Title: Simple principles for a complex output: An experiment in early syntactic development

Christophe Parisse

Institut National de la Santé et de la Recherche Médicale (INSERM), Paris, France

Running head: Simple principles for a complex output

Acknowledgments

This work was supported by a grant from INSERM, France: Contrat de Recherche Inserm (4U009B). Many thanks to Anne Reymond and Zofia Laubitz for proofreading this text.

Corresponding address:

Christophe PARISSE
Laboratoire de neuropsychopathologie de l'enfant
Bâtiment Pharmacie, 3ème étage,
Hôpital de la Salpêtrière
47 Boulevard de l'Hôpital
75651 PARIS CEDEX 13
FRANCE

Title: Simple principles for a complex output: An experiment in early syntactic development

Abstract:

A set of iterative mechanisms, the Three-Step Algorithm, is proposed to account for the burst in the syntactic capacities of children over age two. These mechanisms are based on the children's perception, memory, elementary rule-like behavior and cognitive capacities, and do not require any specific innate grammatical capacities. The relevance of the Three-Step Algorithm is tested, using the large Manchester corpus in the CHILDES database. The results show that 80% of the utterances can be exactly reconstructed and that, when incomplete reconstructions are taken into account, 94% of all utterances are reconstructed. The Three-Step Algorithm should be followed by the progressive acquisition of syntactic categories and use of slot-and-frame structures which lead to a greater and more complex linguistic mastery.

Title: Simple principles for a complex output: An experiment in early syntactic development

The background of the experiment

Between the ages of two and three, children 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 unstructured input data (Pinker, 1994; Ritchie & Bhatia, 1999). How is this done?

In their book Origins of grammar, Hirsh-Pasek and Golinkoff (1996, p. 12), present a set of questions that every theory about language acquisition should address:

1. What is present when grammatical learning begins?
2. What mechanisms are used in the course of acquisition?
3. What types of input drive the language-learning system forward?

Although these three questions cover the subject quite well, a fourth question should be added:

4. What is the structure of the adult grammatical system?

This last question is equally fundamental to the problem, as explained by Chomsky (1959). The point has been made repeatedly in studies about language development (see, for example, Radford, 1990; Wexler, 1982) that children progress from stage (1) to stage (4) and that language acquisition principles should take this fully into account. However, generative grammars do not provide the only proposed answer to question (4) – see Van Valin & La Polla (1997, p. 642 and 675), and alternatives such as cognitive grammars (Tomasello, 1998) exist. Other works have gone further and criticized the classical linguistic (and cognitivist) paradigm (see, for example, Harris, 1990; Shanon, 1993).

The goal of this article is to propose some answers to these questions, especially questions (1) to (3). Question (4) is there only as a reminder that the problem is a very open one. Possible answers to questions (1), (2) and (3) vary a lot from one theory of language acquisition to another. There is not much agreement yet as to whether children have a strong or weak initial knowledge of grammar (Hish-Pasek & Golinkoff, 1996; Tomasello, 2000). The great discrepancies between the various theories are probably a consequence of the youth of this field, but also of the size and the complexity of the phenomenon. Language acquisition takes years, children's input is numbered in the millions of utterances and little enough is known about the brain's properties yet. This shows how difficult it is to simply observe the phenomenon and how important the quality of the tools used for this purpose becomes.

In this article, the tool is a computer experiment of the kind developed by Brent and Cartwright (1997), Elman (1995) and Schütze (1997). It is based on real child language: the Manchester corpus (Theakston, Lieven, Pine, & Rowland, 1999) of the CHILDES system (MacWhinney & Snow, 1985). The underlying idea is to posit certain acquisition mechanisms that answer questions (1) to (3) and to test whether these mechanisms are capable of producing the results actually demonstrated by children. The experiment works as follows. First, the children's output and input are processed up to a given age and the result represents the knowledge acquired until then. Second, utterances produced by the children after that age are matched against a generative algorithm that emulates their output. The quality of the match—how closely the algorithm's output matches the children's real utterances—is deemed to reflect the adequacy of the mechanisms devised.

The principles underlying the mechanisms

We propose that the development of language follows a path that starts from simple linguistic processes, then progressively involves more and more complex and abstract processes. When a new type of linguistic process appears, it does not take the place of the previous processes but comes as an addition to them. This principle is presented in Figure 1, where processes of growing levels of complexity and abstraction (from 1 to 4) add to each other during the course of development. One can see on the figure that the relative import of the more simple processes diminishes with age, but never disappears nor even becomes merely negligible.

include Figure 1 about here

The principles corresponding to the four levels of abstraction shown in Figure 1 would be:

1. Isolated rote items and combinations of rote items: free constant + free constant.
2. Item-based constructions: bound constant + free variable.
3. Rule-based constructions of limited scope: free constant + bound variable.
4. Rule-based constructions of general scope: free variable + free or bound variable.

An example of a level 1 construction is mummy socks, where mummy and socks are free constants. An example of level 2 construction is this is a horse where this is is a bound constant and a horse is an instance of a free variable. An example of level 3 construction is the range of constructions goes, goed, and going, where go is a free constant and –es, -ed, and -ing, are instances of a bound variable. Finally an example of level 4 construction is the man is running where the man is an instance of a free variable noun, and is running is an instance of a free variable verb.

For each level, the order and number of elements involved in the different constructions may vary freely according to each specific construction. Constant may precede variable or the opposite and there is no other limit than processing limit to the number of constants and variables used in a construction (see MacWhinney (1982) for a description of a similar set of strategies, which also progress from simple to complex constructions). In the model presented here, constructions are the result of the development of brain connectivity following natural experience. That which allows a construction to develop is the presence of adequate data in the input, and more importantly, in the brain itself – the patterns of brain connectivity developed during previous experience. There is a logical ordering of the constructions acquired during development, and constructions of abstraction level 2 rely on material of level 1, constructions of level 3 on levels 1 and 2, and constructions of level 4 on levels 1, 2 and 3. However, constructions at level n do not rely on the whole of the material at level n-1, but only on some parts of it. Constructions of level 2, or 3, or 4, may therefore appear before other constructions of level 1, or 2, or 3, provided that these constructions are independent from one another and that the required material is available. The lower the level of abstraction, the smaller the quantity of input data needed to develop a construction. For this reason, constructions of a higher level of abstraction appear after constructions of a lower level of abstraction, but there is no reason why highly frequent constructions of level 2, 3 and 4 should not appear before infrequent constructions of lower levels of complexity.

Item-based patterns –level 2 of abstraction–, also know as formulaic frames (Peters, 1995) or slot-and-frame structures (Lieven, Pine, & Baldwin, 1997) were first described by Braine

(1963; 1976) and were later supported by Tomasello (1992; 2000) as children's privileged way of producing their first complex multi-word utterances. However, constructions such as See ___ or Daddy's ___ (from Tomasello, 2000) are very limited in scope. The items See and Daddy's are fixed and children must have encountered a whole set of input patterns such as See x_1, See x_2, etc., to be able to produce the great number and the large variety of utterances they often already do produce at age two. This is the reason why the existence of a level based on more simple principles has been postulated above.

The current experiments focus on processes of level 1. Whenever confusion between processes of level 1 and 2 is possible, this will be specified as it appears. Level 1 is simple enough so that it works with a very limited quantity of data and knowledge and is thus fully operational at the very beginning of the production of multi-word utterances. Because it is simple and fast, this level is more involved when a child or an adult is subject to a heavy cognitive load or is asked to produce fast responses. Thanks to their simplicity, the set of level 1 mechanisms will also to quicker response and more frequent use when opposed to more complex mechanisms, unless it is superseded by a conscious process for some reason.

The set of mechanisms used in level 1 will be called the Three-Step Algorithm. It is based on the infant's well-attested capacities for perception, memory and rule learning (Jusczyk & Hohne, 1997; Marcus, Vijayan, Bandi Rao, & Vishton, 1999; Saffran, Aslin, & Newport, 1996; Saffran, Johnson, Aslin, & Newport, 1999). At the beginning of multi-word utterance acquisition, children iteratively follow the three steps described below. Two assumptions are made about young children's perceptive and mnemonic capacities: anything they have once produced, they can produce again; when their language exactly reproduces an adult's, this is to be explained as a simple copy of their input.

The Three-Step Algorithm:

Step 1: All single-word utterances produced by children are meaningful to them; they are directly derived from adults' output.

Step 2: Children's multi-word utterances containing only one word already produced in isolation (words produced at step 1), along with other words never produced in isolation (never produced at step 1), are directly derived from adults' output; this is facilitated by the children's knowledge of isolated words. These multi-word utterances are manipulated and understood by children as single blocks, just as isolated words are.

Step 3: Children link utterances produced at steps 1 and 2 to produce multi-word utterances with more than one word already produced in isolation (words produced in step 1). They do this using the meanings of the utterances they link to create new utterances that are meaningful (to themselves at least).

One should note here that the elements identified in Step 2 are not fundamentally different from the elements identified in Step 1, from the child's point of view. The first two steps of the Three-Step algorithm describe the memorization process of the child. We know that children can memorize strings of words containing more than one adult word as a whole. Step 1 and Step 2 of the Three-Step algorithm help us to identify which strings are considered as atomic elements by the child. These atomic elements are then produced and associated by the child following the principles described in Step 3.

