How do children acquire early grammar and build multiword utterances?

A corpus study of French children aged two to four.

M.T. Le Normand, I. Moreno-Torres, C. Parisse, G. Dellatolas

This corpus study examines whether early multiword utterances are built around lexical or grammatical categories. We explored lexical, pragmatic and grammatical word types in relation to a global index of linguistic maturation (Mean Length of Utterance, hereafter MLU). We also investigated whether these associations, when they exist, remain stable across age and socio-economic status (SES), and whether there is a specific set of predictors for French. The data used was a large database of 312 speech samples from French-speaking children aged two to four with high or low SES. All the words in the corpus were annotated with a Part Of Speech Tagger (POS-T), which is part of the CHILDES set of computational tools. While all the linguistic categories correlated with MLU, hierarchical regression analyses showed that mainly the grammatical word types (with positive coefficients) and secondarily the pragmatic word types (with negative coefficients) were the most consistent predictors of MLU. The relationship of MLU to lexical, pragmatic and grammatical word types was quite similar across SES groups, but tended to weaken as MLU increased with age. Three word types - personal pronouns, determiners, and prepositions - each accounted for more than 50% of the variance, and together accounted for 73% of the variance in MLU. These findings support a view of language development according to which children use distributional regularities to build a rudimentary grammar, which in turn provides them with a basic cognitive architecture to learn adult grammar.

Introduction

Over the last 50 years, many researchers have attempted to explain how children learn language so quickly (Brown & Berko, 1960; Brown & Fraser, 1963; Braine, 1963; Brown, 1973; Valian, 1986). Various hypotheses have been proposed that differ crucially in the type of information (e.g., distributional, semantic, grammatical innatism, etc.) that the child uses to start producing multiword utterances. However, methodological and theoretical limitations made it difficult to compare alternative hypotheses.

This long-standing debate has recently been reinvigorated by the design of new experimental methods (Tomasello, 2000; Sebastián-Gallés, 2007), as well as by the existence of large corpora databases (MacWhinney, 2005). While the main aim is the same as 50 years ago, researchers have focused on more detailed aspects of early language development. One central issue today is to determine the time when children begin to use grammatical information, first in perception and later in production. Perception studies have found that before their second birthday (i.e. before they produce multiword utterances), children use grammatical information to process linguistic input (e.g., Shi & Melançon, 2010; Sebastián-Gallés, 2007, for a review).

As regards productive language, the debate is still open. Two alternative hypotheses have recently been advanced. According to Tomasello (2000), early multiword utterances (i.e., produced around the age of two) are based on knowledge of lexical patterns, and creative use of grammar does not start until around the age of four. According to Ninio (2006), children begin to make use of grammatical information during their third year of life.

This study is designed to extend our understanding of the emergence of grammar. Using a large corpus of French children aged two to four, we explored whether there is evidence that children use grammar productively close to their second birthday (what we call the *early-grammar* hypothesis) or whether, alternatively, productive use of grammar is only evident close to their fourth birthday (what we call the *late-grammar* hypothesis).

Early multiword utterances and grammatical development

During the first two years of life, children develop a partial knowledge of grammatical words (van Heugthen & Johnson, 2010; Shi & Melançon, 2010). For instance, by the end of their first year of life, children can categorize novel words into nouns using determiners (Shi, Werker & Morgan, 1999; Hallé, Durand & De Boysson-Bardies, 2008), or based on *frequent frames* (e.g., "the X on"; Mintz, 2003). Then, during the second year, they learn to extract complex patterns across morphophonologically inconsistent natural language (e.g., agreement: Nazzi, Barrière, Goyet, Kresh, & Legendre, 2011; article-noun dependencies: van Heugten & Johnson, 2010). Thus, by their second birthday, they have implicit knowledge of at least surface grammar.

However, there is no agreement on when children begin to make use of this information in language production. Usage-based proponents claim that this does not occur until around the fourth birthday. According to this approach, early multiword utterances are organized totally around individual verbs and other predicative terms (Verb Island hypothesis, Tomasello, 2000; 2003; Lieven, Pine, & Balwin, 1997). This hypothesis has been supported by data showing that children produce the same lexical patterns as those that they hear (Lieven, Salomo, & Tomasello, 2009), and also by experimental studies using *tracer* elements (i.e., a nonce word that the child is expected to produce in different syntactic patterns). In several studies, Tomasello and colleagues have shown that children around two and three years make conservative use of new verbs, while around the age of four, they can use these new items creatively (Akhtar & Tomasello, 1997). Thus, according to this approach (called here the *late-grammar* hypothesis), early multiword production is guided mostly by the knowledge children have of item-based patterns, and grammar develops *after* a period of multiword production. One limitation of these studies is that they do not provide direct evidence that children are not using grammar early in multiword production. Even if children do organize multiword utterances around the lexicon, we cannot rule out that they may also use grammar. This hypothesis has also been criticized for resulting in the following paradox stated by Naigles, (2002):

how is it possible that children learn to process grammar so easily and they need so much time to use it in production? It is important to explore whether grammar, and not only lexicon, helps children to build early utterances.

An alternative proposal has been advanced recently by Ninio (2006, 2011). In agreement with linguistic theories (e.g., Kaplan & Bresnan, 1982; Pollard & Sag, 1994; Chomsky, 1995), Ninio suggests that language development crucially involves the learning of dependency relationships between verbs and complements (e.g., direct object, subject, etc.). This means that multiword utterances are built around verbs, which is compatible with the usage-based approach, but implies that children do encode formal grammatical relationships from the very beginning. To support this approach, Ninio reinterprets the tracer experiments described by Tomasello and others (e.g., Tomasello & Brooks, 1998) and proposes that while they do not use abstract syntactic representations, two-and-a-half-year-old children transfer the syntactic structure of known verbs to newly acquired ones (e.g., transitive to intransitive). This proposal is also compatible with data from Conwell and Demuth (2007) showing that three-year-old children can make dative alternation with newly acquired verbs (e.g., I pilked the cup to Peter, I pilked Peter the cup). Thus, according to this approach (here, referred as to the *early-grammar* hypothesis), children need a rudimentary knowledge of grammatical relationships in order to produce multiword utterances. One appealing aspect of this proposal is that it suggests a solution to the perception-production paradox (Naigles, 2002): indeed, children might benefit very early on from the skills they learnt prelinguistically. However, it is important to note that there are important differences in how researchers interpret the above data. Researchers from the generativist tradition such as Conwell and Demuth (2007) and Valian, Solt, and Stewart (2009) take their results as evidence for abstract grammar. On the contrary, Ninio (2006, 2011) proposes that children merely have surface knowledge of grammatical relationships. Therefore, more empirical research is needed to understand how children build early multiword utterances.

