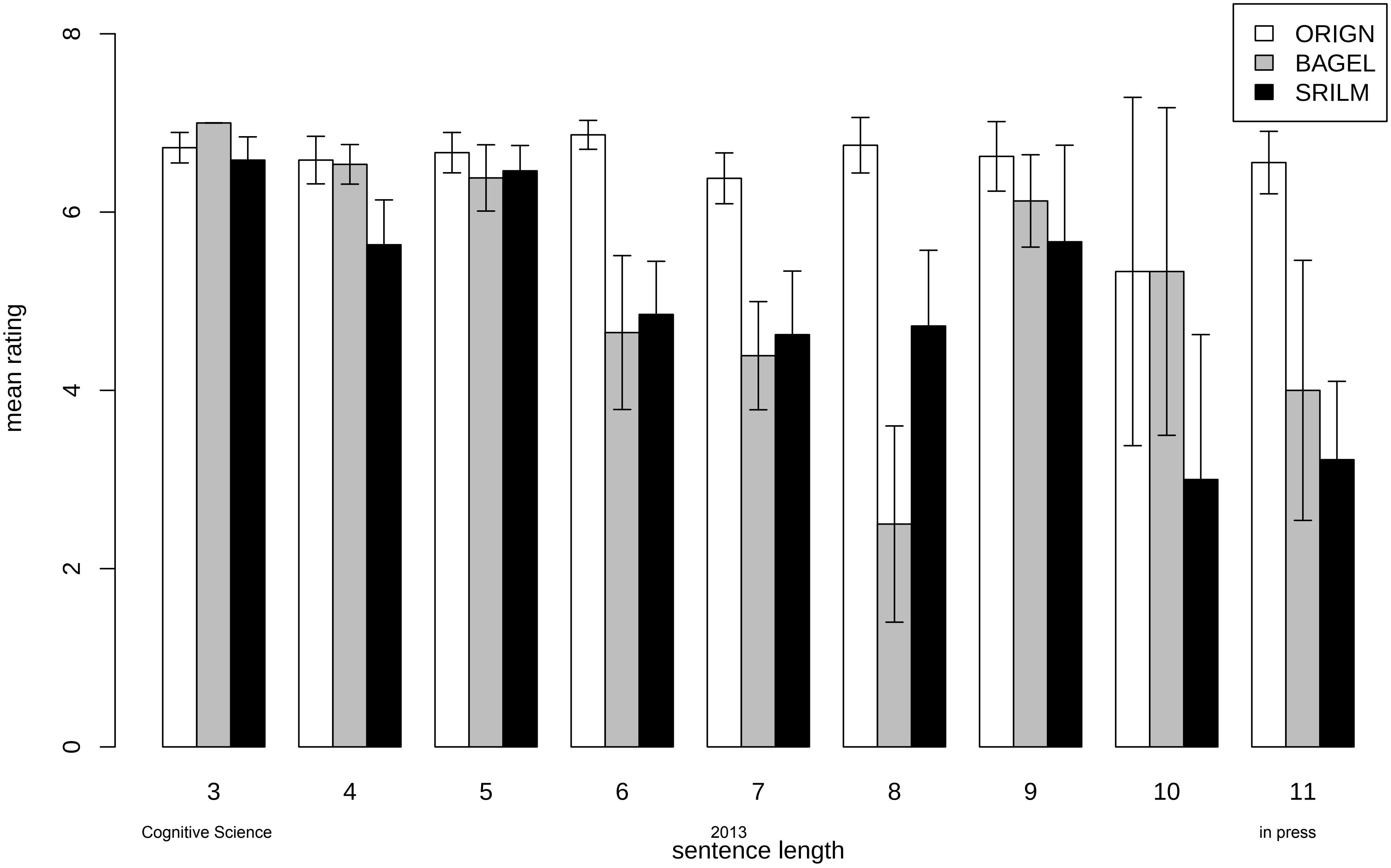
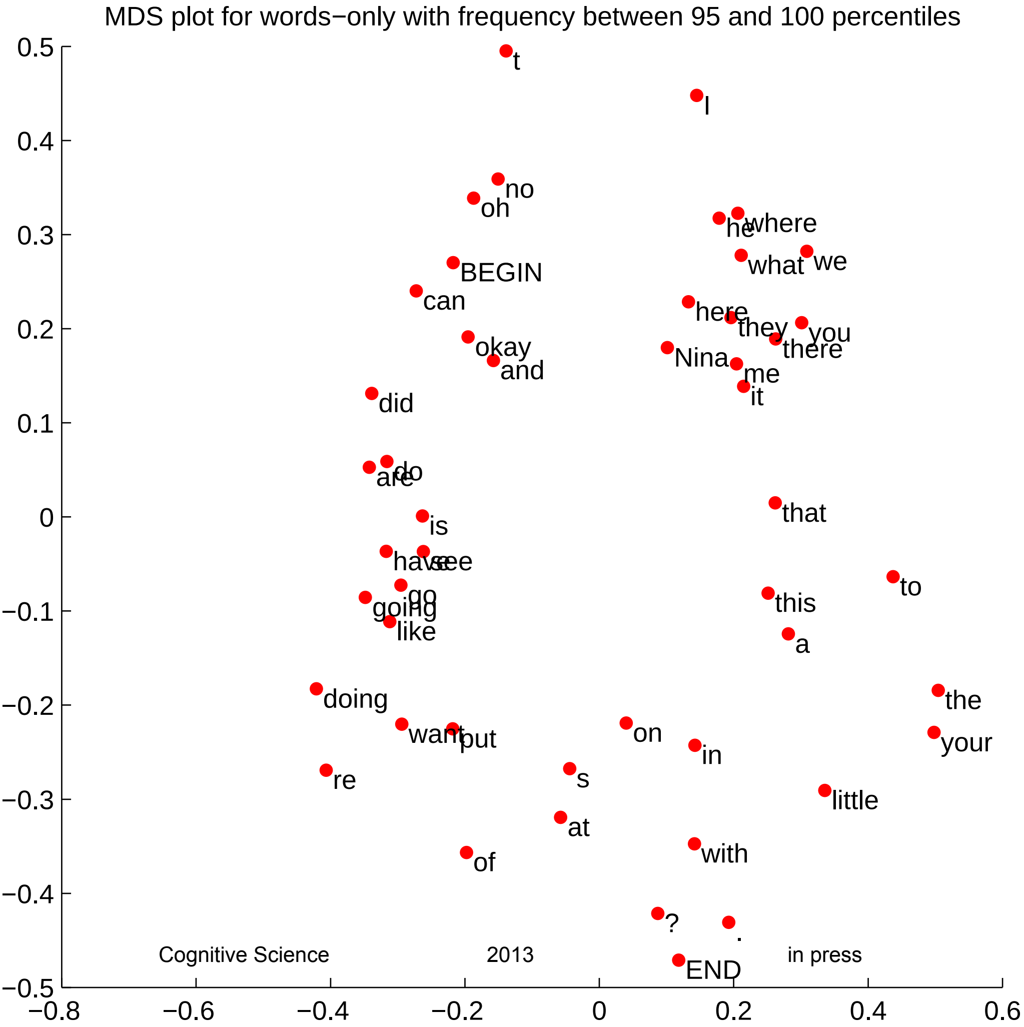


Figure 2B


mean rating
0
2
4
6
8
ORIGN
BAGEL
SRILM
3
4
5
6
7
8
9
10
11
sentence length


MDS plot for words−only with frequency between 95 and 100 percentiles
0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
t
I
no
oh
he
where
what
we
BEGIN
can
here
they
you
there
okay
and
Nina
me
it
did
do
are
that
is
have
see
to
go
going
this
like
a
doing
the
want
put
on
your
in
re
s
little
at
with
of
?
.
END

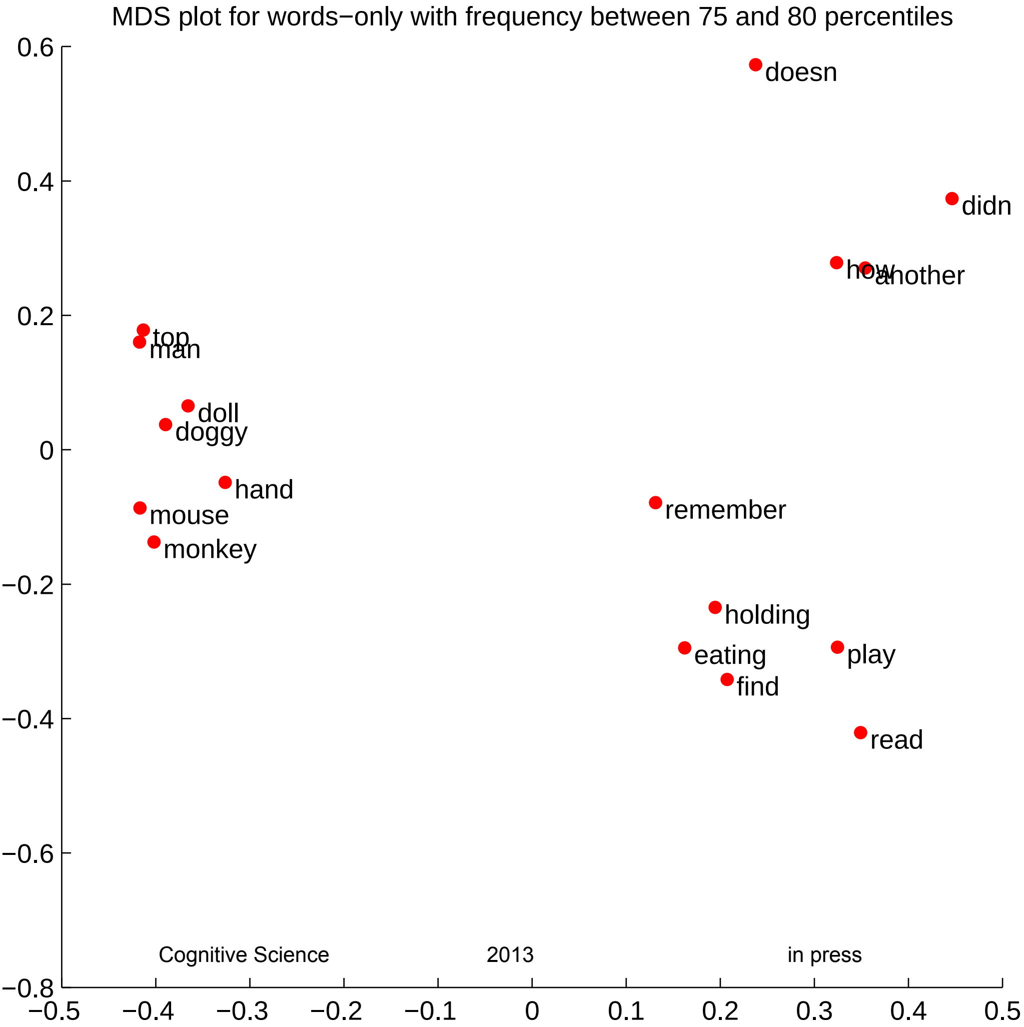

MDS plot for words−only with frequency between 75 and 80 percentiles
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−0.5
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5
doesn
didn
how
another
top
man
doll
doggy
hand
mouse
monkey
remember
holding
eating
play
find
read

Figure 4

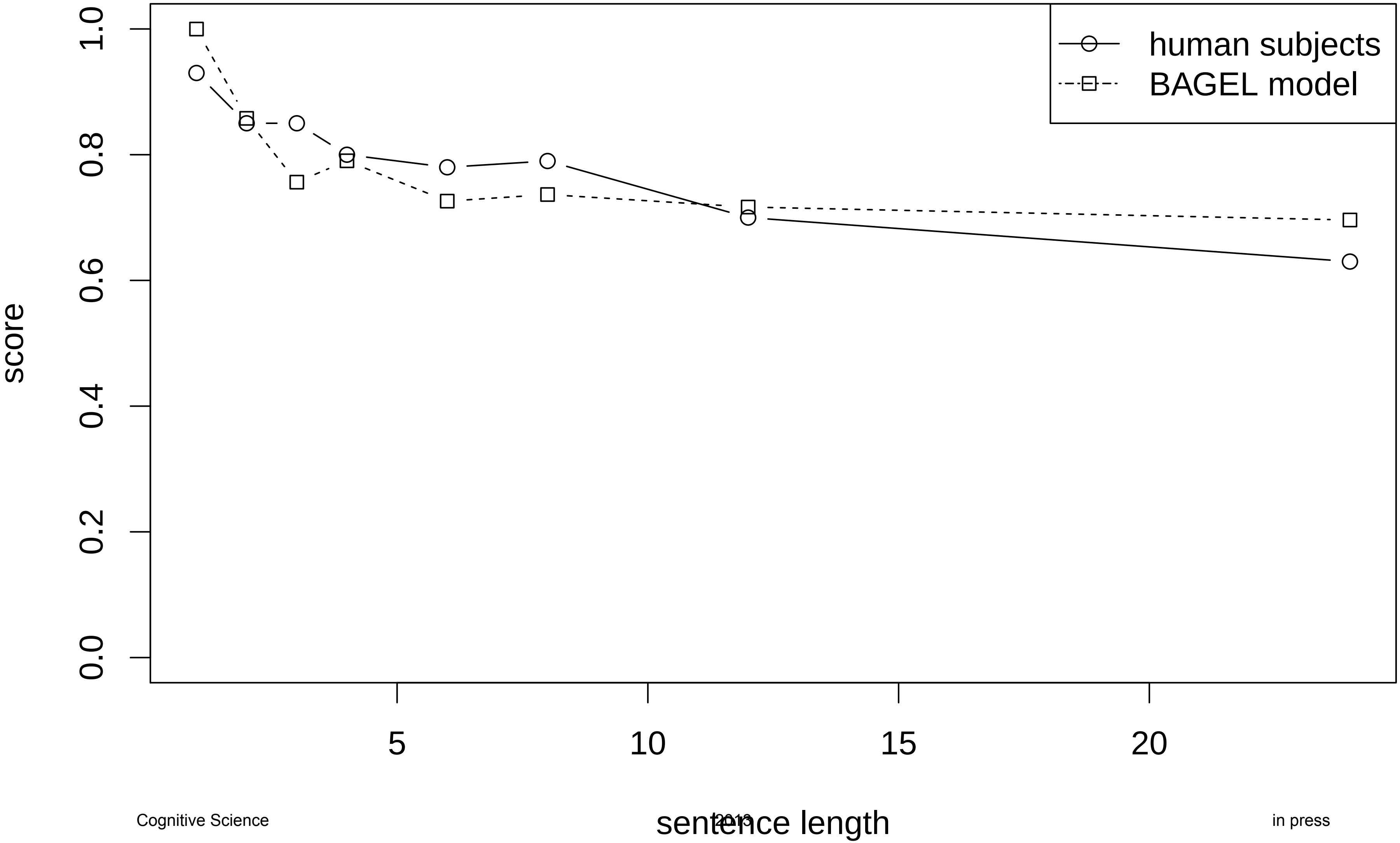

human subjects
BAGEL model
score
1.0
0.8
0.6
0.4
0.2
0.0
5
10
15
20
sentence length

