How children build their morphosyntax: The case of French

Christophe Parisse, Marie-Thérèse Le Normand

This is the penultimate version of an article to appear in Journal of Child Language (2000).

RUNNING head: How children build their morphosyntax

Title: How children build their morphosyntax: The case of French

Christophe PARISSE and Marie-Thérèse LE NORMAND

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

Acknowledgments:
This work was supported by a grant from INSERM, France: 'Contrat de Recherche Inserm (4U009B)'. Grateful thanks to Hrafnhildur Ragnarsdóttir, Thomas Clegg and Anne Reymond for proof-reading this text.

Address for correspondence:
Christophe Parisse
Laboratoire de neuropsychologie de l'enfant
Bâtiment Pharmacie, 3ème étage,
Hôpital de la Salpêtrière
47 Boulevard de l'Hôpital
75651 PARIS CEDEX 13
FRANCE

E:Mail: parisse@ext.jussieu.fr

How children build their morphosyntax: The case of French

Abstract

Early morphosyntax is very rich and uniform in French-speaking young children. The present study aims to give a thorough analysis of the morphosyntax produced at the outset of multi-word speech, with a classification of free language produced at 2;0 by 27 French speaking children. The corpus was fully tagged by an automatic part-of-speech tagger. A classification performed with words taken in isolation shows a clear difference between the categories used in single-word utterances and those used in multi-word utterances. A classification performed with word sequences reveals surprisingly adult-like sequences of syntactic categories and words; the non-adult combinations are few in a French child's language.

The very successful use of the tagger demonstrates the morphosyntactic coherence of the child's speech. When compared with adult language, the quantitative results, and more precisely the data concerning regularity and error types contribute to the documentation of all the specificities of the emerging morphosyntax in normally developing French children.

How children build their morphosyntax: The case of French

Introduction

Many of the studies of young children's acquisition of syntax are based on naturalistic production data. When these data are compared with adult language which is considered to be the children's goal, the standard adult reference is usually a powerful and complete syntactic framework, whether generative, cognitive or lexical-functional. Few studies have tried to use the same kind of naturalistic data as a reference to compare the child's language with. This has been done mainly in studies of imitation, children's errors or negative evidence, and in computational simulations of language acquisition. What has not been done is to try to find similarities and differences between child and adult language with the same tool, the same standpoint and the same type of data. This would allow a quantitative evaluation of how much children really create when they are learning language and how much they reproduce or copy. What looks similar between child and adult is not necessarily copied by the former from the latter, but could also arise from previously-acquired language structures or from the necessities of the situation; in the same ways one adult's language is similar to another's. However, the knowledge of what is different in quality and in quantity between child and adult is necessary to assess and fine-tune language acquisition theories.

The current study is devoted to the beginnings of morphosyntax in young French children, comparing it with the morphosyntax of naturalistic speech by French adults. One of the aims of this study is to start out by limiting definitions of syntax. In order to achieve this, the same tools are used to analyse the productions of children and adults, and automatic comparisons on whole corpora are performed. The goal of the first section of the study is to compare a lexical classification of child language with that of adults. We find that children and adults use the same set of syntactic categories, as classically defined by French grammar. Where children differ from adults is in the distribution of categories, which differs for single-word utterances but not for multi-word utterances. The second section compares the distribution of syntactic categories and words, in pairs or triplets, between children and adults. This yields important information about which types of structures used by children are or are not adult-like and the exact percentage of each. This knowledge is important for building a child's developmental syntax and for measuring the relative importance and influence of the sub-parts of this syntax. The third section will fine-tune the previous analyses with a focus on content words because of their salient characteristics in children's speech.

Seminal studies of morphosyntax

The child's first word combinations have been studied along three main lines: distributional analysis, universal grammar and semantic approaches. These axes can interact with one another as in the case of Pinker (1984), where all three come into play. A description of the various historical works along each of these three axes can be found in Ingram (1989). But distributional analysis and how it evolved from the original work of Braine (1963) until today, is the most relevant for this paper.

Braine (1963) developed a theoretical description of the grammatical structure of early multi-word utterances, the 'pivot grammar', that influenced many of the works that followed and which remains a useful approach. He used a corpus of two hundred utterances from three children for his work. His theory is based on two-word utterances and uses a distributional analysis principle according to which children select certain words according to frequency characteristics and their own phonetic capacities. These 'pivot' words are few and have a fixed position: they always appear before or after a number of words defined as belonging to the 'open' class of words. As the membership of the pivot class is a function of the interactions of a child with his/her linguistic environment, its elements are specific to each

child. The first word pairs produced by the child are of the pivot-class/open-class type. Only later will word pairs made exclusively of open-class words appear. The function of the pivot class is to enable the child to acquire new classes of words. At the onset, the child might simply put any open class word with any pivot word. After a while, the child will recognise that certain open class words only occur with certain pivots. As the pivot-class words are exclusively used to acquire new word classes, the association of two pivot words is held to be unjustified and so non-existent. The criticism of pivot grammar bears mostly on its inability to describe anything but the first word combinations, and its inability to be extended during further language acquisition until complete adult grammar is reached. But it provides a basis for distributional analysis which is still frequently used, along with a tentative but complete theoretical description of the first steps in language learning.

The next major advance in distributional analysis is the work of Maratsos and Chalkley (1980) and Maratsos (1982). They proposed an algorithmic method using distributional analysis for the construction of grammatical categories by the child. Maratsos never rejected, and in fact used, the important influence of semantics in the course of child language acquisition. He described a succession of steps that makes it possible to discover grammatical categories from the regularities of context use. Maratsos (1982: 265) said himself that his studies were still incomplete: '... considerable empirical and theoretical analysis is required before we can be said to have any good idea of plausible complete accounts of formal category formation.' Response to this work came from Pinker (1984) and later Radford (1990). Pinker presented two arguments against the semantic-distributional model of Maratsos and Chalkey: learnability (the problem of negative evidence) and efficiency (the number of possible patterns is too enormous for the model to be efficient). The problem of negative evidence is a complex one, as can be seen in Post (1994) or Saxton (1997), and is still open to controversy. But the responses of Pinker and Radford to the model of Maratsos also raised a problem of efficiency, though in different ways: Pinker (1984) lacked detailed analyses of child data (see Ingram, 1989: 330), and Radford (1990) did not give any numerical assessment of the samples presented. As it is difficult to evaluate how distributional analysis takes place without precise numerical accounts of the evidence available to children, it is not completely satisfactory to propose a model and validate it by presenting some examples without an exact quantitative assessment. This question of quantitative analysis of child language is addressed below, with the use of morphologically labelled (tagged) corpora.

Recent work on first word combinations

More recent works have tried to explore the question of distributional analysis further. In three successive papers (Lieven, Pine & Dresner Barnes, 1992; Pine & Lieven, 1993; Lieven, Pine & Baldwin, 1997), Lieven, Pine, Barnes and Baldwin have tried quite successfully to give a developmental account of distributional analysis at work during child language acquisition. Their work stems from remarks about variations in the output from one child to another. These variations may be interpreted in different ways, referential versus expressive as proposed by Nelson (1973), or holistic versus analytic (see also Bloom, Lightbown & Hood, 1975; Bates & Marchman, 1988). Some children tend to build multi-word utterances from unanalysed chunks of words. Other children use a more analytical procedure, with productive patterns where an empty slot can be filled with one of a class of words similar to those found in pivot-grammar. In order to make a more valid comparison between children, Lieven et al. used longitudinal studies of the first words and patterns produced by children at a fixed point in vocabulary development, trying to compare structural complexities at similar levels of development. The authors sorted multi-word utterances into three categories: frozen phrases, intermediate utterances and constructed utterances. This is somewhat similar to the proposal by Ingram (1989, pp. 332-337) who described a similar classification and emphasised the need for a more quantitative assessment of the data, which

Lieven et al. indeed do perform. In their classification, they showed that frozen sentences are not an inhibiting factor for language development, but are more of a source of data for future analysis. In particular, they theorised that children use construction patterns organised around specific lexical items, in a mechanism that could be at work well beyond the early stages of multi-word utterances.