This study

The main aim of this study is to explore whether early multiword utterances are built on rudimentary knowledge of grammar (the *early-grammar* hypothesis) or whether alternatively they are built around lexical items (the *late-grammar* hypothesis). From a general perspective, we aim to determine how young children acquire grammar. Data for this study was obtained from a large corpus of spontaneous speech samples of French children in the period in which they acquire grammar (between two and four years of age).

The data analysis of this study consisted of two steps. In the first step, using an automatic part-of-speech tagger (Parisse & Le Normand, 2000a, 2000b), words in the corpus were annotated according to a list of linguistically motivated word classes (see Appendix) which were categorized as either lexical, grammatical or pragmatic; MLU was also obtained for each child. In the second step, multiple regression analyses were performed to examine whether MLU variance could be explained by word diversity within each major category (i.e., lexical, grammatical and pragmatic) or within specific categories (e.g., determiners, personal pronouns, etc.). One important feature of this approach (as opposed for instance to Bates & Carnevale, 1993) is that by exploring lexicon, grammar and MLU separately, multiple regressions make it possible to compare the contributions of lexicon and grammar independently i.e., grammar and MLU are not assumed to be equivalent.

The rationale for using this approach is that if children build their multiword utterances around specific word types (e.g., lexical versus grammatical), increases in such word types should explain variation in MLU. For example, if early multiword utterances are item-based, as defended by the *late-grammar* hypothesis, in order for the child to produce longer utterances she/he will have to learn new predicates; therefore increases in lexical diversity should correlate with MLU. On the contrary, if early multiword utterances have a rudimentary grammatical organization (*early-grammar* hypothesis) increases in the grammatical words used to encode such relationships should

predict an increase in MLU. Two methodological issues require further attention: the validity of MLU and the linguistic coding.

MLU has been used since Brown (1973) as a core measure in language acquisition (e.g., Bornstein, Haynes, Painter, & Genevro, 2000; Meline & Meline, 1981; Miller & Chapman, 1981; Miller & Leadholm, 1992, Rice, Redmond & Hoffman, 2006). Its reliability has often been questioned due to its variability within age-groups (Klee & Fitzgerald, 1985). However, MLU has been shown to be highly reliable for children younger than four years (Rondal, Ghiotto, Bredart & Bachelet, 1987), a period in which it is highly correlated with age and with the development of morphological and syntactic skills (Blake, Quartaro, & Onorati 1993; Miller & Chapman, 1981; Rollins, Snow, & Villett, 1996; Rondal et al., 1987; Scarborough, Rescorla, Tager-Flusberg, Fowler, & Sudhalter, 1991). MLU has also been shown to be sensitive to pragmatic influences, such as differences in situation and discourse context (Bornstein, Painter & Park, 2002; Johnston, 2001). This suggests that MLU is reliable when the context of language production is strictly controlled in children up to the age of four. The data from the present study comply with the restrictions necessary to make MLU reliable. Participants were between two and four years old, and data were obtained using the same contextual situation for all children.

The second relevant methodological issue has to do with the identification of lexical, grammatical and pragmatic word types in the corpus. These three major word classes are generally accepted both in linguistic and in developmental studies. Grammatical and lexical word types are the basic building blocks of linguistic referential units (e.g., noun phrases, simple and complex sentences, etc.), with lexical words providing the basic content (e.g., dog, table, run, etc.), and grammatical words providing more abstract information (e.g., aspect, tense) and the formal architecture of the sentences (e.g., case markers). Pragmatic words are relatively independent of the basic syntactic system. They are used to guide the talker in the interpretation of the intended

meaning of the utterances and they can occur either as free-standing words (e.g., interjections) or in combination with relatively large linguistic units (e.g., discourse markers; Schiffrin, 1987). Note also that the distinction between these three word classes is not straightforward, as several words can be classified in more than one group, e.g., many grammatical words can be used as discourse markers such as *alors* (so). Furthermore there are important differences within classes, e.g., verbs versus nouns. However, it was assumed that this classification would provide an adequate starting point for the present study (see the full list in the Appendix). As noted above, annotation of word-class information was made by means of an automatic computer-based part-of-speech tagger (Parisse & Le Normand, 2000b). This automatic analysis is rooted in a distributional principle. This means that the category of a word depends purely on the structure of the language produced –and especially on lexical information and word context– but not on the meaning of the language produced. Note that this type of analysis resembles to a certain extent the type of knowledge that children have been shown to use in order to make early analyses of oral input (e.g., Shi & Melançon, 2010). Note also that the linguistic annotations used in the corpus, which are commonly used to describe adult language, are not meant to imply that the speaker (i.e., the child) uses this type of syntax (see below Method section).

A secondary aim of this study was to examine whether the emergence of grammar was influenced by age and socioeconomic status (SES) of the family. As regards age, one crucial question was to determine whether MLU predictability remains stable as children's language becomes more complex. Two types of variation might be observed: first, different categories might predict MLU at different times (e.g., lexicon initially, and grammar later on); second, the same categories might predict MLU (e.g., grammar), even if global predictability decreases with time (due to more complex language being less predictable). The first type of change would be compatible with the *late-grammar* hypothesis, as according to this approach lexicon should predict increases in MLU, but after some time grammar might take their place. The second possibility

would be compatible with the *early-grammar* hypothesis. As regards SES, it is well known that this determines the rate of development (Vasilyeva, 2008). However, we may ask if apart from rate of development, there are more qualitative differences among children from different SES groups. If language development is viewed as an input-driven construction, then one might expect similar processes to occur in all children. On the contrary, if language is more dependent on maturation, then differences in SES might have other consequences in language development.

