Semantic discrimination of Noun/Verb categories in French children aged 1;6 to 2;11

Christophe Parisse, Caroline Rossi

Semantic discrimination of noun-verb categories in French children aged 1;6 to 2;11

Christophe Parisse[1] and Caroline Rossi[2]

1: Modyco, Inserm, CNRS & Université Paris Ouest Nanterre la Défense, France
2: Université Grenoble Alpes, [ILCEA], F38040

1. Introduction

The knowledge of word categories is fundamental to the use of language by human beings (Lakoff, 1987:5; Tomasello, 2003:4). Without it, it would be impossible to create new linguistic forms and meanings. However, categorization is anything but a straightforward phenomenon, and it should be expected that at least some linguistic categories will have a complex structure and fuzzy boundaries (Lakoff, 1987:154). The aim of the Polycat project was to determine whether there is a clear-cut difference between linguistic categories- especially major ones such as nouns and verbs- or whether there is a continuum between categories so that such things as polycategorial elements (especially elements that can be used as nouns or verbs with an identical form, with no derivation from one to the other) should be found in some languages.

Language acquisition is particularly interesting to study in this context. Children might make no difference between linguistic categories as they start learning their mother tongue and progressively acquire those differences, or their early linguistic productions might demonstrate clear-cut differences. In the latter case, the use of polycategorial elements is possible but would be achieved only by learning the characteristics of a specific language, which children might do in the course of language acquisition. In the former case, children's early use of language might be said to hint at a universal ability to use polycategorial elements, and the use of specific and less versatile elements would result from language acquisition.

Although these issues have already been widely studied, from the point of view of the grammatical and semantic development of the child's language (see especially seminal studies of Bloom, 1970; Brown, 1973; Braine, 1976; Maratsos, 1976, for a more grammar-oriented research, or from a more semantic oriented approach see especially Nelson, 1973), the debate about the emergence of grammatical categories is still open. Indeed, children's verbal knowledge and behavior only partly resemble those of the adults, and with or without a Continuity Assumption (Pinker, 1984), the road to target categories remains a puzzle. What is the nature and range of children's categories? How do children come to grasp the categories of adult language? In order to assess these questions, we studied two basic linguistic categories --nouns and verbs-- and their emergence in three young French children. A first fundamental question is that of category boundaries: is there a clear-cut difference between nouns and verbs, and if so, does it emerge during development or is it present from the production of the very first words? A second, related question is how children acquire the knowledge of both categories and the relationship between them.

To the linguist, the age-old distinction between noun and verb is inevitably complex (Black and Chiat, 2003) and implies semantic classes, phonological properties, as well as grammatical notions such as the distributional and morphological properties of words (e.g. Bassano, 2000; Veneziano, 2003, for extensive work about these properties in French). But how are all these aspects involved in early child language?

In keeping with developmental approaches of categorization, this study tries to determine whether noun and verb categories are different in early child language vs. what is expected in the adult standard lexicon. Based on studies with English-speaking children, Clark (1982:395) suggested that young children might use nouns for lack of the appropriate, specific verbs, e.g. using *car* and *foot* instead of *drive* and *kick*, respectively, but found no evidence of such overextensions. Indeed, studies on the English and French language acquisition (Bernal, Lidz, Millotte & Christophe, 2007; Veneziano, Parisse, & Delacour, 2010) show that nouns and verbs are not mistaken for one another in child speech, and it does appear that children differentiate nouns and verbs very early in comprehension if the words include clear morphosyntactic markings (at least by age two -see e.g. Veneziano, Parisse, & Delacour, 2010). Could children's early differentiation of nouns and verbs be accounted for on the basis of morpho-syntax alone?

Although morphosyntactic markings might be sufficient for children to develop their syntactic categories, an additional element of explanation is that noun and verb semantics may form distinctive sets that children could rely on, in coordination from the information provided by morphosyntactic markings that are available in their own language. Our hypothesis is that there could exist such knowledge, and that it could follow a developmental pattern, as is the case for knowledge about morphosyntactic markings. It does not amount to denying the use of syntactic information. Rather, it implies that semantic properties could be used, even in the absence of morphosyntactic markings, as earlier and more fundamental cues in the acquisition of noun and verb categories. If this is the case, then lexical words would have semantic properties that correlate with morphosyntactic ones.

Mis en forme: Anglais (E.U.)

In Rossi and Parisse (2012), we addressed these two questions. We studied the development of the semantic characteristics of nouns and verbs in three French-speaking children from age 1;6 to 2;6, by coding for six semantic dimensions (corresponding to basic, real world properties) each time a noun or a verb was produced.[1]

It was expected that, if any of the semantic dimensions were a potential candidate to help children characterize nouns and verbs, there would be developmental trends showing that children learned to rely on this dimension. However, no dimension came out clearly enough in our results, which therefore did not seem to provide evidence for the development of semantic characterization of nouns and verbs by the children.

