Rethinking the syntactic burst in young children

Christophe Parisse
INSERM-Modyco
Paris X Nanterre University
CNRS
parisse@vjf.cnrs.fr

Abstract

A testing procedure is proposed to re-evaluate the syntactic burst in children over age two. The experimentation is based on the children's capacities in perception, memory, association and cognition, and does not presuppose any specific innate grammatical capacities. The procedure is tested using the large Manchester corpus in the CHILDES database. The results showed that young children grammatical capabilities (before age three) could be the results of simple mechanisms and that complex linguistic mastery does not need to be available so early in the course of language development.

1 Introduction

Between the ages of two and three, most children go through a syntactic burst. In other words, they progress from uttering one word at a time to constructing utterances with a mean length of more than three words, and frequently longer, and they do this without any negative evidence and with limited input data (Ritchie & Bhatia, 1999). This represents quite a mystery, which is often explained by postulating the existence of innate constraints on the grammar of the human languages and the human mind (Pinker, 1984; Wexler, 1982). This report uses an iterative procedure to demonstrate that what appears to be near magical could result mostly from mechanisms that do not require the existence of innate principles of grammar, as they are based on children's inherent capacities for perception, memory and association (Jusczyk & Hohne, 1997; Saffran, Johnson, Aslin, & Newport, 1999). The acquisition of complex 'across the board' grammar does not appear to be necessary to explain children's behavior before age three or more. At that age, much more complex and structured input data will be available to children, thereby increasing their learning capacities and reducing the limitations on knowledge they may acquire.

2 A testing procedure in three parts

The testing procedure for grammatical development that will be implemented in this report is made of three parts.

The goal of the first and the second part is to determine the basic elements that children use to construct language. Two assumptions are made about young children's perceptive and mnemonic capacities: anything they have once produced, they can produce again; and, when their language exactly reproduces an adult's, this can be explained as a simple copy of their input.

Part 1: All single-word utterances produced by children are meaningful to them; they are directly derived from adults' output. They are the basic elements that children use to build language.

Part 2: Children's multi-word utterances containing only one word already produced in isolation (words produced in part 1), along with other words never produced in isolation (never produced at part 1), are also basic elements that children use to speak. They are also directly derived from children's input; this is facilitated by the children's knowledge of isolated words. These multi-word utterances are manipulated and understood by chil-

Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition, pages 65–72,
Prague, Czech Republic, June 2007 ©2007 Association for Computational Linguistics

dren as single blocks, just as isolated words are. They may also be called frozen forms.

The goal of the third part is to check whether the basic elements identified in part 1 and 2 are sufficient to account for the children's multiword utterances.

Part 3: Children link utterances produced at parts 1 and 2 to produce multi-word utterances with more than one word already produced in isolation (words produced in part 1). They do this using a simple concatenation mechanism and the fact that the utterances they create have a pertinent meaning prevents them from producing aberrant utterances.

Since the productions of children and their adult partners are easy to record, it is possible to test whether the testing procedure has sufficient generative power to account for all children's productions. However, some points could make such a demonstration more difficult than it appears. First of all, the assumption made in part 1 is not always true, as it is quite possible for a child to reproduce any sequence of sounds while playing with language. This uncertainty about part 1 is only important in conjunction with part 2, as isolated words are the key used to parse the elements of part 2. To decide that a word has meaning in isolation for a child, it has been assumed that it must first have meaning in isolation for an adult. Words in the categories of determiner and auxiliary produced in isolation have been considered as not having meaning in isolation and have therefore been removed from the elements gathered at part 1. Analysis of language data demonstrated that this assumption is quite reasonable, as the use of these words in isolation is often the result of unfinished utterances, with incomplete prosody.

Measuring the generative power of the testing procedure implies evaluating the accuracy of the assumptions made in parts 1, 2 and 3. These assumptions are quite easy to accept for very young children, at the time of the first multi-word utterances, i.e. before age two. The question is: to what extent is this true and until what age? Two experiments have been carried out in order to answer this question.

3 Experiment 1

The experiment 1 used a corpus extracted from the CHILDES database (MacWhinney, 2000). It is referred to as the Manchester corpus (Theakston, Lieven, Pine, & Rowland, 1999) and consists of recordings of 12 children from the age of 1;10 to 2;9. Their mean length of utterance varies from 1.5 to 2.9 words. Each child was seen 34 times and each recording lasted one hour. This results in a total production of 537,811 words in token and 7,840 in type. For each child, the average is 44,817 words in token (SD = 9,653) and 1,913 in type (SD = 372).

