Syntactic Generalization in a Connectionist Model of Complex Sentence Production

H. Fitz*

Institute for Logic, Language and Computation, University of Amsterdam,
Nieuwe Doelenstraat 15, 1012 CP Amsterdam, the Netherlands
**E-mail: fitz@science.uva.nl*
www.illc.uva.nl

F. Chang

NTT Communication Science Laboratories,
2-4 Hikari-dai, Seika-cho, Souraku-gun, Kyoto 6190237, Japan
E-mail: chang.franklin@gmail.com

We present a neural-symbolic learning model of sentence production which displays strong semantic systematicity and recursive productivity. Using this model, we provide evidence for the data-driven learnability of complex yes/no-questions.

Keywords: Statistical learning; semantic processing; systematicity; recursion; polar interrogatives.

1. Introduction

Usage-based theories of language acquisition have emphasized the role of experience in the bottom-up construction of language knowledge (Tomasello,[1] Goldberg[2]). But since languages are lexically open and combinatorial in structure, no amount of experience covers their expressivity. These theories must therefore explain how children can generalize properties of their linguistic input to an adult grammar and, ideally, provide evidence that this explanation can be implemented explicitly. Connectionist models of language processing generally align well with fundamental tenets of usage-based theories, but they have frequently been criticized for not generalizing like humans (Marcus[3]). In this paper we present a neural network model of sentence production and syntactic development which generalizes in interesting ways, both lexically and structurally. In the second part of the paper, we will argue that our model might help to explain how complex yes/no-questions can be learned in the absence of direct experience.

2. The Dual-path model

Our modelling work built on the Dual-path model of Chang, Dell and Bock[4] which was adapted for the processing of multi-clause utterances. The model consisted of two pathways (Figure 1). One pathway, the sequencing system, was a standard simple-recurrent network.[5] This system learned distributional regularities over word sequences and developed syntactic categories at the compress-layer. The second pathway, called the message-lexical system, learned to use sentence meaning to activate words. The model learned from exposure to sentences paired with their meaning. Sentence meaning was represented by three components: concepts, thematic roles and event-structure. Concepts represented the meaning of individual words in the what-layer. Units in the where-layer represented thematic roles, such as the agent, patient or recipient of an action. These roles in the where-layer could be bound temporarily to sentence-specific content in the what-layer through dynamic weights. Hence thematic-role units could act like semantic variables. The event-semantics encoded the number and relative prominence of participants in an event. To represent a transitive event, for instance, an agent and patient feature were activated in the event-semantics-layer. Their relative level of activation biased the model towards selecting an active or passive construction. In multi-clause utterances, the event-semantics also encoded the relative prominence of basic events to signal the relation of clauses in the target utterance. Before production began, sentence meaning was activated in the message-lexical system. The model then mapped this message incrementally onto a sentence form. It learned in a standard error-based word-to-word prediction paradigm. The present model had the same architecture as in Chang et al.,[4] but it used extra units in the event-semantics- and where-layers to represent participant roles in relative clauses.

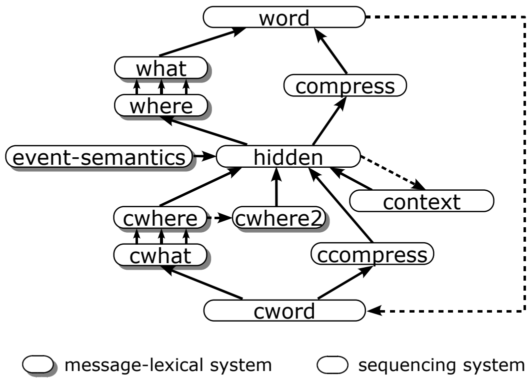

Fig. 1. Dual-path model architecture.