A major limitation of the study was that it started from the premise that any one semantic dimension coded might be enough to help children learn to differentiate nouns and verbs. If this were the case, then a clear differentiation between nouns and verbs would probably have been demonstrated long ago. Besides, linguistic dimensions tend to be multiple and involve multiple parameters. Can nouns and verbs in French be distinguished on the basis of a network of semantic dimensions (or real word properties) and would this differentiation help children understand how these categories are organized in French?

In order to answer this question, a second set of analyses was performed, which is presented below. It is based on multidimensional statistical analysis and presents a more reasonable approximation of what real linguistic processes might imply/be implied. Our hypothesis is that multidimensional statistical analysis will provide clues into noun and verb differentiation that unidimensional statistical analyses used in Rossi and Parisse (2012) could not reveal.

In what follows, we present the theoretical background of our work (sections 2-4), then the children's data (section 5) and the results of our previous study (section 6). Next, we

[1]All ages below are represented as *year*;*months*.*day* – e.g. 2;4.15 = 2 years 4 months 15 days. Day is optional, the other values are mandatory

explain the principle behind the statistical tools that we used in the present analyses (section 7). Finally, we describe the results of this study and discuss the nature of semantic categorizing in young children's language development accordingly (sections 8-11).

2. Learning semantic and syntactic categories

Language productivity is a basic feature of language (Hockett, 1960). As a result of language productivity, it is possible to create new associations between forms and meanings, and to combine forms or meanings in new ways. In this paper, our use of the concepts of form and meaning follows the principles described in Goldberg (1995), Croft and Cruse (2004), and shared across the Cognitive Grammar framework: "All levels of description are understood to involve pairings of form with semantic or discourse function", Goldberg (2003:219). When we refer to 'forms', we refer to structural properties of what is uttered (oral, signed, or written forms). We will refer to 'syntax' as capturing the regularities and generalizations of the organization of phonological and lexical forms. By 'meaning', we refer to semantic functions only, following Croft & Cruse's definition of meaning: "all of the conventionalized aspects of a construction's function, which may include not only properties of the situation described by the utterance but also properties of the discourse in which the utterance is found (…) and of the pragmatic situation of the interlocutors", Croft and Cruse (2004:258). In many ways, this corresponds loosely to the meaning conveyed through language (Croft & Cruse, 2004). In our paper, we focus on a short inventory of semantic dimensions, selected as appropriate for coding children's first words. Most of what we will describe is based on our observations of meaning in child language corpora, however, it may be equally relevant to an interactional or pragmatic analysis. The term 'grammatical categories' will refer to categories that are defined by both syntactic and semantic information.

Although it is difficult and not the goal of the present paper to know how human beings understand these principles and implement them mentally, it is possible to study the characteristics and regularities of their use. Learning and using generalizations about forms and meanings and their relationships allow humans to be creative in their language use. This is true for every human being, from adults with a mature knowledge of language to young children acquiring their mother tongue: syntax and semantics are theoretical notions that go hand in hand throughout language development. However, their evolutions and possible interactions are not fully understood.

The use of word categories, and especially their productive use (i.e. to guess the meaning of word forms heard for the first time, to change the meaning of a word by changing its category), implies being able to characterize both the forms and the meaning of the elements of the categories (see e.g. Olguin and Tomasello, 1993; Tomasello and Olguin, 1993). Classification of forms is not always easy to infer, as many words can have the same forms in both the noun and verb categories. For example, in English, verbs clearly differ from nouns when they are in the past form because of the –ed suffix, mostly used for verbs, but all other verb forms can correspond to several categories (see e.g. Zingeser and Berndt, 1990:16). This difficulty also applies to French, as we will see below. Classification is even more problematic in languages such as Mandarin where the nouns and verbs do not have any morphological variations. And yet, in all these cases, the grammatical context, such as for example preverbal pronouns in English or French verbs, prenominal determiners in English or French, post verbal aspect markers vs. prenominal classifiers in Mandarin, allows speakers to determine the grammatical category of a word, from the point of view of 'form'. This has been confirmed in French where controlled experiments have shown that the sole syntactic

context of the word can be sufficient for children to choose between a noun and a verb interpretation (Bernal, Lidz, Millotte, & Christophe, 2007; Veneziano et al., 2010).