Figure 5A

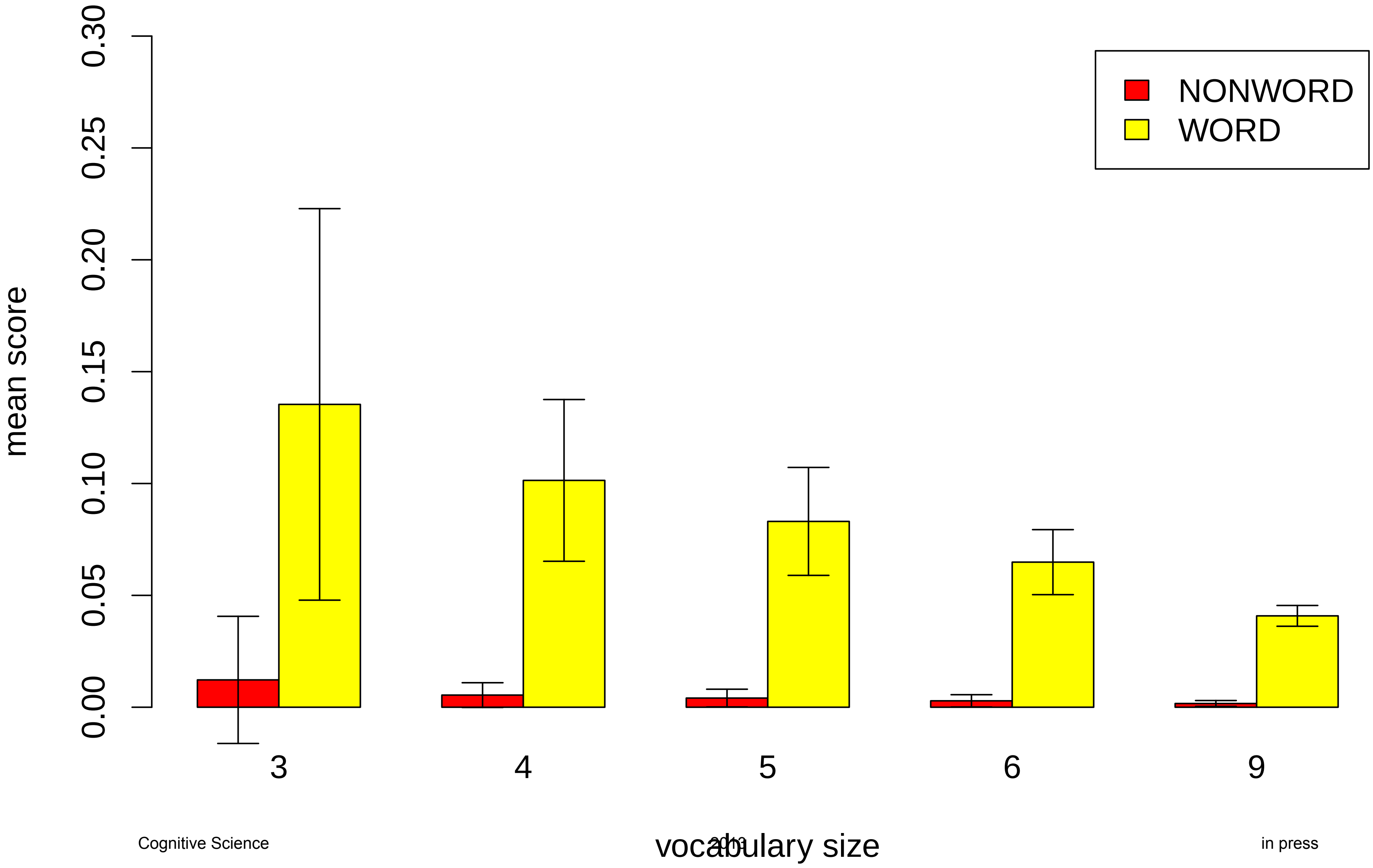

mean score
0.00
0.05
0.10
0.15
0.20
0.25
0.30
🟥 NONWORD
🟨 WORD
3
4
5
6
9
vocabulary size