Previous works describing the acquisition of the French language include Karmiloff-Smith (1979), Clark (1985) and Le Normand (1991, 1996a, 1996b). Special mention should be made of the work of Veneziano, Sinclair and Berthoud (1990), as it bears some relation to the present work. The authors carried out a study of the early transition from single-word to two-word utterances in French children. There were combinations of clearly delineated meanings separated by pauses or stops, but the authors also described the simultaneous appearance of vowels at the beginning of words (in children aged 1;5 to 1;8). In context, these vowels corresponded to proto-articles, proto-pronouns or proto-modal verbs. This showed not only distributional analysis at work, but also an increasing length in the phonological structure of children's language. What is shown here isn't the emergence of full-fledged grammar, such as GB, but a consequence of the purely phonetic and morphologic properties of the French language. The work also presents a dual developmental mechanism, a semantic one with the association of meanings, and a morphologic one with word-lengthening.

Specifics of the current study

The rationale of the present study is to document and keep track of observations on child language, using simple tools and making as few theoretical assumptions as possible, on a large scale, so as to provide not only a qualitative description but also a quantitative one. Qualitative analysis is of course necessary because it can offer valuable insights about child language, but it cannot be separated from quantitative results because a certain linguistic structure will carry a different weight in regards to the theoretical work depending on whether it is rare or very frequent.

The current work is mainly a study of child language output but in order to minimise assumptions, adult oral output will also be studied, as a reference to be compared with child language. The use of adult oral language is necessary here because it is produced in similar conditions to that of children. On the one hand, adult to adult speech may give a better reference point than adult to child speech, especially because the latter can be much simplified or even artificial. On the other hand, adult to child speech is what children really do hear. Using it as a reference makes it possible to address the issue of language input at the same time. For this work, the analysis was made on child directed speech because it was the only available corpus. But to compensate, the child corpus and the adult corpus came from different sources, so that whatever correlation is drawn between the two corpora, it will not be the result of imitations by children or adults, but common linguistic features.

It only remains to choose the appropriate tool for analysing and processing large databases such as those available in the CHILDES project (MacWhinney & Snow, 1985). There is an apparent contradiction between the notions of detailed analysis and large corpora, which makes it necessary to develop tools appropriate for the task. A morphological study of children's productions was decided upon. Morphosyntactically tagged corpora can be quickly created with the help of an automatic part-of-speech tagger (POST) followed by manual control. POST uses a training phase on already tagged data which makes it possible to adjust the system of syntactic categories to the task and to the type of language (child, adult, oral, written). It offers the possibility of carrying out quantitative studies of a great range of fine-grained phenomena, of comparing child and adult syntax on an equal basis, and of obtaining precise numerical data. But it isn't built on, nor does it presuppose, complex syntactic structures and is thus suited to child language study. This does not presuppose that a child has a morphological understanding of language, but that the product of morphological parsing can

be used to compare child and adult language.

The age of the children to be studied is a crucial issue: before children produce their first two-word sentences or their first inflected words, their early production sounds mostly like a collection of lexical items. Usually, around age 1;6 to 1;9, comparison between child and adult language can be said to amount to a lexical comparison. Later, from the age of 2;6 on the average, most children produce complex structures which reflect a fairly advanced mastery of syntax. The period of the first multi-word utterances, around age 2;0, is more appropriate for an attempt at describing early grammatical combinations, as utterances are still simple while already displaying a beginning of linguistic mastery. Furthermore, explaining the beginning of syntax in children seems a natural way of progressing into a more complex understanding of language. Thus, this study looks at productions at 2;0 because this is when multi-word utterances have begun for nearly every child.

Method

Material

The data come from a database created through the direct observation of young children's behaviour (Le Normand, 1986): direct spontaneous speech data produced during symbolic play, always in the same standard situation, always openly video-recorded by the same observer. The play situation allows the children to comment on their own actions, to speak about real or imaginary events and to have some exchanges with a familiar adult partner. The strictly standardised material involves five characters (two adult figurines, two child figurines and one baby), one dog, eleven pieces of furniture (two tables, four chairs, two armchairs and three beds) and five figurative objects (stairs with a mobile door, a garage with a sliding door and a front door bell).

For data collecting, the technique of full sampling of behaviours was used, and the children's speech has been segmented into utterances using the criteria defined by Rondal, Bachelet and Pérée (1985), which allows a standard transcription and the computation of linguistic parameters described in the corpus processing system CLAN (Child Language Analysis, version 2.01, MacWhinney, 1995). The transcription was done using the normal conventions of French orthography and grammar. As a lot of written elements are silent in French, these elements have been written correctly unless the pronunciation of the child shows a clear grammatical error. Standard French interjections have been transcribed conventionally.

Subjects

The corpus used in the current work was produced by 27 children aged 2;0, all with a normal linguistic development pattern. Their mean MLU in words (footnote 1) is 1.63, ranging from 1.10 to 2.88. The number of utterances for each child ranges from 27 to 187, with an average value of 80. The total number of utterances for all 27 children is 2,157. This corpus will be referred to hereafter as the ‘Le Normand corpus’.

In order to compare the lexical classes of young children to those of adults, it was important to use an adult corpus as close as possible to spoken language and, if possible, corresponding to a conversation with or in the presence of a child. The adult reference corpus presented here comes from the CHILDES database (MacWhinney & Snow, 1985). It consists of the whole set of adult data (extracted from conversations with the child Philippe) gathered by Madeleine Léveillé with the participation of Patrick Suppes (Suppes, Smith and Léveillé, 1972; Suppes, Léveillé and Smith, 1974). This corpus corresponds to 33 tape recordings of an hour each, of a child at home, covering a whole year. At first, they were done every week, later with longer gaps in between. The transcripts include both the utterances of the child Philippe and of the adults, namely the mother and father of the child and the field researcher

Madeleine Léveillé. Every sentence has been analysed and the utterances have been divided into child and adult utterances. The adult part of the corpus, referred to hereafter as the 'adult corpus', contains 22,669 adult utterances – 8062 from the mother, 6479 from the father and 8128 from the investigator – corresponding to 130,053 words, not including punctuation, with a MLU of 6.33. The child's part contains 15,150 utterances, and has been divided into two parts: the recording done at 2;1, 953 utterances, referred to hereafter as the 'Philippe corpus at 2;1' and the whole year of recording minus the first month (14,197 utterances), referred to hereafter as the 'Philippe corpus aged 2;2 to 3;2'. Every transcription in the Léveillé corpus follows standard French orthography. Adult language has also been very carefully transcribed according to classic French grammar.

Morphosyntactic analysis

Morphosyntactic analysis consists in looking for the syntactic category and the morphological decomposition of a word. The tags used in this kind of analysis match those one could find in a lexicon, that is, the word class without any semantic or pragmatic context. Thus, a word can have an ambiguous category: it could be a homophone or a homograph, when several lexical entries share the same phonemic or graphemic shape. The analysis has to rely on context in order to determine which class a word really belongs to. For instance, in French, it is necessary to determine if the string of letters 'p o r t e' corresponds to the feminine singular substantive 'porte (door)' or to the conjugated verb 'porter (to open)' in the present tense, either in the first or third person.