To sum things up, the Three-Step is an algorithm that helps the researcher to understand which linguistic structure the child is using. From the point of view of the child, there are only two steps, the first one containing Step 1 and Step 2 of the Three-Step algorithm and the second one corresponding to Step 3 of the Three-Step algorithm. How Step 1 and Step 2 can develop into more complex processes will be discussed in experiment 2 and further in the general discussion.

Since the productions of children and their adult partners are easy to record, it is possible to test whether the Three-Step Algorithm has sufficient generative power to account for all the children's productions. But some factors could make such a demonstration more difficult than it appears at first glance?

First of all, the assumption made in step 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 step 1 is only important in conjunction with step 2, as isolated words are the key used to parse the elements in step 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 produced in isolation and belonging to the categories of determiner or auxiliary have been considered as having no meaning in isolation and have therefore been removed from the elements gathered at step 1.

Measuring the generative power of the Three-Step Algorithm implies evaluating the accuracy of the assumptions made in steps 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 are these assumptions true and until what age? Two experiments have been carried out in order to answer this question.

The goal of this article is to test whether a constructivist theory can explain how children develop language so quickly. It is not to demonstrate the validity of a constructivist approach of language development (see instead Miller & Weinert, 1998, chapter 8; Tomasello, 2000). Many generativist theories explain that since language acquisition is so error free and so fast, the innate features of language have to be quite complex and powerful. The present experiments try to show that not so complex features can produce quite complex language and to add a few stones to constructivist edifice of language development theories.

Experiment 1

Materials

The first experiment used a corpus extracted from the CHILDES database (MacWhinney, 1991; MacWhinney & Snow, 1985). It is referred to as the Manchester corpus (Theakston et al., 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). Because children's production is sometimes phonetically imperfect, all the tests below were done using the %mor line of the Manchester corpus. This is possible because the Manchester corpus has been fully tagged for parts of speech, as described in the MOR section of the CHILDES manual (MacWhinney, 1991). The %mor line contains the parts of speech and the normalized version of the words. For this reason, there are no discrepancies between child and adult language due to phonetics, only the lexical or syntactic ones.

Procedure

The Three-Step Algorithm is tested in an iterative way:

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. At this point, it is possible to measure 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. At this point, it is possible to measure 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, resulting in a percentage of reconstruction covering of 66%. Thus, the percentage of exact reconstruction is the percentage of utterances with a 100% reconstruction covering.

Results

In Step 1 it was found that the percentage of words on L1 present in adult speech has a mean value of 72% (SD = 0.10). Step 2 revealed that the percentage of elements of L2 present in adult speech has a mean value of 58% (SD = 0.07). 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 change much, varying between 6 and 8.

include Figure 1 and Figure 2 about here

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 for each individual than for the children as a group, which implies a size effect.

Discussion

The Three-Step Algorithm 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. With a larger corpus, the results would probably be better, as indicated by the increase in the percentage of reconstruction when one moves from children in isolation to children as a group (see Figures 1 and 2). The global corpus corresponds to a pseudo-corpus of 408 hours, which amounts to 8 to 10 weeks of speech. A corpus covering all productions of a real child would be at least five times bigger. With this limited corpus, the percentage of reconstruction is still quite high, as was the case for results obtained using a similar methodology with Hungarian children (MacWhinney, 1975).

Experiment 2

Materials and procedure

The second experiment uses the same materials and the same basic procedure as the first experiment. Its goal is to test whether higher covering and reconstruction results can be obtained, either with a larger training set, or with more complex mechanisms involved. In a larger corpus, open-class words are more likely to have been produced in isolation by the child. Even if they have not been, it is still perfectly possible that they are nonetheless known to the child, as these words have the same grammatical status as other words produced in isolation. To simulate this knowledge, three options are possible.

The first option is to consider that all words produced by children belonging to grammatical categories usually produced in isolation are known to them. These categories are: nouns, verbs, adjectives, adverbs, exclamations, demonstrative pronouns, and interrogative pronouns. The last two are not open-class categories, but are nonetheless produced in isolation and therefore have meaning when taken in isolation—in other words, they are content words without being open-class words. To simulate this, we need only add all the words produced by the children and belonging to one of the above categories to list L1. The rest of the Three-Step Algorithm does not change. This option tests whether a maximal knowledge of isolated words would generate all the utterances produced by children. This supposes that each word has been heard (and memorized) by the children, which would increase the generative power of the Three-Step algorithm and simulate a bigger corpus.

The second option is to simulate a bigger corpus by multiplying the adult constructions available to the children, and not just the number of free lexical items. To do this, we consider that any construction produced by children containing one noun or one verb could have been heard and produced with any other noun or any other verb. To simulate this, we only have to replace every occurrence of common or proper noun in the Manchester corpus by the symbol "NOUN" and every occurrence of lexical verb by the symbol "VERB". This is easy to do using the %mor line of the Manchester corpus. The result is that list L1 now includes all the words occurring in isolation, as in the first experiment, with the exception of nouns and verbs, plus the two symbols NOUN and VERB. List L2 contains utterances with only one word occurring in isolation, as was the case in the first experiment, but where any occurrence of a word from the category Noun or Verb is replaced by the relevant symbol, such as for example "my + NOUN", stored instead of "my car". So that any noun or verb is considered to be known in isolation as well as in all the different noun or verb contexts known to the child –for example, it is supposed that every noun has been produced with "my" in front of it.

The third option uses the same principle as in the second one, but applied to all syntactic categories instead of just the categories Noun and Verb. This allows us to evaluate how generic the Three-Step Algorithm ultimately is. This option is an artificial means of multiplying the possible construction available to children, with the sole condition that these constructions respect standard syntactic constructions.

When we reproduce the first experiment under those conditions simulating bigger child and adult corpora, the 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.

Results

The results for Step 1 and Step 2 are indeed better than before. Previous results were 80% (SD = 6.9) for exact reconstruction and 94% (SD = 3.6) for reconstruction covering. With the assumption of a knowledge of the Noun and Verb categories, the percentage of words on L1 present in adult speech has a mean value of 86% (SD = 7.9). If one assumes a knowledge of all syntactic categories, the percentage is obviously 100% (SD = 0.0). On the other hand, if all children's content words are included in L1, then the percentage is smaller than in the first experiment, 62% (SD = 4.7). This is normal, as we are now considering more children's words than previously but exactly the same number of adults' words. With the assumption of a knowledge of the Noun and Verb categories, the percentage of utterances on L2 present in adult speech has a mean value of 79% (SD = 6.5); with a knowledge of all syntactic categories, the percentage is 99% (SD = 0.01); if all children's content words are included in L1, then the percentage is slightly higher than in the first experiment, 62% (SD = 5.1).

include Figure 3 and Figure 4 about here

The results for Step 3 are presented in Figure 3 (for exact reconstruction) and Figure 4 (for reconstruction covering). In each of these figures, four results are presented for the whole Manchester corpus: one assuming no category knowledge –which corresponds to the first experiment–, one assuming the knowledge of all content words in isolation, one assuming knowledge of the categories Noun and Verb, and one assuming knowledge of all syntactic categories. The percentages of reconstruction become markedly higher. The mean for exact reconstruction with "no category" knowledge is 80% (SD = 6.9) and 94% (SD = 3.6) for reconstruction covering. These values increase to 89% (SD = 6.6) and 97% (SD = 3.1) for "Noun and Verb" knowledge, and 99% (SD = 4.1) and 100% (SD = 1.4) for knowledge of "all syntactic categories". The result for the knowledge of all content words in isolation is 86% (SD = 5.1) and 96% (SD = 2.3).

Discussion

The assumption of a benefit from either the knowledge of content words or the knowledge of the Noun and Verb categories circumvents the limited size of the corpus. Yet, in both these cases, the 100% level is still not reached and knowledge of all categories is necessary to reach the 100% level. These tests are not a proof that children could produce all their utterances merely by following the Three-Step algorithm, because we have no guaranty that children have access to so much input data. However, these results do give an idea of the maximal generative capacity of the Three-Step algorithm.

Children very probably do not use processes of level 1 abstraction exclusively, but also have to tap into higher levels of abstraction. In some way, this is what is involved in Experiment 2. From the point of view of an adult or of the experimenter, what we did was simulate a large amount of input data. But from the child's point of view, the experiment with the Noun and Verb categories becomes a small experiment in item-based patterns –and thus an experiment of level 2 abstraction. If we assume that children have a knowledge of the categories Noun and Verb, they can then produce a great diversity of patterns in the X + NOUN or X + VERB forms where X are bound patterns and NOUN and VERB free variables. This amounts to having a subset forms that can be produced by item-based patterns, and the results demonstrate that the use of such item-based patterns from a very young age is necessary to account for the full complexity of language development. It would be interesting to test if a full simulation of item-based patterns would generate all the utterances produced by children. but this would mean defining how children regulate which type of elements may fill the free slots of the item-based patterns. This is outside the scope of the present work, which concentrates on processes of level 1 abstraction, but it would be a natural follow-up.

However, it is possible to anticipate this follow-up by looking at the circumstances in which the Three-Step algorithm fails to reconstruct the utterance fully. The type of errors produced at a given age tends to remain the same for at least a few months, so that a really thorough study of the incompleteness of the algorithm as been done for all the children, but only at age 1;11 and at age 2;9.