Figure 5B

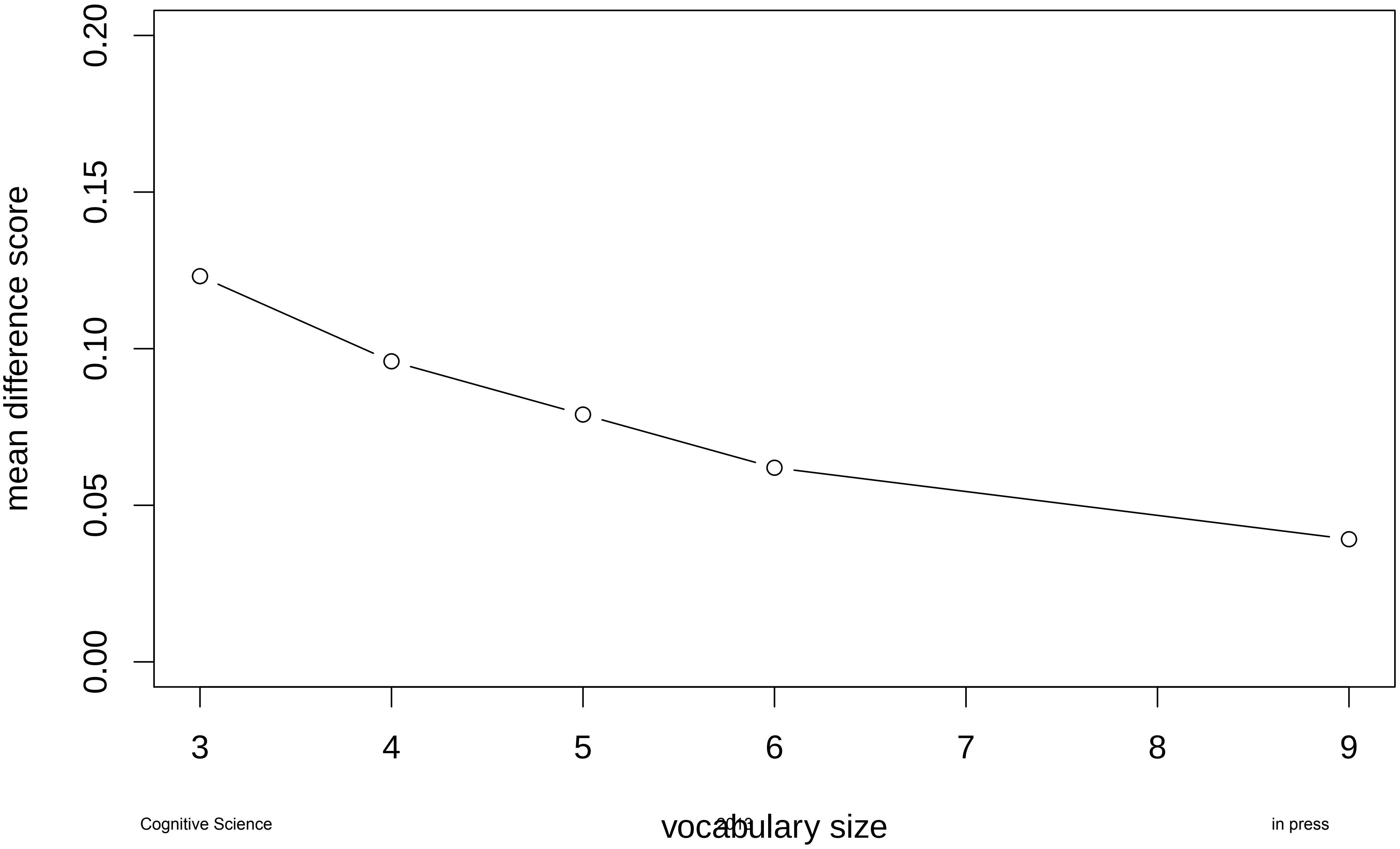

mean difference score
0.00
0.05
0.10
0.15
0.20
3
4
5
6
7
8
9
vocabulary size

Figure 6

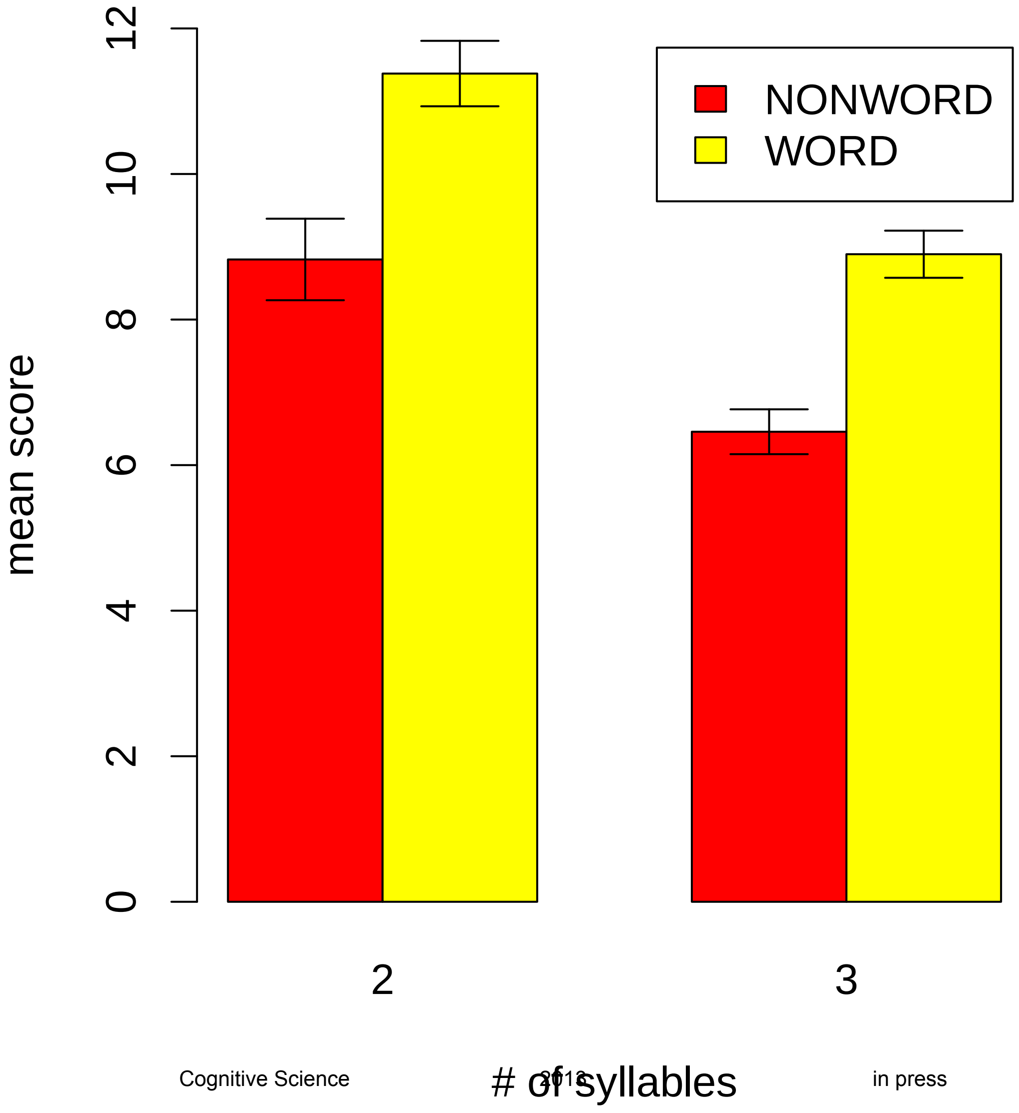


Figure 7

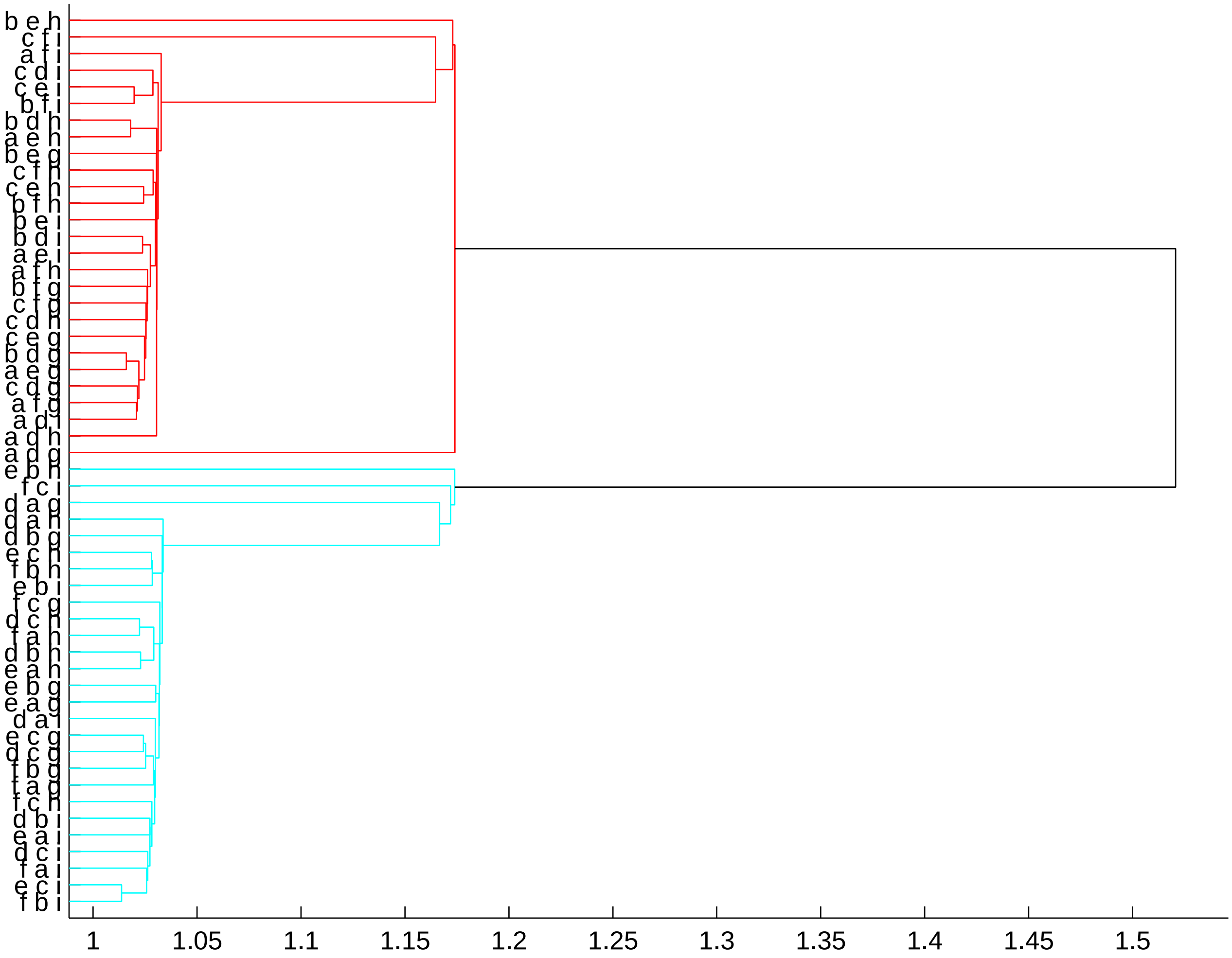

b e h
c f i
a f i
c d i
c e i
b f i
b d h
a e h
b e g
c f h
c e h
b f h
b e i
b d i
a e i
a f h
b f g
c f g
c d h
c e g
b d g
a e g
c d g
a f g
a d i
a d h
a d g
e b h
f c i
d a g
d a h
d b g
e c h
f b h
e b i
f c g
d c h
f a h
d b h
e a h
e b g
e a g
d a i
e c g
d c g
f b g
f a g
f c h
d b i
e a i
d c i
f a i
e c i
f b i
1
1.05
1.1
1.15
1.2
1.25
1.3
1.35
1.4
1.45
1.5

Figure 8

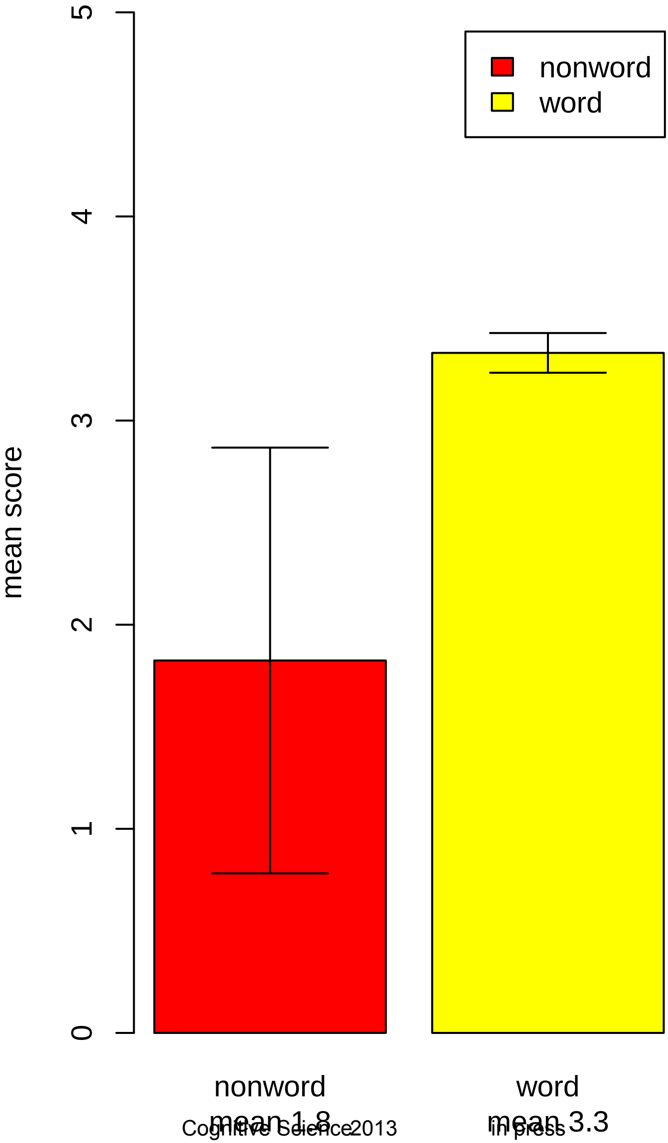


Figure 9

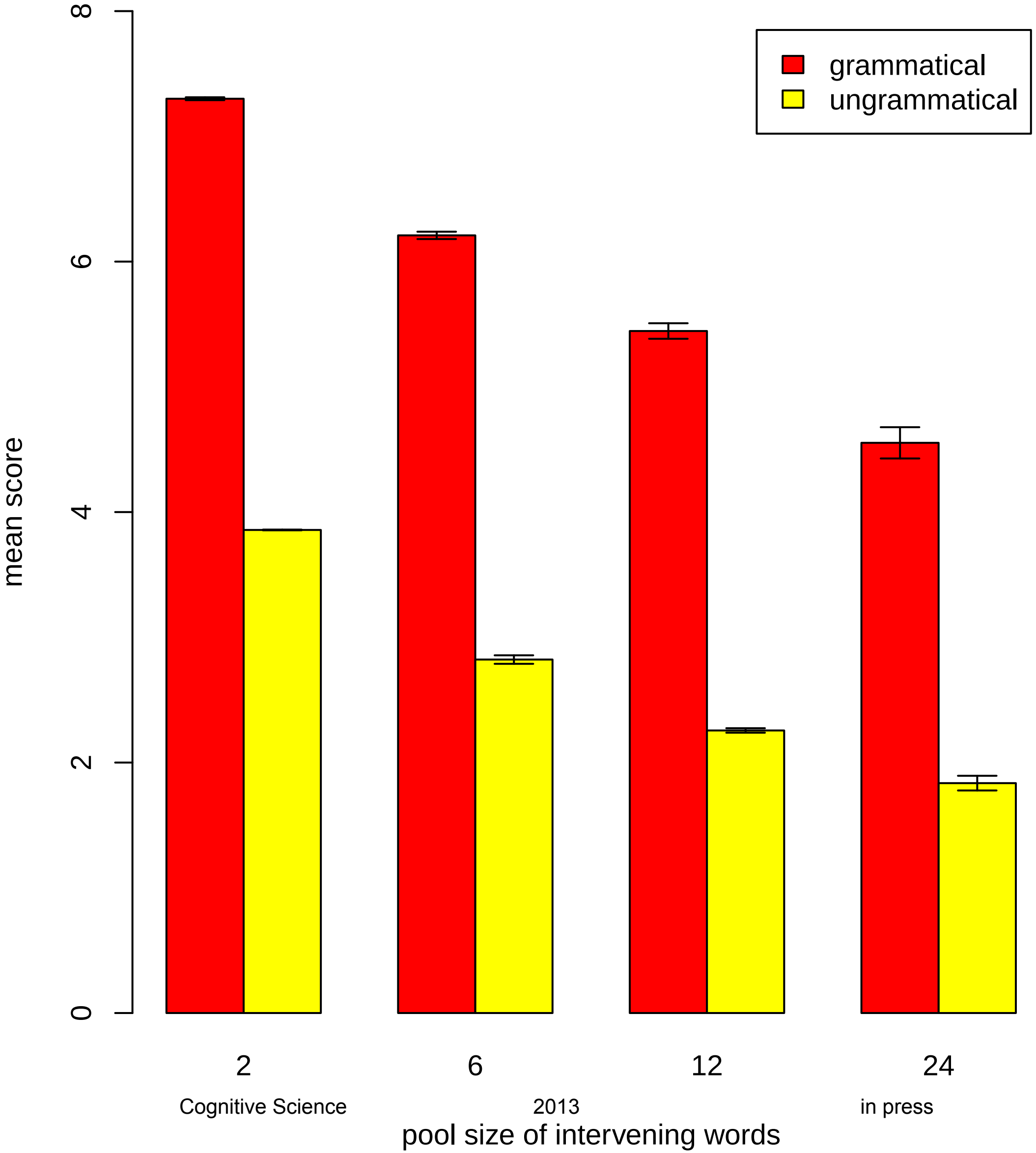

mean score
0
2
4
6
8
grammatical
ungrammatical
2
6
12
24
Cognitive Science
2013
in press
pool size of intervening words

Figure 10

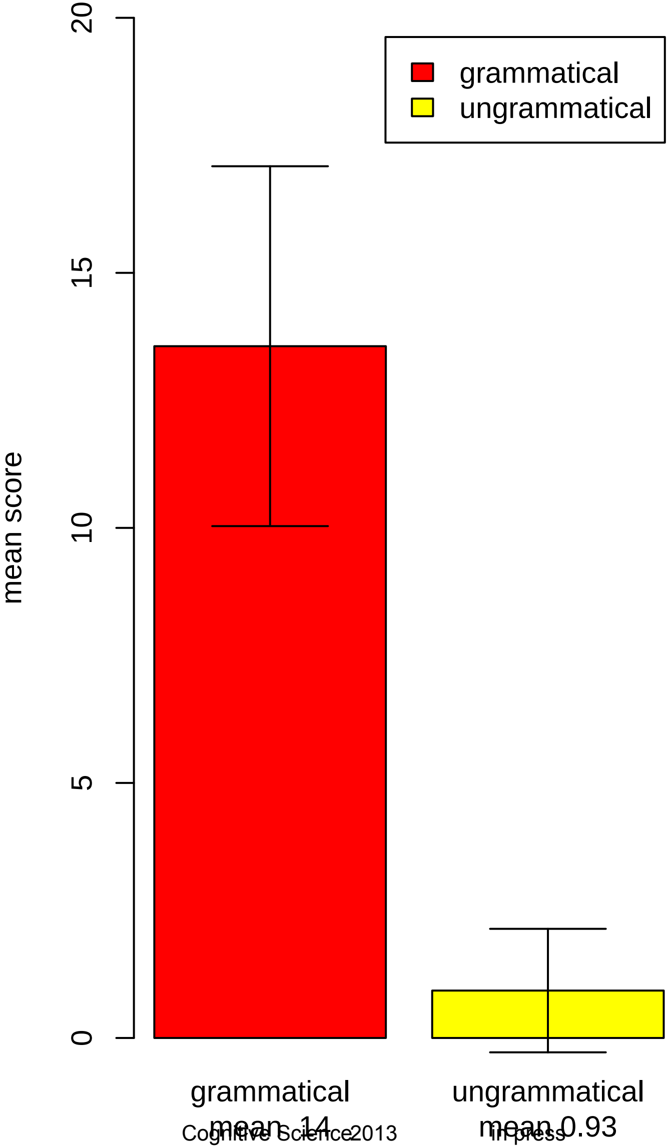
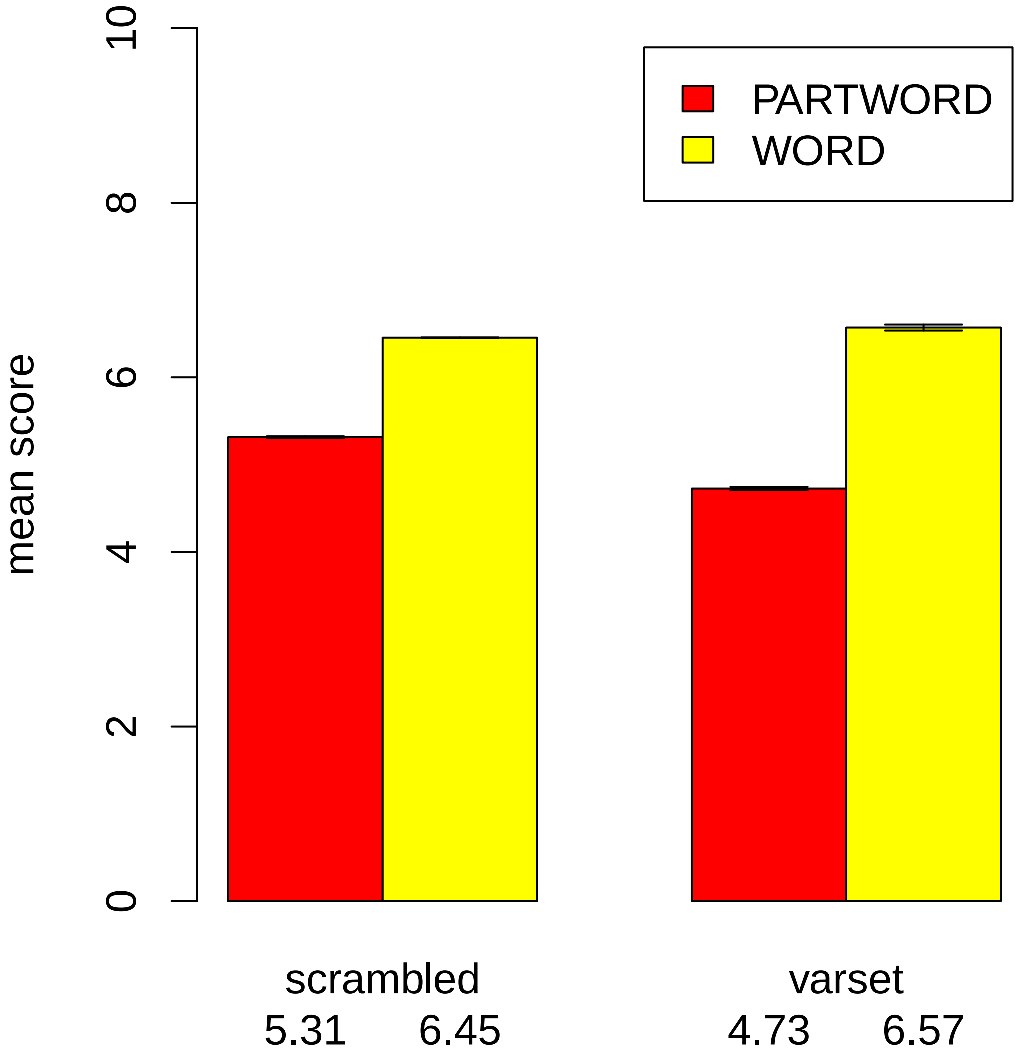