One very important point must be raised at this stage: this sort of morphosyntactic analysis is not based on a theoretical grammar, whether of child or adult language. It is not known whether children are (consciously or unconsciously) using lexical categories or not. The current work is simply a characterisation of texts using a tool that is efficient and appropriate. The use of an automatic tagger is efficient because it reduces, by a factor of at least ten, the time needed to label a corpus (the operator's tiredness not being considered); it is appropriate because it takes three significant linguistic components into account – morphology, syntax and distributional analysis – and thus is well suited to the study of the language of young children. The analysis of text by POST is done in a fashion similar to the way an adult could tag the discourse of a child: mapping it to adult language structure, using an adult interpretation. This is justified inasmuch as every person who converses with a child does the same. The situation is very natural and a mirror to that of the child who is trying to learn language, trying to understand what surrounds him/her, and seeking norms that will enable him/her to communicate with someone else. The aims of the present study can be stated as follows: (1) to look for a morphosyntactic description of children's utterances in order to verify if an adult can interpret them on the basis of morphological criteria; (2) to find common elements between the language of children and of adults; (3) to pinpoint real 'agrammatical' utterances, i.e. children's 'creative' productions, where learning is obviously under way – as opposed to correct productions where it is difficult to determine if learning is under way or if children are only reproducing their input.

POST works with positional or semi-positional languages such as French or English. It has been more fully presented in Parisse and Le Normand (1997) and is based on a Markov model of the resolution of ambiguous bi-class succession rules. It reproduces the initial text, with each word provisionally tagged into one or several categories. The rate of lexical ambiguity in two-year-olds' language is already quite high, ranging between 1.15 and 2.30 possible lexical categories for each word, depending on the richness of the reference lexicon (see footnote 2).

The use of a POST for child language and for adult oral language did raise some specific problems, especially for one-word utterances. It is obviously not possible to build sophisticated context rules for sentences which consist of only one word. Because the only

context is the punctuation (full stop, exclamation or question mark) surrounding the word, no rules can resolve these ambiguities. The difference between types of punctuation have not been taken into account because they give information which is more pragmatic than morphological. Categorisation of one-word utterances must thus be performed manually, sometimes using the context of other sentences. Ambiguities between a noun and an adjective were always resolved as a noun, and those between a noun and an interjection as an interjection. Ambiguities between a noun and a verb were resolved case by case. When there were ambiguities between a content word and a closed class word, the content word was usually opted for. One example is that of 'un' (a/one), which stands in French for the number 1 as well as for the indefinite article. When used in isolation, it was considered to be the number. But when looking at all the instances of '*un*' occurring as an isolated word, it was discovered that in one case it was in fact the article, used by an adult to suggest a word to the child (Philippe in the Léveillé data). This case is exemplary in two ways: first in that it shows that automatic analysis cannot fully replace a manual examination of the data when studying some very specific and localised situations; and secondly in that it shows that there are always 'non-grammatical' utterances which are justified by the pragmatics of the discourse, and that no software will be able to deal with these in the near future.

Lexical categories

The 25 lexical categories used correspond to very general syntactic categories (see Table 1, columns 1 and 6 – punctuation is not included in this table). No tagging effort was made in regard to gender and number as they are easy to analyse in the corpora of two-year-olds, and their study does not justify the development of very sophisticated tools. Although very general, this set of categories reflects the distributional properties of the French language. For example, the three types of pronouns reflect their different contexts of use. If a new category is to be added, one should make sure that this category will be distinguishable from others on the basis of context only.

insert Table 1 about here

Results

The purpose of the present work is threefold:

(1) to compare the lexical classification of children's language with that of an adult;
(2) to show the distribution of sequences of two or three syntactic categories or words in comparison with adults;
(3) to make a detailed analysis focused on an extended concept of content words.

Lexical characteristics

Table 1, column 2, shows the raw numbers of occurrences of the different syntactic categories used by two-year-olds, and column 3 shows the percentage of occurrences of these syntactic categories in reference to the total number of occurrences of all syntactic categories at this age. Parallel percentages are given for Philippe's corpus (2;1 to 3;2) in column 4, and for the adults in the Philippe corpus in column 5. With the exception of categories related to the location of objects (ADV-l, PRN-d, VOILA and I-e) and of interjections (I), there is a great similarity in the percentages of child and adult syntactic categories. A Pearson correlation analysis between the percentages used by the children at 2;0 and by the adults gives a significant result, $r = 0.49$, $p < 0.01$. If interjections are not taken into account because they may be considered as specific to children, the result is even more significant, $r = 0.56$, $p < 0.005$. A control performed by computing the same values for every syntactic category including interjections between Philippe, the child of the Leveillé database (from 2;1 to 3;2 – a 55,616 words corpus) and his parents, gives the same kind of result, again even more

significant, $r = 0.88$, $p < 0.0005$. The result is similar to the one obtained through the comparison of the Le Normand corpus and the Philippe corpus at 2;1, $r = 0.68$, $p < 0.0005$. The correlation between child and adult language production is thus very significant, even without taking into account the specificity of child language. This should not come as a surprise because children get their input from adult language, but it would not have been true at the time of the production of first words. There are however differences in the class occurrence percentages: the children tend to use a higher proportion of substantives and fewer verbs than adults (see Table 1). A more detailed presentation of the syntactic categories is shown in Table 2, where those categories occurring in one word utterances are separated from those occurring in multi-word utterances.

insert Table 2 about here

The values given in Table 2 allow us to compute separate correlations for the one-word and multi-word utterances between the children at 2;0 and the adults. The results confirm what intuition and tradition in child language analysis have suggested: the correlation between adult speech and the children's one-word utterances is not significant, $r = 0.23$, whereas that for multi-word utterances is highly so, $r = 0.66$, $p < 0.0005$. Thus, the correlation previously observed between whole corpora must have been due to the multi-word utterances, where successions of words could provide a close match to adult language. Many of the children's one-word utterances are of a different nature than those of the adults, although both come from the same subset of syntactic categories. The main difference in category use are that adults sometimes utter conjunctions in isolation whereas children never do and that children often produce isolated infinitives but adults don't. Other differences reside only in numbers of occurrences, not in the syntactic categories themselves. The match between syntactic categories used in isolated words reflects general properties of language semantics, not those of language structure. Children's production is interpreted in context by the observer (as would be the case for production by another adult) and expected to make sense. Thus, if a child utters a sound containing only the phoneme /a/, the adult observer is liable to interpret /a/ as a noun, a verb, a demonstrative or a negative adverb whose phonetic form contains this phoneme, or else consider it as uninterpretable. The observer will never interpret this sound as an article, an auxiliary, a subject or object pronoun or a preposition, except in a metalinguistic context such as the repetition of part of the last sentence heard or a suggestion from an adult. Only in such cases would the interpretation of the phoneme as a functional word make sense. For example, in Table 2, although no numbers are shown for ART, ART-g and V-inf in isolated words for adults, the ART and ART-g syntactic categories did in fact appear once each, and the V-inf syntactic category three times. This corresponds to percentages of 0.0001 and 0.0003, which are very low, and are not due to errors but to specific metalinguistic situations.