In the first example, at age 1;11, 731 utterances were reconstructed and out of them 226, 31% of the utterances, were not completely reconstructed. These incomplete reconstructions can be sorted into five different situations:

1. determiner the is not reconstructed – e. g. {oh} {where} the {cat} –elements between curly brackets are reconstructed.
2. determiner a or an is not reconstructed – e.g. a {fish}
3. personal pronoun I is not reconstructed – e.g. I {big} {baby} or I {find} {more}
4. a lexical form (noun, verb, adjective) is not present in the L1+L2 list – e.g. {two} penguin, buy {one}, or fat {sausage}
5. all other cases

They were 54 cases of situation 1, 27 of situation 2, 17 of situation 3, 68 of situation 4, and 81 of situation 5. The items in situation 5 have been checked by searching all adult utterances to determine whether the specific utterance could have been extracted from the adult's output. In 59 cases, it was possible to find the exact utterance or an overlapping utterance produced by an adult in a different recording than the one where the child's utterance occurred. In one case, an utterance, he's all gone, was produced by the adult only as an expansion of the child's utterance, but the large number of occurrences of it's all gone suggests that he's all gone could probably be used by an adult. This left 16 cases out of the 731 child utterances that could not be explained by the repetition of an adult utterance. Still, even in these cases, it was often possible to find close matches–e.g. for farmer ride him, it was possible to find the words farmer riding and the words ride him included in two different utterances, for and that bit out, the words and that bit and bit out. This shows that a large number of incomplete reconstructions can effectively be solved by the use of item-based patterns –namely situations 1, 2 and 3 as well as some cases in situation 5, but it is also true that situations 1, 2 and 3 do not necessarily need item-based patterns to be explained. In situations 1 and 2, it is just as likely that children have heard words such as fish or cat produced with a determiner, so that rote learning of blocks of words is perfectly possible. This also applies to verbs which are often heard following a personal pronoun. Situation 4 cannot be solved without resorting to rote learning and those few errors of the I {big} {baby} type seem more likely to be the product of the Three-Step algorithm than any other algorithm. For all these reasons, item-based patterns may not be a necessary feature for children aged 1;11.

Performing the same analysis at age 2;9 leads to a different conclusion. 1827 utterances were reconstructed, and of these 670, 36% of the utterances, were not completely reconstructed. The incomplete reconstructions could be sorted into the same situations as above, with the difference that situation 3 was extended to all subject personal pronouns and situation 4 and 5 were not differentiated. Situation 1 applies 198 times, situation 2, 75 times, situation 3, 180 times. There were also 95 situations which could be explained by the lack of a short phonological form (‘s, ‘re, ‘ve). The difference with age 1;11 is that in situations 4 and 5, one finds not only missing lexical forms, but also many cases where the unreconstructed form contains a preposition – with, near, to, at, of, in, into, for. If in English, one can argue that determiners are first learned by rote, this does not sound a workable solution for prepositions. At age 2;9, children begin to have a productive use of preposition which argues for other algorithms than the Three-Step one. Item-based strategies are a good candidate for this and will probably be more and more used as children grow older. However, the utterances produced by children aged 2;9 are not very complex yet and do not seem to require the use of the more complex strategies of level 3 and 4.

Experiment 3

The results obtained in the first two experiments do not prove that children really do something that is close to the Three-Step Algorithm. They only prove that the postulated mechanisms, while they do not include any innate specific linguistic mechanism, can generate the type of output produced by young children. In order to better assert the plausibility of the Three-Step Algorithm, it is necessary to find out whether predictions specific to this algorithm are obtained.

1. The characteristics of children’s errors should be compatible with the properties of the Three-Step Algorithm.
2. No ordering of the elements is specified in the Three-Step Algorithm, so it should be possible to find the same elements used in various orders. When children make word order errors, these should entail inversions of words uttered alone or blocs of words grouped around a word uttered in isolation, but not with words that are never uttered in isolation. As children get older, they gradually learn to copy the adult word order, so that word order errors should occur more often in young children.

Results and discussion—Question 1

Given that errors are indicated in the Manchester corpus, it is possible to make a typology of them and see if they match the properties of the Three-Step Algorithm. Two formats for error descriptions exist. If some morphosyntactic or grammatical element is clearly missing, the format will use the “0” notation. This means that the missing element is transcribed, but with a 0 sign before it. For example:

*CHI: what 0is [*] this

*CHI: Warren-0’s [*] hair

In all other cases, the error is only signaled by a “[*]” sign. For example:

*CHI: me [*] play

*CHI: foot-s [*]

There are 12,216 child errors tagged as such in the Manchester corpus. Of these, 9,063 correspond to missing elements. These are 35 different types (see Appendix 1) of missing elements in the corpus, out of a total of 9,253 tokens—there may be more than one missing element per utterance. Examples of the ten most common types of error are, in order of frequency:

*CHI: baby 0is [*] stuck

*CHI: I 0am [*] write-ing

*CHI: they 0have [*] gone

*CHI: all 0are [*] eat-ing table

*CHI: it 0has [*] gone

*CHI: Daddy-0’s [*] thumb

*CHI: Andy want-0es [*] it

*CHI: there two penguin-0s [*]

*CHI: what-’is he do-0ing [*]

*CHI: I bang-0ed [*] it

The examples of each type of error have been randomly chosen from the recordings of the youngest children. In these examples, all words produced are also used by the children in isolation or as a group in a single utterance. In particular, this is the case for I bang, what-’is, and he do. The only exception is the I in the second utterance I writing; all the other utterances could have been produced by the Three-Step Algorithm. I, however, is not found in isolation. As the problem raised by I is also raised by a in the utterances where there is no obvious grammatical element missing, we will first discuss this specific point. The 3,153 errors that do not correspond to missing grammatical elements are more diverse than the errors involving missing elements. One common type is the use of determiner a with a plural noun or other inappropriate word, for example, a car-s, a flower-s, a apple, a people, a same. There 121 such errors.

The problem raised by I and a may give rise to two different interpretations. The first is that it is by no means certain that I and a are never used in isolation. In fact, they are, as I and a may happen to be the last element of an incomplete sentence. This occurred 65 times for I (46 times in isolation) and 227 times for a (29 times in isolation). It may be that what is considered to be an incomplete sentence by an adult is not so from the child’s point of view. If this is the case, then the Three-Step Algorithm should produce utterances such as I writing and a pants. The second possibility is to interpret this phenomenon as the emergence of mechanisms other than the Three-Step Algorithm. I + x and a + x are clearly very productive patterns in young children’s language. There are 1,316 utterances of the type I + x (with 216 different values for x) and 2,030 occurrences of the type a + x (with 552 different values for x). This represents 7.4% of all two-word utterances. These two patterns could be the first slot-and-frame structures used by children (Lieven et al., 1997; Pine & Lieven, 1997). The Three-Step Algorithm is obviously not the only mechanism used to produce language, and this would be an example of another mechanism gradually coming into play.

Another type of error unaccounted for by the Three-Step Algorithm is the overgeneralization of morpholexical constructions. A common example (166 occurrences) is an incorrect use of the plural marker s, for example milk-s, foot-s, smoke-s. This may be explained by the same slot-and-frame mechanism as above.

Most other errors are perfectly accounted for by the Three-Step Algorithm. One of the most common ones is the use of words such me or my as obligatory subject pronouns or existence verbs, for example, me play, me sit down, me egg, me tea, my make a tower, my do that. This occurs 615 times for me and 210 times for my. This is perfectly accounted for by the Three-Step Algorithm, as me and my are both produced in isolation by the children. Other examples of words commonly used to build utterances in a similar fashion are no (167 occurrences) and mine (47 occurrences) in constructions such as no fit, no away, mine doggie, mine water.

For the purposes of this article, it is unnecessary to go through all possible types of errors. After deducting all the errors already accounted for, there are still 429 different words preceding the errors (as marked in the Manchester corpus) and 393 different words following them. It is interesting to note that many errors look as if they result from the concatenation of two elements. For example, mine [*] cover, do it [*] the animal, draw another one [*] fish, or I want [*] need my sock-s on. It is possible to check automatically whether this is true or merely an impression. If those errors come from the simple concatenation of two strings of words, then the pair of words located just at the error (in the examples above, this corresponds to the pairs one fish and want need) would be a creation of the child's, and thus less likely to be found in adult utterances. The other pairs of words (in the two examples above, this means the pairs draw another, another one, I want, need my, my sock-s and sock-s on), because they belong to strings of words extracted from children's input, should be more liable to be found in adults' utterances. All these pairs have been extracted; 1,384 different pairs located at the errors were found, and 3,584 pairs located elsewhere. Of the pairs located at the errors, 674 are found in adult utterances (49%); of those located elsewhere, 2,475 are found in adult utterances (69%). This result confirms the plausibility of children's following the Three-Step Algorithm. However, the same result would also be obtained if children were using an incomplete grammar, so this result is coherent with the principles of the Three-Step Algorithm but does not prove that it is being used.

Results and discussion—Question 2

There are three different types of word inversions. The first is the inversion that occurs between isolated words or words grouped around a word used in isolation (lists L1 and L2 discussed above). For example, baby stuck vs. stuck baby, where both words belong to list L1, or that one there vs. there that one, where that one belongs to list L2 and there to list L1. The second is the inversion that occurs within words grouped around a word used in isolation (within one element of the list L2 above). For example, is it a baby vs. it is a baby, where both groups belong to list L2. The third type occurs anywhere and between any type of words. There is no typical example for this type of inversion, it only means that the words involved do not belong to list L1 or L2.