mean score
0
5
10
15
20
🟥 grammatical
🟨 ungrammatical
grammatical
mean 14
ungrammatical
mean 0.93

Figure 11



mean score
10
8
6
4
2
0
🟥 PARTWORD
🟨 WORD
scrambled
varset
5.31
6.45
4.73
6.57

Figure 12

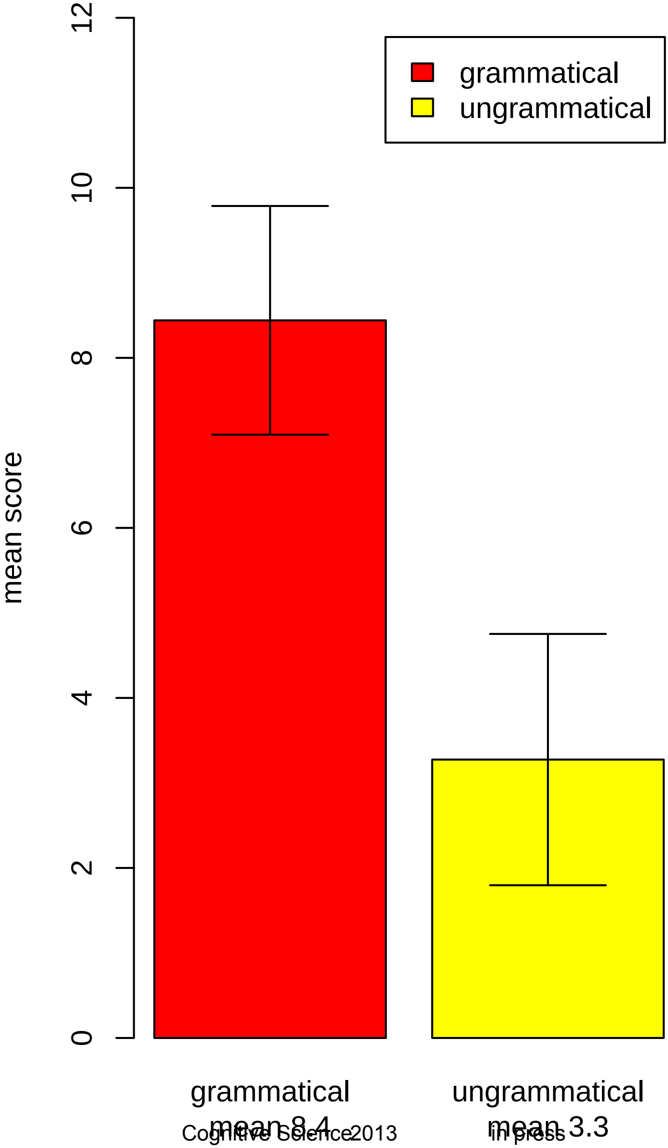
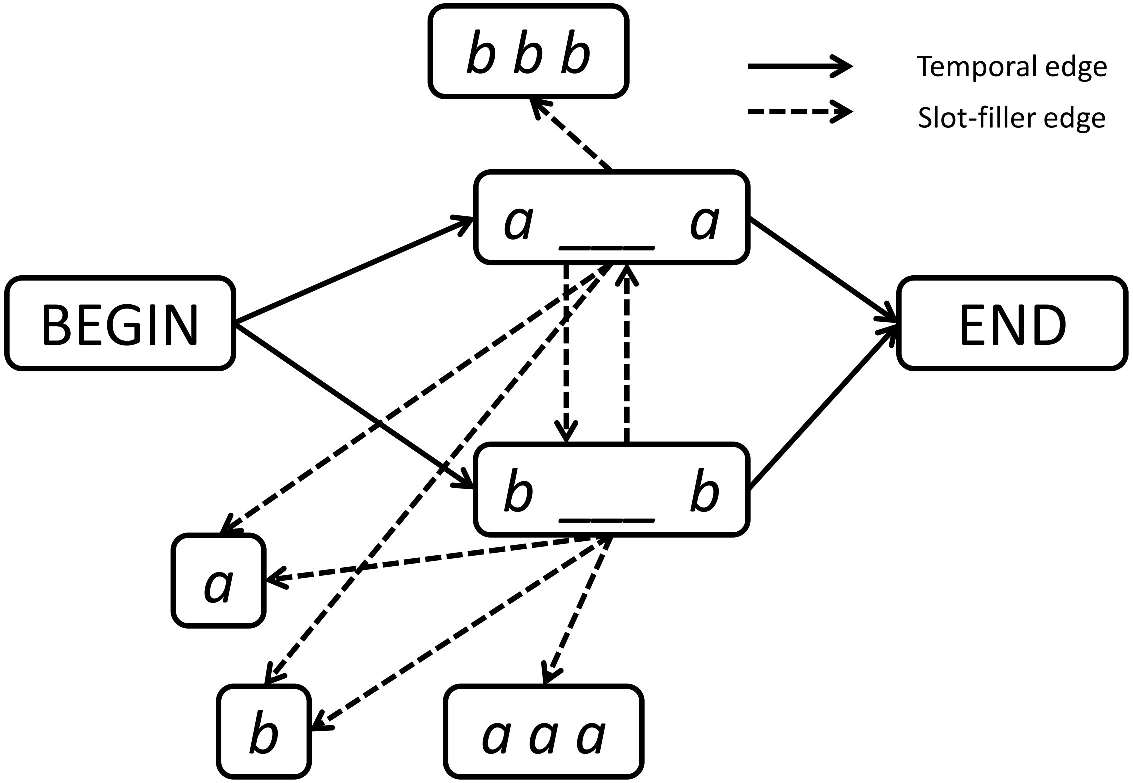

mean score
0
2
4
6
8
10
12
🟥 grammatical
🟨 ungrammatical
grammatical
mean 8.4
ungrammatical
mean 3.3

Figure 13



Examples of output sequences:

BEGIN b a b a b a a a b a b a b END
BEGIN b a b END
BEGIN a b a a a b a END
BEGIN b a b a b a b b b a b a b a b END
BEGIN a b a b a b b b a b a b a END
BEGIN b a b a b a b a b END
BEGIN a b a b b b a b a END
BEGIN a b a b a b a END
BEGIN b b b END
BEGIN a b a b a b a b a b a b a b a END
BEGIN b a b b b a b END
BEGIN a b a b a b a b a b a b a END

The learned grammar is equivalent to the following set of rewrite rules:

BEGIN (a b)^n {a, b, a a a} (b a)^n END
BEGIN (a b)^n a {a, b, b b b} a (b a)^n END
BEGIN (b a)^n {a, b, b b b} (a b)^n END
BEGIN (b a)^n b {a, b, a a a} b (a b)^n END

n ϵ {0,1,2,…}

Table_1

	node	**5 most similar nodes**[1]
	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

[1] 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*.

Table_2

<u>Language 1</u>

$$S \rightarrow \left\{ \begin{array}{l} aXd \\ bXe \\ cXf \end{array} \right\}$$

<u>Language 2</u>

$$S \rightarrow \left\{ \begin{array}{l} aXe \\ bXf \\ cXd \end{array} \right\}$$

$X \rightarrow x_{1,}\, x_{2, \ldots ,}\, x_n$; $n = 2, 6, 12$ or 24

<u>Test strings</u>

<u>Language 1</u>	<u>Language 2</u>
pel wadim rud	*pel wadim jic*
vot wadim jic	*vot wadim tood*
dak wadim tood	*dak wadim rud*
pel kicey rud	*pel kicey jic*
vot kicey jic	*vot kicey tood*
dak kicey tood	*dak kicey rud*

Table_3

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*

An annotated pseudocode of U-MILA's learning process

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):

    2.1 **Insert** the new token into the short term memory queue;

    2.2 **Conduct** a greedy search and create a list of newly-completed alignments within the queue.

    This procedure uncovers both identically repeating sequences and partially-repeating sequences. Partially repeating sequences are those in which the beginning and end are identical but internal parts vary. The former are added to the list as they are and the latter are added as slot collocations. The internally varying sequences are added to a secondary list. Maximum allowed length of internal non-identical section is a model parameter.

    2.3 **Add** each element in the list to the graph with probability proportional to $e^{-c \cdot d}$.[1,2] *Sequences from the secondary list are added if their container sequences are added or if their container sequences are previously known; the slot-interchangeability-within-window fields of the paired sequences are updated accordingly.*

[1] c is 0.15·D_{short_term}, where D_{short_term} is the short-term memory decay parameter and d is the distance between the two overlaps in units of characters or of tokens.

[2] This procedure realizes the idea of searching for a recurrence within a short time window, which is implemented here, with an eye towards biological realism, as a probabilistic event. The recurrence of a sequence has a higher probability of being discovered and the sequence being added to the graph if the two occurrences are near each other. It is customary to view the 'effective distance' in this setting as the distance over which the probability of discovery drops below a certain value , and thus to abstract it to an 'effective window' (cf. Goldstein et al., 2010).

2.4 **Ensure** greedy updating: if a newly-added sequence contains a sequence added following the previous token-reading, remove the shorter sequence. *Thus, for example, a recurrence within the short-term memory of the phrase "dogs and cats" will not lead to inclusion in the graph of "dogs and" as a collocation in itself but only of the complete phrase "dogs and cats".*

3. **Update** temporal relations and **construct** collocations (bottom-up):[3]

3.1 **Create** a list of all nodes in the graph that terminate the short-memory sequence *(i.e., those that have been completed by the recent token's addition; e.g., adding the token "to" to the queue "John is going" completes the possible nodes "is going to", "is ____ to", "going to", and "to")*. **Create** a corresponding secondary list of sequences which fill the slot of slot-collocations in the primary list *(e.g., "going" fills "is ____ to" in the example).*

3.2 **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 ending just before node X's location of beginning *(e.g., "going to" in the previous example begins where "is" and "John is" end)*[4,5] .

[3] Steps 3.1 to 3.3 are a way of updating all temporal relations among known units in the graph that are affected by the encounter with the current token, such as the relation between "*the*" and "*boy*" when "*boy*" is encountered in the input and the previous word in the input was "*the*". Although these steps are highly technical, they are ultimately rooted in associative lookup and in weight updates in neural networks, operations that in turn have biological counterparts.

[4] Note that, for biological plausibility, changes to the graph structure by addition of edges or by update of edge weights, are spatially local. These changes do not trigger secondary effects such as weight normalization.

[5] Note that this update rule treats all nodes equally, increasing all edge weights by an identical increment. Over the course of learning, this causes nodes that encode shorter sequences to

3.3 **Update** slot-candidacy edges of all nodes that are within slot collocations in the primary list.