The testing procedure was run in three steps in an iterative way. Each step from the experiment corresponds to one of the parts described above.

Step 1: For each transcript, the child's single-word utterances are extracted and added to a cumulative list of words uttered in isolation, referred to as L1. It is possible to measure at this point whether the words on L1 can be derived from the adult's output. In order to do this, a cumulative list, L-adult, of all adult utterances is also maintained.

Step 2: For each multi-word utterance in the transcript, the number of words previously uttered in isolation is computed using list L1. Multi-word utterances with only one word uttered in isolation are added to a list called L2. It is possible to measure at this point whether the utterances on L2 can be derived from the adult's output (list L-adult above).

Step 3: the remaining utterances (list L3), which contain more than one word previously uttered in isolation, are used to test the final step of the algorithm. The test consists in trying to reconstruct these utterances using a catenation of the utterances from lists L1 and L2 only. Two measurements can be obtained: the percentage of utterances on list L3 that can be fully reconstructed (referred to below as the 'percentage of exact reconstruction') and the percentage of words in the utterances on list L3 that contribute to a reconstruction (referred to below as the 'percentage of reconstruction covering'). For example, for the utterance 'The boy has gone to school', if L1 and L2 contain 'the boy' and 'has gone' but not 'to school', only 'the boy has gone' can be reconstructed, thus leading to a percentage of reconstruction covering of 66%. Thus, the percentage of exact reconstruction is the percentage of utterances with a 100% reconstruction covering. The percentages of list L3 that are reconstructed or recovered do not include utterances from L1 and L2 lists.

The testing procedure is iterative because it is performed in turn for each of the transcripts of the corpus. List L1, L2 and L-adult are cumulative, which means that the list obtained with transcript 1 are used as a starting point for the analysis of transcript 2, and so on. This presupposes that children can reuse data they heard only once a long time after they heard it.

In Step 1 it was found that the percentage of words on L1 present in adult speech has a mean value of 91% (SD = 0.03). Step 2 revealed that the percentage of elements of L2 present in adult speech has a mean value of 67% (SD = 0.05). These two results are stable across ages—even though lists L1, L2 and L-adult are growing continuously. After two transcripts, for all 12 children, lists L1 + L2 represent 11,979 words in token and L-adult contains 82,255 words in token. After 17 transcripts, these totals are 89,479 and 688,802, respectively. After 34 transcripts, they total 167,149 and 1,370,565. The ratio comparing the size of L1 + L2 and L-adult does not evolve much, varying between 6 and 8.