There are two different ways of computing the proportion of inversions, depending on the number chosen as a reference. The first possible reference number is the number of possible word inversions. The second is the number of possible word inversions, but taking into account only the pairs of words that appear at least twice, in whatever order.

The percentages of inversion can also be computed in two different ways: either for the corpus as a whole or transcript by transcript. The first option is probably felt to be more "fair", because there is no reason why an inversion should occur during one particular recording and not during others. But the second option is the only way to show that the same child is using the same words in a variable order, within a period short enough to judge that the variability is warranted by the child's grammatical knowledge . All values are computed in types.

include Figure 5 and Table 1 about here

All results are presented in Table 1. The results per transcript, for pairs of words that appear at least twice and sorted by age are presented in Figure 5. The percentage of inversion is much larger for the corpus considered as a whole than for single transcripts, which is not surprising as there are many more circumstances where semantics may lead to word reversal in a large transcript. Also, pairs in any order are more frequent when one only considers frequent pairs of words.

When percentages are computed transcript by transcript, it becomes possible to test the significance of the difference between the types of inversions. A t-test computed across children in the frequent pairs case shows that the difference between type 1 and type 2 is highly significant, t(11) = 5.67, $p < 0.00001$, as is the difference between type 2 and type 3, t(11) = 9.40, $p < 0.000001$. The difference between type 1 and type 3 is not significant, t(11) = 1.55, p = 0.07. Results computed across age give exactly the same pattern of results, as do statistics computed using percentages for all pairs of words. In this case, percentages are lower, but the relative ordering of results is the same.

Examples of variable order between groups of words are presented in Appendix 2 for the youngest children, where one can see that there are two main categories of variable orders. The transcript "Warren 01A" gives a perfect representation of these two categories. The first type is Controller gone vs. gone Controller. Here is a pure inversion of two apparently equivalent elements. The second type is there brick there which is a variable order in itself. In this case, it would seem that the two elements produced, there and brick, have no order and that the child repeats them to emphasize what she wants to say.

Inversions within a group of words from list L2 are very unusual, and only two cases seem to occur. The first is the very common pattern of pronoun and auxiliary inversion in questions (I can vs. can I, they are vs. are they). The second includes repetitions (got a got a rabbit, in a in a minute) and coding or segmentation errors. It thus appears that inversions between a function word and a content word are impossible other than in questions.

On average, the percentage of variable word order occurrences within a transcript is not very high. If the order of content word constructions were really free, the percentage of constructions in any order should have reached the 100% level. However, for the pairs that appear twice, there is a non-negligible percentage of words that appear in two different orders, so it is difficult to decide whether word order is chosen on a morphological basis—which would imply a strict respect of word order—or on a semantic one—which would allow more laxity in word order. In any case, it is true that word order is a strong syntactic feature of the English language and that it will appear at some point during the development of syntactic structures.

The most important result here is that inversions are much more frequent between words grouped around a content word than between a functional word –that is a word that never appears in isolation– and a content word. This shows (1) that it is when semantics has the highest content that the word order is the most free; (2) that word order is only meaningful—at first—in the case of words which tend to occur together and are very frequent in the children's input. It is very difficult to discover a word order rule applying to two content words (such as nouns and verbs), unless either the category of these words is known or the words are very frequent. If young children follow word order more in morphological situations—the repetitive ones—than in semantic situations—which are less repetitive—this could just mean than they have not yet learned the syntactic categories and are still learning language on an example-driven basis.

As for the changes in the proportion of the various word orders through time, it does not seem that our hypothesis is confirmed. As can be seen in Figure 5, the number of variable word order elements is stable with age and only a slow decrease in variability is apparent. This would mean that word order inversions are not a developmental feature, but an intrinsic pattern of the English language, and that young children are as sensitive to word order as older children are. It could also mean that the basic characteristic of the Three-Step Algorithm, word order leeway—with the exception of the morphosyntactic derivation of words—holds good until at least age three.

Experiment 4

The analysis of errors offers a better understanding of the characteristics and limits of the Three-Step Algorithm. Another 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.

Results

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. If all children's words are included in L1, results for Step 1 and 2 are, respectively, 41% (SD = 5.1) and 41% (SD = 25.8). 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). If one assumes a knowledge of all syntactic categories, results for Step 1 and 2 are, respectively, 100% (SD = 0.0) and 83% (SD = 8.4).

include Figure 6 and Figure 7 about here

The results for Step 3 are presented in Figure 6 (for exact reconstruction) and Figure 7 (for reconstruction covering). In each of these figures, four results are presented: one assuming no category knowledge; one assuming the knowledge of all content words uttered by the child in isolation; one assuming knowledge of the three categories Proper Noun, Common Noun and Verb; and one assuming knowledge of all syntactic categories. 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 to 98% (SD = 5.9) and 99% (SD = 2.7) for "all content words" knowledge, 72% (SD = 11.9) and 93% (SD = 4.0) for "Noun and Verb" knowledge, and 71% (SD = 12.7) and 92% (SD = 3.1) for "all syntactic categories" knowledge.

Discussion

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 the Three-Step Algorithm by the child.

General discussion

A generative algorithm –the Three-Step– rooted in simple cognitive mechanisms and involving no complex knowledge of grammar has been tested. The first part of this algorithm (Step 1 and 2) allows to specify what type of building blocks are used by children. The creation of these building blocks involves no syntactic knowledge because they are directly extracted from children's input. The blocks are then combined by the simple strategy that consists of catenating them (Step 3). This catenation is controlled by semantic and pragmatic knowledge – no more and no less than in adults uttering more than one phrase or sentence in sequence. It was found that 80 to 85% of all the utterances produced by a child can be fully reconstructed this way. A second experiment was done to try to compensate for the fact that the corpus used in this study does not equal the size of what a child really hears and produces in a year. This provided an increase in the percentage of utterances fully reconstructed, which went up to 90%. As it is difficult, with a corpus limited in size, to know for sure whether a 100% reconstruction is possible, we tried other means to test the algorithm. We investigated children's errors to see whether they confirmed that the catenation in Step 3 is not controlled by syntactic knowledge and found that children do not order their building blocks as randomly as could be expected, even though they do make inversion errors. Finally, a test using a corpus from an older child showed that the Three-Step algorithm may still be at work in older children.

These results are far from trivial for two reasons. First, children have to produce a sufficient number of single-content-word utterances for the algorithm to work. This is what we tried to demonstrate in spite of the fact that the corpus we are working with does not cover the full production of a child since birth. If children did not produce enough variety of single-content-word utterances, they would have to split many complex utterances to obtain more basic linguistic building blocks. This splitting process is highly complex, as it would lead to many over-segmentation errors if unchecked, which type of errors is not produced by children. It sounds possible for a child to extract a content word (Cutler, 1995), or a content word and some the functional elements associated with it, out of an adult sentence, but this process provides only coarse results (otherwise children would learn to use the exact words of their language), and is efficient only when there is a good match between morphological and semantic cues. A precocious fine-grained segmentation seems out of order before children have enough material to work with, which material is what the Three-Step Algorithm provides.

Second and conversely, there is no reason to think that children's utterances with more than one content word are made up of shorter single-content-word utterances because this implies that single-content-word utterances present the same structure as that of multiple-content-word utterances, which is not always true in adult language. One would also suppose that children use only the most simple single-content-word utterances –isolated words only, for example– to build more complex utterances, in which case a dissociation would occur between the single-content-word utterances used to build more complex utterances and those not used so, which is not the case. This can be checked by comparing the ratio between the total number of words and the number of content words in utterances (1) of only one content word, (2) of two content words and (3) of three content words. If the single-content-word utterances used for building multiple-content-word utterances are more simple than the single-content-word utterances produced alone, then the ratio between the number of words in an utterance and the number of content words in the same utterance should differ according to the number of content words in the utterance. Results are presented in Figure 8 for utterance of one, two and three content words. It appears that there are no significant difference between the three cases and therefore that children may indeed be using no more and no less than their own single-content-word utterances when building more complex utterances. This has important repercussions on theories that rely on children's processing limitations to explain their behavior (Valian, Hoeffner, & Aubry, 1996). If children were omitting functional words in single-content-word utterances because of processing limitations, one would expect that they would omit even more functional words in utterances with more than one content word. That this is not the case is surprising because there are reasons other than grammatical ones to argue for the existence of processing limitations in children. Another explanation would be that single-content-word utterances are considered as unanalyzed wholes and thus do not tax the cognitive system when used to produce further utterances. This would then argue against the existence of syntactic processes in single-content-word utterances, and therefore against the precocious existence of syntactic processes in children. If strong nativist hypotheses are not completely ruled out, they should not rely on a processing limitation hypothesis.

include Figure 8 about here

It is very important for any theory which relies strongly on learning algorithms to show that it is indeed possible for children to learn to produce and control complex language very quickly, by using a simple learning procedure and without having to build a complex system of rules. If not, these theories would also have to provide the means to instantaneous acquisition of language structure, which would be quite difficult if not impossible. It has been shown above that it is possible to find a system that explains the initial language production of children without any innate language-specific features, which does not mean that the natural development of language will not induce properties such as language modularity and abstract syntactic competence later on. This system would be fully operative by the age of two and remain for some time the most basic and productive language mechanism. It might also remain active for adults because it is an efficient way of producing language when the cognitive load is heavy or speed is crucial (Peters, 1983, p. 80, pp. 105-106). However, the Three-Step Algorithm does not account for all language acquisition processes before the age of three. First, its rather crude mechanisms would produce many aberrant utterances on their own, if they were not regulated by other mechanisms. Second, some utterances clearly cannot be produced by the Three-Step Algorithm.