3.4 **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 in conjunction with a prior (*Pc*):

$$\frac{P(A \Rightarrow B)}{\sum_{\lambda \in G} P(A \Rightarrow \lambda) \cdot \sum_{\theta \in G} P(\theta \Rightarrow B)} \cdot \frac{Pc}{1-Pc} > 1,$$

Here, '=>' denotes the relation 'is followed by,' and so $P(A \Rightarrow B)$ denotes the estimate of the probability that node A is followed by node B. *Pc* denotes the prior against collocations, which is a parameter of the model. *The higher Pc is, the higher the value of suspicious coincidence index must be to result in concatenating two nodes into a supernode.*

A pair of nodes A,B may be combined to a supernode A+B only if both had occurred in the input at least *MinOccs* times, where *MinOccs* is a model parameter. This ensures that some minimal statistics regarding the nodes' properties are collected before constructing higher hierarchies.

accumulate high weights, compared to those of nodes that represent longer ones. This effect is exacerbated by the fact that it may take a long time for multi-word sequences to be recognized as units and to begin their weight accumulation process. Fine-tuning the weight update parameters may redress this imbalance in future implementations of the model.

Supplementary material 2: results

1.1 Study 1: generative performance

A key purpose of acquiring a grammar-like representation of language is the ability to generate acceptable utterances that transcend the learner's past experience, that is, the corpus of language to which it has been exposed. 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 the field of 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 our estimates both of recall and of precision.

A commonly reported graded counterpart for all-or-none 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 (cf. Goldsmith, 2007). In the experiment described below, we compared the perplexity of our model 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 to 7, how likely they think each test sentence (from a corpus generated by the model) 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 (see below), and included in the test set equal numbers of sentences generated by the two models and sentences taken from the original corpus.

Perplexity and the precision of a model must always be considered together. A model that assigns the same nonzero 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 suppl. material).

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.

1.1.1 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)

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

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 whole test set.

The resulting perplexity was 40.07, for the similarity-based generalization and smoothing parameters used throughout the experiments (see SM1). 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 far 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 $P_{generalize}$=0.2, $\mathrm{P}_{\mathrm{rand}}$=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.

1.1.2 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 asked adult English speakers 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, Sandbank, Onnis, & Edelman, 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). As noted above, the perplexity should not be disregarded when evaluating precision, because of the tradeoff between them. This analysis was thus carried out twice, once with the smoothing parameters in the trigram model set to optimize its perplexity score (ppl=22.43) and once with the parameters set to achieve ppl=40.07, the perplexity achieved by U-MILA with the parameter values used in all runs.

Six subjects participated in the first precision experiment. The results (see Fig. 2A in the main text) indicated an advantage of U-MILA over SRILM ($t = 3.5$, $p < 0.0005$, 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*: D. Bates, 2005) of results broken down by sentence length (see Fig. 2B in the main text) 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 (MCMC) run to estimate the confidence limits on the slope parameters (R procedures *mcmc* and *HPDinterval*), did not yield a significant difference.

[see Fig. 2A and Fig. 2B in the main text]

In the second precision experiment, six subjects (none of whom participated in the first experiment) evaluated the test sentences. The results obtained this time, when SRILM was set to optimize its perplexity, underscore the tradeoff between perplexity and precision: they indicate an even stronger advantage of U-MILA over SRILM (see Fig. S1A; t=14.4, with p effectively equal to 0; R procedure *lme*: D. Bates, 2005). Sentences from the original corpus received a mean score of 6.7; sentences generated by U-MILA, 6.22; sentences generated by SRILM, 4.2. Including all the generated sentences led to a similar outcome (original: 6.7; U-MILA: 5.91; SRILM: 3.78). When broken down and plotted by sentence length, the results (see Fig. S1B) indicated a faster degradation in score for SRILM- than for U-MILA-generated sentences. A mixed-model analysis (R procedure *lmer*: D. Bates, 2005) confirmed a significant interaction between model type and sentence length for both models (U-MILA: t=-2.57; SRILM, t=-4.03). A comparison of the interaction slopes of the two models, for which we used a 10,000-iteration Markov Chain Monte Carlo (MCMC) run to estimate the confidence limits on the slope parameters (R procedures *mcmc* and *HPDinterval*), did not yield a significant difference. Interestingly, however, the same type of analysis indicated that the score vs. sentence length slope for the original corpus sentences did not differ from that of U-MILA, while the slope for SRILM-generated sentences was significantly larger than for the original ones.

[Fig. S1A and S1B should be here]

U-MILA's advantage in precision can be understood in light of its exemplar-based approach: sequences of words picked up on the strength of actual corpus appearance, particularly within a variation set, are guaranteed to belong together, and thus once a collocation is entered, the sentence generation process, like a well-entrenched behavioral habit, will proceed to its end, reducing the probability of sentence fragmentation to which n-gram models are susceptible.

It should be noted that the actual amount of linguistic input that human learners are exposed to during the first years of their life (Bates & Goodman, 1999) is greater by two or three orders of magnitude than the corpus on which U-MILA was trained. Training on such a corpus, which was impossible with the present implementation (aimed at transparency and not optimized for speed and memory usage), should significantly improve U-MILA's perplexity both directly and indirectly, by allowing for better statistics regarding the edge profile of each node, on which the substitutability calculation is based.

1.2 Equivalence-class inference

The U-MILA model does not attempt to cluster words into "crisp," task-independent equivalence classes in which every word is either a member or not a member in any given syntactic or semantic cluster (for classical arguments advocating graded, task-dependent categories, see Barsalou, 1987; Lakoff, 1987; Rosch, 1978). The information accrued in the graph does, however, support ad hoc similarity-based grouping of units, when needed.

To illustrate the model's ability to learn useful similarity relations from a corpus of natural language, we offer two characterizations of such relations, using 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 each of 11 other chosen nodes. Not surprisingly, the most common nodes are all function words or slot collocations built around function words; their similarity neighbors generally make sense. Thus, in example 1, the neighbors of the determiner *the* are all determiners, and the neighbors of the pronoun *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). In some cases, 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). Such errors could be avoided if more contextual data were retained by the learner.

[see Table 1 in the main text]

Our second illustration of the manner in which U-MILA captures similarity among words takes the form of two plots generated from similarity tables by multidimensional scaling (Shepard, 1980). We used the Matlab procedure *mdscale* to reduce the dimensionality of the word similarity space to 2, while preserving as much as possible the interpoint distances. To keep the resulting plots legible, we sorted the words by frequency and generated layouts for two percentile ranges: 95-100 and 75-80 (Fig. 3A and 3B, respectively). As in the previous analysis, the first of these plots, which corresponds to more frequent items, consists mostly of function words and auxiliary verbs, while the second contains open-class words. In both plots, the proximity among word locations in the map generally corresponds to their intuitive similarity.

[see Fig. 3A and 3B in the main text]

For its estimates of similarity, U-MILA presently relies only on "first-order" context data, that is, it retains, in the form of temporal edges, information regarding which nodes occurred immediately before and after which. Many phenomena in sequence processing, including, of course, language, require, however, contextual information that is both wider-ranging and conceptually broader. In particular, such information can take the form of association between nodes that is not necessarily sentence-sequential, such as that between *beach* and *sand*. Such associations are not supported by the present implementation, although the model can be easily extended to incorporate them, by adding extra fields for various types of similarity bookkeeping to each node, ultimately implementing the idea that the similarity of words should be defined in terms of the similarities of their sentential contexts and vice versa (Karov & Edelman, 1996).

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

Our next set of studies has been inspired by a recent paper by French, Addyman & Mareschal (2011) that describes a connectionist model of unsupervised sequence segmentation and chunk extraction, TRACX, and compares 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). Although sequence segmentation is only one of the many aspects of language acquisition that our approach (but not TRACX or similar models) can address, we found the collection of tests described by French et al. (2011) extremely useful in positioning U-MILA in a burgeoning research field where comparison with existing models is important.

1.3.1 Saffran, Aslin & Newport (1996), experiment 1

In a groundbreaking study, Saffran, Aslin & Newport (1996) showed that 8-month old infants can learn to segment a continuous stream of syllables into "words" using only the transition probabilities between adjacent syllables as cues. A set of 12 syllables was used to construct six tri-syllabic words. In the training phase, subjects were exposed to a sequence of 90 words, composed from this six-word cadre, randomly selected and ordered so as to avoid immediate repeats; there were no pauses anywhere between syllables. Subjects were then tested for ability to discriminate between words (syllable sequences with high transitional probability) and non-words (sequences consisting of the last syllable of one word and the first two syllables of the next one), using a preferential looking paradigm.

Following Saffran et al. (1996), French et al. (2011) created a language of four tri-syllabic words and trained their model learner 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 used by French et al. All test words were assigned higher probability scores (see section 2.7) than non-words, achieving perfect discrimination between the two groups, with the difference approaching significance despite the small number of items (Wilcoxon signed rank test, one-sided; $V = 10$, $p < 0.0625$). Running the model in the flat-Markov mode (by disabling the acquisition of hierarchical representations) led to perfect discrimination between words and non-words. 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.

1.3.2 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 remove this potential frequency confound, Aslin, Saffran & Newport (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. In a testing paradigm similar to that used by Saffran et al. (1996), infants discriminated successfully between the low-frequency words and the non-words that spanned high-frequency word boundaries.

French et al. (2011) constructed a dataset analogous to that of Aslin et al. (1998), composed of 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 and test sets, U-MILA performed perfectly, always assigning a higher probability score to low-frequency words than to non-words (Wilcoxon signed rank test, one-sided; $V = 10$, $p < 0.0625$; the seemingly low significance value despite the perfect discrimination is determined by the small number of items in the test set). As in the previous experiment, using our model in the flat-Markov mode achieved similar results.

1.3.3 Perruchet and Desaulty (2008), experiment 2: forward transition probabilities

In the study by Perruchet and Desaulty (2008), adult subjects listened to a training sequence of 1035 words, each occurring 115 times in a random order. The nine bi-syllabic words were constructed from syllables drawn from a 12-syllable set. Words and non-words appeared in the training set with the same frequency, and differed in that within-word transition probabilities were equal to 1, while transition probabilities within non-words was lower. Following training, the subjects had to indicate for each test item whether or not it was a word that appeared during training. The subjects performed at a level significantly better than chance. French et al. (2011) constructed an analogous dataset and testing scheme and showed that both the TRACX and the SRN models 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; $V = 21$, $p < 0.016$), assigning words a mean probability more than four times greater than the probability assigned to non-words. Unlike in the previous experiments, running the model in its flat-Markov mode did not lead to successful discrimination.[1]

To tease apart the effects of the two mechanisms by which U-MILA detects "words" in this task, we trained and tested two "lesioned" versions of the model. In the first instance, we disabled the alignment procedure (using run-mode *"phonoloop*

[1] 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.

collocation"; see section 2.4); second, we disabled statistical collocation detection (run-mode *"bottom-up collocation"*). Surprisingly, both modified versions of the model performed well. This was expected with alignment disabled and collocation detection in place, because the collocation procedure is based on the relative probability of joint occurrence, of which high forward transition probabilities are a special case. Good performance was, however, unexpected in the second instance, in which chunking occurred solely by alignment. An inspection of the training set resolved this conundrum: it turns out that the distribution of adjacent repeats (cases in which a two-syllable sequence occurs in succession) was different for words and for non-words. For example, all the test words occurred repeatedly at least three times, while three of the non-words did not occur repeatedly at all. If this was also true of Perruchet & Desaulty's (2008) original training set, it may have played a part in their findings.[2]

1.3.4 Perruchet and Desaulty (2008), experiment 2: backward transition probabilities

The second experiment reported by Perruchet & Desaulty (2008) was the first to show that adults can segment chunks out of continuous auditory input on the basis of backward transition probabilities. In the training sequence, for each syllable pair defined as a word (e.g., *X A*), the second syllable's occurrences were always preceded by those of the first (*X* occurs before every occurrence of *A*), but not vice versa: the first syllable (*X*) could be followed by any of a number of syllables (*A*, *B,* and *C*). The overall frequency of test words and test non-words was identical. Forward transition

[2] Note that U-MILA seeks alignments not only when these are adjacent; we report here adjacency of repeated occurrences in order to illustrate that the distribution of words in the dataset was very different from that of non-words.

probabilities within words were lower than those straddling word boundaries (0.2 as opposed to 0.33), while backward transition probabilities were significantly higher within words as opposed to within non-words (i.e., across word boundaries; 1 as opposed to 0.2). In the forced choice test, adult participants chose words significantly more than they chose non-words, as did the TRACX model of French et al. (2011). The SRN model failed in this task.

In our replication of this experiment, which used the same training and test sets, U-MILA successfully assigned significantly higher scores to words than to non-words: the mean probability assigned to words was almost six times greater than that assigned to non-words (Wilcoxon signed rank test, one-sided; V = 21, $p < 0.016$). As expected, the run in a flat-Markov mode did not differentiate between words and non-words. An exploration similar to the one described in the previous experiment revealed that the model, when run in *phonoloop collocation* mode, assigns higher mean probabilities to words than to non-words, contrary to what could have been expected. Here, too, this stems from a different distribution of adjacent occurrences of words as opposed to that of non-words.

1.3.5 Giroux and Rey (2009)

The chunking experiment of Giroux & Rey (2009) is the only one taken up by French et al. (2011) that we did not replicate in full. As explained below, this is due to the functional need to address issues in derivational morphology that are beyond the scope of models that focus merely on sequence chunking, yet are of central concern to a comprehensive model of language such as U-MILA.

Giroux & Rey (2009) showed that once a lexical unit ("sub-chunk") is assimilated into a larger one ("chunk"), it becomes progressively harder to recognize.

French et al. (2011) constructed a training corpus composed of two-, three- and four-syllable words, including the word *klmn*, and repeatedly exposed the TRACX model to this corpus. 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 by the model.