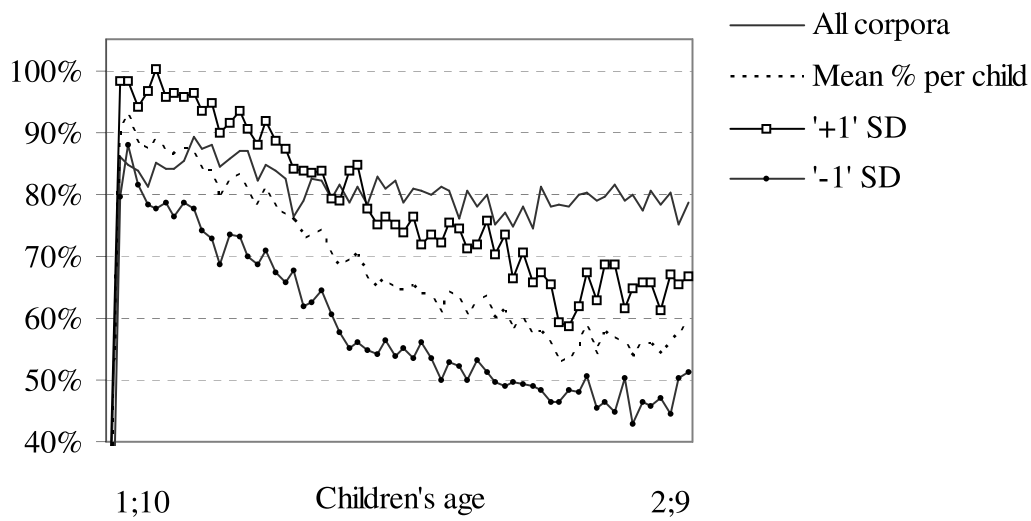

Figure 1: Percentage of utterances exactly reconstructed

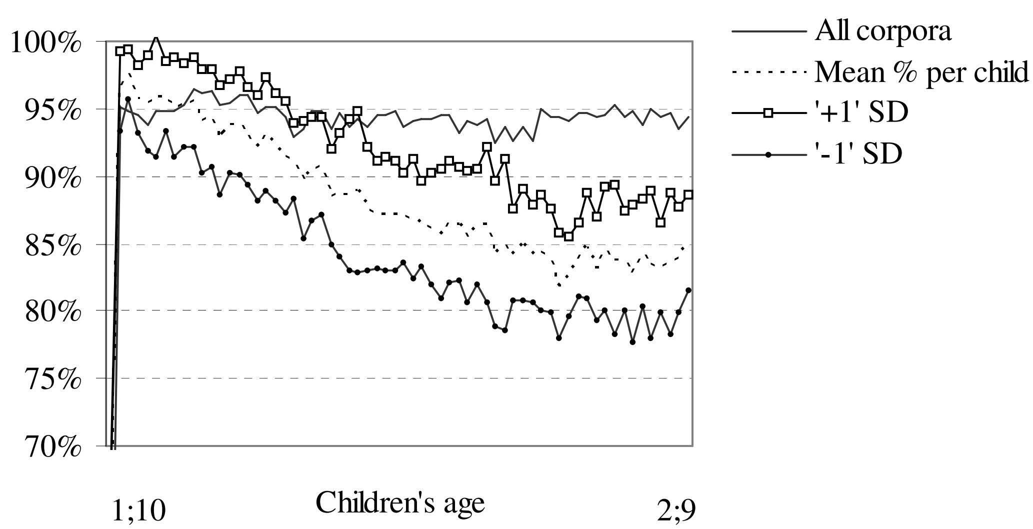

Figure 2: Percentage of reconstruction covering in all utterances

The results for Step 3 are presented in Figures 1 and 2. Each point in the series corresponds to the nth iteration performed with the nth transcript. The mean value is the mean of the percentage for all children considered as individuals (reconstruction between a child's corpus and his/her parents' corpus only). The algorithm is also applied to all corpora: for each point in the series of recordings, the 12 files corresponding to 12 children are gathered into a single file used to run the nth iteration of the algorithm. Percentages for all corpora are shown with a bold line. The percentages are clearly higher for the aggregated corpora, although the number of unknown utterances (list L3) increased more than the number of known utterances (lists L1 and L2). After two transcripts, there are half as many elements in list L3 as in L1 + L2. But after 17 transcripts, L3 is 42% larger than L1 + L2, and after 34 transcripts, it is 127% larger. As children grow older, there is a decrease in the scores for exact reconstruction and reconstruction covering. This decrease is greater in individuals than for the children as a group, which suggests a size effect.

4 Experiment 2

The second experiment uses the same corpus and reproduces the same tests but assumes that children have knowledge of the syntactic categories Noun and Verb. The conditions of step 2 and step 3 are more easily fulfilled if the children have a certain amount of syntactic class knowledge. As described by Maratsos and Chalkley (1980), it is possible for children to learn syntactic classes from the contexts in which words occur. However, knowledge of part of speech is unlikely in very young children on the basis of syntactic distribution. Semantic knowledge can also help to construct syntactic knowledge (Bloom, 1999) for classes such as common nouns, proper nouns and verbs, and perhaps also adjectives and adverbs. To simulate the fact that children are able to construct the classes of common nouns, proper nouns and non-auxiliary verbs, it suffices to substitute every occurrence of common or proper nouns in the Manchester corpus by the symbol 'noun' and every occurrence of non-auxiliary verbs by the symbol 'verb'. This is easy to realize because the Manchester corpus has been fully tagged for part of speech, as described in the MOR section of the CHILDES manual (MacWhinney, 2000). The result is that list L1 now includes all nouns, all verbs plus all words occurring in isolation, as in the first experiment. In list L2, in utterances that include a word from the categories Noun or Verb, this word is substituted by the symbol 'noun' or 'verb'. These utterances now form rule-like productive patterns known as formulaic frames (Peters, 1995) or slot-and-frame structures (Lieven, Pine, & Baldwin, 1997) — for example, 'my + NOUN'.