The first of those 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 to master 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 undergo representational redescription (Karmiloff-Smith, 1992). Their semantic values allow one to make semantic combinations.

Utterances that are produced at age two and that cannot be produced by the Three-Step Algorithm are utterances such as <u>a pants</u> because <u>a</u> is never found in isolation. This construction calls for more complex mechanisms such as the productive use of bound patterns. If this strategy is used in production, it is most probable that it can be used in comprehension, which increases the efficiency of the child's lexical acquisition. The existence of more complex mechanisms was hypothesized in the introduction (see Figure 1) and has been proposed elsewhere in the literature. For example, MacWhinney (1982) described a set of six strategies to account for the development of morphosyntax: rote, analogy, feature-based patterns, bound patterns, free patterns, and class-bound patterns. He presented many arguments showing the importance of rote in the production of young children but could not evaluate exactly how much rote children might be using, because this would require some procedure similar to the one presented in this article. Two other strategies described by MacWhinney, predispositions and bound patterns, are relevant to the current study. Predispositions correspond to the fact that children sometimes appear to combine words using various semantic or pragmatic strategies such as agency, salience, informativeness, etc. This is what children use to combine words or word groups in any order before they acquire linearization rules. Bound patterns are constructions like the pivot-constructions of Braine (1963; 1976), the patterns used by MacWhinney (1975) to reconstruct the production of Hungarian children, the pattern-filling slots of Lieven et al. (1997), Peters (1983), and Pine and Lieven (1997), or again the verb-island hypothesis of Tomasello (1992). All these patterns are of the <u>X + too</u>, <u>no + X</u>, <u>where + X</u>, etc., type. Their level of abstraction is 2. The predispositions strategy is the backbone of Step 3 in the Three-Step algorithm, although no limits are set as to what type of strategy the child uses. Bound patterns are not used productively in the Three-Step algorithm, but they account for the construction of the elements of the L2 List at Step 2. These elements are built in the same way as bound patterns are, but are used in comprehension to recognize which patterns are interesting, and not yet in production. In other words, children are not yet able to construct new elements following the bound patterns principle, but they are able to recognize the occurrence of such patterns when they come across them in their input. Other strategies that can provide access to levels of abstraction 3 or 4 are the discovery of syntactic categories using distributional analysis (Maratsos & Chalkley, 1980) and the induction of grammatical knowledge using non language-specific innate mechanisms (O'Grady, 1997). The use of all these learning strategies is all the more possible because the Three-Step algorithm provides children with a reservoir of rich morphological forms with multiples semantic associations. It is important for language development theories that children be demonstrably able to use language with some efficiency before having a full grammatical mastery, otherwise language development would never be explained.

Since Braine posited his pivot-constructions (Braine, 1963), many studies have tried to present the structure of the child's language using rules or principles simple enough to be acquired or readily explained by biological mechanisms. One of the major arguments against these approaches is that they do not describe mechanisms powerful enough to last beyond the very earliest stages of language acquisition. Our work shows that simple mechanisms can indeed handle the task of language acquisition much longer than is generally believed. Children do not begin to master the more complex generative rules governing noun and verb constructions until the age of at least four, as shown by the age of the subjects in studies such as Berko (1958), Pinker, Lebeaux, and Frost (1987), or Gropen, Pinker, Hollander, and Goldberg (1994) –see also Tomasello (2000). This casts a new light on the theories of language acquisition. If children acquire high-level grammatical rules at a later period of their development than is usually admitted in theories such as those expressed by Pinker (1984; 1987), Wexler and Culicover (1980), or Wexler (1982), then the structure of children's input—the couple "base phrase marker" plus "surface sentence" (Wexler, 1982)—will be more complex. The more complex these structures are, the lower the innate conditions on grammars. It would then be possible to progress from a simple rule-abiding system such as the Three-Step Algorithm to a more complex one.

It has always been emphasized that children learn syntax very quickly and with very few errors (see, for example, Pinker, 1994, pp. 269-273). This quick and impressive feat—"the child is a grammatical genius" (Pinker, 1994, p. 273)—is the prime reason for considering either that an incredibly powerful learning system is operating or that most of the knowledge is innate. Thus, Chomsky (1959) writes in the conclusion to his article: "The fact that all normal children acquire essentially comparable grammars of great complexity with remarkable rapidity suggests that human beings are somehow specially designed to do this, with data-handling or 'hypothesis-formulating' ability of unknown character and complexity." However, just because children's output appears to be of great complexity does not mean that it is complex from their point of view. If, as suggested by Peters (1983), children use long units when they learn to speak, their production may give the impression, from the linguist's point of view, of being better than it really is, and children's language could be built with much simpler rules than adults' language.

The current work suggests that some simple generative mechanisms can explain the explosive acquisition and 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 results of the current study could not have been obtained without using large data sets of children's language to test our hypotheses. This emphasizes the need for large corpora, of the magnitude of what a child produces and hears in a year, and the importance of computer simulations that go beyond what can be tested by hand alone. Work in this direction (see, for example, Brent, 1997; Elman, 1995; Ritchie & Bhatia, 1999; Rumelhart & McClelland, 1987), using real data as much as possible, will prove invaluable in finding the answer to the three questions from Hirsh-Pasek and Golinkoff (1996) quoted at the beginning of this paper.

References

Berko, J. (1958). The child's learning of English morphology. Word, 14, 150-177.

Braine, M. D. S. (1963). The ontogeny of English phrase structure: The first phase. Language, 39, 3-13.

Braine, M. D. S. (1976). Children's first word combinations. With Commentary by Melissa Bowerman. Monographs of the Society for Research in Child Development, 41.

Brent, M. (Ed.). (1997). Computational approaches to language acquisition. Cambridge, MA: MIT Press.

Brent, M., & Cartwright, T. (1997). Distributional regularity and phonotactic constraints are useful for segmentation. In M. Brent (Ed.), Computational approaches to language acquisition . Cambridge, MA: MIT Press.

Brown, R. W. (1973). A first language: The early stages. Cambridge, Mass.: Harvard University Press.

Chomsky, N. (1959). A review of verbal behavior, by B. F. Skinner. Language, 35, 26-58.

Cutler, A. (1995). Prosody and the word boundary problem. In J. L. Morgan & K. Demuth (Eds.), Signal to Syntax: Bootstrapping from speech to grammar in early acquisition . Mahwah, NJ: Lawrence Erlbaum Associates.

Elman, J. L. (1995). Language as a dynamical system. In R. F. Port & T. van Gelder (Eds.), Mind as Motion: Explorations in the Dynamics of Cognition (pp. 195-223). Cambridge, MA: MIT 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.

Gropen, J., Pinker, S., Hollander, M., & Goldberg, R. (1994). Affectedness and direct objects: The role of lexical semantics in the acquisition of verb argument structure. In P. Bloom (Ed.), Language Acquisition: Core Readings (pp. 285-328). Cambridge, MA: MIT Press.

Harris, R. (1990). On redefining linguistics. In H. G. Davis & T. J. Taylor (Eds.), Redefining linguistics . London: Routledge.

Hish-Pasek, K., & Golinkoff, R. M. (1996). The origins of grammar. Cambridge: MA: MIT Press.

Jusczyk, P. W., & Hohne, E. A. (1997). Infants' memory for spoken words. 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. J Child Lang, 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. (1982). Basic syntactic processes. In S. A. Kuczaj (Ed.), Language development . Hillsdale, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (1991). The CHILDES project - Computational tools for analyzing talk. Hillsdale, NJ: Lawrence Erlbaum Associates.

MacWhinney, B., & Snow, C. E. (1985). The child language data exchange system.

Journal of Child Language, 12, 271-296.

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.

Marcus, G. F., Vijayan, S., Bandi Rao, G. F., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. Science, 283, 77-9.

Miller, J., & Weinert, R. (1998). Spontaneous Spoken Language. Oxford: Clarendon Press.

O'Grady, W. (1997). Syntactic development: The University of Chicago Press.

Peters, A. M. (1983). The units of language acquisition. New York, NY: Cambridge University Press.

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

Pine, J. M., & Lieven, E. V. M. (1997). Slot and frame patterns and the development of the determiner category. Applied Psycholinguistics, 18(2), 123-138.

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

Pinker, S. (1987). The bootstrapping problems in language acquisition. In B. MacWhinney (Ed.), Mechanisms of language acquisition . New York, NY.: Springer-Verlag.

Pinker, S. (1994). The language instinct. New York: William Morrow.

Pinker, S., Lebeaux, D. S., & Frost, L. A. (1987). Productivity and constraints in the acquisition of the passive. Cognition, 26, 195-267.

Radford, A. (1990). Syntactic theory and the acquisition of english syntax. Oxford: Blackwell.

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.

Rumelhart, D. E., & McClelland, J. L. (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processing? In B. MacWhinney (Ed.), Mechanisms of language acquisition . New York, NY.: Springer-Verlag.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. Science, 274(5294), 1926-8.

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.

Schütze, M. (1997). Ambiguity resolution in language learning. Stanford: CSLI Publications.

Shanon, B. (1993). The representational and the presentational. London: Harvester Wheatsheaf.

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. (1992). First verbs: A case study of early grammatical development. Cambridge: Cambridge University Press.