When trained on this training corpus, U-MILA recognized all chunks and sub-chunks (*kl, lm, mn, klm, lmn, klmn*) as independent units. The reason that our language-oriented version of the model does not eliminate sub-chunks even after they are incorporated into larger units is that this step would be counterproductive in many cases; for instance, it would cause the word *dead* to be forgotten after learning the word *deadline*.[3]

1.3.6 Frank, Goldwater, Griffiths & Tenenbaum (2010), experiment 1

In their first experiment, Frank, Goldwater, Griffiths & Tenenbaum (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

[3] In contrast, the version of the model that was applied to birdsong (Menyhart, Kolodny, Goldstein, DeVoogd, & Edelman, submitted) 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.

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) composed a similar dataset, with the words *ab, cd, efg, hij, klmn, opqr*, and trained and tested TRACX in each of the conditions, presenting each sentence during the training six times. Similarly to the subjects of Frank et al. (2010), TRACX discriminated between words and part-words (i.e. non-words composed of the last and the first parts of adjacent words) better as the sentence length 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 (see Fig. 4). This result holds for a range of parameters, with correlation with the human data ranging from 0.49 to 0.87. We attribute little importance to the specific correlation value, as this measure is quite variable and is sensitive to multiple assumptions and design choices that are rather arbitrary.[4] For instance, using as the measure of discrimination success the difference between the mean probability assigned to words and that assigned to non-words leads to a correlation of 0.81, while analyzing these same results using the proportion-better score (French et al., 2011) leads to a correlation of 0.49. Notably, even running the model in its flat-Markov mode yields the same trend in discrimination ability among the different experimental conditions, with a correlation coefficient of 0.82 for the proportion-better score. In

[4] Our model assigns higher scores to words than to part-words in all cases, and so we cannot use a percentage of correct classifications as our measure of discrimination ability. Different measures of discrimination reflect different possible mechanisms of stochastic decision-making regarding such discrimination; we remain agnostic as to which is the most realistic.

any case, improved discrimination for shorter sentences may be explained by the fact that the splitting of the dataset into more sentences is not done at random locations, but at the borders between words. Thus, the overall number of occurrences of each word in the dataset does not change from one condition to another, while the number of part-word occurrences changes significantly, becoming smaller as the sentence gets shorter.

[see Fig. 4 in the main text]

1.3.7 Frank, Goldwater, Griffiths & Tenenbaum (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 the difficulty of differentiating between words and non-words in it. 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, depending on condition. Words varied in length 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) found that as the word cadre got smaller, the subjects' ability to discriminate between words and non-words following the training 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 Frank 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 to occur only under a specific range of parameters. Specifically, it was obtained when the model was run in the flat-Markov mode, or when the prior against creating collocations was strong and the phonological loop coefficient of decay was very large or the alignment module disabled. The score used in all cases was the difference between the mean probability assigned to words and that assigned to non-words. An analysis of covariance (R procedure *lm*) applied to a typical finding (see Fig. 5A, 5B) yielded significant effects of word-hood ($t = 3.0$, $p < 0.0039$) and vocabulary size ($t = -5.46$, $p < 0.0000015$) and a significant interaction ($t = 2.1$, $p < 0.04$). 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.

[see Fig. 5A and 5B in the main text]

1.3.8 French, Addyman & Mareschal (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). This corpus of 9,800 sentences contains 1321 distinct words and a total of 33,400 words and 95,800 phonemes. Sentences in the corpus have no pauses between words, the objective of the model being to extract the words from the continuous stream of phonemes. During the learning phase, French et al. (2011) presented their model with each sentence six times in succession, completing five passes through the corpus. The test phase was patterned after the tests in the previously described simulations: equal numbers (496) of bi- and tri-syllabic words and part-words from the corpus were individually presented to the model, and the average error over the model's output layer units was used as a measure of the model's error. Their TRACX model showed significantly better learning of words than of non-words.

We trained the U-MILA model 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, for both bi- and tri-syllabic words/part-words (see Fig. 6). An analysis of covariance (R procedure *lm*) yielded significant effects of word-hood ($t = 2.1$, $p < 0.035$) and number of syllables ($t = -7.08$, $p < 2.9\text{X}10^{-12}$) and no interaction.

Of the 496 words in the test set, 362 were recognized by U-MILA as units, and received a status of a supernode in the graph. Of these, 271 were uncovered by both collocation mechanisms (top-down segmentation, which recognizes recurring sequences through alignment, and bottom-up segmentation based on repeated co-occurrence of the syllables), 24 were uncovered by the alignment mechanism only, and 67 were found by the bottom-up mechanism only. Thus it seems that on this short natural language corpus, the mechanisms complete each other to some extent, with a significant overlap among them.

[see Fig. 6 in the main text]

1.3.9 French, Addyman & Mareschal (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 the original experiment 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). French et al. (2011) trained the TRACX model on this corpus, then tested it on all possible words in both languages. To examine the model's ability to categorize words by language and to generalize category labels to new words that had not appeared in the training set, the activities in the hidden layer of the TRACX model evoked by each test word were clustered. The resulting clustering contained very few mistakes, both on familiar and on previously unseen words.

Although our model 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[5]. 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)+sim(c,f))/3.[6] A clustering algorithm (Matlab

[5] All 48 words were uncovered by both collocation mechanisms (top-down and bottom-up, see section 2).

[6] This is equivalent to using Levenshtein distance over strings (e.g., Ristad & Yianilos, 1998).

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.

It is important to point out that similarity between nodes in U-MILA is estimated based on the edge profiles of the entire nodes, and not on the similarity of their internal structure, if any. Only when required to categorize units that had not been encountered previously, and thus were not granted a node status in the graph, does it resort to categorization based on the edge profile of the unit's constituents. U-MILA's success in categorizing both previously encountered and novel words shows that it can incorporate both approaches and to do so in a transparent manner that allows clear understanding of the underlying causes in each case (see also Lotem & Halpern's discussion of generalization and concept formation; Lotem & Halpern, 2008).

[see Fig. 7 in the main text]

1.3.10 French, Addyman & Mareschal (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, but 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 ($t = 2.213$, $p < 0.04$, see Fig. 8). All words were recognized as units by both of U-MILA's collocation mechanisms.

[see Fig. 8 in the main text]

1.4 Gomez (2002), experiment 1

Gomez (2002) reported that both adults and infants can learn nonadjacent 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. The participants in her study were exposed to 3-word sequences (e.g., *pel wadim rud*) from one of two languages, which differed in the dependencies between the first and third words (e.g., if *pel* is the first word in the sentence, then in L1 *rud* must be the third, while in L2 *jic* must be the third), but were identical in their dependencies between adjacent words. They were then tested for acceptance of 3-word sentences from both languages. The size of the pool from which the middle words were chosen was varied systematically (2, 6, 12 or 24 words) in order to determine how the variability in the statistics of adjacent dependencies affects learning of nonadjacent dependencies.

While participants showed preference to sentences from the language they had learned in all cases, the difference was significant only in the case of the largest pool

(24 words). Gomez suggested that this result may be due to the participants' default tendency to search for adjacent dependencies (which are useless in this discrimination task), and to shift the focus to nonadjacent dependencies only when the adjacent dependencies seem unreliable.

We trained one instance of the U-MILA model on each of the datasets used by Gomez (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.

[see Table 2 in the main text]

The results are summarized in Fig. 9. Nonadjacent 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 the sentences were in accord with the rules of the training set's language) and pool size ($t = -22.7$, $p < 2X10^{-16}$; $t = -14.4$, $p < 2X10^{-16}$) and a significant interaction ($t=3.0$, $p < 0.0045$).[7] There was, however, no abrupt change in performance between pool sizes 12 and 24, contrary to the effect reported by Gomez (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).

[7] This interaction amounted to a small (in absolute terms) difference in the slopes of the grammaticality effect, rather than in a change of the sign of the effect. As such, it does not reflect on the rest of the discussion of this experiment.

In further support of this explanation, the model fails 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.

[see Fig. 9 in the main text]

1.5 Gomez and Lakusta (2004), experiment 1

Gomez and Lakusta (2004) showed that infants are capable of unsupervised learning of syntactic categories and rules in an artificial language. Infants were exposed to a 3-minute auditory sequence of 72 sentences in one of two artificial languages and were then tested using a preferential looking paradigm (Saffran, et al., 1996) to determine whether they developed sensitivity to the regularities in the language they had been trained in. Sentences in language L1 were composed of two phrases, each of the form *aX* or *bY*, where *a*, *b*, *X*, and *Y* were word categories; *a*, *b*, and *Y* words were monosyllabic, and *X* words disyllabic (Table 3). Language L2 sentences were composed of *aY* and *bX* phrases.

The test sentences contained words from categories *a* and *b* that had appeared in the training sequence, but all the *X* and *Y* words were novel. Infants attended significantly longer to sentences from the language they had been trained on than to sentences from the other language.

[see Table 3 in the main text]

We trained a U-MILA instance on an L1 training set, patterned after that of Gomez and Lakusta (2004), with spaces inserted between each two consecutive syllables and a random ordering of the sentences. The learner then assigned a probability score to each of the test sentences in Gomez and Lakusta (2004). The model's parameter that controls its sensitivity to slot filler length, $B_{FillerSetSizeSensitivity}$ (see section 2.2), was set so the learner would be sensitive to the filler set size, measured in syllables.

Sentences from 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 < 8.9X10^{-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.

[see Fig. 10 in the main text]

1.6 Onnis, Waterfall, & Edelman (2008), experiment 1

The first experiment in the study of Onnis, Waterfall & Edelman (2008) examined the effects of variation sets[8] on artificial grammar learning in adult human subjects. 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.

[8] 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).

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 < 0.0001; 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 < 0.0001). Crucially, the word-hood × condition interaction was significant (t = 57.8, p < 0.0001).

A further exploration revealed that, 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 form a variation set being simultaneously present within the loop (in addition to not decaying too quickly).

[see Fig. 11 in the main text]

1.7 Reali & Christiansen, (2005)

The experiment of Reali & Christiansen (2005) replicated here is the first of two test cases in which we examine the ability of U-MILA to deal with "structure dependence" — a general characteristic of the human language faculty that, according to some theorists, cannot be due exclusively to statistical learning from unlabeled examples, requiring instead that the structure in question be built into the learner as an "innate" constraint (Chomsky, 1980). Reali & Christiansen (2005) set out to demonstrate that one of the poster cases of the Poverty of the Stimulus Argument for innateness in linguistics (Chomsky, 1980) — 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?*

Their training corpus is composed of sentences uttered by nine mothers addressing their children, recorded over a period of 4 to 5 months, while the children were of ages 1:1 to 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 & 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 the U-MILA model on all the sentences made available to us by Reali & 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, $p < 7.08X10^{-07}$).

[see Fig. 12 in the main text]

1.8 Pearl & Sprouse (2012): Island constraints and long-term dependencies

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, Wagers, & Phillips, 2012a; see footnote for an example). Recently, Sprouse, Fukuda, Ono & Kluender (2011) conducted a quantitative study of the interaction between grammatical island constraints and short- and long-term dependencies in determining sentence acceptability. They used a factorial design, with four types of sentences: (i) short-term dependency + no island, (ii) long-term dependency + no island, (iii) short-term dependency + island, (iv) long-term dependency + island.[9] The pattern of acceptability judgments exhibited the signature of the island effect: an interaction between the two variables, 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 on reductionist theories and others (Hofmeister, Casasanto, & Sag, 2012a, 2012b; Sprouse, et al., 2012a; Sprouse, Wagers, & Phillips, 2012b).

[9] 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.

In an attempt to account for this finding by a statistical learning model, Pearl & Sprouse (2012) trained a parser to recognize shallow phrasal constituents in sentences represented as sequences of part of speech (POS) tags, while collecting the statistics of POS trigrams covering these parses. 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 different 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 by our model, hypothesizing that the collocations that it learns, which in some sense are analogous to POS n-grams, may lead to the emergence of an interaction between islands and dependency length. 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 type of test set 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 & Sprouse (2012); words that did not occur in the training corpus were replaced with words of the same part of speech that did. The trained instance of U-MILA assigned probabilities to each of the test sentences, which we then analyzed and plotted as in Pearl & 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.

Bibliography for SM2

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.

Barsalou, L. W. (1987). The instability of graded structure: implications for the nature of concepts*. Concepts and conceptual development,* pp. 101-140.

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

Bates, E., & Goodman, J. C. (1999). On the Emergence of Grammar From the Lexicon*. Emergence of Language,* pp. 29-79.

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.

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

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

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

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

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*.

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.

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.

Jelinek, F. (1990). Self-organized language modeling for speech recognition*. Readings in Speech Recognition,* pp. 450-506.

Karov, Y., & Edelman, S. (1996). *Learning similarity-based word sense disambiguation from sparse data* (CS-TR): The Weizmann Institute of Science.

Küntay, A., & Slobin, D. (1996). Listening to a Turkish mother: Some puzzles for acquisition. *Social interaction, social context, and language: Essays in honor of Susan Ervin-Tripp,* pp. 265-286.

Lakoff, G. (1987). *Women, Fire and Dangerous Things: What Categories Reveal about the Mind*. Chicago, IL: University of Chicago Press.

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.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Erlbaum.

Menyhart, O., Kolodny, O., Goldstein, M. H., DeVoogd, T. J., & Edelman, S. Like father , like son: zebra finches learn structural regularities in their tutors' song.

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*: Cambridge University Press.

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.

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

Ristad, E. S., & Yianilos, P. N. (1998). Learning String-Edit Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*, 522-532.

Rosch, E. (1978). Principles of categorization*. Cognition and Categorization,* pp. 27-48.

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

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.

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 Proc. Intl. Conf. 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.