Distribution of sequences of syntactic categories

Correlations between different syntactic categories, interesting as they may be, remain suggestive and non-conclusive. Where exactly does the difference between one-word and multi-word utterances lie? Substantives are the most frequent category produced by children. Is the use of substantives in multi-word utterances really different from their use in isolation? The answer to this question calls for a careful analysis of children's multi-word utterances. and from here on, the current article will only deal with multi-word utterances unless otherwise specified. The term 'bi-tags' will be used for a sequence of two syntactic categories in a given utterance, 'tri-tags' for a sequence of three, and the term 'bi-words' for a sequence of two words. The study of children's multi-word constructions is related to that of frozen utterances. Which are constructed by children and which are formulaic expressions? To show some of the specificity of children's constructions, Table 3 presents the fifteen most frequent

bi-tags produced by children.

insert Table 3 about here

The total number of occurrences of the fifteen bi-tags shown in Table 3 corresponds, in tokens, to exactly half the total number of occurrences of all the children's bi-tags (812 out of 1608), and to 23% of the adults' (27663 out of 120843), whereas these fifteen correspond, in types, to only 6.8% of the children's possible bi-tags (15 out of 218) and to 2.4% of the adults' (15 out of 624). The four most common adult bi-tags (article + noun; pronoun + finite verb; pronoun + auxiliary be; relative pronoun + pronoun) are among the nine most frequently used by the children. This clearly reinforces the previous findings that the distribution of syntactic categories is similar in both the children and the adults. Furthermore it emphasises the fact that it is not only the relative numbers of occurrences which are similar, as was shown earlier, but also the order of the syntactic categories.

A correlation value between sets of bi-tags cannot be computed because the sets of values are completely different – there are 520 adult bi-tags as compared to 188 child bi-tags. But it is still possible to look at the percentage of bi-tags produced by children which are also produced by adults. Bi-tags produced by children but not by adults are very few: 1% in tokens, 4.5% in types – these figures are computable from Table 4 which shows the number of coinciding bi-tags and tri-tags (99% and 95%). Adult produced bi-tags represent 77% of the possible bi-tags in types (520 out of 676, i.e. 26 tags times 26 tags). If children were producing bi-tags independently of adult input, their production would be randomly distributed and cover both adult bi-tags and non-adult bi-tags. If so they would produce only 77% of adult bi-tags in types, whereas they produce 95.5%.

The same computation performed on tri-tag values shows a similar tendency. The tri-tags produced by children but not by adults represent 17% in types and 7% in tokens – these figures are also computable from Table 4 (83% and 93%). As above, adult produced tri-tags represent 20.4% of the possible tri-tags in types (3586 out of 17576, i.e. 26 tags times 26 tags times 26 tags). If children were producing tri-tags at random, 79.4% of their production should be non-adult, not only 17%.

Bloom (1970) proposed disregarding child productions occurring less than five times. Applying this criterion to the adult corpus allows us to strip away small tagging errors and metalinguistic phenomena such as repetitions of child errors or suggestions to a child – which are often incomplete sentences. Thus, the number of bi-tags produced by children but not by adults goes up to 6% in tokens and 13.8% in types, 16% and 34% respectively for tri-tags.

A last confirmation of the similarity between children's and adults' productions can be performed using bi-word occurrences instead of bi-tags, that is by finding the number of word-pairs produced by the children which exactly match, including order, word-pairs produced by some adult, even though they are not from the same corpora. This analysis shows 44% exact coincidence in types (61% in tokens) – see Table 4 – between the two-words sequences of the Le Normand children and the 33 hours of adult speech in the Léveillé database. This demonstrates that even if the figures obtained with bi-tags and tri-tags are due to an oversimplification resulting from POST's tagging, the tendency they exhibit is still valid when considering the raw lexical forms of the words. A manual check of the list of the children's specific bi-words shows that 36% may perfectly well occur in an adult sentence. This evaluation comes up with a value (80%, i.e. 44%+36%) very close to that obtained for tri-tags (83%) and frequent bi-tags (86%), but which lacks the reliability and repeatability of the previous measures. An even stronger result is obtained from a comparison between the corpus of Philippe at 2;1 and that of the adults surrounding him: 72% of the types are exact matches (82% of the tokens). Manual verification yields the same proportion of at least 86% of bi-words (in types) that would be perfectly correct if uttered by an adult. All the results

above are summed up in Table 4.

insert Table 4 about here

A detailed study of the bi-tags occurring more than five times in the children's corpus but not produced by the adults is very interesting because it represents a qualitative analysis of a quantitative account. These bi-tag sequences are certainly specific to the children and most revealing of their syntactic command. The full list is quite short: I-e/PP (Interjection of exclamation followed by Past participle, 24 items), S/I (Substantive followed by Interjection, 11 items), I-e/S (Interjection of exclamation followed by Substantive, 10 items), I-e/VOILA (Interjection of exclamation followed by Locution of place, 8 items), ADV-l/S (Adverb of place followed by Substantive, 7 items), Y/Y (only represented by the formulaic expression 'y en' for the present corpus, 5 items).

insert Table 5 about here

S/I and Y/Y could have been produced by an adult, but they correspond to colloquial language that was not encountered in the Léveillé corpus. All the other cases above, as well as 70% of the bi-tags not found in the adult corpus and occurring less than five times, correspond to a very specific feature of child language which is also, perhaps, specific to the task performed by the children during the recordings. This feature is the use of 'object-focus' words such as *'là'* (there), *'ça'* (this), *'voilà'* (there it is), *'oh!'* (oh!). The first three words are used to pinpoint the presence and sometimes the location of a object, whereas the fourth word is only used to point out a presence. The use of these words is very consistent among the different children. There are other words or word combinations with the same functions which belong to four categories: ADV-l (Adverb of place), PRN-d (Demonstrative pronoun), VOILA (Locution of place 'voici' and 'voilà') and I-e (Interjection of exclamation). The category differences correspond to different morphosyntactic properties, but the semantic values of these words are difficult to differentiate. For example, the cognitive difference expressed by 'là' (adverb of place) and 'ça' (pronoun) is very small for children in isolated contexts. There are only 4 contexts in the Le Normand corpus where 'ça' cannot be considered to be synonymous to 'là', and they are problematic because they correspond to subject contexts where 'ça' is almost never used as a referential pronoun by adults, but more as an obligatory impersonal subject pronoun (example: 'ça tourne' which means either 'this turns' or 'it turns'). Thus semantic function and syntactic function may be very different from each other. These four classes are very frequent in the Le Normand corpus, representing 20% of all words, 25% of the one-word utterances; they appear in 30% of all utterances and 40% of utterances of more than one word.

Content words and functional words

It has been shown above that a large number of children's combinations are adult-like, but there is as yet no indication of how this process works, and no way of distinguishing what comes from adult syntax from what doesn't. A more thorough description of the morphosyntactic structures used by children is needed.

Although the distinction between open-class words and closed-class words is fundamental in any study of language, it does not seem to be fully satisfactory for the study of child language. Instead, Braine (1963) used a dichotomy between pivot words and open-class words on the basis of distributional characteristics. Radford (1990) suggested that because early utterances showed no evidence of functional categories, early structures produced by English speakers are exclusively lexico-thematic structures. Yet another dichotomy can be studied: content words vs. functional words. As almost all children's utterances make sense,

most single-word productions should consist of a content word and most multi-word productions should contain at least one.

In the following section, the content words have first been separated into five subsets: interjections (I), object-focus (I-e, VOILA, PRN-d, ADV-l), adjectives (A), substantives (S, NP) and verbs (PP, V, V-inf, V-ppre). The other classes are considered to be functional classes. This is not a standard division: interjections and object-focus words would usually have been included in the functional categories, but this has been done because interjections and object-focus words have content for children.

