PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

How semantic biases in simple adjacencies affect learning a complex structure with non-adjacencies in AGL: a statistical account

Fenna H. Poletiek and Jun Lai

Phil. Trans. R. Soc. B 2012 **367**, 2046-2054
doi: 10.1098/rstb.2012.0100

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2012) **367**, 2046–2054
doi:10.1098/rstb.2012.0100

Research

How semantic biases in simple adjacencies affect learning a complex structure with non-adjacencies in AGL: a statistical account

Fenna H. Poletiek[1,2,*] **and Jun Lai**[1]

[1]*Cognitive Psychology Department, Leiden University, Pieter de la Court building, PO Box 9555, 2300 Leiden, The Netherlands*
[2]*Max Planck Institute of Psycholinguistics, Nijmegen, The Netherlands*

A major theoretical debate in language acquisition research regards the learnability of hierarchical structures. The artificial grammar learning methodology is increasingly influential in approaching this question. Studies using an artificial centre-embedded A^nB^n grammar without semantics draw conflicting conclusions. This study investigates the facilitating effect of distributional biases in simple AB adjacencies in the input sample—caused in natural languages, among others, by semantic biases—on learning a centre-embedded structure. A mathematical simulation of the linguistic input and the learning, comparing various distributional biases in AB pairs, suggests that strong distributional biases might help us to grasp the complex A^nB^n hierarchical structure in a later stage. This theoretical investigation might contribute to our understanding of how distributional features of the input—including those caused by semantic variation—help learning complex structures in natural languages.

Keywords: language acquisition; semantic biases; statistical learning

1. INTRODUCTION

Recursion has been argued to be a crucial property of human language [1]. However, hierarchical recursive structures with non-adjacent dependencies are known to be difficult to process, even for native speakers [2–5]. More generally, the learnability of recursive non-linear hierarchical structures is subject to a long running debate in language research. *The rat the cat the dog chased killed ate the malt* [6] is a typical hierarchical sentence with two centre-embedded sub-clauses. Formally, these phrase structure grammars [7] are more complex than linear grammars that can be implemented with a simple finite-state system. In particular, centre embedding structures entail long distances between related dependencies (e.g. *rat* and *malt*), causing processing and learning difficulties. This raises the question as to how humans learn and can use these structures after all.

Fitch and colleagues [7,8] proposed that the capability of mastering supra-regular hierarchical structures was critical to distinguish human and non-human primates, indicating a possible innate faculty for complex language in humans. This proposition has boosted a renewed interest in processing and learning complex hierarchical grammars, especially supra-regular grammars, in the formal hierarchy of grammars proposed by Chomsky [9]. A new approach to the question of learnability of hierarchical structures, comes from experimental studies using artificial grammars and the artificial grammar learning (AGL) paradigm [7,10–15].

In the classical AGL paradigm [16–18], the grammar learning process is studied in a controlled laboratory environment in which participants are first exposed to exemplars of an artificial grammar. After the training phase, participants are informed that the strings they studied obeyed an underlying grammar. Finally, in the test phase, participants give grammaticality judgements about new strings similar to those studied during training, being either grammatical or not. The proportion of correct judgements is an indication of how much of the grammar has been learned. Typically, participants perform above chance, often without having any explicit knowledge of the grammar [17]. In the early period of this paradigm, these artificial grammars typically are simple finite-state systems with a limited number of symbols (e.g. five letters) generating sequences of these symbols. By manipulating various aspects of the experimental setting separately (e.g. the grammatical rules, the training input sample, learning conditions and instructions), controlled tests can be performed of specific influences, possibly at play in the natural situation [12].

Whereas during the first decades of the AGL paradigm, mostly regular grammars were used, the focus has shifted recently towards grammars with more hierarchical complexity [10,19–21]. The results with

* Author for correspondence (poletiek@fsw.leidenuniv.nl).

One contribution of 13 to a Theme Issue 'Pattern perception and computational complexity'.

artificial grammars are inconclusive, however, as to whether and how this complex type of structures can be learned by induction from stimulus exemplars only. Friederici and colleagues [19,22] carried out functional magnetic resonance imaging (fMRI) research into the neural basis of processing hierarchical structures. Significantly greater blood flow was observed in Broca's area during processing of hierarchical strings generated by a context-free grammar than of linear strings generated by a finite-state grammar, supporting the possibility that a specific neural circuit is engaged by this type of complex structures.

However, subsequent critical studies argued that only superficial learning of the centre-embedding structure had been demonstrated in the fMRI results. These studies [10,13] showed that the specific hierarchical mapping of dependent words in centre-embedded constructions (like mapping *rat* on *ate*, *cat* on *killed* and *dog* on *chased* in '*The rat the cat the dog chased killed ate the malt*'), being an essential characteristic of centre-embedded structures, had not yet been demonstrated in AGL research. The critical experiments suggested that what is learned in these tasks is merely a simple counting mechanism keeping track of the number of levels of recursion, but with no knowledge of the hierarchical pattern of correspondences between interdependent elements in non-adjacent positions. Going back to natural language, learners would understand that an equal number of subject noun phrases and verb phrases are needed for a sentence like '*The rat the cat the dog chased killed ate the malt*' to be acceptable. However, each particular subject noun would not be mapped on one particular verb as prescribed by the hierarchical centre-embedding rule, and the meaning of the sentence possibly misunderstood. Similarly, in the AGL paradigm, the counting could be demonstrated experimentally, but not the hierarchical mapping rule. When the test materials required detection of the hierarchical mapping pattern of non-adjacent elements in an artificial grammar task, participants failed to learn [10].