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

Figure S1A

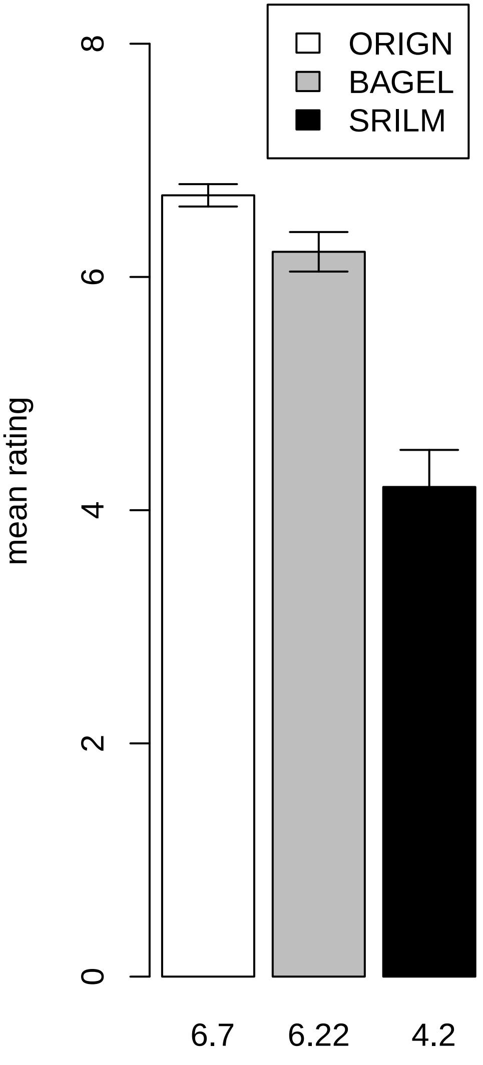
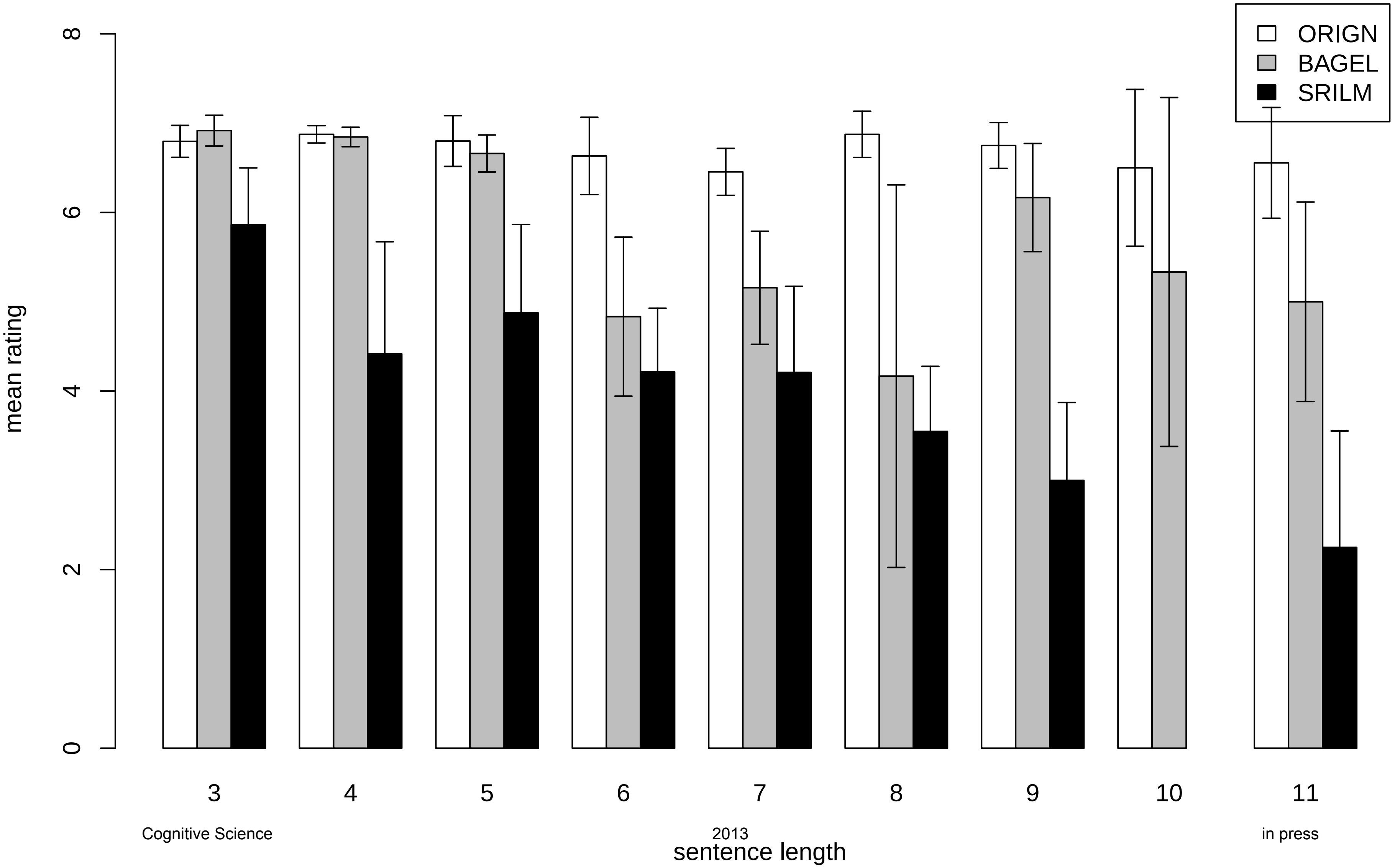


Figure S1B


□ ORIGN
▨ BAGEL
■ SRILM
mean rating
0
2
4
6
8
3
4
5
6
7
8
9
10
11
sentence length

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

Attempting a reduction of U-MILA to another formalism would take us too far away from the main thrust of the present project, and so we offer instead some informal analogies and observations.

On the face of it, the U-MILA graph looks like a finite state automaton (FSA). In the light of the classical arguments that invoke the Chomsky hierarchy (Hopcroft & Ullman, 1979), this would seem like a severe limitation. In practice, however, it is not: if realistic limits on center embedding are assumed (Christiansen & Chater, 1999), the class of sentences that needs to be represented can be readily represented within a finite-state framework (Roche & Schabes, 1997). That said, it should be noted that the power of a FSA can be easily extended by adding a push/pop operation that temporarily shifts activation from one part of the graph to another and eventually returns it to the originating node — an operation not unlike a shift of perceptual attention, for which neuromorphic architectures have been proposed (Itti, Koch, & Niebur, 1998). The result is a Recursive Transition Network or RTN (Woods, 1970) — a CFG-equivalent automaton that supported the first practical natural-language question answering system (Woods, Kaplan, & Nash-Webber, 1972). Allowing feature checking and side effects on transitions turns RTN into Augmented Transition Network, or ATN (Wanner & Maratsos, 1978), which has the formal power of a Turing Machine and can therefore accept recursively enumerable languages, a family of which context-free languages are a proper subset.

As noted in the main text, U-MILA can, by virtue of its ability to learn slot-collocations, learn and represent infinite central embedded recursions. Fig. 13 and

Fig. S1 illustrate two graphs which represent such grammars and some outputs produced by these graphs. For the sake of clarity, the graphs shown in both cases are simplified versions of the graphs learned by the model. See SM4 for an extensive corpus of sequences produced by these graphs.

[Fig. S2 should be here]

The current implementation of the model was not aimed specifically at learning such grammars. In order for it to learn a PCFG and produce only sequences that cannot be accounted for by a finite state automaton, specific parameter values were used: no smoothing was applied to the graph, and in producing sequences (by following possible trajectories along the graph, see section 2), longer sequences were given strict preference. Also, the learning of grammars that would clearly illustrate the model's ability to learn a PCFG depended crucially on the structure of the training set. For this purpose, we engineered the training set in a way that would avoid the construction of collocations and links that might obscure the recursive central embedding in the learned grammar. Notably, this gives rise to a large difference between the training corpus and the target corpus that the learner eventually produces.

As is always the case with formal representational schemes, the real challenge lies, of course, not in endowing one's model with sufficient power to accept/generate languages from some desirable family (which is all too easy), but rather in shaping its power in just the right way so as to accept/generate the right structures and reject all others. As Chomsky (2004, p. 92) commented, "It is obvious, in some sense, that processing systems are going to be represented by finite state transducers. That has got to be the case, and it is possible that they are represented by finite state transducers with a pushdown tape. [...] But that leaves quite open the question of what is the internal organization of the system of knowledge."

We note that the present version of U-MILA is not aimed at learning context free grammars. Its learning process gives rise to extensive redundancy in accounting for input sequences and in generation of sequences. Thus, training U-MILA on a typical corpus produced by a recursive rewrite rule usually leads to the learning of a grammar that accepts and generates recursions but whose set of outputs can also be accounted for by a finite state grammar. As a result, demonstrating U-MILA's ability to learn PCFGs required the construction of specific training sets as noted above. Further exploration is required to follow up on the promise that U-MILA may hold for the learning of context free regularities in natural language and other behavioral modalities.

Bibliography for SM3

Chomsky, N. (2004). *The Generative Enterprise Revisited*. Berlin: Mouton de Gruyter.

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

Hopcroft, J. E., & Ullman, J. D. (1979). *Introduction to Automata Theory, Languages, and Computation*. Reading, MA: Addison-Wesley.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*, 1254-1259.

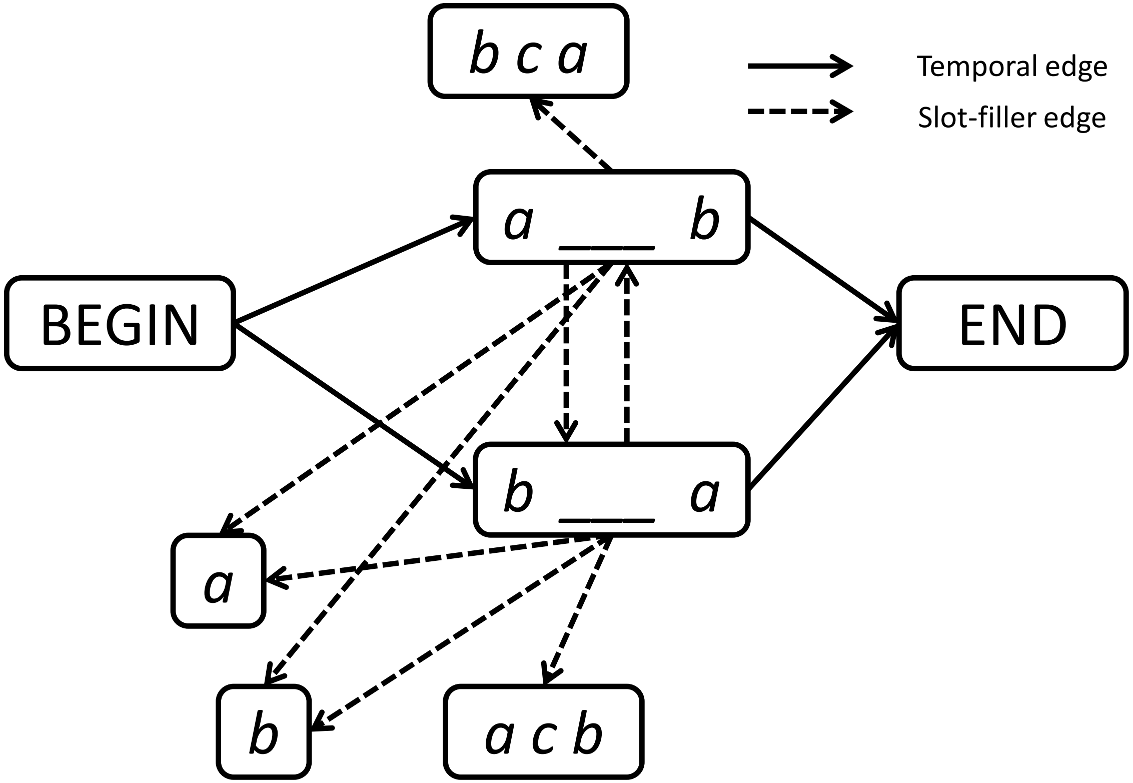
Roche, E., & Schabes, Y. (1997). *Finite-State Language Processing*. Cambridge, MA: MIT press.

Wanner, E., & Maratsos, M. (1978). An ATN approach to comprehension. *Linguistic theory and psychological reality,* pp. 119-161.

Woods, W. A. (1970). Transition Network Grammars of Natural Language Analysis. *Communications of the ACM, 13*, 591-606.

Woods, W. A., Kaplan, R., & Nash-Webber, B. (1972). *The LUNAR sciences natural language information system: Final report* (BBN Report 2378): Bolt Beranek and Newman.