Tomasello, M. (1998). The new psychology of language: Cognitive and functional approaches. Mahwah: NJ: Lawrence Erlbaum Associates.

Tomasello, M. (2000). Do young children have adult syntactic competence? Cognition, 74, 209-253.

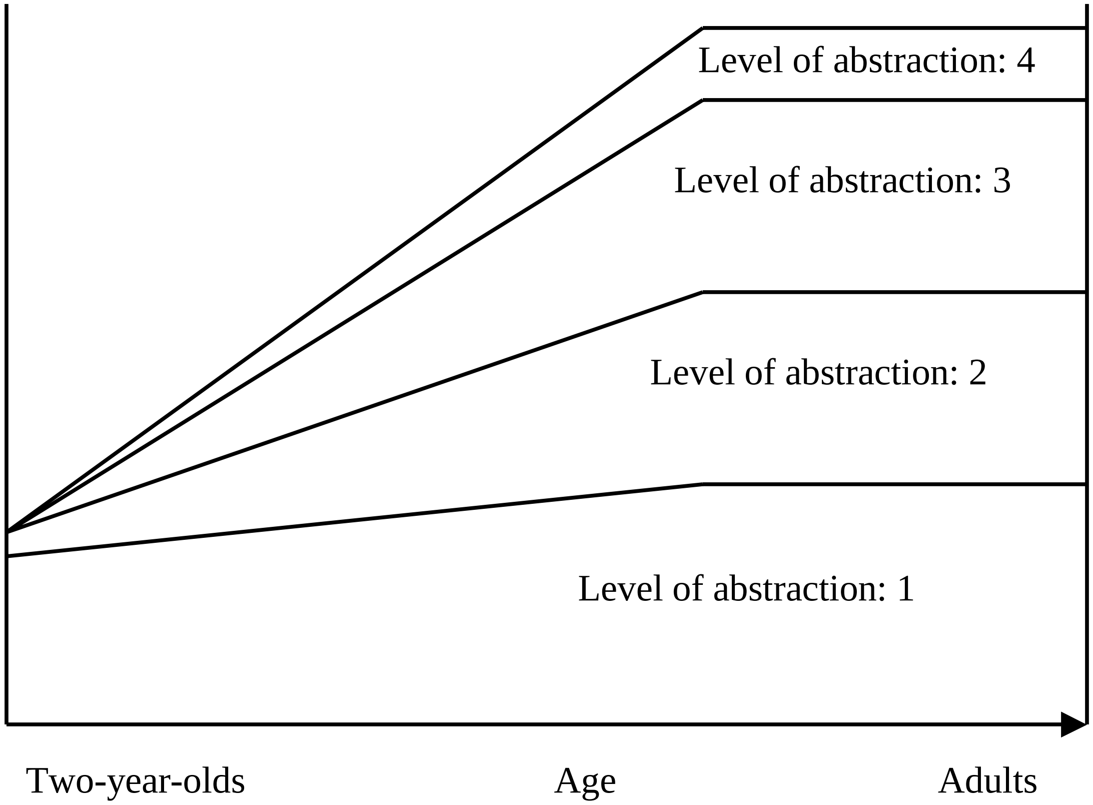
Valian, V., Hoeffner, J., & Aubry, S. (1996). Young children's imitation of sentence subjects: Evidence of processing limitations. Developmental Psychology, Developmental-Psychology.

Van Valin, R. D., & LaPolla, R. J. (1997). Syntax: Structure, meaning and function. Cambridge: CUP.

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.

Wexler, K., & Culicover, P. W. (1980). Formal principles of language acquisition. Cambridge, MA: MIT Press.

Figure 1: Increase of the number of levels of abstraction and of complexity with age



Note: To each level of abstraction corresponds different linguistic processes.

Figure 1: Percentage of utterances exactly reconstructed

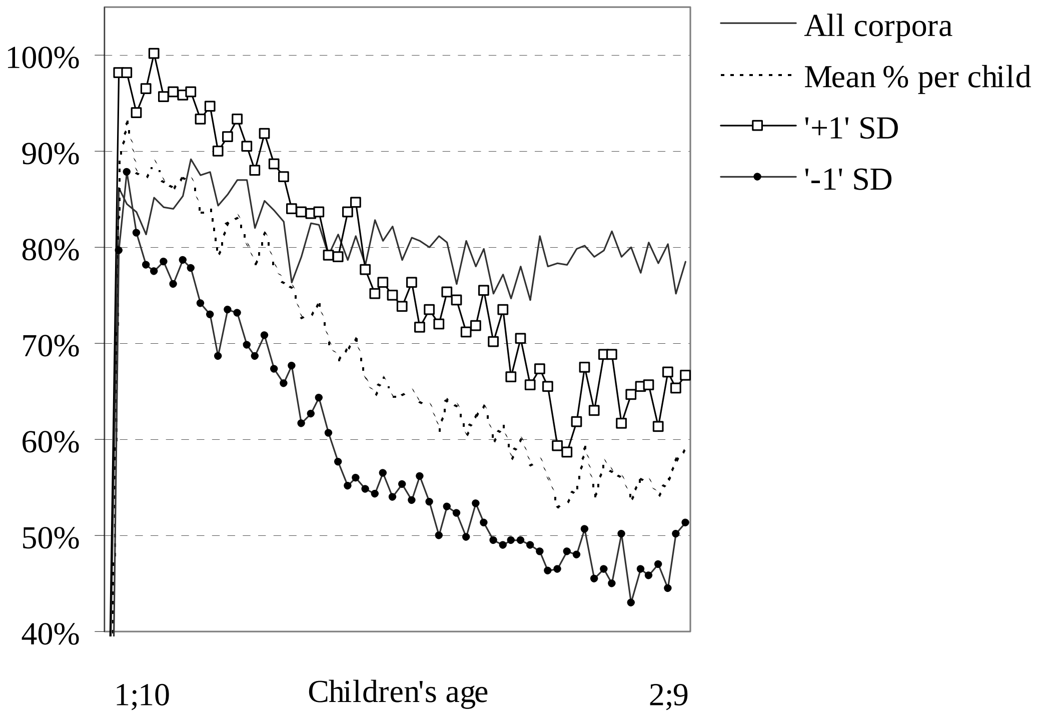


Figure 2: Percentage of reconstruction covering in all utterances

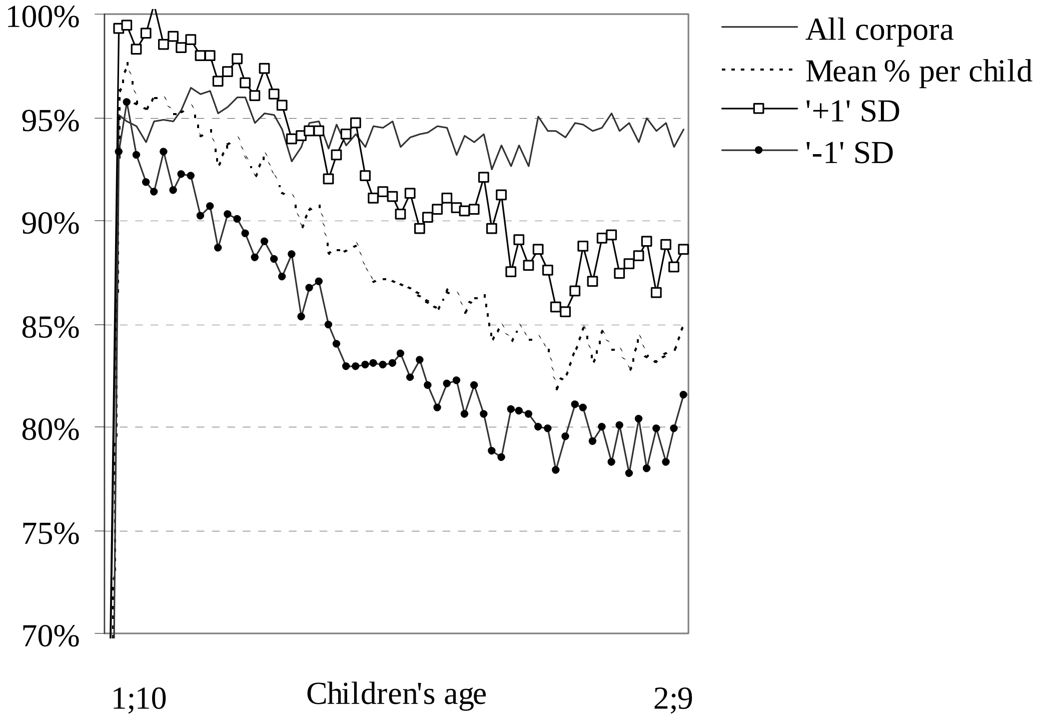


Figure 3: Percentage of utterances exactly reconstructed, depending on the degree of knowledge of vocabulary and syntactic categories