3. Strong systematicity

Children learn words in one semantic/syntactic context and reuse them in another. Strong lexical generalization requires that familiar words can be used correctly in novel sentences, at novel levels of embedding, in novel thematic roles. For example, children might learn the meaning of the word `cat` from simple sentences in which a cat is the agent of a transitive action and generalize its use to the recipient role of a dative embedding (Figure 2). This property of the human language faculty has been called strong

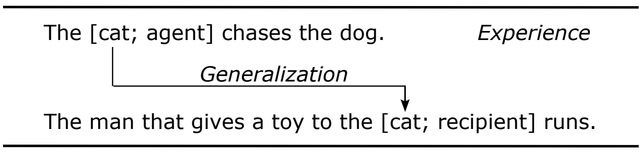

Fig. 2. Lexical generalization.

semantic systematicity.[6] We trained the Dual-path model on an artificial English-like language with up to three nested relative clauses. This language contained intransitives, active and passive transitives, prepositional datives and obliques as basic constructions from which sentences with relative clauses were assembled. Over a lexicon of 48 words, particles and inflectional morphemes, it allowed the creation of 2.49×10^{18} distinct sentence tokens. 10.000 tokens were randomly selected for training out of which 40% were single-clause sentences and this proportion was decremented by 10% for each additional level of embedding. In training, the word `cat` only occurred in the agent slot of single-clause, active transitive sentences. The model was then tested on novel sentences with various numbers of relative clauses in which the word `cat` always occurred as a dative recipient in the deepest embedding. Model behavior on these items was compared with performance for the exact same sentences in which `cat` did not occur (Figure 3). The x-axis shows the amount of training, the y-axis measures performance in terms of 'perfect match' with the target utterance. The model learned to produce sentences with one and two relative clauses to perfection, sentences with three nested relative clauses were more difficult as the model reached only around 70% accuracy. For each level of embedding, however, the model reached comparable levels of accuracy whether the word `cat` filled the recipient slot of the deepest dative embedding or not. Since in training the word `cat` was not experienced in recipient slots, dative constructions, or relative clauses, this suggests that the Dual-path model displayed strong semantic systematicity in Hadley's sense.

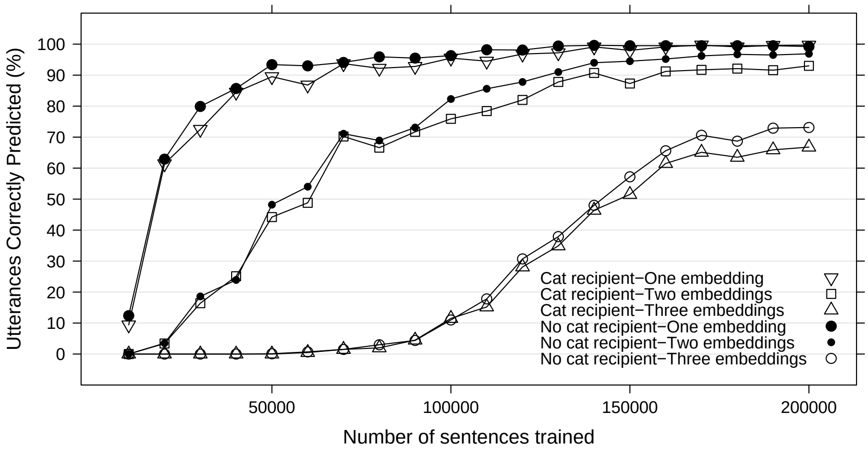

Fig. 3. Strong semantic systematicity in the Dual-path model.

How is systematicity achieved? The key component in the model is its weight-based message. To understand how this works, let's examine how the model would deal with the generalization in Figure 2. In the Dual-path model, this generalization has two parts. One part is learning the concept-word association, which in this case involves learning that the concept CAT (what-layer) maps to the English word cat (word-layer). This association can be learned from any input sentence about cats. The second part involves learning how to activate the appropriate recipient role Z (where-layer) at the embedded clause position where the word cat is supposed to be produced (i.e., gives the toy to the...). The model learned to activate this role from exposure to other sentences which contained embedded clause recipients (e.g., The woman that gave a stick to the dog jumped). This experience was sufficient because the message for the novel generalization in Figure 2 had a message link between the recipient role and the concept CAT (Figure 4). If

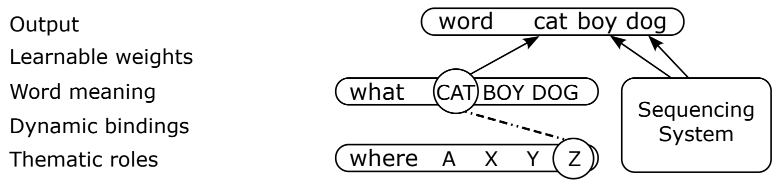

Fig. 4. Dynamic bindings (dashed line) enforce systematicity.

the recipient was activated, then the concept CAT was activated, and hence the word cat was produced. Thus, the Dual-path model could generalize systematically, because it could combine knowledge learned from different input utterances by means of its dynamic message.[7]

4. Recursive productivity

In structural generalization, familiar constructions are recombined into sentences with a novel hierarchical organization. By means of relativization, for example, the sentences

(1) The dog gave a toy to the cat.