Determining the category of a word according to regularities of form is not sufficient for using language productively. To do this, it is also necessary to learn to categorize the word according to regularities of meanings (i.e. semantic, pragmatic, and discourse functions). It is the regularities of form/meaning associations that allow for the creation of new words, and enable quick learning of the meaning and use of a new word (Slobin, 2001:442).

Children and adults must have a means to determine functional categories so that they will be able to generalize form/ meaning associations. This is true for all linguistic theories, from generative to functional linguistics. One major problem is that there are no obvious characteristics for categorizing meanings. This does not prevent children from using verbs and nouns early on, without confusing one form with another.

(1) *Madeleine* 2;4.15[2].

CHI: On lit sur mon lit ! [We read on my bed!]

Here Madeleine uses and clearly distinguishes two homophones: the verb *lit* (read), and the noun *lit* (bed) despite the phonological identity, /li/.

However, as one- to two-year-olds utter their first words, they start by producing forms that are devoid of syntactic marking, or incorporate marks that are ambiguous in the child's input and sometimes also in the child's own production (for example, '/la/' that corresponds to a noun marker (determiner), a verb marker (object pronoun), and a locative adverb). How can they categorize these early words? In what follows, we focus on the categorisation of meaning, which is part and parcel of the functional analysis of language.

3. Young children's acquisition of grammatical categories

Two views have been proposed to explain children's initial acquisition of word categories. The emergentist view represented by Tomasello's (2003) usage-based and construction grammar approach – together with its functionalist precedents (notably Bates and MacWhinney, 1982; MacWhinney, 1999) stands apart from the generativist, or nativist view, represented e.g. by Pinker's (1984) semantic bootstrapping theory. According to nativist views, there is no discontinuity between children's grammatical categories and those categories in adult speech. But the continuity assumption does not explain how children discover innate grammatical categories. The semantic bootstrapping view posits that 'grammatical categories have an intuitive basis, namely a semantic one' (Macnamara, 1984:126). According to Pinker (1994:385) this can eventually 'help the child get syntax acquisition started'. Indeed, Pinker posits that children have innate knowledge of the semantic properties of nouns and verbs, as well as syntactic knowledge of noun and verb categories. The 'intuitive basis' of semantic categories would then allow the children to link semantic and syntactic information in their input. Pinker thus suggests that children have innate knowledge of a relation between the noun word class and reference to a person or thing on the one hand, and between the verb word class and such semantic categories as 'action' or 'change of state' on the other hand (Pinker 1984:41).

One of the major problems with the semantic bootstrapping approach is that it relies heavily on innate knowledge, which in recent years proved to be very elusive (Tomasello, 2004;

[2] The corpus examples presented in this paper can all be accessed from the CHILDES website, in the Paris Corpus (Morgenstern & Parisse, 2012).

Evans & Levinson, 2009). Another problem is that variability in human languages (Croft, 2001) goes against such clear-cut associations between forms and meanings. Theories in generative linguistics tend to be very abstract (Evans & Levinson, 2009) and theories such as Pinker's semantic bootstrapping would need to be backed up by strong observations in child language. One important consequence of Pinker's approach that still awaits empirical verification is that children's first utterances should contain mostly prototypical nouns (nouns that correspond to objects) and true verbs (verbs that correspond to actions) and that this should be true for all children since a very young age. More complex semantic noun and verb associations should appear only later.

The other approach, the so-called usage-based and construction-based approach or cognitive linguistics approach (see Tomasello, 2003; Goldberg, 2006), holds that linguistic knowledge is built gradually, based on constructions, i.e. form-meaning associations with varying degrees of openness. The most specific constructions correspond to words of the lexicon and lexical collocations: both forms and meanings are fixed elements. The other constructions are open ones: they correspond to grammatical rules such as for example 'subject + verb' constructions (in languages in which this applies). The construction is open because a very large set of elements can fill the first slot (subject) and the second slot (verb). A key feature of this approach is that there is a continuum between the most fixed constructions and the most open ones. A semi-open construction in French is for example '*il y a + X*' (there is + X) where '*il y a',* the first slot, can take several forms – '*il y avait'* (there was), '*il y aura'* (there will), '*il y aurait'* (there would) – and 'X', the second slot, can take many forms (even more than 'subject' and 'verb' in the previous example). A second key feature of the usage-based approach is that the knowledge of each individual can contain a slightly different set of constructions, even though adults tend to share a large number of constructions. This leaves room for larger differences across children, which are not comprised in the generativist, nativist approach. Another unique consequence of usage-based approaches is that categorical knowledge and construction openness (or generalisation) are understood as a gradual and piecemeal development in young children (see Tomasello, 2000).