Out of 1,215 single-word utterances, 1,067 (88%) corresponded to content words and 148 (12%) to functional words. 114 of these 148 utterances correspond to the 3 words 'oui' (yes), 'non' (no) and 'encore' (again). The others were mostly interrogative pronouns (questions put by the children) and adverbs. They were perfectly justified in isolated contexts, and had in this situation a content-word value that they could come to lose in sentence contexts. The same thing holds for adult language where the four words 'oui', 'non', 'quoi' (what) and 'pourquoi' (why) correspond to 83% of the occurrences of functional single-word utterances. For these reasons, it has been decided to extend the first list of content words to two supplementary subsets: negation (ADV-n, 'oui') and interrogation (PRN-r). This definition of content words may seem counter-intuitive to classical grammars and differs from the open-class definition, but it reflects the cognitive characteristics of two-year-olds' language. At this age, negation is not a modifier for another word, as it will later become (Gopnik & Metzloff, 1985). It stands alone and has a different function in single-word utterances and with a verb. This is a semantic categorisation of child language, as well as a morphological one. The whole syntactic and semantic framework presented here does not try to fit a classic adult grammatical description, but tries to be a tool for describing and understanding the characteristics of child language and its evolution. However, the word for negation 'non' is also used in isolation in French adult language and this has to be taken into account in adult grammars.

insert Table 6 about here

Table 6 presents the percentage of multi-word utterances containing 0, 1, 2, 3 and more content words, for both the children and the adults, and for both types of classifications: the 7 content word classes and the 3 content word classes (see Table 7). The 7 content word classes give a very interesting result. Almost all multi-word utterances contain a content word and very few utterances are composed of functional words only. This was not predictable because those categories were chosen on the basis of isolated words, not multi-word utterances. This would suggest that isolated words do not belong to special categories, but are subjected to the same semantic and pragmatic principles as connected words. This is true for children as well as for adults and could represent one of the first elements learned by children, an automatic by-product of learning language, or a universal of language.

insert Table 7 about here

Table 7 presents the use of content words in children's sentences. The percentages in the subsets of content words, for single-word utterances and multi-word utterances, are broken down by content word types. Table 7a corresponds to the 7 subsets of content words, Table 7b to the 3 subsets. The distributions are roughly similar in shape across the different situations, although there are some notable differences, and the adults' pattern is more stable than that of the children. Some of the differences between adults and children are striking, all the more so because the general tendencies are very similar. There are also great differences between the numbers obtained with the classification made on the 7 subsets of content words and that made on the 3 subsets, due to the fact that the totals of the sets of one, of two and of three content words groups are different in the two classifications. In the case of 7 subsets,

there are many utterances with two content words (440: 46.7%) whereas they are much less numerous in the 3 subsets case (120: 12.8%). So that, if an utterance with two content words is a child's semantic and pragmatic creation, then 46.7% of the multi-words utterances are children's semantic and pragmatic creations, a percentage reduced to 12.8% when content words are limited to nouns, verbs and adjectives. With 7 classes of content words, many children's productions are semantic and pragmatic creations whereas with 3 classes of content words most children's productions are purely grammatical creations.

The main differences between child and adult productions are as follows:

1. Interjections: this class is much more used by children than adults in multi-word utterances. When an interjection is present in a multi-word utterance, there will always be at least one other content word in the utterance. The proportion of interjections by children and adults in one-word utterances is nearly identical. This reflects a morphosyntactic property. When interjections are used, they are either in isolation, or at the beginning or the end of a sentence, and require no morphosyntactic complements. Thus, when they are the only content word of a sentence, they are likely to be the only element in it.
2. Object-focus: this class is also used much more by children than adults. Its use by adults is not uncommon, however, and follows the same syntactic, semantic and pragmatic structure as the children.
3. Verbs: the children use verbs less frequently than adults, except as isolated words.
4. Substantives: the use of isolated substantives is higher for the children.
5. Interrogations: the children use them less frequently than the adults.

Content words and morphosyntax

The foregoing results characterise global differences between French children and adults. A clear convergence has been demonstrated between children's and adults' speech. However, this convergence should be smaller for the children's productions which are innovative and not the simple reproduction – complete or incomplete – of adult input. We suggested above that utterances with more than one content word are likely to be children's semantic and pragmatic creations. If this is true, then the convergence between child and adults should be smaller for this type of utterances.

insert Table 8 about here

The Le Normand corpus was classified into utterances with one, two or three content words. Characteristics of the sub-corpora resulting from this classification are given in Table 8. All the statistical computations performed previously (see Table 4) have been applied separately to the results of this classification. Computation results of the percentages of coincidence between bi-tags, tri-tags and bi-words produced by children and adults are presented in Table 9. All the values in Table 9, with the exception of the bi-word values, are computed in types with very infrequent cases, those occurring less than five times, eliminated. Bi-word values are computed in types with infrequent cases taken into account. The reason for this decision is to present clear-cut results and avoid ceiling effects. It doesn't change the significance of the results, because all results come from comparing values of the same kind, and not from absolute values.

insert Table 9 about here

The main result in Table 9 is that there are more adult bi-tags, tri-tags and bi-words in single content-word utterances than in two content-word utterances. From the results presented in Table 9, it could be said that, if there is an imitation of adult language or merely a respect of the morphosyntactic properties of adult language, this is less frequent in utterances

with more than one content word. These results are statistically significant for the Student t test: for bi-tags, $t(54) = 3.1$, $p = 0.003$; for tri-tags, $t(46) = 2.84$, $p = 0.008$; for bi-words, $t(54) = 3.64$, $p = 0.0006$. The number of tri-tag samples is smaller because tri-tag values cannot be computed for 4 children with very low MLU. The difference between multi-word utterances with a single content word and multi-word utterances with two content words cannot be fully accounted for by the greater complexity of the latter because the difference between the MLUs of each is only marginally significant: $t(54) = 1.88$, $p = 0.06$. The difference between types/tokens ratios is also not significant: $t(54) = 0.22$, p = ns.

Conclusion

From the results section evidence was found for three points:

1) The distributional characteristics of child and adult language were shown to be very similar. There is a significant correlation in the number of occurrences of syntactic categories. This correlation, however, finds support only in multi-word utterances (see Table 2). The correlation value obtained between a written corpus of 192,000 words from newspapers and juridical accounts – coming from a previous work of Parisse (1989) – and the oral adult Léveillé corpus of the CHILDES database was only $r = 0.43$, $p < 0.05$, whereas the correlation between child and adult oral language is $r = 0.66$, $p < 0.0005$; this comparison only makes sense for multi-word utterances, as there are no one-word utterances in the written corpus. Although imprecise, lexical correlation reflects the existence of common linguistic patterns. The lexical correlation between adult and child syntactic category use should not be surprising, as children takes their examples from adults.

2) Experiments using bi-tags (two successive syntactic categories), tri-tags (three successive syntactic categories) and bi-words (two successive words) demonstrated a close relationship between child and adult morphosyntax. A correlation value cannot be obtained here because the sets of child and adult bi-tags are too different. Adult bi-tags are much more numerous and this reflects the greater complexity of adult language. However, many of the children's bi-tags correspond to adults'. This match between child and adult has been evaluated and the same evaluation performed with tri-tags and bi-words instead of bi-tags. The results are shown in Table 4.

3) In order to pinpoint the syntactic structure of children's first multi-word utterances, a study of content versus function word use was performed. Content word categories were considered to correspond to the complete list of isolated word categories. Results showed that nearly all children's multi-word utterances contained a content word (as previously understood.) Finally, the syntactical correctness of utterances with one, two and three content words was investigated. There was a higher tendency toward errors, in comparison to adult morphosyntax, in utterances with more than one content word.

Discussion

The present study used texts tagged by a stochastic morphosyntactic parser. This parser can in no way be taken as a model of the language acquisition in children. It was a means of characterising the language of children in a morphosyntactic dimension, using adult knowledge and interpretation. The analyses above show that the distributional characteristics of children's multi-word utterances match those of the adult's output. This match is not limited to the lexicon but covers word and morpheme order as well. This seems to reflect some deep characteristics of language acquisition by French children. First of all, morphology and functional words appear at an early age and this is probably related to the phonetic characteristics of French as a syllable-timed language (Peters, 1995). Secondly, syntactic markers like articles, pronouns (subject or object), prepositions, auxiliaries, and modals are made of words that can be separated from their syntactic head, and it is this particular construction which is reflected in word order. In languages where articles – gender and

number markers in French – and prepositions are not entities separated from the noun, one will probably not find the same regularities in word order but morpheme order regularities instead.