(2) The girl that is chasing a dog was hit by the boy.

can form a novel structure with an additional embedding:

(3) The girl that is chasing a dog that gave a toy to the cat was hit by the boy.

To see whether the Dual-path model could generalize structurally, it was trained on a language with at most two relative clauses. Then it was tested on novel structures with three and four nested relative clauses (Figure 5). The model learned the training language with at most two embeddings

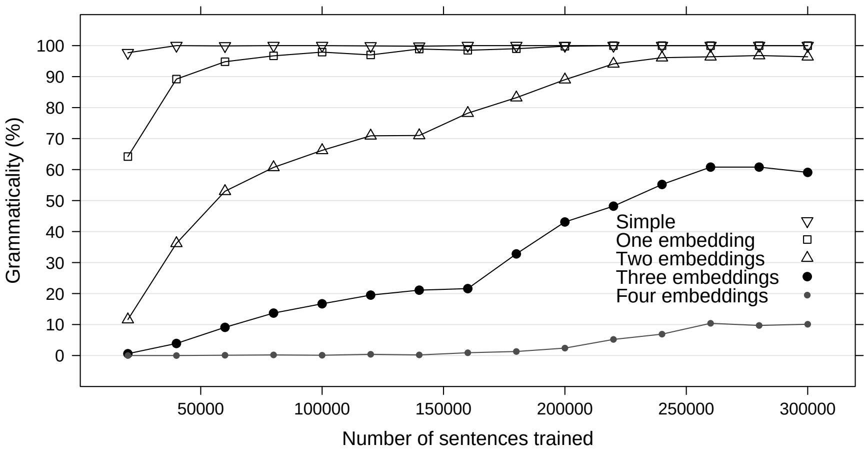

Fig. 5. Recursive productivity in the Dual-path model.

to perfection. In addition it produced 60% grammatical utterances with three embeddings and reached 10% grammaticality on sentences with four embeddings. For example, the model correctly produced sentences such as

(4) A dog that a boy that a teacher that gave the orange to a cat is carrying was attacked by is running with the toy.

without exposure to the syntax of three nested relative clauses in learning. The degradation of performance with depth of embedding is in line with human data.

Unlike systematicity, which depended on the role-concept weights, recursive productivity depended on another part of the message, the event-semantics. In training, the model learned to associate subparts of a sentence with the event-semantics of the proposition that controlled it (Figure 6). The model learned from simple messages how to sequence participants in single-clause transfer events (dog give toy to cat). Other features of the event-semantics controlled the position of relative clauses (X that) and the thematic role of the head noun in the relative clause (that *gap* VERB). When

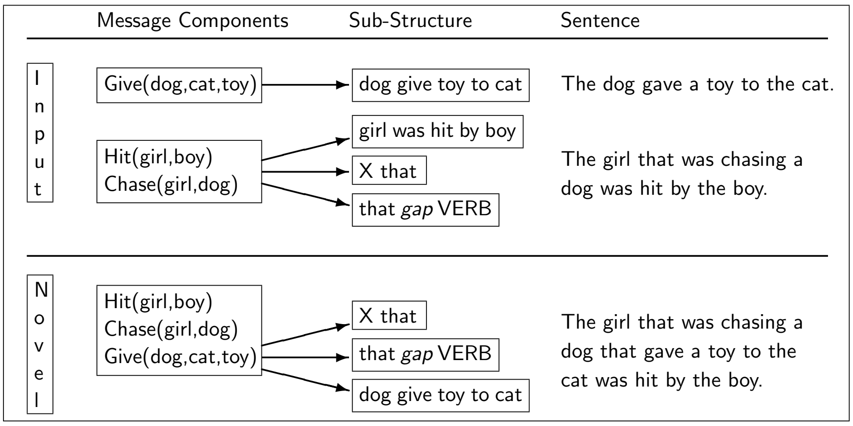

Fig. 6. Different components of the message control different subsequences of words in the target structure.

presented with a message for a novel construction, the model could use semantic regularities in the conceptual structure of the event-semantics and combine these regularities to generate additional embeddings.