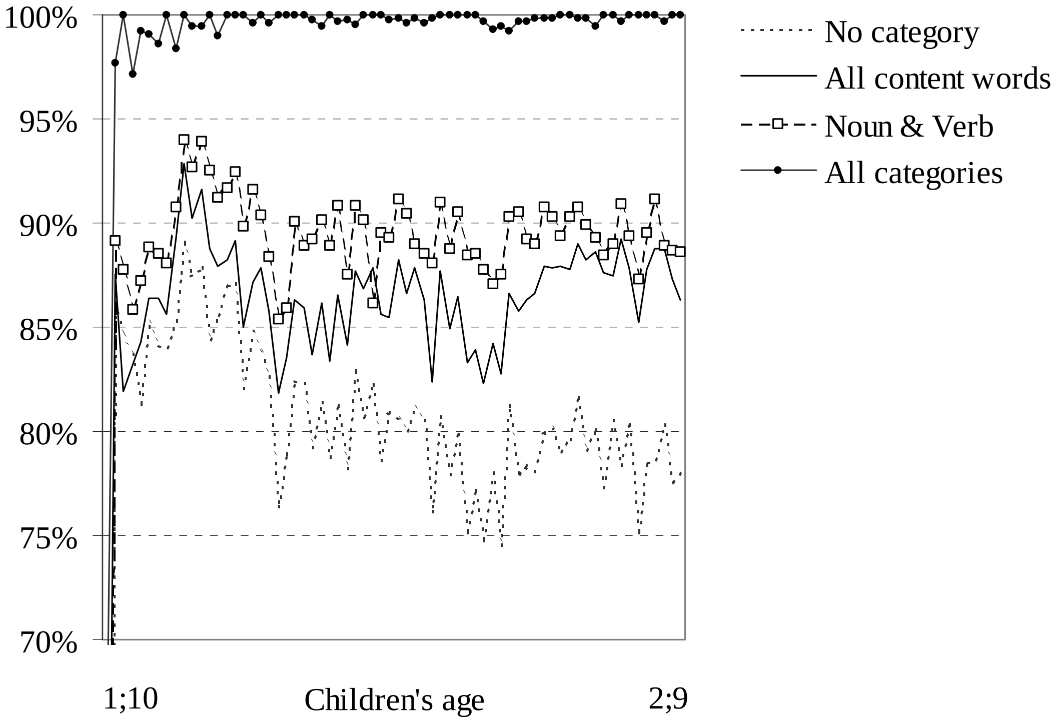


Figure 4: Percentage of reconstruction covering in all utterances, depending on the degree of knowledge of vocabulary and syntactic categories

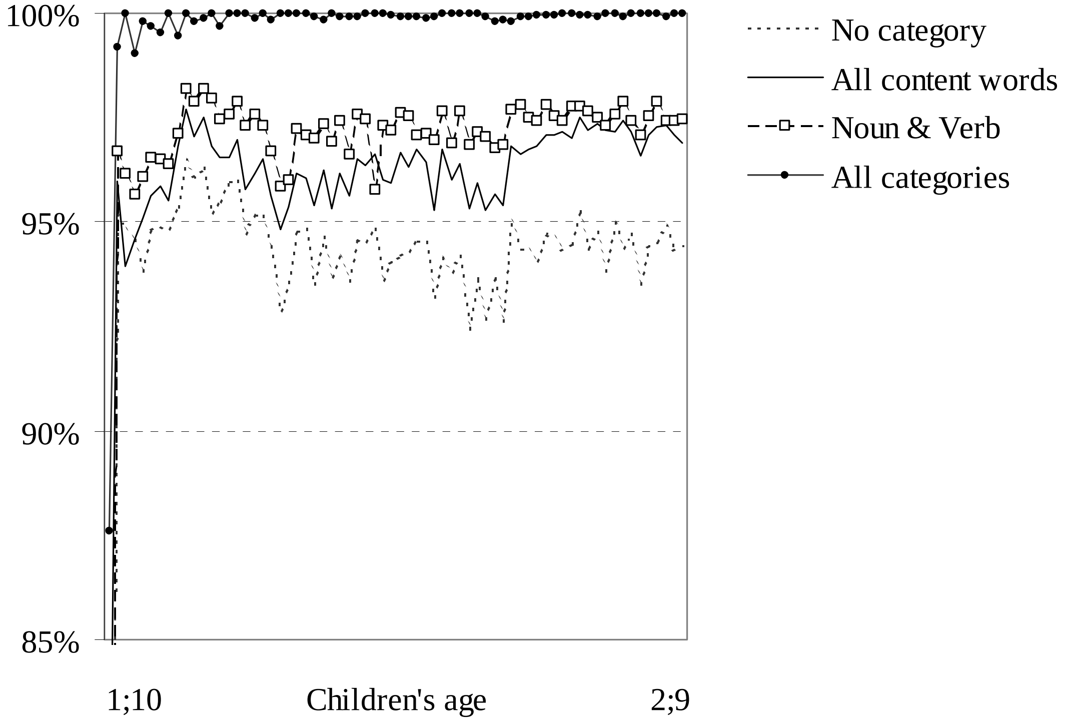


Figure 5: Percentages of inversions for pairs found at least twice in the corpus, computed per transcript and per child

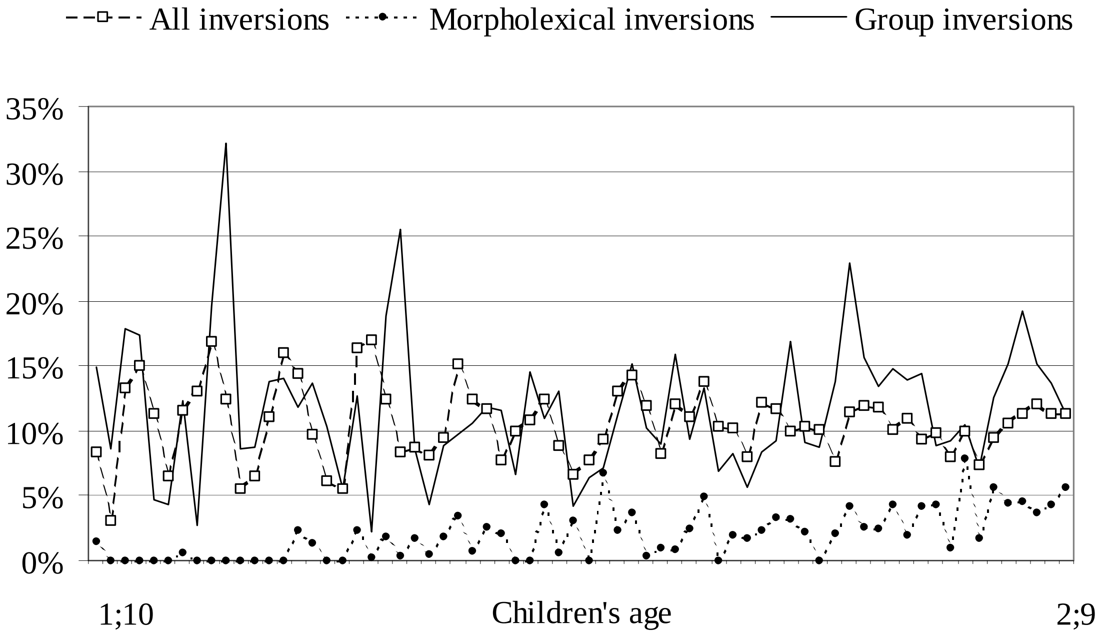


Figure 6: Percentage of utterances in the Sarah corpus exactly reconstructed, depending on the degree of knowledge of vocabulary and syntactic categories

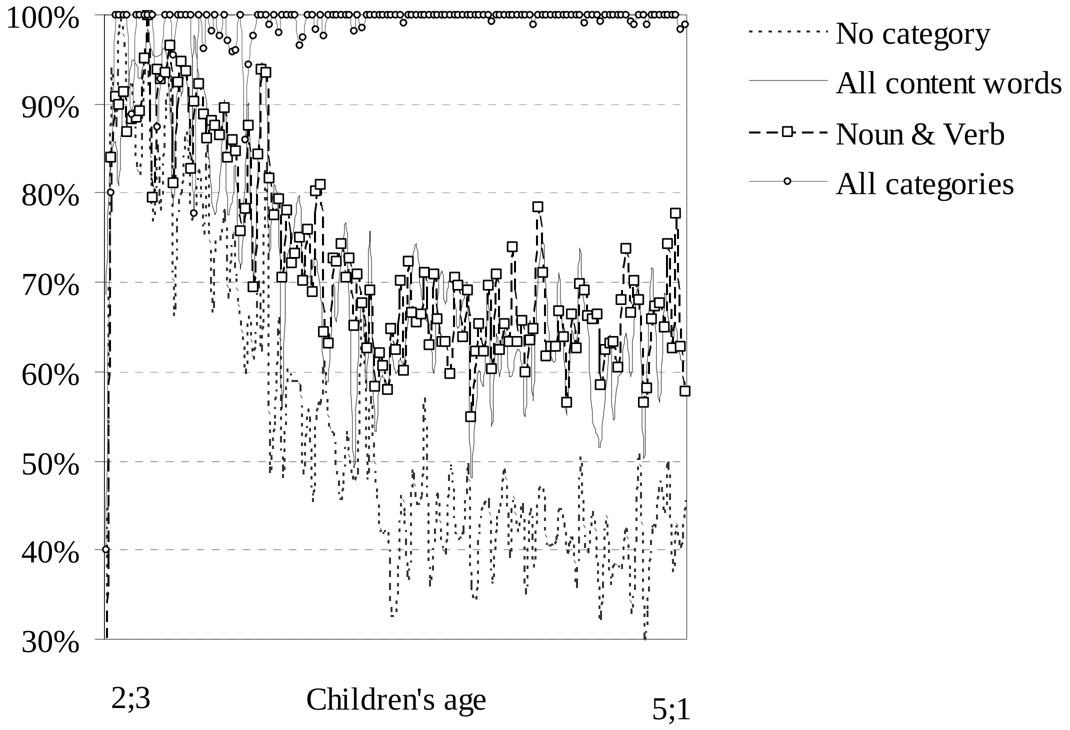


Figure 7: Percentage of reconstruction covering in all utterances in the Sarah corpus, depending on the degree of knowledge of vocabulary and syntactic categories