The analyses presented in the current paper follow the usage-based approach and are aimed at understanding language development as a gradual process. Leaving aside the nativist vs. emergentist debate, however, the approaches are not necessarily as mutually exclusive as was previously thought. Our suggestions can also be seen as an attempt to rethink the semantic bootstrapping with a usage-based model (see Rossi and Parisse, 2012)

4. Selecting semantic characteristics

A number of arguments have been produced to justify the gradual emergence of forms, but little work has been conducted into first categories of meaning and how they are related to forms. Tomasello (2000, 2003) argues that children learn the category of nouns earlier than the category of verbs and that this takes place in a piecemeal fashion. What is unclear is the status of semantics at each developmental step. Indeed, the semantics of early words is difficult to delineate (Bloom, 1991), and verb semantics is even harder to grasp (Gleitman, 1990; Golinkoff et al., 1995) which could account for their later acquisition (Gentner, 2006).

Starting from simple definitions, we coded 'semantic dimensions' to provide semantic information about the first words or word groups produced by a child, and see what pieces of information could help sorting these productions into nouns and verbs. We looked at children's first linguistic productions, i.e. mainly one-word utterances or unanalyzable groups of words or chunks. We therefore chose not to tear apart the elements that the child encoded as one entity –for instance filler syllables, but also clitic pronouns, determiners, were not coded separately. We wanted to see whether there were 'noun-like' and 'verb-like' elements

in children's first utterances, without bringing adult grammatical categories to the task of analyzing those utterances. We analyzed the children's 'lexicon', which here will refer to elements produced by children that are close enough to an adult lexical target to be recognized and coded.

In order to take into account the ambiguity and richness of children's early productions, our semantic dimensions can take several discrete values (two to three values) plus the value 'not coded', which refers to cases where a semantic dimension does not have apply. For instance at 1;6 Anaé refers to her grandmother almost unwittingly, so that at least some semantic dimensions are difficult to code:

(2) *Anaé 1;6.8* : line 2196
*MOT: c'est qui ce bébé ? [Who's that baby ?]
*CHI: ma mamie ! [My grandma!]
*MOT: non c'est pas mamie ! [No, it's not grandma !]

Anaé's grandmother is not there as Anaé refers to her, and could hardly be assimilated to a baby. We coded "ma mamie" as being absent, singular, animate and specific, but it was impossible to tell whether the referent was static or in motion, so we gave this feature a 'not coded' value The 'not coded' value is a piece of information in itself, so it adds to the two or three possible values of the semantic dimension.

One of the main problems with Pinker's approach, and one that is also found in some cognitive linguistics research, is that they use semantic characteristics that are so specific to the properties of nouns and verbs that the knowledge of the noun/verb distinction is implicitly posited, though at another level. For example the proposal that nouns are often linked to physical objects and verbs to actions gives no clue as to how children may learn to distinguish (or discover the difference between) actions from objects.

To avoid this pitfall, we chose to use semantic dimensions that are shared by nouns and verbs, and that can be identified by children by non-linguistic means (as proven by previous developmental studies). For example, any word used by the child and which refers to something in the environment can be characterized as referring to something moving or immobile. To do this, children do not need to develop any linguistic knowledge, as basic properties of perception (sensitivity to change) are sufficient to characterize motion, and motion is indeed one of the earliest categories of perception (see e.g. Slater, 1989; Casasola & al., 2006). The list of the categories coded and the values available for each category is presented in Table 1.

	1st value	2nd value	3rd value	Default value
Animacy	Animate	Inanimate		Not coded
Concreteness	Concrete	Abstract		Not coded
Determination	Specific	Generic		Not coded
Distance	Touching	Visible (Audible)	Absent	Not coded
Motion	Movement	Static		Not coded
Number	Singular	Plural	Uncountable	Not coded

Table 1: List of all semantic categories used to code nouns and verbs

All these notions or dimensions can be applied to nouns and verbs and are found in all languages and cultures. For all of them there are examples of languages where the notion is grammaticalized, which shows that they are all rather basic semantic dimensions. However, the notions are often lexicalized rather than grammaticalized. In French, only number and determination are grammaticalized: they both apply to nouns and verbs alike. However, as

shown by Macnamara (1984), number is closely related to sortals (i.e. concepts underlying the logical work of identifying and individuating count nouns), which Macnamara puts forward to explain how children learn the very concept of object. Determination (specific vs. generic values) also applies preferably or more prototypically to nouns (Macnamara, 1984: 144–156). These dimensions could help characterize and discriminate children's first productions: number should be used mainly with first nouns (especially reference to multiple or massive elements which should apply to nouns before being used with verbs), and first nouns are known to be specific and context-bound, as opposed to first verbs, which should be more generic (due to the predominance of "light verbs" such as *go*, *do*, *make* or *give* (Ninio, 2006), referring to broad categories of actions in early productions).