Figure S2



Examples of output sequences:

BEGIN a b a b a b b a b a b END
BEGIN b a b a a b a END
BEGIN b a b a b c a b a b a END
BEGIN b a b a b a b a a b a b a b a END
BEGIN a b a b a a b a b END
BEGIN b a b a b b a b a END
BEGIN a b a c b a b END
BEGIN b a a b a END
BEGIN b a b a b a b b a b a b a END
BEGIN a b c a b END
BEGIN a b a b a b a b c a b a b a b a b END
BEGIN a b a a b END

The learned grammar is equivalent to the following set of rewrite rules:

BEGIN (a b)^n {a, b, a c b} (a b)^n END
BEGIN (a b)^n a {a, b, b c a} b (a b)^n END
BEGIN (b a)^n {a, b, b c a} (b a)^n END
BEGIN (b a)^n b {a, b, a c b} a (b a)^n END

n ϵ {0,1,2,…}

Supplementary Material 4 : output sequences from two PCFG grammars learned by U-MILA

After training U-MILA on a short corpus, we used it to produce multiple output sequences. These appear below in the order of their production, with repeating sequences omitted. To facilitate the interpretation of the output in terms of the graph nodes, each set of output sentences is shown twice, with brackets in the second occurrence marking the structural parse of each sequence, so as to expose its recursive structure.

```
The grammar presented in Fig. 13
--------------------------------
The grammar is equivalent to the set of rewrite rules:

           n                    n
BEGIN (a b)  {a, b, a a a} (b a)  END
           n                       n
BEGIN (a b)  a {a, b, b b b} a (b a)  END
           n                    n
BEGIN (b a)  {a, b, b b b} (a b)  END
           n                       n
BEGIN (b a)  b {a, b, a a a} b (a b)  END

                 n = {0,1,2,…}

Output (without brackets):

BEGIN b a b a b a a a b a b a b END
BEGIN b a b END
BEGIN a b a a a b a END
BEGIN b a b a b a b b b a b a b a b END
BEGIN a b a b a b b b a b a b a END
BEGIN b a b a b a b a b END
BEGIN a b a b b b a b a END
BEGIN a b a b a b a END
BEGIN b b b END
BEGIN a b a b a b a b a b a b a b a END
BEGIN b a b b b a b END
BEGIN b a b a b END
BEGIN b a b a b a b a b a b a b END
BEGIN a b a b a b a b a b a b a END
BEGIN b a b a b a b a b a b a b a b END
BEGIN a b a a a b a END
BEGIN a b a b a b a b b b a b a b a b a END
BEGIN a b a b a b b b a b a b a END
BEGIN a b a b a b a b a b a b a b a b a END
BEGIN a b a b a a a b a b a END
BEGIN b a b a b b b a b a b END
BEGIN b a b a b a b b b a b a b a b END
BEGIN a b a b a b a a a b a b a b a END
BEGIN b a a a b END
BEGIN b a b a b a b END
BEGIN a b a END
BEGIN b a b a b a a a b a b a b END
BEGIN b a b a a a b a b END
BEGIN a b a b a b a b a b a END
BEGIN a b a b b b a b a END
BEGIN a b a b a b a b b b a b a b a b a END
BEGIN b a b a b a b a a a b a b a b a b END
BEGIN a b b b a END
BEGIN a b a b a a a b a b a END
BEGIN b a b a b a b a a a b a b a b a b END
BEGIN b a b b b a b END
BEGIN a a a END
BEGIN a b a b a b a b a END
BEGIN b a b a b a b a b a b END
BEGIN b a b a b b b a b a b END
BEGIN a b b b a END
BEGIN a b a b a END
BEGIN b a a a b END
BEGIN b a b a a a b a b END
```

```
Output (with brackets marking node structure):
```

```
BEGIN b [a [b [a [b [a [a] a] b] a] b] a] b END
BEGIN b [a] b END
BEGIN a [b [a [a] a] b] a END
BEGIN b [a [b [a [b [a [b [b] b] a] b] a] b] a] b END
BEGIN a [b [a [b [a [b [b] b] a] b] a] b] a END
BEGIN b [a [b [a [b] a] b] a] b END
BEGIN a [b [a [b [b] b] a] b] a END
BEGIN a [b [a [b] a] b] a END
BEGIN b [b] b END
BEGIN a [b [a [b [a [b [a [b] a] b] a] b] a] b] a END
BEGIN b [a [b b b] a] b END
BEGIN b [a [b] a] b END
BEGIN b [a [b [a [b [a [b] a] b] a] b] a] b END
BEGIN a [b [a [b [a [b [a] b] a] b] a] b] a END
BEGIN b [a [b [a [b [a [b [a] b] a] b] a] b] a] b END
BEGIN a [b [a a a] b] a END
BEGIN a [b [a [b [a [b [a [b b b] a] b] a] b] a] b] a END
BEGIN a [b [a [b [a [b b b] a] b] a] b] a END
BEGIN a [b [a [b [a [b [a [b [a] b] a] b] a] b] a] b] a END
BEGIN a [b [a [b [a a a] b] a] b] a END
BEGIN b [a [b [a [b b b] a] b] a] b END
BEGIN b [a [b [a [b [a [b b b] a] b] a] b] a] b END
BEGIN a [b [a [b [a [b [a a a] b] a] b] a] b] a END
BEGIN b [a a a] b END
BEGIN b [a [b [a] b] a] b END
BEGIN a [b] a END
BEGIN b [a [b [a [b [a a a] b] a] b] a] b END
BEGIN b [a [b [a a a] b] a] b END
BEGIN a [b [a [b [a [b] a] b] a] b] a END
BEGIN a [b [a [b b b] a] b] a END
BEGIN a [b [a [b [a [b [a [b [b] b] a] b] a] b] a] b] a END
BEGIN b [a [b [a [b [a [b [a a a] b] a] b] a] b] a] b END
BEGIN a [b b b] a END
BEGIN a [b [a [b [a [a] a] b] a] b] a END
BEGIN b [a [b [a [b [a [b [a [a] a] b] a] b] a] b] a] b END
BEGIN b [a [b [b] b] a] b END
BEGIN a [a] a END
BEGIN a [b [a [b [a] b] a] b] a END
BEGIN b [a [b [a [b [a] b] a] b] a] b END
BEGIN b [a [b [a [b [b] b] a] b] a] b END
BEGIN a [b [b] b] a END
BEGIN a [b [a] b] a END
BEGIN b [a [a] a] b END
BEGIN b [a [b [a [a] a] b] a] b END
```

The grammar presented in Figure S1

The grammar is equivalent to the set of rewrite rules:

BEGIN $(a\ b)^{n}$ {a, b, a c b} $(a\ b)^{n}$ END

BEGIN $(a\ b)^{n}$ a {a, b, b c a} b $(a\ b)^{n}$ END

BEGIN $(b\ a)^{n}$ {a, b, b c a} $(b\ a)^{n}$ END

BEGIN $(b\ a)^{n}$ b {a, b, a c b} a $(b\ a)^{n}$ END

n = {0,1,2,…}

Output (without brackets):

```
BEGIN a b a b a b b a b a b END
BEGIN b a b a a b a END
BEGIN b a b b a END
BEGIN b a b a b c a b a b a END
BEGIN b a b a b a b a a b a b a b a END
BEGIN b a b a b a b a b b a b a b a b a b a END
BEGIN a b a b a a b a b END
BEGIN b a b a b b a b a END
BEGIN a b a c b a b END
BEGIN b a b a b a b a b c a b a b a b a b a END
BEGIN b a a b a END
BEGIN b a b a b a b b a b a b a END
BEGIN a b c a b END
BEGIN a b a b a b a b c a b a b a b a b END
BEGIN a b a a b END
BEGIN a b a b c a b a b END
BEGIN b a b a c b a b a END
BEGIN a b a a b a b END
BEGIN a b a b a b a b c a b a b a b a b END
BEGIN a b a b a b a b a c b a b a b a b a b END
BEGIN b a b a b a b a c b a b a b a b a END
BEGIN b a b a a b a b a END
BEGIN b a b a b a b a b a a b a b a b a b a END
BEGIN a b a c b a b END
BEGIN a b a b a b a c b a b a b a b END
BEGIN a b c a b END
BEGIN b a a b END
BEGIN a b a b a b a c b a b a b a b END
BEGIN a b a b a b a a b a b a b END
BEGIN a b a b a b c a b a b a b END
BEGIN a b b END
BEGIN b a c b a END
BEGIN a b a b a c b a b a b END
BEGIN b a b a c b a b a END
BEGIN b a b a b c a b a b a END
BEGIN a b a b a b a b b a b a b a b END
BEGIN b a c b a END
BEGIN b a b a b a a b a b a END
BEGIN a b a b a b c a b a b a b END
BEGIN b a b a b a c b a b a b a END
BEGIN a b a b c a b a b END
BEGIN a c b END
BEGIN a b a b a c b a b a b END
BEGIN b c a END
BEGIN b a a END
BEGIN a b a b b a b END
BEGIN b a b c a b a END
BEGIN a a b END
BEGIN b a b a b a c b a b a b a END
BEGIN b a b c a b a END
BEGIN a b a b a b a a b a b a b a b END
BEGIN a b a b a a b a b a b END
BEGIN b a b a b a a b a b a b a END
BEGIN b a b a b a b c a b a b a b a END

Output (with brackets marking node structure):

BEGIN a [b [a [b [a [b] b] a] b] a] b END
BEGIN b [a [b [a] a] b] a END
BEGIN b [a [b] b] a END
BEGIN b [a [b [a [b c a] b] a] b] a END
BEGIN b [a [b [a [b [a [b [a] a] b] a] b] a] b] a END
BEGIN b [a [b [a [b [a [b [a [b [a [b] b] a] b] a] b] a] b] a] b] a END
BEGIN a [b [a [b [a] a] b] a] b END
BEGIN b [a [b [a [b] b] a] b] a END
BEGIN a [b [a c b] a] b END
BEGIN b [a [b [a [b [a [b [a [b c a] b] a] b] a] b] a] b] a END
BEGIN b [a [a] b] a END
BEGIN b [a [b [a [b [a [b] b] a] b] a] b] a END
BEGIN a [b [c] a] b END
BEGIN a [b [a [b [a [b [a [b c a] b] a] b] a] b] a] b END
BEGIN a [b [a] a] b END
BEGIN a [b [a [b c a] b] a] b END
BEGIN b [a [b [a c b] a] b] a END
BEGIN a [b [a [a] b] a] b END
BEGIN a [b [a [b [a [b [a [b [c] a] b] a] b] a] b] a] b END
BEGIN a [b [a [b [a [b [a [b [a c b] a] b] a] b] a] b] a] b END
BEGIN b [a [b [a [b [a [b [a c b] a] b] a] b] a] b] a END
BEGIN b [a [b [a [a] b] a] b] a END
BEGIN b [a [b [a [b [a [b [a [b [a] a] b] a] b] a] b] a] b] a END
BEGIN a [b [a [c] b] a] b END
BEGIN a [b [a [b [a [b [a c b] a] b] a] b] a] b END
BEGIN a [b c a] b END
BEGIN b [a] a b END
BEGIN a [b [a [b [a [b [a [c] b] a] b] a] b] a] b END
BEGIN a [b [a [b [a [b [a] a] b] a] b] a] b END
BEGIN a [b [a [b [a [b [c] a] b] a] b] a] b END
BEGIN a [b] b END
BEGIN b [a [c] b] a END
BEGIN a [b [a [b [a [c] b] a] b] a] b END
BEGIN b [a [b [a [c] b] a] b] a END
BEGIN b [a [b [a [b [c] a] b] a] b] a END
BEGIN a [b [a [b [a [b [a [b] b] a] b] a] b] a] b END
BEGIN b [a c b] a END
BEGIN b [a [b [a [b [a] a] b] a] b] a END
BEGIN a [b [a [b [a [b c a] b] a] b] a] b END
BEGIN b [a [b [a [b [a [c] b] a] b] a] b] a END
BEGIN a [b [a [b [c] a] b] a] b END
BEGIN a [c] b END
BEGIN a [b [a [b [a c b] a] b] a] b END
BEGIN b [c] a END
BEGIN b [a] a END
BEGIN a [b [a [b] b] a] b END
BEGIN b [a [b [c] a] b] a END
BEGIN a [a] b END
BEGIN b [a [b [a [b [a c b] a] b] a] b] a END
BEGIN b [a [b c a] b] a END
BEGIN a [b [a [b [a [b [a [a] b] a] b] a] b] a] b END
BEGIN a [b [a [b [a [a] b] a] b] a] b END
BEGIN b [a [b [a [b [a [a] b] a] b] a] b] a END
BEGIN b [a [b [a [b [a [b c a] b] a] b] a] b] a END
```

Supplementary material 5: simulation parameters

	Analysis Mode	D_{short_term}	Slot Collocations allowed?	$B_{FillerSetSizeSensitivity}$	$P_{generalize}$	Pc	$MinOccs$	P_{rand}
3.1–3.2 Generativity & Categorization	normal	0.01	Yes	false	0.05	0.5	2	0.001
3.3.1 Saffran et al. (1996)	normal / flat markov	0.01	Yes	false	0.05	0.5	2	0.001
3.3.2 Aslin et al. (1998)	normal / flat markov	0.01	Yes	false	0.05	0.5	2	0.001
3.3.3 Perruchet Desaulty Forward TPs	all	0.01	No	false	0.05	0.25	2	0.001
3.3.4 Perruchet & Desaulty Backward TPs	all	0.01	No	false	0.05	0.25	2	0.001
3.3.6 Frank et al., Exp 1	normal	0.01	No	false	0.05	0.5	2	0.001
3.3.7 Frank et al., Exp. 3	bottom-up colocation	0.01	No	false	0.05	0.1	4	0.001
3.3.8 French et al., Sim 8	normal	0.01	No	false	0.05	0.25	2	0.001
3.3.9 French et al., Sim. 10	normal	0.01	No	false	0.05	0.25	4	0.001
3.3.10 French et al., Sim. 11	normal	0.01	No	false	0.05	0.25	2	0.001
3.4 Gomez (2002)	normal	0.01	Yes	false	0.05	0.25	2	0.001

3.5 Gomez-Lakusta (2004)	normal	0.01	Yes	True	0.05	0.25	2	0.001
3.6 Onnis et al. (2008)	normal	0.035, 0.055, 0.075	No	false	0.05	0.25	30	0.001
3.7 Reali & Christiansen 2005	normal	0.01	Yes	false	0.05	0.5	2	0.001