Finally, we aimed to explore whether there is a subset of grammatical categories (e.g., determiners, prepositions, pronouns) that are more closely associated with increases in language complexity. This question is relevant because, as noted above, the list of grammatical words includes various subcategories which are linguistically and cognitively different, and which are also acquired in different time periods. For instance, determiners are acquired earlier than conjunctions (Demuth, 1996, 2006), the latter being linguistically more complex; determiners are also acquired earlier than time or space prepositions, which are conceptually more elaborate (Hickman & Robert, 2006). Thus, by exploring this issue we expected to gain a better understanding of the processes underlying the emergence of grammar.

Three questions are addressed in this study:

1) **Is there a relationship between MLU and the three major categories (lexical, pragmatic and grammatical word types)?** In accordance with perception studies providing evidence that children develop sensitivity to grammatical information before the age of two, it is hypothesized that an increase in the number of grammatical word types should be the best predictor of MLU.

2) **Does the relationship between the three major linguistic categories and MLU vary across age and SES groups?** In agreement with the *early- grammar* approach, it is hypothesized that an increase in grammatical diversity should be the best predictor of MLU for the three age groups. As for SES level, it is hypothesized to have an impact on the rate of development, but not on the relationship between linguistic categories and MLU (i.e., it is the same across SES groups).

3) **Is there a subset of different grammatical categories which is more strongly related to MLU for French language?** Given the evidence that frequency plays a major role in early language development, it is hypothesized that a subset of grammatical words, namely the most frequent ones, should be the best predictors of utterance length.

Method

Participants

312 typically developing children (142 girls and 170 boys) ranging in age from 24 to 48 months participated in this study. Participants were recruited from homes and nurseries in the Paris area, France. Selection criteria were as follows: normal hearing in an auditory screening test, scoring in the normal range on an age-appropriate nonverbal cognitive test (Symbolic Play Test; Lowe & Costello, 1976) and being a monolingual native speaker of French. The participants' SES level was assessed using the classification developed by Desrosières, Goy and Thévenot (1983), taking into account family income, father's occupation and mother's level of education, and categorized as low or median-high.

Corpus

Traditionally, two approaches have been used to assess language samples in preschoolers: one focuses on a sample of 50 consecutive utterances minimum (e.g., Templin 1957; Rondal & Defay, 1978), the other on the speech produced during a specific amount of time (e.g., Crystal, Fletcher, &

Garman, 1976; Tyack & Gottsleban, 1977). We adopted the second approach, with a 20-minute sample-time. This corpus, comprising 104 hours of free interactions, was considered a representative sample of French-speaking children of the relevant age group within each SES level.

Procedure

Each child participated in a dyadic interaction with a familiar adult partner (parent or nursery teacher) either in the child's home, nursery or school. The child and adult were seated at a small table, and the same standardized set of 22 Fisher-Price toys was used with all children: one house with five family members, one dog, four beds, four chairs, two armchairs, two tables, one rocking horse, one stroller, two cars and one staircase. In this conversational context, the children could be expected to engage in talking and sharing experiences (Le Normand, 1986; Le Normand et al., 2008).

Transcription and analysis of recorded language samples

Two trained assistants transcribed the recorded language samples following the transcription and segmentation conventions for spoken French (Rondal et al., 1985; Le Normand, 1986), allowing for the computation of linguistic production as described in the corpus processing system CLAN (Child Language ANalysis; Mac Whinney, 2000). The entire corpus of the children's productions was fully tagged by POS-T, a fully automatic parser developed by Parisse and Le Normand (2000b). The parser is freely available as part of the CLAN program, which can be found on the CHILDES website (http://childes.psy.cmu.edu/morgrams/).

The system has two main components: the MOR analyzer and the POS-T disambiguator. The MOR system automatically creates, for all the words in a transcript, the set of all possible categories for the words. For example, in English, MOR provides two morphological forms, for the word ‘play’: ‘v|play’ (v stands for verb, e.g. ‘I play’) and ‘n|play’ (n stands for noun, e.g. ‘the play’). The function of the POS-T tool is to take into account the context and automatically provide the most

suitable category for this context. Some analysis errors remain and manual checking is necessary in some cases. However, depending on the syntactic complexity of the language to be processed, the error rate, which is in principle about 4%, may be as low as 1%. For French, the set of syntactic categories implemented in MOR and POS-T is large. As noted above, this automatic analysis is rooted in a distributional approach, with categorization being dependent on context and not on the meaning of the language produced. All the syntactic categories used in the current analysis correspond to oral adult language categories. This means that the result of the analysis is described on the basis of ADULT knowledge, but it does NOT mean that the child under study has knowledge of this type of syntax. On the contrary, this provides a reference that allows comparison between children with different levels of grammatical knowledge. Tagging quality was checked by hand, as the corpus is intended to become a reference for future syntactic analyses of children's French language corpora. The effective tagging quality of the present corpus after checking by hand averages 97%. Word types were calculated as the number of different word forms. For instance, in this utterance: *"oh le bébé, maman veut le promener dans le jardin!"* (English: "oh the baby, mummy wants to take him to the garden!") There are nine word types:

- One pragmatic type: *oh*|co;
- Five lexical types: *bébé*|n, *maman*|n:prop, *veut*|v, *promener*|vinf , *jardin*|n;
- Three grammatical types: *le*|det, *le*|pro:obj, *dans*|prep.

Note that the French word "*le*" occurs three times, which correspond to two different word types. So, in this example, the frequency of word type *le*|det is two and the frequency of word type *le*|pro:obj is one .

This study calculated MLU in words. Previous studies have calculated MLU in words (MLUw) or in morphemes (MLUm). While it is often assumed that MLUm is more reliable than MLUw, it has been observed that there is a very high correlation between the two measures both in English (.998, Parker & Brorson, 2005) and in French (.990, Parisse & Le Normand, 2006).

Statistical Analysis

Correlations and hierarchical linear regression analyses were performed using the SAS 9.1 software (CORR, REG, and GLM procedures). In all regressions, MLU was the dependent variable; the number of lexical, pragmatic and grammatical word types, age, and SES, were the independent variables. The distribution of the 312 samples included two SES levels (Low: 141 samples High: 171 samples), and three age-groups (24-30 months: 105 samples; 33-39 months: 110 samples; 42-48 months: 97 samples).