Data	Utterance	Noun	Verb	1	2	3	4	5	6
Antoine 1;5	Bain (bath)	bain (bath)		abs	sing.	abst.	inanim	static	specif.
Théophile 2;2	Il est cassé (it's broken)		est cassé (is broken)	touch	sing.	concr.	anim	static	specif
Madeleine 2;4	Je vais te raconter un livre (I'm going to read you a book)		vais raconter (going to tell)	touch	sing.	concr.	anim	mobile	specif.
Madeleine 2;4	Je vais te raconter un livre (I'm going to read you a book)	livre (book)		abs	sing.	concr.	inanim.	-	generic

Table 2: Coding examples
Note: See Table 1 for codes of the semantic dimensions (column 1 to 6)

Transcription extracts for the above coding examples:

(1) *Antoine: 1;5*
MOT: oui on va aller dans le bain maint(e)nant ? [yes we are going to take a bath now, aren't we?]
CHI: bain . [bath]
MOT: bain ? [bath?]
MOT: on va dans le bain ? [shall we go and take a bath?]

Here the mother and the child are playing in dinning room. The mother tries to encourage the child to stop playing and to get ready to take a bath. The bath is considered as absent because the object (bathtub) or the room (bathroom) are not directly visible to the child. It is considered as abstract because it does not, in this context, designate a specific visible or manipulated object, but a whole script or situation.

(2) *Théophile: 2;2*
OBS: ouh@i la la que c'est gros ça [oh la la this is really quite big].
OBS: c'est joli hein [that's pretty isn't it].

CHI: il est cassé [it's broken].
OBS: il est cassé ? [is it broken?]

Here the observer is talking to the child and saying that she is impressed with the toy the child is playing with. The child responds by showing the toy and giving it to the observer, and at the same saying that the toy is broken and pointing and touching the broken part with his thumb.

(3) *Madeleine: 2;4*
MOT: tu veux montrer un livre ? [do you want to show me a book?]
MOT: non tu devrais raconter un livre à Martine ! [no you should tell a story (from a book) to Martine]
CHI: vais te raconter un livre. [I'm going to tell you a story (from a book)].

Here the child is playing and talking to her mother and the observer. She is asked by her mother to tell a story from one of her books to the observer. The child stops what she is doing and goes to another room to find a book all the while explaining what she will do. We chose, for all verbs, to code dimensions such as distance, number, animacy, and movement according to the agent if there is one or to the subject otherwise. So, the verb 'vais raconter [going to tell]' is described as touching because the agent is a first person (the child), and moving because the child is moving when she produces this utterance. The argument 'livre [book]' is coded independently from the verb.

All six dimensions (see Table 2 for examples of the coding material) were chosen for their discriminating potential. Concreteness should be biased towards nouns if verbs are more abstract labels (Bird, Lambon Ralph, Patterson, & Hodges, 2000; Breedin, Saffran, & Coslett, 1994; Vigliocco et al., 2004), with an inherently relational content (Gentner, 1982; Black and Chiat, 2003). Animacy applies to both nouns and verbs, but just like motion, the dimension should apply more often to verbs in young children's universe, where the concept of animacy appears to develop in a piecemeal fashion (Rakison, 2005:187), and is primarily based on a notion of self-initiated vs. caused motion (Gelman & Gottfried, 1996). Reference to inanimate beings should also be a characteristic of first nouns, if they prototypically refer to physical objects, but not of first (presumably egocentric) verbs. It could be argued that animacy refers mainly to subject or the patient rather than to the verb. However, this is not obvious when verbs are produced in isolation (imperative in French), or with clitic pronouns that could arguably be considered as inflexions of the verb, or with filler syllables. In all these cases, coding the verb and coding the whole utterance are similar, and coding animacy can be applied.

Finally for distance, three levels of coding were used: touching (proximal), visible (distal) or absent (distal + invisible). Although all three values can be grammaticalized on both nouns and verbs (at least in some languages, see Payne, 1997), distance (and especially visible vs. absent) is more closely associated with objects, for at least three reasons: distance anchors definite vs. indefinite reference on nouns (see Maratsos, 1976), it is also at the root of the concept of object permanence, and objects are more likely to be in the child's hands or close to him when he talks about them. Differences, however, might be toned down as the first actions named have also been analyzed as 'ego-centered' (see e.g. Clark, 1987:27), i.e. as referring to the child's own action.