From message-sentence pairs in training, the model learned which features of the event-semantics controlled which aspects of the hierarchical organization of complex sentences. Since novel messages shared features in the event-semantics with input messages, the model could generalize its learned subpart mappings and built novel structures from relevant message components. In this way, productivity was enabled by similarity-based meaning-to-form transduction.

5. The problem of auxiliary fronting in polar interrogatives

A major controversy in language acquisition revolves around the question which aspects of language, and syntax in particular, can be learned from experience and which aspects require some kind of language-specific biological endowment. Arguably, one of the most prominent issues in this debate concerns the learnability of yes/no-questions with relative clauses ('complex polar interrogatives'). A single-clause declarative such as

(5) The dog is barking.

can be turned into a yes/no-question by inverting the auxiliary is and the subject NP the dog:

(6) Is the dog barking?

Now consider the relative-clause sentence

(7) The dog that is chasing the cat is barking.

An ungrammatical question is obtained if the sequentially first auxiliary is moved to the front:

(8) *Is the dog that chasing the cat is barking?

This rule of forming complex questions disregards the hierarchical organization of the declarative (7) into main and subordinate clause. It is therefore a *structure-independent* rule. The correct rule requires that the main clause auxiliary be fronted across the relative clause; it is *structure-dependent*:

(9) Is the dog that is chasing the cat barking?

Simple questions such as (6) are quite frequent in child-directed speech. Chomsky argued that these questions support the structure-independent rule (8) because in both cases the auxiliary which is closest to the subject NP is placed in front. Complex yes/no-questions such as (9), on the other hand, seem to be virtually absent from child-directed speech. Consequently, a child has no inductive basis to infer the correct rule for question formation from the linguistic input. To explain why children acquire the syntax of yes/no-questions nonetheless, Chomsky proposed that children have an innate bias to induce hierarchical structures which allow for appropriate fronting constraints to be learned.[8]

Others have pointed out that structure-dependent auxiliary displacement also occurs in many other sentence types with subordinate or complement clauses[9,10] such as

(10) a. Could I have your French fries, if you're done with eating?
b. Why couldn't anyone who was at home close the window?

If these structures are sufficiently frequent in child-directed speech they might support learning the correct rule for yes/no-questions. A third approach comes from statistical learning with neural networks.[11,12] Reali & Christiansen, for instance, trained a simple-recurrent network on the Bernstein-Ratner corpus of mother-child interaction. In a grammaticality judgement task their trained model displayed a strong bias towards grammatical over ungrammatical yes/no-questions. Since these questions did not occur in the input to the model, this suggests that the linguistic environment of children might be rich enough for them to induce the correct syntax. The results of Reali & Christiansen were obtained by tagging the training corpus with parts of speech and they did not distinguish between verbs and auxiliaries or between different kinds of pronouns. Their account of question learning might only work under these assumptions and it remains to be seen whether children induce statistical constraints over similar types of categories.[a]

6. Question learning in the Dual-path model

In contrast to earlier approaches which emphasized the structural nature of the learning problem, our account of this generalization is based on meaning. We assume that children and adults who produce polar interrogatives represent a message that is made up of two propositions. We will show that a model which is given input messages with one and two propositions can learn proposition-specific syntactic constraints that allow it to generalize appropriately for polar interrogatives.

To describe our approach, we will first characterize the language that the Dual-path model was exposed to. This language contained basic single-clause constructions (intransitives, active/passive transitives, prepositional/double-object datives, and obliques), the combinatorially complete set of sentences with one relative clause composed from these constructions, simple yes/no-questions, and complex wh-questions:

[a]Similar results obtained with a more general n-gram model proposed in their paper, however, do not depend on these assumptions.

(11) Who is the cat that was chasing the dog playing with?

As many have argued, we suggest that the syntax of complex yes/no-questions can be assembled piecemeal from simpler and similar constructions which are warranted in a child's linguistic environment. Subject-auxiliary inversion might be learned from simple yes/no-questions in the input, and auxiliary displacement across a relative clause might be learned from complex wh-questions such as (11). In contrast to other approaches, our approach assumes that children use language-independent message information to help them produce polar interrogatives.