RESULTS

Predicting MLU from Lexical, Pragmatic, and Grammatical word types

Before testing the predictive value of the three major language categories, we conducted a series of correlational analyses. Figures 1a. b. c. show that lexical, pragmatic and grammatical word types were positively correlated with MLU ($r = .76$, .59, and .86 respectively, $p<.001$). When two word types were entered as predictors, lexical and grammatical word types remained significant. By contrast, when lexical word types were partialed out, the effect of pragmatic word types was marginal ($p = .07$). Finally, when the three word types were entered in the regression, grammatical word types remained strongly significant (the greater the number of different grammatical words, the higher the MLU), lexical word types were not significant, and pragmatic word types had a significant negative coefficient. This indicates that when grammatical word types were taken into account, lexical word types were independent of MLU and that the children using more pragmatic word types had the lowest MLU scores.

(Insert Figure 1a b c and Table 1 here)

One potential limitation of this regression analysis is circularity: as the dependent variable (MLU) and the independent variables (word types) were obtained from the same corpus, they might be mathematically related, which would result in a circularity effect. To ensure the independence between MLU and the three categories (lexical, pragmatic and grammatical word types), we first split the corpus into two halves and used one half to compute MLU and the other half to compute word types. We performed the same regression analysis on both subcorpora. The pattern of results

in the whole corpus and in the split analysis was similar for lexical and grammatical word types (i.e., lexical word types were non significant, and grammatical word types strongly significant). In the case of pragmatic words, the pattern of results in the whole corpus was also similar to those of split 2 (i.e., the coefficients were negative and significantly different from 0, $p<.001$) but not to those of split 1 (i.e., the coefficient was also negative, but it was not significantly different from 0, $p=.12$). These results confirmed that MLU was not mathematically related to lexical and grammatical word types. At the same time, the split-half analysis indicated that the observed association of pragmatic words with MLU is less reliable than the association of lexical and grammatical words with MLU.

(Insert Table 2 here)

Secondly, we carried out a Monte-Carlo analysis, generating 10000 random speech samples from normal distributions (with the observed mean and SD) for MLU, and lexical, pragmatic and grammatical word types. Then, we performed a multiple regression on this random corpus. The results of this regression were quite different from those observed in the actual sample: the coefficient of lexical word types was significant and the coefficients of grammatical and pragmatic word types were non significant. Such results argue against a "circularity" problem and spurious results.

Predicting MLU from Age, SES, Lexical, Pragmatic, and Grammatical word types

In order to understand the contribution made by age and SES in predicting MLU, we first examined the correlations between all independent variables: the three major categories were positively correlated with age ($r = .57$, $.44$, and $.67$ respectively, $p<.001$), and negatively correlated with SES ($r =-.37$, $-.30$, and $-.35$ respectively, $p<.001$ in all three cases). The three major categories were strongly intercorrelated (lexical-pragmatic: $r =.82$, lexical-grammatical: $r =.92$, pragmatic-grammatical: $r =.80$, $p<.001$), and these correlations remained high when age and SES were partialed out ($r =.74$, $.85$, and $.73$, respectively, $p<.001$).

Univariate regression analysis (Table 3) showed that all five predictors were associated with significant changes: lexical, pragmatic and grammatical word types accounted for 58%, 35% and 75% of the variance in MLU respectively ($p < .001$ in the three cases). Age and SES accounted for 52% and 13% of the variance in MLU respectively ($p < .001$ in both cases). Multivariate regression analysis (Table 4) showed that when age and SES factors were entered first as predictors, these two variables accounted for 63% of the variance in MLU, $p < .001$ (step1). Their effect remained significant when lexical, pragmatic and grammatical word types were entered as predictors. The regression accounted for 82% of the variance in MLU, with all independent variables significant ($p < .0001$) except for lexical word types (step2), which further confirms the result of previous analyses with the three linguistic categories. Finally, when interactions between age, SES and word types were entered, the regression accounted for 75% of the variance in MLU with lexical word types; 68% with pragmatic word types; and 80% with grammatical word types (all $p<.001$). All interactions between age and word types were significant, but interactions between SES and word types were not significant. This indicates that the correlations between MLU and word types vary across age groups, but not across SES groups (step3).

(Insert Tables 3 and 4 here)

Predicting MLU from Lexical, Pragmatic, and Grammatical Word Types by Age groups

Because the interactions between age and the three major linguistic categories were significant, we performed another set of analyses across the three age groups. An ANOVA of SES, age on MLU showed that MLU increased with age ($F (1, 310) = 202.6$, $p<.001$) and was lower in children from low SES than in children from high SES ($F (1, 310) = 90.7$, $p <.001$) (see Figure 2). There were no statistically significant interactions between age and SES ($F (2,306) = .96$, $p= .38$). Subsequently, we performed regression analyses for the three age groups (see Table 5). The differences were found to be significant for grammatical word types and non significant for lexical word types in the three age groups. As expected, there was a negative and significant coefficient in the youngest and the oldest group for pragmatic words, but this was marginally significant ($p<.06$) in the intermediate

age group (i.e., 33-39 months of age). Word types accounted for 80% of the variance in MLU for children aged 24-30 months, 57% for children aged 33-39 months and 49% for children aged 42-48 months.

(Insert Figure 2 and Table 5 here)

Predicting MLU from Grammatical Categories

Because grammatical word types were the best predictors of MLU, we performed a correlational analysis of the 18 grammatical categories with MLU. As shown in Figures 3a, b, c the highest correlations were observed for subject personal pronouns (r =.81), prepositions (r =.79), and determiners (r =.75). These correlations showed that personal pronouns accounted for 66% of the variance in MLU, prepositions for 62%, and determiners for 56%. Multivariate regression analysis showed that the above three grammatical word types accounted for 73% of variance in MLU with all three coefficients positive and significantly different from 0 ($p<.0001$).