5. Analysis of children data

We coded the semantic characteristics of all the potential noun and verb elements used in the productions of three children aged 1;6 to 2;6 (exact description of the children's data is

presented in table 2). A word was a potential noun or verb if it was a noun or a verb in the target adult language. However, we did not use this piece of information to evaluate children's categorization into nouns or verbs. In order to quantify similarities and differences in children's use of elements that could be classified as nouns or verbs in the target language, we used semantic dimensions that can be found grammaticalized in different lexical categories across languages such as distance, number, animacy, movement, concreteness, and definiteness. We then sought to assess how the use of these semantic dimensions by the children matched the adult categories of nouns and verbs. All children were recorded in spontaneous speech situations for one-hour sessions.

Child		Age	MLU	Number of Nouns	Number of Verbs	Number of Words
Anaé	First session	1;6	1.38	194	30	390
(9 sessions)	Middle session	2;0	2.44	246	154	1129
	Last session	2;6	3.02	199	169	1397
Antoine	First session	1;5	1	3	0	23
(20 sessions)	Middle session	1;11	1.48	71	51	330
	Last session	2;6	2.72	183	187	354
Théophile	First session	1;5	1.05	4	0	59
(20 sessions)	Middle session	1;11	1.60	41	10	323
	Last session	2;8	2.61	124	47	1297

Table 2: Data coded and analysed for all three children
Note: MLU represents the Mean Length of Utterances (in words, calculated for the whole session). Utterance length is the number of words in the utterance. Words are assumed to be either noun or either verb according to the target adult language.

All coding was done by one trained coder and checked by the authors until achieving full agreement on the coding (see also, Rossi & Parisse, 2012). Coding was performed using the context provided by the video, and according to the interpretation of the child's intentions by the adult as a conversational partner. The coder had to code as if she were the adult interacting with the child, i.e.: the coder had to interpret the situation and decide what was the referent of the child's word, and code this referent according to the above-mentioned six semantic dimensions. We did not presuppose that children knew the noun and verb categories nor the six semantic dimensions, but we coded according to the natural reaction of any adult to the child's production. We presupposed that our coding came close to the information that the child receives in natural conditions: indeed, the adult interacting with the child constantly interprets and makes decisions about potential referents –employing strategies to respond to children's unintelligible utterances (Ochs and Schieffelin, 1984; Budwig, 1995; Chapman, 2000).

6. First results using unidimensional analysis

Results computed for each semantic dimension separately did not provide clear differentiation between noun and verb categories (see Rossi & Parisse, 2012 for more details about one-dimensional analyses). As shown in Table 3, there were few semantic dimensions

according to which a clear-cut difference could be found between nouns and verbs. These include the mobile value of the motion dimension, the animate and inanimate values for animacy and the absent value for distance. However, in all cases, but for verbs with the 'animate' value, there was no single dimension that characterized both noun and verb grammatical categories. For example, the 'animate' value in our data characterized most of what is a verb in the adult language, but words considered as nouns were nearly half animate and half inanimate. So using the animate vs. inanimate dimension as a dividing line between nouns and verbs, half of the nouns might be confused with verbs.

	Noun	Verb	Noun	Verb	Noun	Verb	Noun	Verb
Animacy	not coded		Animate		Inanimate			
	1	0	40	95	59	1		
Concreteness	not coded		Abstract		Concrete			
	2	2	23	13	75	82		
Determination	not coded		Generic		Specific			
	5	8	8	4	87	85		
Distance	not coded		Absent		Touch		Visible	
	3	9	24	7	23	44	51	37
Motion	not coded		Mobile		Static			
	31	7	43	74	27	16		
Number	not coded		Plural		Singular		Uncountable	
	1	2	6	0	84	93	8	0

Table 3: Percentages of noun and verb words for each semantic dimension
Note: The percentages correspond to the average percentages of the three children's results.

Little evidence could be found in Parisse and Rossi (2012) to confirm the principle of gradual development of semantic categories. Anaé showed a trend for verbs to become more often abstract as she got older. Verbs were not used to denote abstract reference before age two in Anaé's data. Antoine did not use verbs with visible elements first, and then learned to do it.

7. Multiple correspondence analysis and hierarchical clustering

A major weakness of our first results is that our statistical tools were not able to grasp the possibility of an interaction between different semantic dimensions. We therefore thought it would be interesting to use tools that are efficient in dealing with multiple dimensions, or in statistical parlance to use multivariate analysis.

One of these tools is Multiple Correspondence Analysis (MCA). It takes as input a list of individuals (which correspond in our case to each occurrence of a noun or a verb produced by the child) and a set of categorical variables. For each individual, a set of responses is gathered that corresponds to the set of variables. These responses (in our case the set of values coded for each of the six semantic dimensions) must be categorical (i.e. a value in a finite set of possible values). The output of the statistical tool is a representation of individuals (which corresponds to words here), of variables (which correspond to the values of our semantic dimensions), and of categories (which correspond to dimensions themselves).