The Dual-path model was trained on message-sentence pairs from the language described above, and tested on novel sentences from this language and complex yes/no-questions (which were not in the training language). We obtained the learning curves of Figure 7. The x-axis represents the

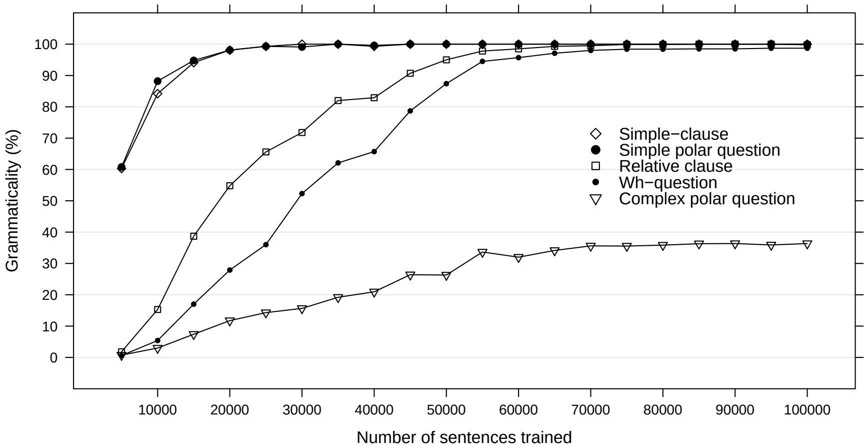

Fig. 7. Learning complex yes/no-questions.

number of sentences the model was trained on, the y-axis measures the grammaticality of the model's productions. Structures which were attested in the input were learned very well. The model correctly produced single-clause utterances and simple yes/no-questions quickly, followed by declaratives with relative clauses and wh-questions. When tested on the novel complex yes/no-questions, the model reached nearly 40% grammaticality. This shows that the model was able to use the two proposition message to help it learn the right generalization. Moreover, the model generalized in desirable ways in that it showed a clear preference for right-branching over center-embedded yes/no-questions and a preference for subject-relativized over object-relativized yes/no-questions (Figure 8).

Sentence meaning helped the model to segment utterances into the parts that correspond to the main clause, e.g., is the dog barking?, and the parts related to the embedded clause (that is chasing the cat, see Figure 6). Thus, the auxiliary in the main clause was controlled by a different part of the message than the auxiliary in the embedded clause. Questions were signaled in the event-semantics by question features which were neutral between clauses. Hence, the message for complex yes/no-questions did not bias the model towards selecting the main clause auxiliary. But the model learned to associate the question feature with the main clause auxiliary because it experienced simple yes/no-questions in training and because sentence-initial auxiliaries in complex wh-questions were never extracted from the embedded clause. These two types of information lead the model to shift the auxiliary that was controlled by the main clause message to the front when tested on complex yes/no-questions. In this way the system learned that picking the auxiliary closest to the subject NP was not appropriate.

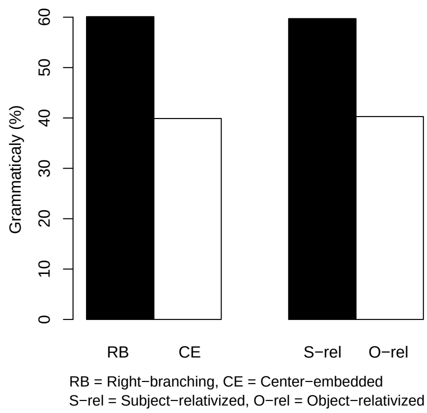

Fig. 8. Differential generalization for yes/no-questions.

Previous modelling work in this domain did not explain how question production could be achieved, and this is the first explicit model that can generate correct complex yes/no-questions from semantic representations, in the absence of these structures in the training corpus. The model did not reach an adult level of performance in which questions are produced flawlessly. However, the model's behavior is consistent with error levels found in English-speaking children age [4;7–5:7] in a study by Ambridge et al.[13] (mean correct production of center-embedded questions: children 27%, model 29%). Furthermore, errors that the model made did not result from structure-dependent auxiliary fronting. This can be verified by examining the initial segments of complex yes/no-questions the model produced in testing. Figure 9 compares the results for two training conditions, one in which the only questions the model received in training were simple yes/no-questions (left bars), the other in which the model was also exposed to complex wh-questions (right bars). The y-axis shows the percentage of productions that perfectly matched either a structure-dependent initial segment (Is the cat that was chasing) or a structure-independent initial segment (*Was the cat that chasing) for 1000 test questions. In both conditions the model never extracted auxiliaries from the embedded clause. This indicates that errors in the Dual-path model's productions did not reflect a structure-independent hypothesis about auxiliary fronting in the absence of complex yes/no-questions in the input.