When we reproduce the first experiment under these conditions, the new results obtained at steps 2 and 3 should be better, in the sense that they should correspond more closely to the adult input, and should hold up longer on the age scale.

The results for Step 1 and Step 2 are indeed better than before. The percentage of utterances on L2 present in adult speech has a mean value of 91% (SD = 0.02).

The results for Step 3 are presented in Figure 3 (for exact reconstruction) and Figure 4 (for reconstruction covering). In each of these figures, two results are presented for the whole Manchester corpus: one assuming no category knowledge, and one assuming the knowledge of the three categories proper noun, common noun and verb. The percentages of reconstruction become markedly higher, as any combination that contains some of three categories proper noun, common noun and verb is known for all occurrences of words from these categories. The mean for exact reconstruction with 'no category' knowledge is 67% (SD = 5.7) and 87% (SD = 2.0) for reconstruction covering. These values increase to 83% (SD = 5.2) and 95% (SD = 2.6) for 'noun and verb' knowledge.

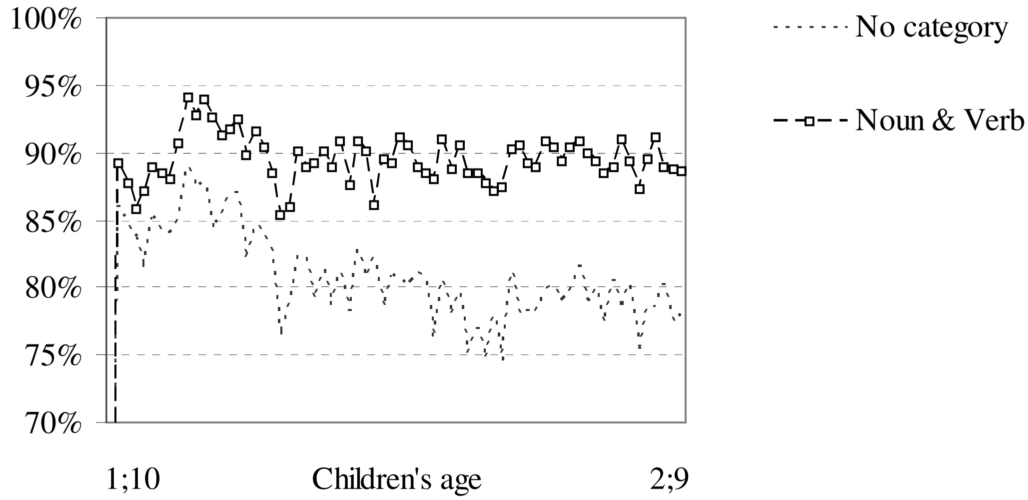

Figure 3: Percentage of utterances exactly reconstructed, depending on the degree of knowledge of noun and verb categories

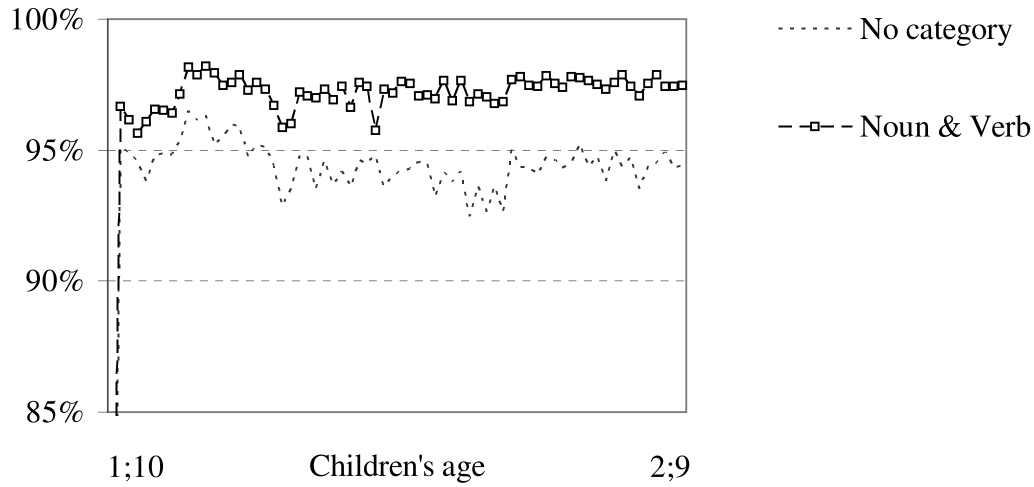

Figure 4: Percentage of reconstruction covering in all utterances, depending on the degree of knowledge of noun and verb categories