MCA is an extension of Correspondence Analysis and belongs to a family of tools for Multidimensional Data Analysis (MDA) – see Husson, Lê and Pagès (2011). It provides a synthetic view of the data, represented in Euclidean space (Euclidean here refers to the usual notion of distance between points in space). The new representation is built through a

transformation of the original data into new statistical dimensions that represent the most informative topological features of the data. The first two or three dimensions can usually represent most of the information in the original set, providing new insights in the organization of the data. Especially, the information represented in these first dimensions, usually represented by a 2D drawing, gives precious clues to how the individual observations are related to one another. Although there is a small loss of information, this new representation is clearer and easier to use than the representation of the original data. It has been widely used in many scientific fields, including humanities and social science with pioneer work by Bourdieu (1984) – see Divjak and Gries (2006), Desagulier (in press), for examples of the use of MCA and HC in linguistics.

One of the limitations of MCA is that it only outputs representations of the positions of the individuals or categories that were provided in the input. As a result, there is no way to compute a probabilistic evaluation of the categorization provided by the multidimensional analysis. The spatial configuration obtained is, however, very informative. Points that are spatially close one to another are likely to correspond to similar situations. Another weakness is that, although the first and second dimensions resulting from the MCA are very informative, they do not represent all the information provided. There is still information in the other less informative dimension that would need to be taken in account in an objective measure.

It is possible to go beyond this last limitation of the MCA by using Hierarchical Clustering (HC) – a method that consists in automatically computing clusters of individuals. The difference with MCA is that HC computes categories that can be compared to the predefined categories (such as noun and verb in the target language). Performing HC without using an MCA first would have been a valid option. However, we chose to use the two operations one after the other for three reasons. First, this allows performing an analysis of the data (thanks to MCA) before trying to perform a quantitative analysis. Second, this allows us to check the quality of the visual interpretation of the MCA results. Third, MCA makes HC more powerful because most of the information is better organized (Husson et al., 2011).

In the work below we will use successively MCA and HC for the three children's data. In a first step, MCA allows us to provide a good representation of which individuals are close together and which semantic categories are most powerful to describe our data. This classification (MCA) is done without taking into account any information about which word is a noun or a verb. Once the classification is done, it is possible to use grammatical knowledge on which word is a noun or a verb in the adult target language to find out how the individuals within each category of child language are organized by the semantic information available.

In a second step, HC is performed on the results of MCA. This allows us to find out the best possible clustering of the six semantic dimensions in our data. The set of clusters obtained can be analyzed with a view to understanding how it is constructed. It can also be compared with noun and verb adult categories to find out whether the child's semantic categories can be mapped onto the adult's syntactic categories. For all children we will study the characteristics of the initial production (the first half of the child's data) and the later production (the second half of the child's data).

8. Analysis for Anae

MCA was performed for Anaé for all words recognized by the coder as targeting either a noun or a verb in the adult language, but this knowledge was not used in the analysis. Out of the 2859 occurrences coded, words that were ambiguous[3] (191 occurrences) or imitations (31 occurrences) were not analyzed (imitations are considered as such only when the full utterance is subject to immediate imitation), so that our coding conveyed a picture of the semantic properties of all words that had a straightforward orientation towards noun or verb. The words analyzed with MCA and HC correspond to 92.2% of the total. The HC analysis was performed by categorizing the data in two sets, so that we could compare directly blind clustering with the noun and verb coding of our data (i.e. the adult lexical categorization into noun and verb).

Anaé up to age 2;0

Figure 1 presents the results for MCA (left) and HC (right) for Anae for age 1;6 to 2;0 (five sessions). The left image is a topological representation of all the nouns and verbs produced by Anae up to age 2;0, according to the six semantic dimensions coded. Each word coded in the corpus is represented on the figure by a point and a number. The numbers help to identify which word is exactly represented at a specific position in the figure. The representation of the black points (the nouns) and the grey points (the verbs) is independent from the actual semantic data coded. It is clear, however, that most of the time, points that are very close in the MCA two-dimensional representation (which means that their description in term of the six semantic dimensions is very similar) belong to the same syntactic category.

Insert figure 1 about here

For example, the dot indicated by the left arrow in the left part of figure 1 corresponds to a cluster of 26 nouns (4 different nouns) and 69 verbs (30 different verbs) and all of them were tagged as [touch, singular, concrete, animate, mobile, specific]. The 26 nouns all correspond to proper nouns and the rest of the items are verbs. This conflation stems from the absence of difference in our semantic coding scheme of ‘a person performing an action’ and ‘an action performed by someone’.