Recently, however, positive results have been reported. In one study [21], hierarchical structures in an artificial grammar could be learned in the presence of prosodic cues. Participants encoded the centre-embedded structure when the stimulus strings were naturally spoken, and pauses were added that marked the boundaries of the embedded clauses. In a more recent AGL study in our laboratory [20], hierarchical learning was also found, under two conditions of exposure. First, learning was facilitated when the input exemplars were presented 'starting small', i.e. in increasing order of complexity (first the exemplars without embeddings, followed by one level of embedding items, etc.). Second, elaborate training with short sequences containing *no* embedded clauses was a prerequisite for any learning of the centre-embedded structure. Hence, learning hierarchical structures from exemplars might not only depend on structural complexity of the grammar, but also on conditions of exposure, and non-linguistic features of the training sample.

The focus of this paper is on the information present in the very first training sample of exemplars presented to the learner, and on how these early simple sequences without embeddings, can enhance eventual learning of a structure *with* embeddings. After having summarized the experimental results by Lai & Poletiek [20], we will propose a statistical framework for AGL. Within this framework, the effect of distributional biases in the training sample on learning of a hierarchical pattern is investigated. We explore the hypothesis that better learning occurs for strongly biased distributions of early simple input sequences. Importantly, we propose that these distributional biases may be used as proxy for semantic variation in the linguistic input. Finally, this hypothesis is tested in a mathematical model with simulated data, for an artificial language with centre-embedded structure.

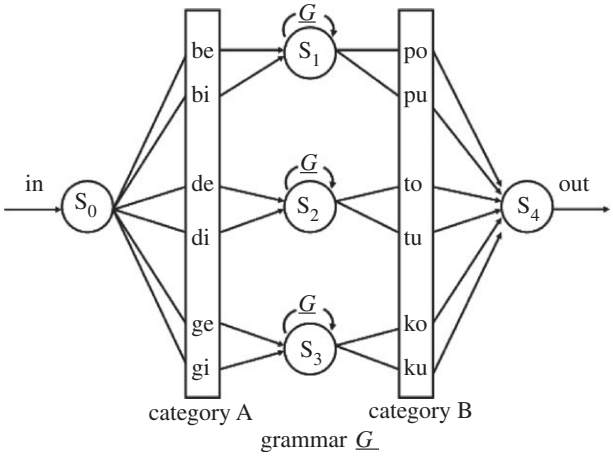

Figure 1. Artificial centre-embedded A^nB^n grammar $\underline{G}$ used in Lai & Poletiek [20]. Artificial syntactical categories A and B contain six non-words. The context-free grammar $\underline{G}$ refers to itself at state S_1, S_2 and S_3, inserting a grammatical string within a grammatical string, to form a new grammatical string.

2. THE IMPORTANCE OF SIMPLE STRUCTURES FOR LEARNING COMPLEX ONES

Research on the learnability of centre-embedded structures with artificial materials typically uses an A^nB^n grammar with two word categories A and B, basic rules producing specific AB pairs, and a centre-embedding operation that inserts a grammatical AB pair within a grammatical AB pair (etc.) to form a new grammatical sequence. In this manner, an infinitely productive system generates an unbounded number of sequences (n being unbounded) with a sequence of 'A's followed by an equal number of 'B's, each A mapping on one particular B in the sequence, according to the centre-embedding rule. Following previous AGL studies [22], De Vries *et al.* [10] and Lai & Poletiek [20] used six one-syllable non-words as category A words (words A1–A6), and six non-words as category B words (B1–B6) (figure 1).

Phonetic cues indicated category membership of a non-word, and grammatical mapping between A and B words: category A syllables contained the vowels -e/-i, i.e. {be, bi, de, di, ge, gi}, whereas category B contained -o/-u, i.e. {po, pu, to, tu, ko, ku}. The mapping rule to connect A syllables with a counterpart in

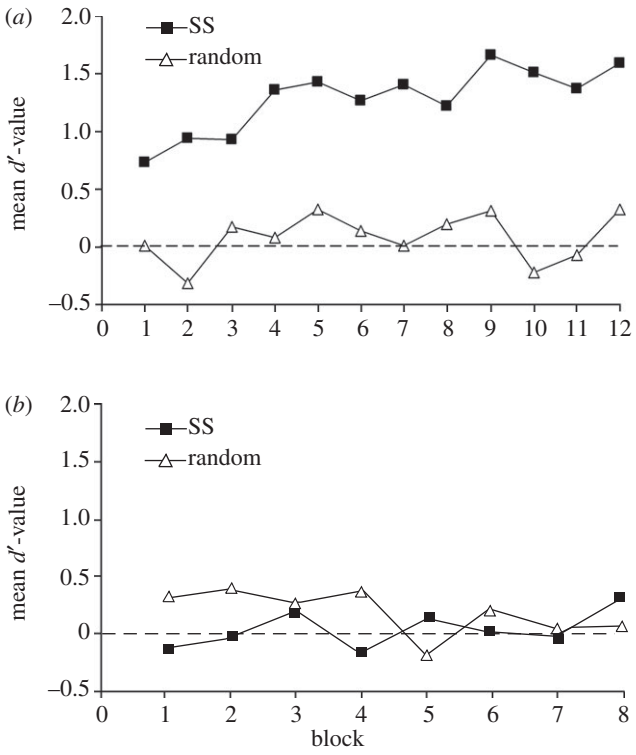

Figure 2. (*a*) Performance in an AGL task after exposure to training items with 0, 1 and 2 levels of embedding [20]. (*b*) Performance in an AGL task after exposure to training items with 1 and 2 levels of embedding [20].

category B was based on the consonants of both syllables: A syllables with b as a first consonant could be paired with B words with a p in first position, A words with d could be mapped on B words with t, and A words with g on B non-words with k (i.e. {be/bi-po/pu}, {de/di-to/tu} and {ge/gi-ko/ku}). For example, *be po* and *de be po tu* are grammatical sequences; *ge di po ku* is ungrammatical, because *di po* violates the pairing rule.