5 Experiment 3

A limit of experiments 1 and 2 is that nothing indicates how long the three-step mechanisms would remain efficient and appropriate. We supposed that these mechanisms would remain operational at an older age. This can be checked using other material from the CHILDES database with recordings spanning a longer period. The corpus chosen for the test is Brown's (1973) Sarah corpus, which ranges from age 2;3 to age 5;1; with its 139 different transcripts, it follows the development of the child's language quite well and is well suited for the purposes of this study, which requires lengthy corpora. The mean length of utterance varies from 1.47 to 4.85 words. This results in a total production of 99,918 words in token and 3,990 in type.

Step 1 found the percentage of words on L1 present in adult speech to have a mean value of 77% (SD = 14.5). Step 2 revealed that the percentage of elements of L2 present in adult speech had a mean value of 38% (SD = 11.5). These two results are stable across ages. With the assumption of a knowledge of the Noun and Verb categories, results for Step 1 and 2 are, respectively, 83% (SD = 13.8) and 55% (SD = 16.6).

The results for Step 3 are presented in Figure 5 (for exact reconstruction) and Figure 6 (for reconstruction covering). In each of these figures, two results are presented: one assuming no category knowledge and one assuming knowledge of the three categories Proper Noun, Common Noun and Verb. The mean for exact reconstruction with “no category” knowledge is 54% (SD = 17.6) and 84% (SD = 6.6) for reconstruction covering. These values increase 72% (SD = 11.9) and 93% (SD = 4.0) for “Noun and Verb” knowledge.

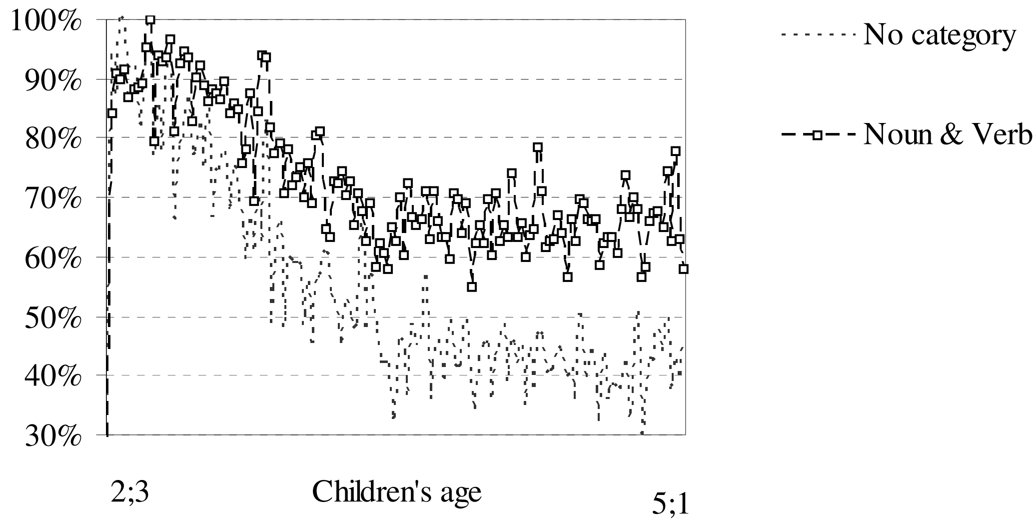

Figure 5: Percentage of utterances in the Sarah corpus exactly reconstructed, depending on the degree of knowledge of vocabulary and syntactic categories