(Insert Figures 3 a.b.c. and Tables 6 and 7 here)

DISCUSSION

The main aim of this study was to explore how children build early multiword utterances and acquire grammar. Two alternative proposals have been explored, referred to in this study as the *late-* and *early- grammar* hypotheses respectively. According to the first hypothesis (i.e., *late-grammar*), early multiword utterances are organized around lexical items. According to the second hypothesis (i.e., *early-grammar*), by the time children begin to produce multiword utterances, they have some rudimentary knowledge of grammar, enabling them to transfer the formal grammatical features of known predicates to the new verbs they learn. In order to evaluate these two hypotheses, we analyzed a large annotated corpus of French children aged two to four. Specifically, we examined whether increases in language complexity (as measured by MLU) could be predicted on the basis of lexical, pragmatic and grammatical diversity (word types). We postulated that if, as proposed by Tomasello (2000), early multiword utterances are lexically driven, then increases in

lexical diversity should predict increases in MLU. On the contrary, if, as proposed by Ninio (2006), early multiword utterances are organized grammatically, increases in the number of grammatical word types should predict increases in MLU. In order to arrive at a better understanding of the processes underlying early language development, two further questions were examined. First, we analyzed whether age and Socio-Economic Status (SES) influenced results. Second, we inquired whether there was a subset of grammatical categories that is more strongly associated with MLU increases.

Lexical, Pragmatic and Grammatical Word Types and MLU

One important finding in this corpus study is the different patterns of association derived from simple regressions and multiple regressions. On simple regressions, the correlations of lexical, pragmatic and grammatical word types with MLU are .76, .59, .86 respectively. When the three major categories are submitted to a multiple regression analysis, the pattern of results is very different: the number of grammatical word types shows a positive correlation with MLU, the number of pragmatic word types shows a negative correlation, and the number of lexical word types shows no effect on MLU. Thus, even if language development involves a series of changes that affect all aspects of the language system, it is grammatical word types that have the most significant effect on MLU growth. The children's inventory of 18 grammatical types accounts for 75% of the variance in MLU. This result confirms our hypothesis that grammatical types are the best predictor of MLU and leads to the conclusion that children use grammar productively from a very early age.

The number of lexical types correlated with MLU, which is in accordance with previous studies (Devescovi, Caselli, Marchione, Pasqualetti, Reilly & Bates, 2005). However, it was not a predictor of MLU when grammatical and pragmatic types were taken into account. This suggests that in this period, an increase in lexical diversity does not have a direct impact on language development as measured by MLU. Thus, this result is not in agreement with the *late-grammar* hypothesis, nor with the Verb Island hypothesis, according to which children's early multiword utterances are organized

around specific predicates. It could be thought however that the association between lexicon and global complexity might still be relevant. It is quite possible that children aged between two and four develop other features of the lexicon, not captured by our part-of speech-tagger (e.g., elaborate lexical representations such as those described by Pustejovsky, Bergler & Anick, 1993).

The negative coefficient of pragmatic word types to MLU suggests that these word types occur mostly in short utterances. This is not surprising: pragmatic words are fixed forms and therefore are not constrained by any morpho-syntactic markings. This makes these expressions easy to use by less advanced children. Furthermore, producing a variety of pragmatic words may be due to an inability to use other language resources (i.e., grammar). As children begin to use more grammatical types, the need to use pragmatic words decreases, which, together with the fact that these expressions can lead to very short utterances, explains the impact on MLU. We know that pragmatic knowledge plays a role in early language acquisition as Veneziano, (1999; 2001); Clark and Amaral, (2010); Herr-Israel and McCune, (2011) pointed out, but this was partially captured by our distributional part-of speech-tagger.

In order to ensure that these results are reliable, Monte-Carlo and split-half analyses were performed. Monte-Carlo analysis showed that the associations between MLU and word types calculated from a random corpus were different from those obtained from the actual corpus, which argues against circularity. The split-half analysis confirmed the reliability of the associations of MLU with lexical and grammatical word types, but not with pragmatic words. The low reliability of the association between pragmatic words and MLU might be attributable to the context-sensitive nature of pragmatic words, which make them distributionally less predictable than lexical and grammatical words. Furthermore, as noted above, pragmatic words include two very different subtypes. On the one hand, there is a large number of pragmatic words which tend to occur as free-standing words. (e.g., interjections, onomatopoeias). On the other hand, many pragmatic words and expressions can be used to organize linguistic units at the narrative level (e.g., discourse markers,

Schiffrin, 1987), for which they should be associated with high MLU. As our annotation system did not explicitly separate these two different subtypes, it was not possible to examine the role of different pragmatic categories independently, as we did in the case of grammatical words. However, some pragmatic words such as là (*there*), voilà (*here is*), encore (*more*), are amongst the most frequent ones in our corpus (see Appendix). In sum, the heterogeneous nature of pragmatic words may explain why their association with MLU was not fully reliable. Future studies using a finer-grained annotation system for pragmatic words should explore to what extent children use pragmatic words to build long multiword utterances.

Age, SES, major linguistic categories and MLU

Another important finding in this corpus study is that while both age and SES contribute significantly to MLU, their relationship to the three linguistic categories is qualitatively different. The contribution of age and SES to MLU was confirmed both in univariate and in multivariate analyses. And their effect remained significant when lexical, pragmatic and grammatical word types were entered as predictors. However, results for age and SES factors were clearly different in terms of the interactions with word types.

As regards SES, there was no significant interaction between SES and linguistic categories. This means that the relationships between linguistic categories and MLU are not influenced significantly by SES level. In other words, input is crucial because it accelerates or decelerates developmental processes, but the processes themselves remain the same. This result is important because it confirms that even if grammar is the best predictor of MLU, this does not mean that some innate processes are taking place but rather that language development is input-driven.

As regards age, the interaction with linguistic categories was significant, which shows that the correlations between MLU and word types vary across age groups. Multiple regression analyses across the three age groups provided further details about these differences: (i) the relationships

between word types and MLU for the three groups were almost similar to those observed in the full corpus; the only exception was pragmatic word types, in which the association was marginally significant in the 33-39 months group; (ii) The percentage of variance of linguistic categories decreased across age groups (from 80% to 49%). In other words, the association between MLU and the grammatical categories is the strongest in the youngest group. These results confirm our prediction that grammatical words are the best predictor of MLU even in the youngest children, which provides further support for the *early-grammar* hypothesis.