It is possible that the high correlation between child and adult language comes from a common system of semantic-thematic rules, or schemata. This would explain the correlation between adults. Of course, this implies that these rules have either already been acquired or that they are innate. A previous acquisition would be difficult for two reasons. First, children in this study are very young and they are producing their first combinations. The small number of obligatory pronouns and articles at that age makes it very unlikely that children have already mastered these rules when they begin to produce pronouns and articles. Secondly, a test can be made using the data of this study. The percentage of bi-tags, tri-tags and bi-words can be computed separately for the children with the lowest MLUs and the children with the highest MLUs. No significant difference obtains. The children with the simplest language do not differ from the children with the most complex language. In bi-words, for example, the fourteen children with the lowest MLUs ($M = 1.32$, $S.D. = 0.13$) present a percentage of coincidence with adults of 61%. The thirteen children with the highest MLUs ($M = 1.87$, $S.D. = 0.30$) present a percentage of 53%. The difference is not significant, $t(23) = 1.33$, $p = 0.097$ and it is the youngest children that follow the adults' production best. The absence of correlation between MLU and the coincidence between child and adult is clearly visible in Figure 1 where the MLUs, the bi-tag, tri-tag and bi-word coincidences are plotted one above the other, and thus every point in the same vertical line corresponds to one child. The result of the Pearson correlation analysis between MLU and bi-tags coincidence is $r = 0.11$, between MLU and tri-tags, is $r = 0.01$ and between MLU and bi-words, is $r = 0.15$. This makes the case for a knowledge of rules at the very beginning of production of multi-utterances harder to defend. It is possible that rules can be learned very quickly after a first short period of adult language reproduction. However, the apparent grammatical proficiency of young children may be an overestimation of their real knowledge.

Figure 1 about here

Since Chomsky's first works, it has often been pointed out that adult production is poor and does not provide enough material for a child to learn language. Following this tradition, Pinker, for example (1990:360-361), says: 'Similarly the crucial input to language acquisition — parent's sentences — can be easily characterised, at least in its essentials. Thus both the input and output to language acquisition can be specified precisely...'. The results presented above show that this poverty of input should be reconsidered. The present data does not prove that children borrow chunks of input, but if a comparison between Philippe and the adults he is talking to (during the 33 hours of recorded speech of the Léveillé database which corresponds roughly to a mere week of parent's speech) shows that 72% of the bi-words produced by Philippe at 2;1 (in type, 82% in tokens) correspond exactly to adult bi-words, the quality of the match between child and adult is very high indeed and it might be even higher over a longer observation time. Of course, as the number of different adult sentences increases with the observation length, so will the number of the children's new combinations, and some combinations used by children will never be produced by adults. Still, it is plausible that up to 90% of the combinations used by children have been heard at least once. A complete demonstration of this, which could not be done with the technique followed here, would be hard to arrange: technically, it would be necessary to have full recordings of the surroundings of a child during several years and then to transcribe all the resulting tapes; and it is as yet impossible to decide how long the interval between the first actual hearing of a word by a child and its first production can be. There exists another way of trying to demonstrate the truth of this assertion, by the consequences that it should have on language acquisition by children.

A proposal would be that children begin to purely copy adult production in several classes of content words (belonging to 7 different types, in French). They will later extend these words to include co-occurring functional words. These groups of single or multiple words are all built around a content word. As these groups begin to make sense to the children, they manipulate them and in particular string them together as whole units, either following some semantic order and/or using phonetic or syntactic regularities. This would be the reason why multi-word utterances with more than one content word were less adult like than utterances with a single content word (see Table 9). As the semantic combination must make sense to the children, it will prevent the production of semantically incoherent sentences. It is when the mastery of small morphosyntactic groups is well under way and the semantic knowledge getting more complex that most of the children's syntactic errors will be found.

Recent advances in distributional and stochastic knowledge acquisition (Redington & Chater, 1997; Schütze, 1997; Seidenberg, 1997) make learning through regularities more credible now, and all the more since over-generalisation – i.e. a clear use of morphological rules – by children does not usually appear very early, thus giving them sufficient time to learn regularities. The initial spurt of language, and not only of vocabulary, may be explained by – a lot – of memory, perception and classification of regularities, and by the mastery of some fundamental cognitive categories that are reflected in language. More work has to be done to study the constructions used by children and the mechanisms thus displayed. It also remains to be seen whether adult language performance can possibly evolve from these same mechanisms.

References

Bates, E. & Marchman, V. A. (1988). What is and is not universal in language acquisition. In F. Plum (ed.), Language, communication, and the brain, New York: Raven Press.

Bloom, L. (1970). Language development: form and function in emerging grammars. Cambridge, MA: MIT Press.

Bloom, L., Lightbown, P. & Hood, L. (1975). Structure and variation in child language. Monographs of the Society for Research in Child Development, **41**, 164.

Braine, M. D. S. (1963). The ontogeny of English phrase structure: The first phrase. Language, **39**, 1-14.

Clark, E. V. (1985). The acquisition of Romance, with a special reference to French. In The crosslinguistic study of language: acquisition, Volune1: The data. Hillsdale, NJ: Erlbaum.

Gopnik, A. & Metzloff, N. (1985). From people to plans to objects: changes in the meaning of early words and their relation to cognitive development. Journal of Pragmatics, **9**, 495-512.

Ingram, D. (1989). First language acquisition. Cambridge: C.U.P.

Karmiloff-Smith, A. (1979). A functional approach to child language: a study of determiners and reference. Cambridge: C.U.P.

Le Normand, M.T. (1986). A developmental exploration of language used to accompany symbolic play in young, normal children (2-4 years old). Child, Care, Health and Development, **12**, 121-134.

Le Normand, M. T. (1991). La démarche de l'évaluation psycholinguistique chez l'enfant de moins de 3 ans, Glossa, **26**, 14-21.

Le Normand, M. T. (1996a). Early morphological development in French children. Proceedings on Learning disorders as a barrier to human development, COST A8, Stockholm, February 29- March 1.

Le Normand, M. T. (1996b). Les modèles psycholinguistiques de développement. In C. Chevrie-Muller & J. Narbona (eds.), Le langage de l'enfant, Aspects normaux et pathologiques, Paris: Masson.

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

Lieven, E. V., Pine, J. M. & Dresner Barnes, H. (1992). Individual differences in early vocabulary development: redefining the referential-expressive distinction. Journal of Child Language, **19,** 287-310

MacWhinney, B. & Snow, C.E. (1985). The Child Language Data Exchange System. Journal of Child Language, **12**, 271-296.

MacWhinney, B. (1995). The CHILDES project: Computational tools for analyzing talk (2nd edition). Hillsdale, NJ: Erlbaum.

Maratsos, M. P. (1982). The child's construction of grammatical categories. In L. Gleitman & E. Wanner (eds.), Language acquisition: The state of the art, Cambridge: C.U.P.

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

Nelson, K. (1973). Structure and strategy in learning to talk. Monograph of the Society for Research in Child Development, **34** (1-2).