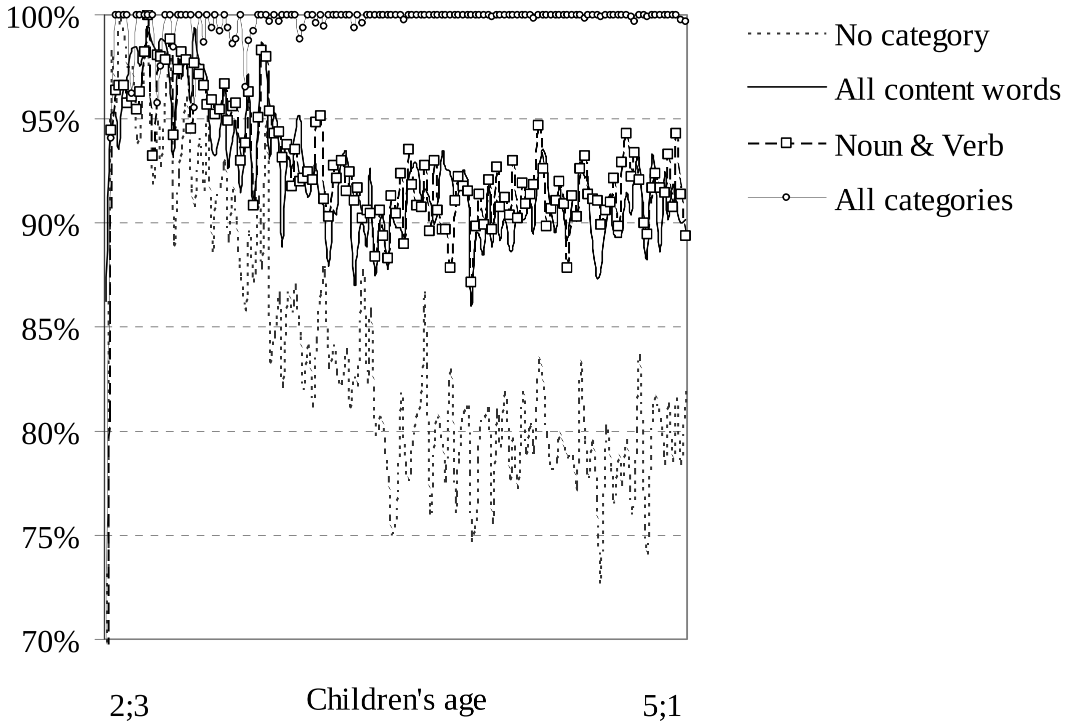


Figure 8: Ratio between the total number of words and the number of content words

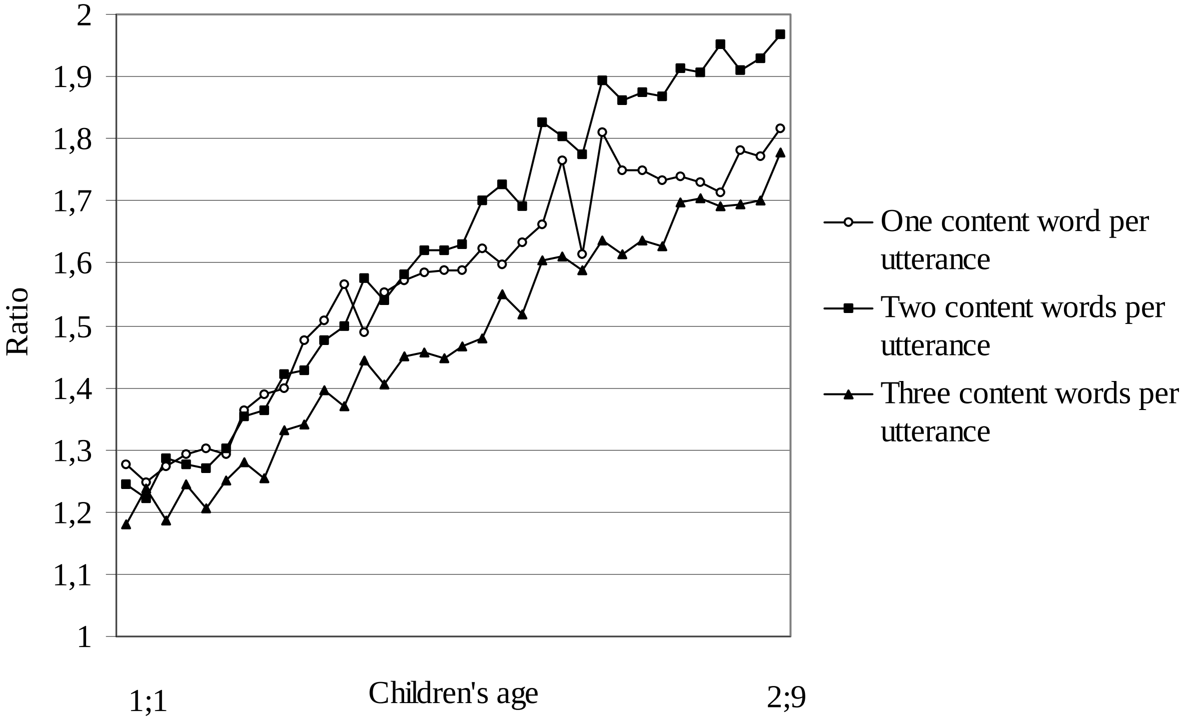


Table 1: Percentages of inversions involving pairs of words

	All inversions	Morpholexical inversions	Group inversions
For the corpus as a whole			
any pairs	8.35	3.68	7.76
pairs occurring twice	23.60	5.91	22.72
Transcript by transcript			
any pairs	1.75 (1.43)	0.52 (1.34)	1.59 (2.03)
pairs occurring twice	10.78 (10.87)	2.05 (5.58)	12.54 (17.19)

Note: Standard deviations are given in parentheses

Appendix 1: List of missing grammatical elements in the Manchester corpus

3,543	0is	Verb to be, contractions included
1,384	0am	
1,351	0have	
784	0are	
643	0has	
491	0’s	Possessive
437	0es	Verb third person singular
160	0s	Plural
121	0ing	
112	0ed	
47	0do	
42	0does	
33	0was	
21	0had	
13	0to	
12	0what	
10	0did	
8	0were	
6	0it	
6	0a	
5	0where	
4	0of	
3	0the	
2	0put	
2	0on	
2	0in	
2	0for	
2	0would	
1	0will	
1	0us	
1	0know	
1	0get	
1	0as	
1	0and	
1	0I	
9,253	Total	

Appendix 2: Examples of words used in any order within the same recording (children's age ranging from 1;10 to 2;2)

Anne	01B	baby stuck
Anne	01B	stuck baby
John	01A	bang bang snail
John	01A	snail bang bang bang
John	01B	go swim-ing
John	01B	swim-ing go
Warren	01A	Controller gone
Warren	01A	gone Controller
Warren	01A	there brick there
Aran	02A	Daddy truck
Aran	02A	truck Daddy
Aran	02B	toy oh toy there
Carl	02A	birdie there no
Carl	02A	no there sheep
Dominic	02b	gone train
Dominic	02b	train gone
Joel	02A	no Mummy
Joel	02A	no Mummy no
John	02B	this it
John	02B	do it this dolly
Ruth	02B	baba in there
Ruth	02B	in there baba
Warren	02A	there red there
Warren	02B	broken it
Warren	02B	it broken
Warren	02B	Warren broken it
Carl	03A	elephant on Thomas
Carl	03A	there cow on elephant
Carl	03A	elephant on train
Carl	03A	hat on man
Carl	03A	man on horse
Carl	03A	man on train
Carl	03A	man on a pink one

John	03B	do sock-s
John	03B	want sock-s do
Liz	03A	that mine that
Anne	04A	fit there down here
Anne	04A	no that fit down there
Aran	04A	pipe got burst
Aran	04A	pipe got wet
Aran	04A	look got pipe burst
Aran	04A	a man there
Aran	04A	there a man
Aran	04A	me sit there
Aran	04A	sit there me
Aran	04A	and me sit there
Aran	04A	it put sand
Aran	04A	put it that
Carl	04B	car fish
Carl	04B	fish car
Carl	04B	it dog it eat
Warren	04A	that one there
Warren	04A	there that one
Aran	05A	like that
Aran	05A	that like that
Aran	05A	that one
Aran	05A	Daddy get another one that door
Aran	05A	get get Daddy
Carl	05A	Percy no
Carl	05A	no Percy
Carl	05A	six seven six
Nic	05A	Mummy no
Nic	05A	no Mummy
Ruth	05A	baba eye
Ruth	05A	eye baba
Ruth	05B	baba on there
Ruth	05B	on there baba

Carl	03A	man on a train
Carl	03A	man in there man
Carl	03B	ooh whee
Carl	03B	whee ooh

Ruth	05B	Mama baba on there
Warren	05A	Mummy look
Warren	05A	look Mummy
Warren	05A	a sleep Mummy
Warren	05A	Mummy sleep Mummy