Grammatical categories and MLU for French

Three grammatical types (subject-pronoun, determiners and prepositions) were the best predictors of MLU for French and were almost sufficient to determine initial grammatical development. This confirmed our hypothesis that some grammatical word classes are more strongly associated with MLU. These results are in agreement with our previous studies (Le Normand, Parisse & Dellatolas, 2010) showing that third person singular pronouns "*il/elle/on*" (he/she/impersonal pronoun) and first person singular "*je*" (I) were the best predictors of MLU: third person singular was used by 100% of children at age two and nine months and "*je*" (English: I) was used by 97% of children at age three and nine months. At age two, 55% of children having an average MLU of 1.3 omitted all personal pronouns whereas 45% of children of the same age having an average MLU of 1.7 started to produce subject pronouns. This suggests that children gradually start using these grammatical categories from very early on.

It could be argued why this should be the case for these three categories, and not for the remaining 15 grammatical categories. According to Valian et al. (2009), given that these function words are very abstract, children should learn them very late, for which the only explanation is that grammatical development is guided by innate grammatical representations. It may be, however, that some features of these words make them easier to learn. Three features of these grammatical words support this possibility. First, all determiners, and at least some subject pronouns (e.g., *je*, (I) *tu*

(you) are clitic words. Clitics are function words, which are prosodically constrained (i.e., they must be produced in the context of a lexical word), with a highly predictable distribution. Similarly, some prepositions (e.g., *de*, (from) *à*, (to) etc.) are most frequently used as formal case markers that identify predicate-argument relationships (e.g., *Il est allé de la maison à l'école*, He went from home to school) in which they are also highly constrained and predictable. Second, determiners, prepositions and subject pronouns are among the most frequent words in adult French, which is highly relevant from a language learning perspective. Finally, these function words, and particularly determiners and personal pronouns, do not convey any conceptually complex content. Thus, if one supports Naigles' claim, (2002) that "form is easy" but "meaning is hard", it is not surprising that children learn function words more easily.

How do children acquire early grammar?

The traditional debate between constructivist and innatist approaches to language acquisition has been transformed more recently into a debate between *late- versus early-grammar* or into lexically- versus grammatically-driven language development. The results of this study provide strong support for the latter hypothesis, and offer some clues as to the processes that might underlie the emergence of grammar. To summarize, three main results of our study have provided support for the *early-grammar* hypothesis: (i) grammatical diversity is the best predictor of general language complexity between two and four years of age, in contrast to lexical diversity; (ii) SES level may accelerate language development but the process remains the same; and (iii) more frequent and prosodically constrained word types are the best predictors of language development in French.

It could be argued that one should not infer from the fact that children produce grammatical words that they already have some grammatical knowledge. However, several sources of evidence contradict such an interpretation of our data. First, the results are compatible with data from perception studies showing that younger children use function words to process auditory input. Importantly, in general terms the same word types appear in perception studies and in this study,

supporting the idea that children do have surface knowledge of function words from very early on. Second, the contrast between a highly predictable MLU in less advanced children, and a less predictable MLU in more advanced children suggests that what all children have in common is a core language which is characterized precisely by the systematic use of basic grammatical words, which suggests that these words must be easily learned. Finally, and contrary to what has often been observed in the early language of several atypical populations (e.g., deaf or SLI children), the children in this study very rarely made grammatical errors (except for omissions), which further supports the proposal that children use grammar productively from very early on. Thus, we may conclude that these results reflect actual knowledge of grammar rather than an apparent statistical effect.

The evidence that children use grammar in early language has been interpreted in two different ways until now. For some researchers (Conwell & Demuth, 2007; Valian et al., 2009) it confirms that children are born with innate grammatical categories. For others (Ninio, 2006, 2011) this merely confirms the fact that children make use of formal grammatical relationships (while at the conceptual level they may only know the specific items). Thus, for one view, two-year-old children have full knowledge of the abstract grammatical categories, while for the second, children may only have surface knowledge of grammatical organization. One of the problems of the abstract category interpretation is that it is based on distributional information (e.g., determiner-noun overlapping in Valian et al.'s 2009 study). That is to say, evidence of any abstract categories is very indirect. It seems more adequate to make a less speculative interpretation of the data.

Such an interpretation of our results is compatible with the view of grammatical development as a long and slow process in which both social experience and cognitive skills are basic pillars. An important part of that process is distributional learning. As noted above, some of the basic features of a number of grammatical words (frequency, distributional restrictions, and formal nature) explain why they are easier to learn. At the same time, there is an interesting overlap between these words

and the ones children use to process early speech, which most probably may facilitate learning even more. Once children have these basic formal structures they can use them to make more elaborate form-meaning mapping. For example, the personal pronoun data show that children need some time to acquire this set of grammatical items, and that acquisition is an ongoing process which takes place gradually as shown in this study (see Figure 3a). Initially, personal pronouns occur as fillers, which means that children know the position of these particles despite having no or very limited grammatical knowledge. Later on, the paradigm of different personal pronouns emerges. This gradual progression in grammatical development is compatible with the fact that other general skills might have to be acquired before the full set of personal pronouns can be used. For instance, a basic contrast between *je/il* (English: I/he) may be recognized by merely establishing a contrast between self/others, a skill which children younger than two-year-old can clearly make. However, the subtle distinctions associated to the full set of personal pronouns (e.g., anaphoric reference, polite forms, etc.) are beyond the maturity of two-year-old children. A similar argument might be applied in the case of determiners or prepositions, the other two categories that are highly predictive. Considering that grammar is the formalization of a huge variety of abstract concepts (place, movement, anaphora, deixis, etc.), which the child acquires in interaction with others, a consequence of this is that grammar is acquired gradually.

In sum, in order to build the cognitive architecture of language, children start by using the formal distributional features of language from very early on in language production. More specifically, they use the most basic grammatical words, which happen to be the easiest to learn and are sufficient to encode basic grammatical relationships. Then, on the basis of these formal structures and various other cognitive skills, children can learn to encode increasingly complex form-meaning relationships until they arrive at adult language learning. Being input driven and based on social interaction, the rate of the process depends on the actual social context. However, in very broad terms the process of building the cognitive architecture of language is identical in all children (i.e., form precedes meaning), for which social context differences have no qualitative effect. Finally, the

specific stones that children use to build the cognitive architecture of language should depend on the particular language they are learning.