Parisse, C. (1989). Reconnaissance de l'écriture manuscrite: analyse de la forme globale des mots et utilisation de la morpho-syntaxe. Unpublished doctoral dissertation,

Université de Paris-Sud, Orsay, France.
Parisse, C. & Le Normand, M. T. (1997) Etude des catégories lexicales chez le jeune enfant à partir de deux ans à l'aide d'un traitement automatique de la morphosyntaxe, Bulletin d'Audiophonologie, **XIII**, 6, 305-328.
Peters, A. M. (1995). Strategies in acquisition of syntax. In P. Fletcher & B. MacWhinney (eds.), The Handbook of Child Language, Cambridge: Blackwell.
Pine, J.M. & Lieven, E.V.M. (1993). Reanalysing rote-learned phrases: individual differences in the transition to multi-word speech, Journal of Child Language, **20**, 551-571.
Pinker, S. (1984). Language learnability and language development. Cambridge: Havard University Press.
Pinker, S. (1990). Language acquisition. In D. N. Oshernon & H. Lasnik (eds.), Language: An invitation to cognitive science: Vol 1, Cambridge, MA: MIT Press.
Post, K. N. (1994). Negative evidence in the language learning environment of laterborns in a rural Florida community. In J. L. Sokolov & C. E. Snow (eds.), Handbook of research in language development using CHILDES. Hillsdale, NJ: Erlbaum.
Radford, A. (1990). Syntactic theory and the acquisition of english syntax. Oxford: Blackwell.
Redington, M. & Chater, N. (1997). Probabilistic and distributional approaches to language acquisition. Trends in Cognitive Science, **1**, 7, 273-281.
Rondal, J.A., Bachelet, J.F. & Pérée, F. (1985). Analyse du langage et des interactions verbales adulte-enfant. Bulletin d'Audiophonologie, **5/6**, 507-536.
Saxton, M. (1997). The contrast theory of negative input. Journal of Child Language, **24**, 139-161.
Schütze, M. (1997). Ambiguity resolution in language learning. Stanford: CSLI Publications.
Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. Science, **275**, 1599-1603.
Suppes, P., Léveillé, M. & Smith, R. (1974). Developmental models of child's French syntax. Technical report # 243, Standford: Stanford University.
Suppes, P., Smith, R. & Léveillé, M. (1972). The French syntax and semantics of Philippe, part 1: noun phrases. Technical report # 195. Stanford: Stanford University.
Veneziano, E., Sinclair, H. & Berthoud, I. (1990). From one word to two words: repetition patterns on the way to structured speech. Journal of Child Language, **17**, 633-650.

Table 1: List of 25 morphosyntactic categories used by two-year-olds and by adults

Tag for the category	Number of occurr. at 2;0	% of occurr. at 2;0	% of occurr. for Philippe	% of occurr. for the adults	Description of the morphosyntactic class
A	87	2.31	2.75	3.19	Verb 'to have'
ADJ	105	2.78	3.13	3.46	Adjective
ADV	159	4.22	4.67	5.54	Adverb
ADV-l	273	7.25	2.10	1.21	Adverb of place
ADV-n	130	3.45	4.24	3.64	Adverb of negation
ART	181	4.80	11.53	6.75	Article
ART-g	9	0.23	1.97	2.10	Generalized article
COJ	31	0.82	1.79	4.24	Conjunction
E	246	6.53	4.60	4.16	Verb 'to be'
I	284	7.54	1.47	1.83	Interjection
I-e	253	6.71	0.86	1.29	Interjection of exclamation
NB	6	0.15	0.43	0.40	Number
NP	156	4.14	2.15	1.62	Last name, proper name
PP	199	5.28	2.22	2.58	Past participle
PREP	17	0.45	5.14	5.02	Preposition
PREP-a	39	1.03	1.85	2.16	Preposition article
PRN	296	7.86	14.31	17.90	Pronoun
PRN-d	122	3.24	2.36	2.03	Demonstrative pronoun
PRN-r	72	1.91	3.00	5.64	Relative or interrogative pronoun
S	671	17.82	16.75	10.65	Noun
V	129	3.42	5.06	7.64	Verb
V-inf	114	3.02	3.20	2.83	Infinitive
V-m	40	1.06	3.07	3.11	Modal verb
V-ppre	--	0	0.16	0.06	Present participle
VOILA	100	2.65	0.31	0.21	Locution 'voici', 'voilà'
Y	46	1.22	0.76	1.01	Pronouns 'Y', 'EN'
Total number of occurr.		3,765	55,616	131,354	

Table 2: Syntactic categories used by two-year-olds and by adults, for one-word and more than one word utterances

1 word				multi-word			
2;0		adult		2;0		adult	
Tag	% of tokens	Tag	% of tokens	Tag	% of tokens	Tag	% of tokens
				A	3.41	A	3.26
ADJ	1.23	ADJ	3.32	ADJ	3.53	ADJ	3.45
ADV	8.15	ADV	31.98	ADV	2.35	ADV	4.87
ADV-l	7.74	ADV-l	0.85	ADV-l	7.02	ADV-l	1.22
ADV-n	2.39	ADV-n	8.07	ADV-n	3.96	ADV-n	3.52
				ART	7.1	ART	6.89
				ART-g	0.35	ART-g	2.15
		COJ	0.35	COJ	1.22	COJ	4.32
				E	9.65	E	4.26
I	15.23	I	17.05	I	3.88	I	1.45
I-e	9.22	I-e	6.52	I-e	5.53	I-e	1.16
NB	0.41	NB	0.35	NB	0.04	NB	0.40
NP	6.01	NP	2.44	NP	3.25	NP	1.60
PP	8.31	PP	0.19	PP	3.84	PP	2.63
				PREP	0.67	PREP	5.13
				PREP-a	1.53	PREP-a	2.21
PRN	0.33	PRN	0.41	PRN	11.45	PRN	18.27
PRN-d	1.98	PRN-d	0.79	PRN-d	3.84	PRN-d	2.06
PRN-r	0.74	PRN-r	15.34	PRN-r	2.47	PRN-r	5.38
S	22.30	S	2.85	S	15.69	S	10.81
V	4.69	V	5.73	V	2.82	V	7.41
V-inf	4.69	V-inf	0.09	V-inf	2.24	V-inf	2.89
V-m	0.16	V-m	0.44	V-m	1.49	V-m	3.43
						V-ppre	0.07
VOILA	6.42	VOILA	3.23	VOILA	0.86	Voilà	0.14
				Y	1.8	Y	1.04
Number of words	1,215		3,160		2,550		128,194

Table 3: Most frequent occurrences of two successive syntactic categories for two-year-olds and their frequency in adult use

Children at 2;0				Adults	
Rank	Tokens	Tag 1	Tag 2	Rank	Tokens
1	197	PRN	E	3	4949
2	168	ART	S	1	6866
3	46	E	ADV-l	131	122
4	46	E	ADJ	29	727
5	44	PRN-r	PRN	4	3292
6	40	A	ADV-n	39	530
7	39	PREP-a	S	8	2340
8	39	E	PP	30	706
9	35	PRN	V	2	5200
10	31	Y	A	23	877
11	29	ADV-n	ADV-l	207	56
12	29	ADJ	S	16	1224
13	25	I-e	PRN	118	146
14	24	I-e	PP	548	1
15	20	E	ADV-n	36	627
Total	812				27663

Table 4: Percentages of coincidence between children's and adults' bi-tags, tri-tags and bi-words in the whole Le Normand corpus.

	tokens (%)	types (%)
bi-tags	99 (94)	95 (86)
tri-tags	93 (84)	83 (66)
bi-words	61	44 (80)

Note: For bi-tags and tri-tags, the values in parentheses represent the percentage without having taken into account the less frequent adult bi- and tri-tags (number of occurrences 5). For bi-words, the value in parentheses represents the evaluation after a manual addition of the correct forms not encountered in the Léveillé corpus.