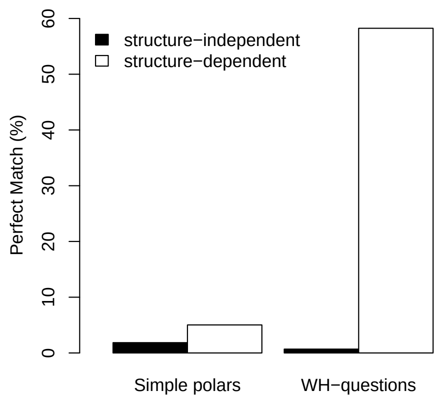

Fig. 9. Initial segments of yes/no-questions.

To compare our results with those of Reali & Christiansen, we tested the trained model on pairs of grammatical and ungrammatical center-embedded yes/no-questions. The model received a message input which was neutral between the two forms. The output was then compared to both targets, and classified as either grammatical or ungrammatical based on a graded performance measure (Figure 10). In 88% of the tested pairs the model's output was closer to the grammatical question. Quantitatively, these results are similar to those of Reali & Christiansen. Our test set, however, contained a considerable amount of structural variation[b] and our results did not depend on tagging the input in a specific way. These results demonstrate that structure-dependent auxiliary fronting can be learned

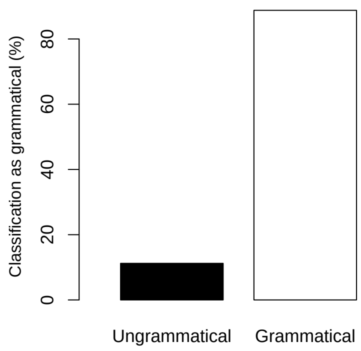

Fig. 10. Grammaticality judgement.

[b]Subject-relativized intransitive, transitive (active/passive), dative (prepositional/ditransitive) and oblique embeddings, and object-relativized embeddings when permitted.

from the structures that occur in child-directed speech as long as one assumes that children link syntax to meaning representations that distinguish different propositions in complex messages.

7. Conclusions

Adult speakers use language to convey meaning, and it has been argued that children must also use meaning in syntactic development if they are to acquire adult-like linguistic representations (MacNamara,[14] Pinker,[15] Tomasello[1]). We presented one of the few explicit models that uses meaning for syntax development. The model learned associations between parts of the message and subsequences of words, and it could combine these regularities in novel ways. This mechanism could explain generalization of words to novel slots and generalization of subsequences to novel embeddings. The Dual-path model could even generate polar questions without having experienced the target structure in training. Therefore, the structure of meaning may obviate the need for innate syntax-specific knowledge in the acquisition of adult-like language abilities.

References

1. M. Tomasello, *Constructing a Language: A Usage-Based Theory of Language Acquisition* (Harvard University Press, 2003).
2. A. Goldberg, *Constructions at Work: The Nature of Generalization in Language* (Oxford University Press, 2006).
3. G. F. Marcus, *Cognitive Psychology* **37**, 243 (1998).
4. F. Chang, G. S. Dell and K. Bock, *Psychological Review* **113**, 234 (2006).
5. J. L. Elman, *Cognitive Science* **14**, 179 (1990).
6. R. F. Hadley, *Mind and Language* **9**, 247 (1994).
7. F. Chang, *Cognitive Science* **26**, 609 (2002).
8. N. Chomsky, The linguistic approach, in *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*, ed. M. Piatelli-Palmarini (Harvard University Press, 1980)
9. G. K. Pullum and B. C. Scholz, *The Linguistic Review* **19**, 9 (2002).
10. B. MacWhinney, *Journal of Child Language* **31**, 883 (2004).
11. J. Lewis and J. L. Elman, Learnability and the statistical structure of language: Poverty of stimulus arguments revisited, in *Proceedings of the 26th Boston University Conference on Language Development*, (Cascadilla, 2001).
12. F. Reali and M. H. Christiansen, *Cognitive Science* **29**, 1007 (2005).
13. B. Ambridge, C. E. Rowland and J. M. Pine, *Cognitive Science* **32**, 222 (2008).
14. J. MacNamara, *Psychological Review* **79**, 1 (1972).
15. S. Pinker, *Language Learnability and Language Development: The Acquisition of Argument Structure* (Harvard University Press, 1984).