In an adapted version of the typical AGL procedure, Lai & Poletiek [20] compared a group of participants exposed to sequences of grammar $\underline{G}$ with zero, one and two levels of embedding, with a group exposed to sequences with the more complex sequences only, i.e. sequences with one and two levels of embedding. The training items were presented in blocks with items having an equal number of levels of embedding. In the condition with zero level of embedding items, four blocks with zero levels of embedding were followed by four blocks with one level of embedding items and four blocks with two levels of embedding items (figure 2*a*). In the condition without the zero level of embedding items, only two times four blocks were presented (figure 2*b*). In addition, the starting small regimen was compared with a random ordering (all items being presented in blocks as well, but in a random order). As can be seen in figure 2*a*,*b* (displaying the accuracy of grammaticality judgements (with d′)), starting small substantially helped learning a centre-embedded A^nB^n grammar, but only when, in the first stage of exposure to the language, participants had the opportunity to learn the simple AB structures without embeddings [20].

The results suggest that information about the grammar present in the basic structures is crucial to learn the more complex parts of the grammar later on. This raises the question of what information in the first subset of simple items makes them so crucially helpful for learning. In the study of Lai & Poletiek, the information the learner can infer from what he or she is presented with is simply which AB pairs are grammatical and which are not, ungrammatical pairs having a probability of zero to occur in the input, and grammatical unique AB pairs occurring with probabilities 1/12—since the grammar generates twelve grammatical pairs of non-words (see figure 1). Hence, in statistical terms, the learner infers from the early input a dichotomous distribution of AB sequences with probability either zero or non-zero, corresponding to ungrammatical and grammatical pairs, respectively. This distribution of adjacencies, in turn, may support the learning of positional structure of 'A's and 'B's in sequences with higher levels of embedding, by facilitating recognition of non-adjacent but legally associated AB pairs that were encoded as adjacencies in the first stage of learning. For example, when a learner is presented with *be gi ko pu*, prior knowledge about the adjacent pairs *be pu* and *gi ko* being grammatical might facilitate associating the 'A's and 'B's—now distant—in the complex embedded sequence. In summary, the study of Lai and Poletiek point to the importance of both a starting small-training regimen and robust learning of the grammaticality status of adjacent dependencies.

These conclusions, however, rely on the assumption that the to-be-learned A^nB^n grammar (in which A and B represent word categories) specifies grammatical (for example *be pu*) and ungrammatical (for example, *be ko*) associations between individual A-words and B-words by occurrences ($(p(B|A) > 0)$ and non-occurrences ($(p(B|A) = 0)$ in the output. For natural language, this assumption does not hold. Syntactical constraints typically apply to the position of syntactical categories, in natural languages, not to individual words belonging to these categories. Every word satisfying the syntactical constraints of a given word category (and additional rules for its pairing with a word of another category, e.g. number agreement in subject–verb pairings), may occupy the position dedicated to that word category. Hence, though constraints for relating a given A word to a given B word partially apply in semantics (e.g. number agreement), no such constraints apply to syntax in natural languages; every specific word of a given (sub)category, in the appropriate form, can be inserted in a location dedicated to that syntactical category. In contrast, typical artificial A^nB^n grammars put grammatical constraints on which A (e.g. noun) can legally precede which B (verb). Thus, the typical grammars used in experimental research with artificial hierarchical A^nB^n grammars seem not to be in line with natural language. This difference between natural and experimental A^nB^n grammars may affect the relevancy of AGL studies for natural language learnability.

In natural language, recursive constructions like centre embeddings owe their powerful semantic productivity to the fact that *every* specific word belonging to one of the categories can be inserted in a location of that category word. Therefore, *the dog the man chases barks* is

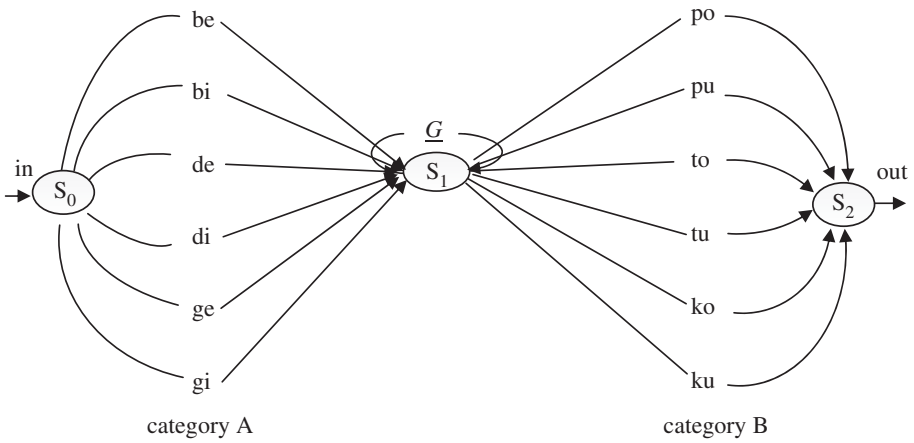

Figure 3. Artificial centre-embedded grammar A^nB^n $\underline{G}$, without grammatical constraints on specific AB sequences. Artificial categories A and B correspond to syntactical (sub)categories in natural grammars like nouns (*A*) and verbs (*B*). It is assumed that the A and B associations satisfy rules of agreement.

syntactically as correct as *the man the dog chases barks*, though the latter pattern-specific noun–verb pairings will rarely occur in real-world situations, and therefore may be expected to be rare in the linguistic input. How could a learner, then, if all possible AB mapping are grammatical, induce a centre-embedded grammar from the input? In a theoretical AGL study, we explore the possibility that distributional characteristics of the AB pairings in the input serve as a cue for this learning.