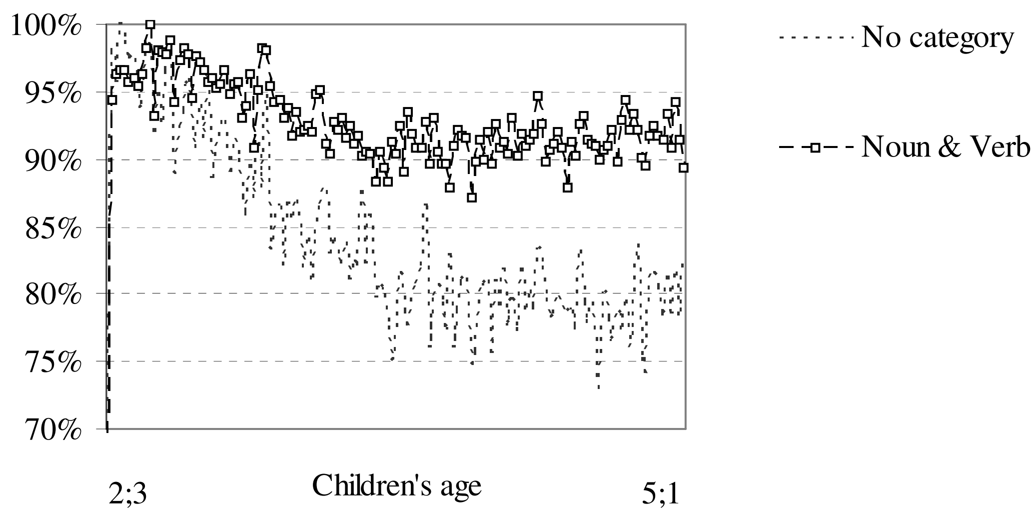

Figure 6: Percentage of reconstruction covering in all utterances in the Sarah corpus, depending on the degree of knowledge of vocabulary and syntactic categories

The average percentages of reconstruction are lower for the Sarah corpus than for the Manchester corpus. Comparing Figures 3 and 6 and Figures 4 and 7, one can see that there is a drop in the reconstruction performances in the third year. The percentages for Sarah in her second year were as high as those for the Manchester corpus children. Part of this drop in performance may be attributed to the smaller corpus. Indeed, comparing Figures 1 and 3 and Figures 2 and 4, it appears that the drop in performance that became visible when single child corpora were used was not in evidence when all the corpora were amalgamated into one big corpus. It is also possible that the drop in performance found in the Sarah corpus reflects a progressive decrease in the systematic use of a simple concatenation procedure by the child.

6 Discussion

The testing procedure does not achieve a full 100% reconstruction in the test conditions described above, where the database consists of only 34 one-hour recordings for each of the 12 children in the corpus. This corresponds globally to a pseudo-corpus of 408 hours, which amounts to 8 to 10 weeks of speech. With a larger corpus, the results would probably be better, as indicated by the increase in percentage of recovery when one moves from children in isolation to children as a group (see Figures 1 and 2). In addition, there are bound to be words that children utter for the first time in multi-word utterances even though they could have been produced as isolated utterances. The percentage of reconstruction, however, is still quite high, as was the case for results obtained using a similar methodology with Hungarian children (MacWhinney, 1975). With the assumption of a benefit from the use of the Noun and Verb categories, which somewhat circumvents the limited size of the corpus, the results are very high.

A problem with the second experiment is that it is not sure that children can have a knowledge of part of speech (even very general part of speech such as noun and verb) with semantic knowledge only. However, the experiment 2 is interesting as it can be viewed as a way to extend artificially a limited corpus. Instead of saying that children have the knowledge of part of speech, we propose that noun and verb as so common in adult speech that an extended corpus will contain all basic utterances with a single content word and the appropriate grammatical context. In other words, list L2 will contain all the most basic syntactic constructions. Although this will not be the case in reality, it is indeed possible that a full corpus covering all utterances produced by adults will contains a very large number of L2 structures. In this way, experiment 2 provides a measure of the upper limit that can be reached by the crude mechanism presented in this article (L3 constructions).

The testing procedure does not cover all language acquisition processes before the age of three. Its rather crude mechanisms would, on their own, produce many aberrant utterances if they were not regulated by other mechanisms. The first of these regulatory mechanisms is semantics, as children produce language that, for them, makes sense. They will articulate thoughts with two or three elements that complement each other logically and thus create utterances interpretable by adults. Strange utterances may be produced on occasion but none will sound alien. Secondly, even though children sometimes join words or groups of words randomly when very young, they soon start to follow a systematic order probably copied from adults' utterances (Sinclair & Bronckart, 1972). To do this, they merely have to concentrate on the words or groups of words that they already master, having previously uttered them as single words. Indeed, form-function mapping is easier with single-word utterances than with multi-word utterances and this helps to manipulate single-word forms consciously. Thus, single-word utterances are better candidates than most to become the first elements in a combinatorial system and to undergo representational redescription (Karmiloff-Smith, 1992). Their semantic values allow one to perform semantic combinations. By the age of two, associations words or frozen forms may be sufficient to allow children to produce and control language.