Our results show the value of using a large corpus to answer theoretical questions in developmental studies. This cross-sectional approach cannot provide a full description of language acquisition, and especially of qualitative information that is not annotated (e.g., complex lexical representations). However, it is particularly suitable to explore the use that children make of distributional regularities. At the same time, it shows the importance of annotation tools. It seems evident that the specific results obtained in this study reflect the annotations used. The comparison of results obtained with different annotation criteria might provide interesting information regarding grammatical development. For that, we need not only large databases, but also efficient and dynamic coding systems.

Conclusion

This study used a part-of-speech-tagged corpus as a model of young children's language. Using hierarchical regression analyses, we have explored the relation of MLU to lexical, pragmatic and grammatical diversity (number of word types), and shown that MLU growth in French depends primarily on the number of grammatical word types and secondarily on the number of pragmatic word types. The results confirm the value of our methodological approach and suggest several directions for future research. First, it seems highly relevant to explore whether the same correlations observed in this study hold across atypical populations (e.g., SLI, Cochlear Implant, etc.). By exploring such correlations we might determine whether these atypical children can make use of distributional regularities or whether they use alternative routes to build complex utterances. Second, future studies should explore whether the general pattern observed for French is also present in children from other linguistic backgrounds. Comparison of these results across languages would help to generalize our findings as a language-independent phenomenon. Finally, future

studies should explore the role of pragmatic word types, and particularly of discourse markers, to build long utterances.

Acknowledgements: We would like to thank one reviewer for constructive comments on earlier versions of the manuscript and Zacharie Esmili for great help with the Monte Carlo analysis.

References

Akhtar, N., & Tomasello, M. (1997). Young children's productivity with word order and verb morphology. *Developmental Psychology*, 33(6), 952-965.

Bates, E., & Carnevale, G. F. (1993). New directions in research on language development. *Developmental Review, 13*(4), 436-470.

Blake, J., Quartaro, G., & Onorati, S. (1993). Evaluating quantitative measures of grammatical complexity in spontaneous speech samples. *Journal of Child Language, 20*(1), 139-152.

Bornstein, M. H., Haynes, O. M., Painter, K. M., & Genevro, J. L. (2000). Child language with mother and with stranger at home and in the laboratory: a methodological study. *Journal of Child Language, 27*(2), 407-420.

Bornstein, M. H., Painter, K. M., & Park, J. (2002). Naturalistic language sampling in typically developing children. *Journal of Child Language, 29*(3), 687-699. doi: 10.1017/s03050090200524x

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

Brown, R.W. & Berko J (1960): Word association and the acquisition of grammar, *Child Development* 31, 1-14.

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

Brown, R.W. & Fraser, C. (1963). *The acquisition of syntax*. In C. Cofer & B. Musgrave (eds.), Verbal behavior and learning: Problems and processes. New York: McGraw Hill. 158-201.

Chomsky, N. (1995). The Minimalist Program. Cambridge, Mass: MIT Press.

Clark, E. V., & Amaral, P. M. (2010). Children build on pragmatic information in language acquisition. *Language & Linguistic Compass* 4(7), 445-457.

Conwell, E., & Demuth, K. (2007). Early syntactic productivity: Evidence from dative shift. *Cognition*, 103(2), 163-179. doi:10.1016/j.cognition.2006.03.003

Crystal, D., Fletcher, P., & Garman, M. (1976). *The grammatical analysis of language disability. A procedure for assessment and remediation.* NewYork: Elsevier.

Demuth, K. 1996. The Prosodic Structure of Early Words. In J. Morgan and K. Demuth (eds.), *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. Mahwah, N.J.: Lawrence Erlbaum Associates, 171-184

Demuth, K. 2006. Crosslinguistic perspectives on the development of prosodic words. *Language and Speech* 49 (2), 129-297.

Desrosières, A., Goy, A., & Thévenot, L. (1983). L'identité sociale dans le travail statistique: la nouvelle nomenclature des professions et catégories socioprofessionnelles. *Economie et Statistique*, 152, 55–81.

Devescovi, A., Caselli, M. C., Marchione, D., Pasqualetti, P., Reilly, J., & Bates, E. (2005). A Crosslinguistic Study of the Relationship Between Grammar and Lexical Development. *Journal of Child Language*, 32(04), 759-786.

Hallé, P. A., Durand, C., & de Boysson-Bardies, B. (2008). Do 11-month-old French infants process articles? *Language and Speech*, 51, 23-44.

Herr-Israel, E., & McCune, L. (2009). Successive single-word utterances and use of conversational input: a pre-syntactic route to multiword utterances. *Journal of Child Language,* 38(01), 166-180. doi:10.1017/S0305000909990237

Hickmann, M., & Robert, S. (2006). *Space in Languages: Linguistic Systems and Cognitive Categories.* Philadelphia, PA: John Benjamins

Johnston, J. R. (2001). An alternate MLU calculation: Magnitude and variability of effects. *Journal of Speech Language and Hearing Research, 44*(1), 156-164.

Kaplan, R., & Bresnan J. (1982) *Lexical-functional grammar: A formal system for grammatical representation*. In Joan Bresnan, editor, The Mental Representation of Grammatical Relations, pages 173–281. The MIT Press, Cambridge, Mass.

Klee, T., & Fitzgerald M. D. (1985). The relation between grammatical development and mean length of utterance in morphemes. *Journal of Child Language,* 12, 251-69.

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*(2), 121-134.

Le Normand, M. T., Parisse, C., & Cohen, H. (2008). Lexical diversity and productivity in French preschooolers: developmental, gender and sociocultural factors. *Clinical Linguistics & Phonetics, 22*(1), 47-58. doi: 10.1080/02699200701669945

Le Normand, M. T., Parisse, C., & Dellatolas G. (2010). *Third personal singular pronoun is a developmental marker of early grammar: the case of French*, in COSTA33,

Let the children speak, Learning of critical language skills across 28 languages, 22-24 January, London.