3. THE INFLUENCE OF DISTRIBUTIONAL BIASES IN THE SIMPLE STRUCTURES FOR LEARNING A HIERARCHICAL ARTIFICIAL GRAMMAR

In figure 3, the artificial grammar $\underline{G}$ from figure 1 is adapted without any grammatical constraints on specific AB mappings.

It is hard to explain how the centre-embedding pattern of correspondences between 'A's and 'B's in the unconstrained grammar in figure 3 could be learned from exposure only. On the basis of an input with occurrences of every possible AB pairings, learners might only grasp the principle that sequences always have an equal number of 'A's and 'B's. A simple counting algorithm is enough to detect that regularity. This learning, which involves detecting and counting A-words and B-words, has indeed been demonstrated in humans [7,10,13,20]. However, knowing $\underline{G}$ requires in addition knowing its centre-embedding rule inserting new AB pairs within AB pairs resulting in shifting related A words and B words one position to the left and to the right, respectively. Hence, accurately parsing exemplars of this grammar requires understanding more than counting and comparing A and B category words. It requires knowing the hierarchical pattern of correspondences between the 'A's and 'B's. This part of the grammar has been shown to be very hard to learn, even in studies providing cues about which AB pairs are legal like in the classical AGL studies [10,13,20]. Without such syntactical cues (figure 3) the linking pattern of specific 'A's and 'B's can be expected to be even more difficult, since learners of the grammar in figure 3 would have no cue about the pattern determining the positions of associated individual 'A's and 'B's in the sentence. If no grammatical constraints on the occurrence of specific A-word and B-word pairings in natural language can help the learner (the following sentences being both perfectly grammatical: *The girl the dog bites shouts*, and *The dog the girl bites shouts*.) how could these type of grammars be learned and used?

The hypothesis we explore in the present analysis is that learners rely on distributional biases in simple AB structures. Translating these biases to natural language, consider the following two English sentences: *The dog* (*A*) *barks* (*B*) and *The dog* (*A*) *talks* (*B*). Due to different occurrences in the real world, reflected in semantic variation in speech, the first sentence will be more probable in the linguistic input of a natural language learner than the second one. As a result, the probabilities $p(B|A)$ of AB pairs will be unequally distributed. Note that many factors may affect distributional variation of individual word sequences in natural language, for example, the type of corpus assumed (e.g. child directed speech versus adult speech; literature), individual word frequency, animacy or phonology. Here, the focus is on distributional biases caused by semantic variation, because they may provide an account of semantic influences on the acquisition of complex syntactical structures in the natural situation. Indeed, these semantic influences in the real world may be approximated in a controlled artificial environment by merely manipulating distributional features of the input in the artificial environment.

In AGL, semantics are generally avoided. In general, artificial grammar laboratory studies typically aim at studying grammar learning in the absence of semantic influences [12]. Also, semantic influences on grammar learning have hardly been addressed in studies using non-natural language learning methodologies. In a computational study with simple recurrent networks [23], the influence of semantic biases in an artificial language was represented in a similar way as proposed here, by variations in transitional probabilities. Substantial facilitation of these biases on the model's learning was found. In this study, we explore how transitional probabilities of simple adjacencies in the language input might affect learning a centre-embedded A^nB^n grammar. Learning with inputs with and without biases is compared in a simple mathematical simulation model.

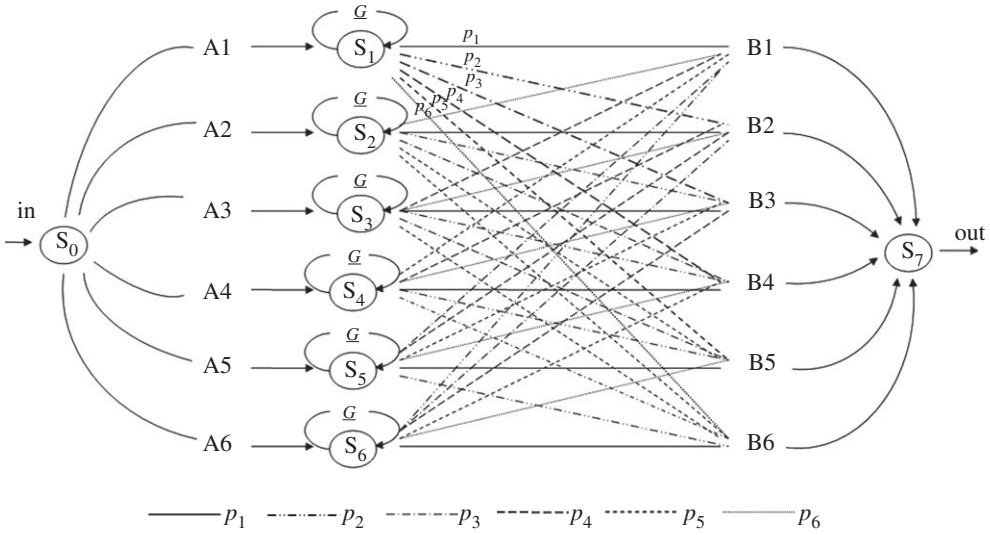

Figure 4. Artificial centre-embedded grammar A^nB^n $\underline{G}$, with transitional probabilities specified for each p_i(B|A)in the basic AB structure, with $\sum p_i(B|A) = 1$. A1–A6 and B1–B6 are the members of category A and category B, respectively (see figure 3).