The fact that children can learn to produce complex speech patterns quickly without complex grammatical knowledge casts a whole new light on the problem of the acquisition of syntax. The testing procedure relies heavily on semantics because it is assumed that what children understand, they will remember and manipulate. This does not necessarily contradict all the theories that claim that there are some innate principles specific to grammar acquisition (Pinker, 1984; Wexler, 1982). If children acquire high-level grammatical rules at a later period of their development than is usually admitted in these theories, then the structure of their input—the couple 'base phrase marker' plus 'surface sentence' (Wexler, 1982) — will be more complex. The more complex these structures, the lower the innate conditions on grammars. It would then be possible to progress from a simple system such as the association of frozen elements to a more complex one. Late grammatical acquisition is a very important notion as it goes a long way towards explaining why there do not seem to be any neuronal structures specific to language or grammar (Elman et al., 1996; Muller, 1996). Late grammatical acquisition is also highly compatible with constructivist proposals such as Tomasello's (2003) and Goldberg's (2006).

It has often been said that children already master syntax by the age of three, which is quite remarkable considering the complexity of what they are acquiring. This report suggests that some simple generative mechanisms can explain the explosive acquisition of an apparent mastery of language observed in young children. It demonstrates once again that, as already shown for other linguistic developmental features (Elman et al., 1996), an apparently complex output may be the product of a simple system. The need for large-scale corpora to better tackle the problem of language acquisition with improved tools is also highlighted here.

References

Bloom, P. (1999). Theories of word learning: Rationalist alternatives to associationism. In W. C. Ritchie & T. K. Bhatia (Eds.), Handbook of language acquisition . San Diego: Academic Press.

Elman, J. L., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). Rethinking innateness: A connectionist perspective on development. Cambridge, MA: MIT Press/Bradford Books.

Goldberg, A. (2006). Constructions at Work: the nature of generalization in language. Oxford University Press.

Jusczyk, P. W., & Hohne, E. A. (1997). Infants' memory for spoken words [see comments]. Science, 277(5334), 1984-6.

Karmiloff-Smith, A. (1992). Beyond modularity: a developmental perspective on cognitive science. Cambridge, Mass.: MIT Press/Bradford Books.

Lieven, E. V., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. Journal of Child Language, 24(1), 187-219.

MacWhinney, B. (1975). Rules, rote, and analogy in morphological formations by Hungarian children. Journal of Child Language, 2, 65-77.

MacWhinney, B. (2000). The CHILDES project : Tools for analyzing talk (3rd). Hillsdale, N.J, Lawrence Erlbaum.

Maratsos, M. P., & Chalkley, M. A. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K. E. Nelson (Ed.), Children's language. Vol: 2 . New York, NY: Gardner Press.

Muller, R.-A. (1996). Innateness, autonomy, universality? Neurobiological approaches to language. Behavioral and Brain Sciences, 19(4), 611-675.

Peters, A. M. (1995). Strategies in the acquisition of syntax. In P. Fletcher & B. MacWhinney (Eds.), The handbook of child language . Oxford, UK: Blackwell.

Pinker, S. (1984). Language learnability and language development. Cambridge, MA: Harvard University Press.

Ritchie, W. C., & Bhatia, T. K. (1999). Child language acquisition: Introduction, foundations, and overview. In W. C. Ritchie & T. K. Bhatia (Eds.), Handbook of language acquisition . San Diego: Academic Press.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. Cognition, 70(1), 27-52.

Sinclair, H., & Bronckart, J. P. (1972). S.V.O. A linguistic universal? A study in developmental psycholinguistics. Journal of Experimental Psychology, 14(3), 329-348.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (1999). The role of performance limitations in the acquisition of 'mixed' verb-argument structure at stage 1. In M. Perkins & S. Howard (Eds.), New directions in language development and disorders : Plenum Press.

Tomasello, M. (2003). Constructing a language: A usage-based theory of language acquisition. Cambridge: MA, Harvard.

Wexler, K. (1982). A principle theory for language acquisition. In E. Wanner & L. R. Gleitman (Eds.), Language acquisition - the state of the art. New York: Cambridge University Press.