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

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20(3), 481-507. doi:10.1515/COGL.2009.022

Lowe, M., & Costello, A. J. (1976). *Manual for the symbolic play test*, experimental edition. Windsor: NFER.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*, Third edition. Volume I: Transcription format and programs. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2005) The emergence of grammar from perspective, in D. Pecher and R.A. Zwaan, (Eds), *The Grounding of Cognition: The Role of Perception and Action in Memory, Language, and Thinking,* Mahwah, NJ: Lawrence Erlbaum.

Meline, T. J., & Meline, N. C. (1981). Normal variation and prediction of mean length of utterance from chronological age. *Perceptual and Motor Skills*, 53, 376-378.

Miller, J.F., & Chapman, R.S. (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech and Hearing Research*, 24, 154–161.

Miller, J., & Leadholm, B. (1992). *Language Sample Analysis Guide: The Wisconsin Guide for the Identification and Description of Language Impairment in Children.* Wisconsin Department of Public Instruction.

Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1), 91-117. doi:10.1016/S0010-0277(03)00140-9

Naigles, L. R. (2002). Form is easy, meaning is hard: resolving a paradox in early child language. *Cognition*, *86*(2), 157-199.

Nazzi, T., Barrière, I., Goyet, L., Kresh, S., & Legendre, G. (2011). Tracking irregular morphophonological dependencies in natural language: Evidence from the acquisition of subject-verb agreement in French. *Cognition*, 120(1), 119-135. doi:10.1016/j.cognition.2011.03.004

Ninio A. (2006). *Language and the learning curve: a new theory of syntactic development.* Oxford University Press, Oxford.

Ninio, A. (2011). *Syntactic development, its input and output* Oxford University Press, Oxford

Parker, M. D., & Brorson, K. (2005). A comparative study between mean length of utterance in morphemes (MLUm) and mean length of utterance in words (MLUw). *First Language*, 25, 365-376.

Parisse, C., & Le Normand, M. T. (2000a). How children build their morphosyntax: the case of French. *Journal of Child Language*, 27(2), 267-292.

Parisse, C., & Le Normand, M. T. (2000b). Automatic disambiguation of morphosyntax in spoken language corpora. *Behavior Research Methods Instruments & Computers, 32*(3), 468-481.

Parisse, C., & Le Normand, M.T. (2006) Une méthode pour évaluer la production du langage spontané chez l'enfant de 2 à 4 ans, *Glossa*, 97, 20-41

Pollard, C., & I.A. Sag (1994), *Head-driven Phrase Structure Grammar*, Chicago: University of Chicago Press

Pustejovsky, J., Bergler, S., & Anick, P. (1993). Lexical Semantics Techniques for Corpus Analysis. *Computational Linguistics*, 19, 331-358.

Rice, M. L., Redmond, S. M., & Hoffman, L. (2006). Mean length of utterance in children with specific language impairment and in younger control children shows concurrent validity and stable and parallel growth trajectories. *Journal of Speech, Language, and Hearing Research*, 49, 793–808.

Rollins, P. R., Snow, C. E., & Willett, J. B. (1996). Predictors of MLU: Semantic and morphological developments. *First Language*, 16, 243-259.

Rondal, J. A., & Defay, D. (1978). Reliability of mean length of utterance as a function of sample size in early language development. *Journal of Genetic Psychology*, 133, 305–306.

Rondal, J. A., Bachelet, J.-F., & Pérée, F. (1985). Analyse du langage et des interactions verbales adulte-enfant. *Bulletin d'Audiophonologie*, 5, 507–535.

Rondal, J. A., Ghiotto, M., Bredart, S., & Bachelet, J. (1987). Age-relation, reliability, and grammatical validity of measures of utterance length. *Journal of Child Language*, 14, 433-446.

Scarborough, H. S., Rescorla, L., Tager-Flusberg, H., Fowler, A. E., & Sudhalter, V. (1991). The relation of utterance length to grammatical complexity in normal and language-disordered groups. *Applied Psycholinguistics, 12*, 23-45.

Sebastián-Gallés, N. (2007). Biased to learn language. *Developmental Science*, 10(6), 713-718. doi:10.1111/j.1467-7687.2007.00649.x

Shi, R, Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, 72(2), B11-21.

Shi, R., & Melançon, A. (2010). Syntactic Categorization in French-Learning Infants. *Infancy*, 15(5), 517-533. doi:10.1111/j.1532-7078.2009.00022.x

Schiffrin, D. (1987). Discourse Markers. Cambridge: Cambridge University Press.

Templin, M. C. (1957). *Certain language skills in children: Their development and interrelationships*. Child Welfare Monograph. N.26. Minnesota, MN: University of Minnesota Press.

Tomasello, M, & Brooks, P. J. (1998). Young children's earliest transitive and intransitive constructions. *Cognitive Linguistics*, 9(4), 379-396. doi:10.1515/cogl.1998.9.4.379

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, 74(3), 209-253.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Tyack, D., & Gottsleban, R. (1977). *Language sampling, Analysis and training.* Palo Alto, CA: Consulting Psychologists Press.

Valian, V., Solt, S., & Stewart, J. (2009). Abstract categories or limited-scope formulae? The case of children's determiners. Journal of Child Language, 36(4), 743-778. doi:10.1017/S0305000908009082

Valian, V. (1986). Syntactic categories in the speech of young children. *Developmental Psychology* 22, 562–79.

van Heugten, M., & Johnson, E. K. (2010). Linking infants' distributional learning abilities to natural language acquisition. *Journal of Memory and Language*, 63(2), 197-209. doi:10.1016/j.jml.2010.04.001

Vasilyeva, M., Waterfall, H., & Huttenlocher, J. (2008). Emergence of syntax: commonalities and differences across children. Developmental Science, 11(1), 84-97. doi:10.1111/j.1467-7687.2007.00656.x

Veneziano, E. (1999). Early lexical, morphological and syntactic development in French : Some complex relations. *International Journal of Bilingualism*, 3 (2), 183-217.

Veneziano, E. (2001). The importance of studying filler-producing children. *Journal of Child Language*, 28, 275-278.