4. A MODEL FOR THE EFFECT OF DISTRIBUTIONAL BIASES IN SIMPLE PAIRS ON GRAMMAR LEARNING

To model the role of semantic biases as distributional biases in AGL, first measures for semantic bias and for learning need to be defined. Consider a miniature A^nB^n language with two category A words (nouns) *dog* and *girl*, and two category B words (verbs) *barks* and *loves*. The grammar generates an infinite number of unique sentences with or without embedded relative clauses: four sentences without embeddings, *the dog bar*ks, *the dog loves*, *the girl love*s and *the girl bark*s; 16 unique sentences with one centre-embedded clause, e.g. *the dog the girl loves, barks*; 64 sentences with two levels of embedding, etc. The centre-embedding rule mapping 'A's on 'B's is 'is subject of'. Furthermore, assuming that the transitional probabilities of a specific B-word given an A-word reflect occurrences in the world of that agent–action combination, in the present example, a possible distribution would assign $p(B|A) = 0.70$ to *the dog bark*s, 0.30 to *the dog loves,* 0.95 to *the girl loves* and 0.05 to *the girl barks*. A neutral distribution would assign equal probabilities to each AB transition. In the present example, any $p(B|A) = 0.50$ without any bias. For the purpose of the present analysis, an artificial grammar was made without grammatical constraints but with various probability distributions for individual AB transitions.

The exemplars of the grammar and their probabilities were generated with a simple computer program GenAutom. The program takes as an input the production rules describing the grammar. Production rules describe the grammar's transitions (indicated by arrows in figure 4) from state to state, and their corresponding labels (A or B). The output of the program is *n* unique exemplars of the grammar together with their probabilities to be generated by $\underline{G}$ (p(exemplar|$\underline{G}$)). The exemplar probability is the product of path probabilities 'run through' by the exemplar. In figure 4, the artificial grammar $\underline{G}$ is displayed, based on the grammar in figure 3, adapted by adding the transitional probabilities p_i ($p_{1,\ldots,6}$) p(B|A) of an A word to be followed by a specific B word. Distributional biases in the AB pairs are defined by the distribution of p(B|A).

In the present artificial language, we assume that the paths starting from node S_0 to nodes S_1–S_6 have all equal probabilities ($p = 1/6$), resulting in equal frequencies of occurrence of each A-word in the language. For a given distribution of the transitional probabilities p_i(B|A), the probability of each unique full string in a random output of the grammar can be calculated by multiplying the probabilities of all paths run through to generate that string [24,25]. Note that the distribution of p_i applies to all AB pairs at any level of embedding. Hence, a sequence with low probable pairs at each level of embedding will occur less frequently in a sample generated by *G* than a sequence with high probable pairs. In this manner, a distribution of string probabilities can be specified reflecting the occurrences of exemplars in the language.

The sum of the probabilities of all unique strings generated by a grammar adds up to one, or approximates one for grammars with an infinite number of unique strings (like the grammar considered here). One implication of the summed probabilities of an input varying from zero (for an empty set) to one (for the full language) is that the summed probabilities of a given subset of unique sequences reflect the *proportion* of the full language output that is generated by that grammar. This proportion (called here the *coverage* of the grammar by the sample [25,26]) may vary over samples with equal sizes, depending on the probabilities of the sampled exemplars. Within this framework, the coverage of a sample of exemplars reflects how much information about the grammar is in the sample, and may be used as a model to predict how much can be learned about the grammar after training with that particular sample. In the model of the learning process proposed here, coverage is indeed taken as an indicator of how much of the grammar can be learned after exposure to the particular set of exemplars. Moreover, coverage is used as a criterion for comparing learning with input samples having various p(B|A) distributions.

Table 1. Parameters of the output sets of 1000 exemplars, for grammar $\underline{G}$, assuming no versus three levels of p(B|A) biases.

semantic bias in grammar $\underline{G}$	p(B|A) distribution $p_1/p_2/p_3/p_4/p_5/p_6$	standard deviation SD p(string)	coverage Σ p(string)
no	0.166/0.166/0.166/0.166/0.166/0.166	0.0025	0.68
weak	0.450/0.450/0.025/0.025/0.025/0.025	0.0040	0.76
strong	0.900/0.020/0.020/0.020/0.020/0.020	0.0058	0.85
very strong	0.980/0.010/0.0025/0.0025/0.0025/0.0025	0.0064	0.89

Four simulated input samples generated by a grammar with and without distributional biases are compared; i.e. four versions of $\underline{G}$ are considered with varying distributions of p_i: $\underline{G}_{nb}$, with no bias, $\underline{G}_{wb}$ with a weak bias, $\underline{G}_{sb}$ with a strong bias and $\underline{G}_{vsb}$ with a very strong bias (see figure 4). Next, four random samples of $n = 1000$ unique exemplars were obtained with GenAutom, one for each version of $\underline{G}$. Random sampling results in output samples containing the most probable unique sequences generated by the grammar. As a consequence, small samples will, in general, consist of relatively higher probability exemplars. Also, since, in general, short sequences have a higher probability than long ones and less embeddings, small sets will contain relatively more zero or low levels of embeddings items. The dependency between sequence length and sequence probability may be loosened however, with varying p(B|A) distributions assumed. Indeed, a grammar with a strongly biased p(B|A) might generate sequences with one level of embedding A1A2B2B1 made of two high probable pairs A1B1 and A2B2 more often than low probable strings with no levels of embedding (i.e. simple AB strings). This is true for $\underline{G}_{vsb}$ in our example. In analogy, in the example discussed above with natural language, the sentence *the dog the girl loves barks* might occur more often than the more simple zero level-of-embedding construction *the girl barks*.

GenAutom was run with the four versions of $\underline{G}$ to produce output samples of $n = 1000$ unique strings each. Each version was characterized by the distribution of p_1 to p_6 (see figure 3). For each sample, the standard deviation of the string probabilities and the sample's coverage were computed (table 1).