Table 5: Examples of utterances with bi-tags specific to children aged 2;0

Bi-tag specific to children	Examples of full utterances
I-e/PP	oh caché, ah tombé, oh assis (oh hidden, ah fall, oh sit)
S/I	joujou hein, poussette boum (toy hey, push chair boum)
I-e/S	oh camion, ah nounours (oh truck, ah teddy)
I-e/VOILA	ah voilà, oh voilà, ah voilà chapeau (ah here it is, oh here it is, ah there a hat)
ADV-l/S	là bobo, dedans chien, (there hurt, in dog)
Y/Y	y'en a plus (there's no more)

Table 6: Distribution of utterances in relation to their number of content words.

number of content words		0	1	2	3	4
3 subsets of content words	children (%)	30	56	13	1	0
	adults (%)	12	33	27	14	14
7 subsets of content words	children (%)	1	45	47	7	0
	adults (%)	2	19	26	20	33

Note: There is a total of 942 utterances for children and 18,509 for adults.

Table 7: Percentages of content word types for single-word utterances and multi-word utterances for two-year-olds.

number of content words	single word utterances 1	multi-word utterances 1	2	3
substantive (%)	29 (5)	44 (41)	59 (55)	70 (62)
verb (%)	18 (6)	22 (30)	30 (49)	27 (63)
adjective (%)	1 (3)	8 (9)	10 (17)	27 (19)
interjection (%)	16 (17)	1 (1)	13 (3)	17 (5)
object-focus (%)	26 (12)	13 (4)	53 (14)	72 (21)
negation (%)	7 (38)	6 (3)	10 (13)	29 (18)
interrogation (%)	1 (16)	3 (5)	5 (15)	10 (24)

Table 7a: 7 subsets of content words (figures for adults are given in parentheses)

number of content words	single word utterances 1	multi-word utterances 1	2	3
substantive (%)	59 (36)	61 (39)	98 (66)	100 (77)
verb (%)	37 (41)	28 (43)	50 (64)	83 (74)
adjective (%)	2 (22)	9 (10)	26 (18)	66 (24)

Table 7b: 3 subsets of content words (figures for adults are given in parentheses)

Table 8: Distribution of the Le Normand children's corpus into multi-word utterances with one, two and three content words and characteristics of the resulting sub-corpora.

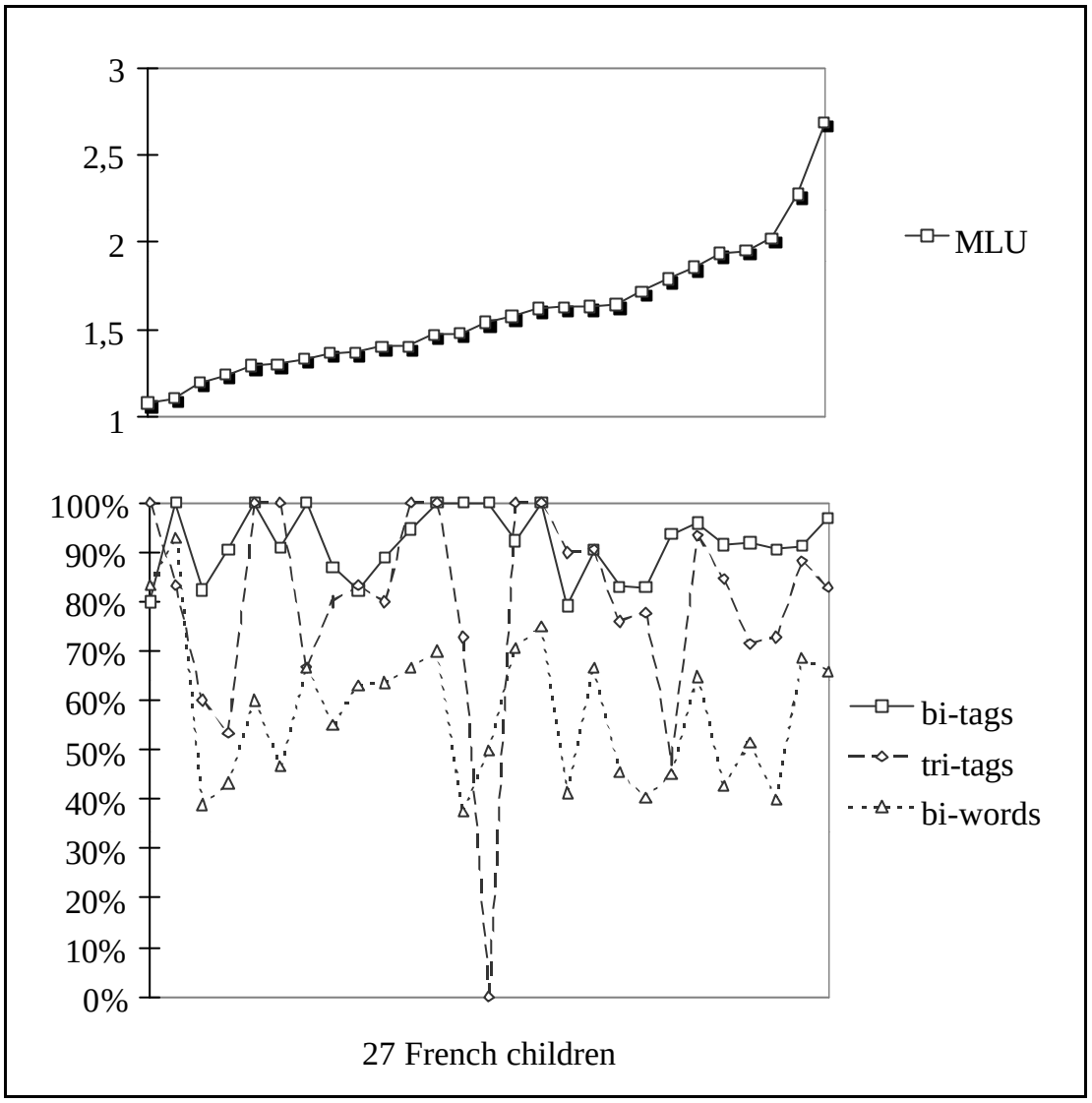
number of content words	1-3	1	2	3
number of utterances	926	44.9%	46.7%	6.7%
MLU	2.70	2.41	2.74	4.42
tokens/types words ratio	8.64	6.03	5.89	3.03

Table 9: Percentages of children's bi-tags, tri-tags, and bi-words corresponding to the adults' in multi-word utterances, in relation to the number of content words.

number of content words	1-3	1	2	3
bi-tags (%)	85	92	86	89
tri-tags (%)	66	86	70	67
bi-words (%)	42	57	38	52

Note: All figures computed in types – for syntactic categories, infrequent occurrences (5) are not taken into account.

Figure 1: MLU for each of the 27 children and coincidence percentages of bi-tags, tri-tags and bi-words production between each of the 27 children and the adults.



Note: 1. Points in the same vertical line correspond to one child, be they on the MLU, or the bi-tags, tri-tags or bi-words plot graphs. The two figures are separated because the scales are different.
2. MLU value is given in words per utterance.

Footnotes:

1. All MLUs in this article have been computed in words. In French, the apostrophe is always considered as a word separator. Thus, ‘j’ai’ (I’ve) is counted as two words as would be ‘je suis’ (I am). The only exception to this rule is ‘aujourd’hui’ (today) and some very infrequent words such as ‘entr’apercevoir’ (to catch of brief glimpse of). Otherwise, white space is the only word separator used.
2. The reference lexicon can be limited to the lexicon of the corpus itself or cover the whole lexicon of French language. In the first case, the possible categories of each word are highly constrained by the knowledge of the situation, and ambiguity is minimised. No so in the second where ambiguity is maximised: for the utterance ‘this book’, in a children’s corpus, ‘this’ can be a determiner or a pronoun and ‘book’ can only be a noun whereas in an adult corpus ‘book’ could also be a verb. In the Le Normand corpus, words produced by children have an average of 1.15 possible lexical categories in the first case and an average of 2.30 in the second case. For adults, the values are 1.79 and 2.52, respectively.