The model clearly shows an increase of coverage with more skewed distributions of AB word-pairs. Also, variance in the transitional path probabilities in the grammar correspond with higher variance in the frequencies of unique strings in the output, as can be seen in table 1. Under the assumption that coverage indicates the amount of information about the grammar in a specific input, strongly biased distributions of the simple AB pairs of a centre embedded grammar entail more informative input samples about the grammar than weakly biased distributions, for samples of equal size.

Not only the total coverage after a given number of exemplars varies with the distribution of 'AB's, but also how the coverage of the grammar *develops* as the sample of exemplars 'grows' over time. To look at this development, we need to specify how the exemplars are ordered during the period of exposure. If we assume that the most frequent exemplars in the language are presented early, a 'growing' sample is defined as a sample of exemplars ordered over time according to their decreasing probabilities. This is the 'starting small' ordering [20]. Assuming a starting small ordering, the function describing the development of the coverage is the cumulative sum of string probabilities of the sample presented at each point in time of exposure (figure 5). In figure 5, the input sets ($n = 1000$) described in table 1 are used, and the exemplars displayed on the x-axis ordered over time according to decreasing string probabilities. For coverage having value one (which is approximated but not achieved for the output of unbounded grammars), all information that can be displayed about the grammar is displayed in the presented exemplars.

Comparing the shapes of the curves, two differences stand out. Depending on the level of bias, exemplars with one or more levels of embedding may occur in the sample. Hence, without bias (equal p(B|A) probabilities), $\underline{G}$ first generates all unique zero level-of-embedding strings (36) followed by 964 (of the 1296 possible unique) one level-of-embedding strings. For stronger biases in the distribution of p(B|A), however, higher levels-of-embedding strings are figuring in the first 1000 exemplars generated. In addition, higher level-of-embedding strings may come in front of lower level-of-embedding ones, for more strongly biased distributions.

Secondly, for strongly biased distributions of p(B|A), the coverage *develops* faster in the beginning. For example, after 300 exemplars, 0.55 of the unbiased grammar is covered, in contrast to 0.86 of the very strongly biased grammar. Thus, most information (as we assumed here to be expressed by the coverage value of the sample) about the grammar is provided in an earlier stage of exposure, as the underlying grammar has a more skewed distribution of AB pairs. As the set grows further with more complex exemplars, biased curves become quickly asymptotical, suggesting that new exemplars provide little additional information about $\underline{G}$. This characteristic of the learning situation, i.e. the concentration of coverage in the early stage of exposure, and the decreasing contribution of later linguistic input, may be important for infinite grammars that can in principle never be 'learned' fully, since not all unique strings can be presented to the learner. What our analysis suggests is that this 'inconvenience' of limited exposure for inductive learning is compensated by biasing the distribution of simple pairs. Exemplars encountered in the first stage of exposure provide most information, whereas stimuli presented later add much less, and hence, matter less for learning. In contrast, more flat distributions of the AB pairs, corresponding to any possible AB pair in the input being equally 'plausible', will result in a curve that is insensitive to the

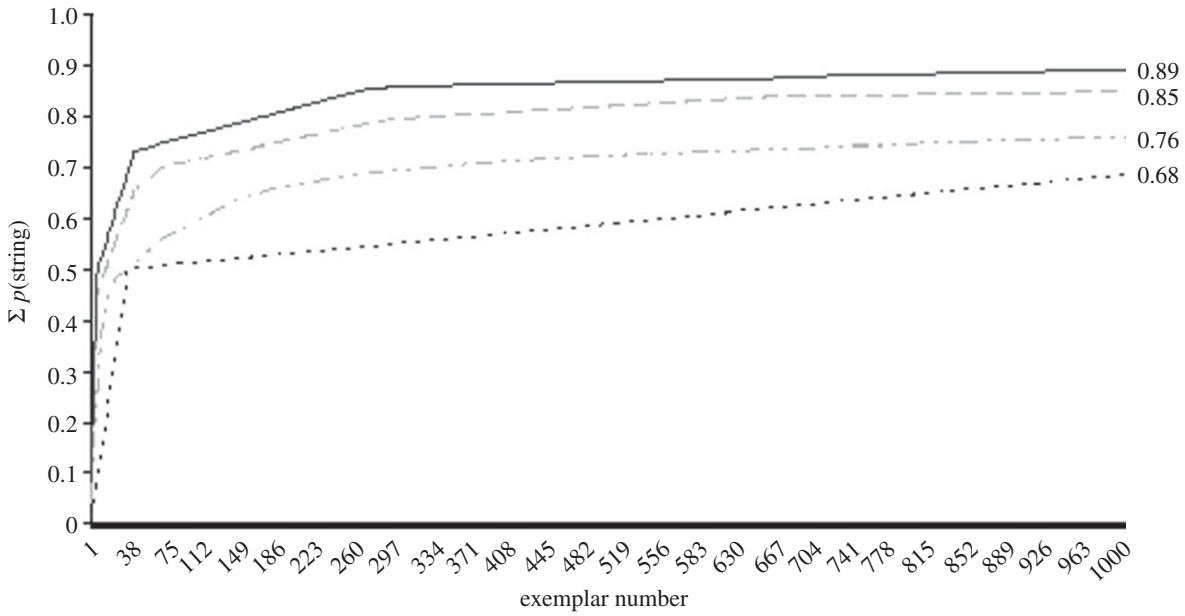

Figure 5. Cumulative Σ p(string) for output sets (1000 strings) of A^nB^n grammar $\underline{G}$ assuming various distributional biases of p_i(B|A). The exemplars (numbered on the x-axis) are ordered according to decreasing probabilities, and may represent different exemplars for different versions of $\underline{G}$ (dashed line, no bias; dash-dotted line, small bias; long dashed line, strong bias; solid line, very strong bias).

concentration of information in the beginning, suggesting the necessity of ongoing learning over a longer period of time. Note that the benefit of early learning for highly biased sets relies on the presumption that the exemplars are presented in a starting small fashion.

5. DISCUSSION

This study investigated the role of simple AB structures in learning a complex hierarchical A^nB^n structure. Earlier experimental research showed that participants could learn such hierarchical structure only after having been exposed to a large number of basic AB exemplars. A statistical account of the impact of distributional biases in these basic structures for learning the complex grammar is proposed. Distributional biases in simple structures occur in natural language as a result of, among other factors, semantic variations related to real-world knowledge. For example, the noun–verb sequence *dogs*(*A*) *laugh*(*B*) is expected to occur less often than *dogs*(*A*) *bark*(*B*), hence p(laugh|dogs) < p(bark|dogs).

In the analysis proposed, the sum of all unique exemplar probabilities in the sample was taken as a measure of the proportion of the grammar that is covered by the sample. The sample's coverage, then, was assumed to indicate how much could be learned about a grammar from exposure to that sample. This assumption has been verified previously in our laboratory with AGL experiments, showing better grammaticality judgement performance after training with input samples with higher coverage, matched for size [26].

In a mathematical simulation of the learning process, coverage of the centre-embedded grammar by a sample was predicted from the distributional biases in the grammar's simple AB constructions. The first striking conclusion is that everything else being equal (equal sample size and equal structure) grammar coverage is higher for stronger distributional biases in the simple AB pairs. That is, as the distribution of p(B|A) is more skewed, more information about the centre-embedded grammar is displayed in a random output of that grammar. Since one of the major factors affecting this distribution in natural language is semantics, the implication of this result is that inductive learning of hierarchical structures is facilitated by semantic biases.

The second result of our simulation model is that most information on the grammar is concentrated in the early phase of exposure to the input, as semantic biases are stronger. This finding regarding the time course of information displayed in the input became apparent when the virtual learner was exposed to a 'starting small' training regimen (displaying the basic structures in the language first, followed by increasingly more complex structures with embedded clauses). This result conveys an interesting perspective on the phenomenon of the critical language learning period. It suggests a concerted action of distributional features of the early language input—correlating with semantic features—to give a young learner both very simple *and* very informative cues about the complex language system that is eventually to be mastered.

Though the present proposal makes a first step to a new understanding of the role of the early linguistic input on learning complex structure, further theoretical and empirical research is needed to evaluate its significance and its limits for natural language learning. Note, first, that the results are at odds with one previous conclusion from a computer simulation study that semantic biases help learning hierarchical grammars, but starting small does not [23]. In the model presented here, the effect of semantic biases on the spreading of the information over time could be made apparent *because* a starting small organization of the input was assumed. The validity of the argument that semantic biases also operate in natural language

learning, therefore, relies on whether the natural language input is actually organized in a starting small fashion. It is a topic of debate whether children learning language make use of a restricted type of language utterances, i.e. child directed speech, or whether they 'use' every language input that reaches their ears equally [27]. Though recent experiments evidenced a strong main effect of starting small [20] in the artificial AGL setting, the implications of this finding for the natural setting are still speculative.

Other assumptions proposed here might be investigated further empirically. We do not know, for example, what the statistical coverage measure corresponds to from a cognitive point of view. The proportion of the full output of a grammar displayed in a sample might correlate with proficiency, but this relation remains to be specified. Learners having seen high coverage samples might be better at judging the grammaticality of new sentences, or alternatively, at producing new grammatical sentences. Another related question is whether exposure to a high coverage sample affects the level of processing of hierarchical exemplars and insight into the hierarchical rule. Note also that the present analysis shows parallels with linguistic ideas on language learning. The coverage curve approximating 100 per cent quite early under conditions of strong distributional bias supports the linguistic assumption that learners are *fully* competent at the end of the critical learning period.

Another question raised by the coverage measure [26,28] refers to the paradox that *low* coverage sets, in general, contain *longer* exemplars, and therefore it might be argued display more discrete information about the rules of the grammar. If low coverage input exemplars display more rules of the grammar, how can they nonetheless lead to poor learning (as was found in our laboratory [26])? And how can learners take advantage of short sentences displaying fewer rules of the grammar? One possible answer is that hierarchical structures with a basic self-embedding pattern are better learned in a staged learning process organizing learning of the basic rule first (short sentences with high coverage), and after that is learned, the self-embedding operation is presented to the learner [29].

By using a methodology that reduces complex factors involved in human natural language learning to simple manipulations of artificial grammars, the influence of semantic biases on complex grammar learning could be traced. Also, the present statistical modelling with artificial language raised new empirical questions. From a theoretical point of view, the present study illustrates the potential benefit of combining into one account both formal properties of language and cognitive developmental learning mechanisms [30]. More specifically, the present results suggest that semantic biases in very simple linguistic constructions may be one of the many useful extra-syntactical cues provided at exactly the right moment in the learning process, to ultimately acquire the very complex hierarchical rules of language.

REFERENCES

1 Fitch, W. T., Hauser, M. D. & Chomsky, N. 2005 The evolution of the language faculty: clarifications and implications. *Cognition* **97**, 179–210. (doi:10.1016/j.cognition.2005.02.005)

2 Bach, E., Brown, C. & Marslen-Wilson, W. 1986 Crossed and nested dependencies in German and Dutch: a psycholinguistic study. *Lang. Cogn. Process.* **1**, 249–262. (doi:10.1080/01690968608404677)

3 Hudson, R. 1996 The difficulty of (so-called) self-embedded structures. *UCLWorking Papers in Linguistics* **8**, 283–314.

4 Newmeyer, F. 1988 Extensions and implications of linguistic theory: an overview. In *Linguistics: the Cambridge survey 2. Linguistic theory: extensions and implications* (ed. F. Newmeyer), pp. 1–14. Cambridge, UK: Cambridge University Press.

5 Vasishth, S. 2001 An empirical evaluation of sentence processing models: center embeddings in Hindi. In *OSUWPL*, vol. 56 (eds M. Daniels, D. Dowty, A. Feldman & V. Metcalf), pp. 159–181. Columbus, Ohio: Ohio State University Department of Linguistics.

6 Chomsky, N. & Miller, G. 1963 Introduction to the formal analysis of natural languages. In *Handbook of mathematical psychology* (eds R. Luce *et al.)*, pp. 286–287. New York, NY: Wiley.

7 Fitch, W. T. & Hauser, M. D. 2004 Computational constraints on syntactic processing in a nonhuman primate. *Science* **303**, 377–380. (doi:10.1126/science.1089401)

8 Hauser, M. D., Chomsky, N. & Fitch, W. T. 2002 The faculty of language: what is it, who has it, and how did it evolve? *Science* **298**, 1569–1579. (doi:10.1126/science.298.5598.1569)

9 Chomsky, N. 1959 On certain formal properties of grammars. *Inform. Control* **2**, 137–167. (doi:10.1016/S0019-9958(59)90362-6)

10 De Vries, M. H., Monaghan, P., Knecht, S. & Zwitserlood, P. 2008 Syntactic structure and artificial grammar learning: the learnability of embedded hierarchical structures. *Cognition* **107**, 763–774. (doi:10.1016/j.cognition.2007.09.002)

11 Saffran, J. R., Aslin, R. N. & Newport, E. L. 1996 Statistical learning by 8-month-old infants. *Science* **274**, 1926–1928. (doi:10.1126/science.274.5294.1926)

12 Gómez, R. & Gerken, L. A. 2000 Infant artificial language learning and language acquisition. *Trends Cogn. Sci.* **4**, 178–186. (doi:10.1016/S1364-6613(00) 01467-4)

13 Perruchet, P. & Rey, A. 2005 Does the mastery of center-embedded linguistic structures distinguish humans from non-human primates? *Psychon. Bull. Rev.* **12**, 307–313. (doi:10.3758/BF03196377)

14 Gentner, T. Q., Fenn, K. M., Margoliash, D. & Nusbaum, H. C. 2006 Recursive syntactic pattern learning by songbirds. *Nature* **440**, 1204–1207. (doi:10.1038/nature04675)

15 Corballis, M. C. 2007 Recursion, language, and starlings. *Cogn. Sci.* **31**, 697–704. (doi:10.1080/15326900701399947)

16 Reber, A. S. 1967 Implicit learning of artificial grammars. *J. Verb. Learn. Verb. Behav.* **6**, 855–863. (doi:10.1016/S0022-5371(67)80149-X)

17 Reber, A. S. 1993 *Implicit learning and tacit knowledge*. New York, NY: Oxford University Press.

18 Pothos, E. M. 2007 Theories of artificial grammar learning. *Psychol. Bull.* **133**, 227–244. (doi:10.1037/0033-2909.133.2.227)

19 Friederici, A. D., Bahlmann, J., Heim, S., Schubotz, R. I. & Anwander, A. 2006 The brain differentiates human and non-human grammars: functional localization and structural connectivity. *Proc. Natl Acad. Sci. USA* **103**, 2458–2463. (doi:10.1073/pnas.0509389103)

20 Lai, J. & Poletiek, F. H. 2011 The impact of adjacent-dependencies and staged-input on the learnability of

centre-embedded hierarchical structures. *Cognition* **118**, 265–273. (doi:10.1016/j.cognition.2010.11.011)
21 Mueller, J. L., Bahlmann, J. & Friederici, A. D. 2010 Learnability of embedded syntactic structures depends on prosodic cues. *Cogn. Sci.* **34**, 338–349. (doi:10.1111/j.1551-6709.2009.01093.x)
22 Bahlmann, J., Schubotz, R. I. & Friederici, A. D. 2008 Hierarchical artificial grammar processing engages Broca's area. *NeuroImage* **42**, 525–534. (doi:10.1016/j.neuroimage.2008.04.249)
23 Rohde, D. L. T. & Plaut, D. C. 1999 Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition* **72**, 67–109. (doi:10.1016/S0010-0277(99)00031-1)
24 Charniak, E. 1993 *Statistical language learning*. Cambridge, MA: MIT Press.
25 Poletiek, F. H. & Wolters, G. 2009 What is learned about fragments in artificial grammar learning? A transitional probabilities approach. *Q. J. Exp. Psychol.* **62**, 868–876. (doi:10.1080/174702108025 11188)
26 Poletiek, F. H. & van Schijndel, T. J. P. 2009 Stimulus set size and grammar coverage in artificial grammar learning. *Psychon. Bull. Rev.* **16**, 1058–1064. (doi:10.3758/PBR.16.6.1058)
27 Pine, J. M. 1994 The language of primary caregivers. In *Input and interaction in language acquisition* (eds C. Gallaway & B. J. Richards), pp. 109–149. Cambridge, UK: Cambridge University Press.
28 Poletiek, F. H. & Chater, N. 2006 Grammar induction benefits from representative sampling. In *Proc. of the 28th Annual Conf. of the Cognitive Science Society* (ed. R. Sun), pp. 1968–1973. Mahwah, NJ: Lawrence Erlbaum.
29 Poletiek, F. H. 2011 What in the world makes recursion so easy to learn? A statistical account of the staged input effect on learning a center embedded hierarchical structure in AGL. *Biolinguistics* **5**, 36–42.
30 Christiansen, M. H. & Chater, N. 2008 Language as shaped by the brain. *Behav. Brain Sci.* **31**, 489–558. (doi:10.1017/S0140525X08004998)