The second arrow, starting from the upper left part of figure 1, points to two very close points, both of them close to the previous one. The first of the two points corresponds to 3 nouns (3 different nouns) and 4 verbs: the nouns are again proper nouns and the verbs are *regarder* (to look) and *regarde* (look). These items are tagged [touch, singular, concrete, animate, static, specific]. The coding is the same here because for verbs, with distance and number, our coding actually refers to the object or person doing the action if there is more than one argument to the verb (see examples 3 & 4 in Table 2). So in the case of a noun referring to a person and a noun referring to an action performed by that person, our coding schema results in the same value. The second point corresponds to 32 nouns (5 different nouns) and 71 verbs (30 different verbs): nouns are once again proper nouns. These items are tagged [visible, singular, concrete, animate, mobile, specific]. In these three examples, the values of the semantic dimensions are very close. The sole difference between the first and

[3] For instance « peur » is a noun but it is likely to be used as a verb construction (as in “elle a peur”: she is afraid): whenever context was unclear we did not analyze it. The same goes for adjectives, which were coded but not taken into account in the analyses discussed here.

last sets of elements is the value of the first dimension, 'touch' vs 'visible'. The second set only differs from the first one in one respect: the value of 'mobile' vs. 'static'. All the items in the three sets contain the values singular, concrete, animate, and specific, which appear to designate mainly verbs and proper nouns.

Sets of elements that are far from the first two sets have distinct characteristics. For example, the third arrow, starting from the left part of figure 1, points to a set composed of two different nouns (both occurring once), which are coded [touch, plural, concrete, inanimate, not coded, specific]. The fourth arrow (the right arrow) points to a set composed of 6 different nouns (all occurring once), coded [visible, plural, concrete, inanimate, not coded, specific]. It seems that this part of space contains mostly plurals and elements that are not coded for movement (because it is not applicable in this case), which differs from the part of space pointed by the first two arrows. There are less elements here than in other sets, simply because there are less plurals than singulars in the data.

The right image of figure 1 is a representation of an HC analysis based on the results of the MCA. The number of clusters asked for was two (the HC statistical tool in R allows declaring the exact number of categories to be expected –in the present case we asked for two categories but we also tried other values, see below) so that we could compare the results with the adult target categorization into nouns and verbs. The points in the figure correspond to the individuals of the MCA analysis (the words of the child) and they are located at the same place in the MCA and HC analysis. However the grey nuances in HC correspond to the two clusters automatically computed. So there is not direct color correspondence between the two images (MCA and HC). Each cluster (cluster 1 and cluster 2) contains a percentage of nouns and a percentage of verbs. The aim of the comparison is to test the degree of overlap between the division in the two clusters and the division into nouns and verbs. The results for Anaé up to age 2;0 are presented in Table 4. In this table, 37% of the words in cluster 1 are nouns and 63% are verbs, so we can conclude that cluster 1 would correspond to 'VERB' and that there is 63% of correct classification in this cluster. Cluster 2, with more nouns than verbs would correspond to 'NOUN' and contains 94% of correct classification. On the whole, classification by HC is 79.1% correct for Anaé up to age 2;0.

	1	2	Cluster	1	2
NOUN	92	249		37%	94%
VERB	156	15		63%	6%
Category	Nb of words			% of words up to 100%	

Table 4: Comparison between categories in MCA and cluster in HC for Anaé up to age 2;0. Right figures correspond to absolute number of occurrences; left figures correspond to percentages.

Anaé from age 2;1

Insert figure 2 about here

Figure 2 presents the results for MCA (left) and HC (right) for Anaé for age 2;1 to 2;6 (five sessions). The main difference with figure 1 is that there is a much larger number of individuals, so that there is also a larger number of small subsets of individual words. The comparison between MCA and HC shows a greater similitude between the categories in the MCA and the HC clusters than for Anaé up to age 2;0. The results of the comparison between MCA and HC are presented in Table 5. Classification by HC is 95.6% the same as the

classification done by hand for Anaé after age 2;1 whereas it was only 79.1% up to age 2;0. On the whole, the results are very good and better than before age 2;0. Some words, however, were not categorized into noun or verb by the HC computation in the same way than our initial coding. This corresponds to words that do not have the same grey nuance in the two parts of figure 2. For example, the individual token 2598, indicated by the left arrows in the two images of figure 2, corresponds to the word *Jules* (a proper noun) and is coded [visible, singular, not coded, animate, not coded, specific]. In the MCA categories, it is a noun, and it is considered close to a verb in the HC clusters. Another example is the individual token 960, indicated by the right arrows in the two images of figure 2, which corresponds to a verb, *sais* (I or you know), and is coded [visible, singular, abstract, not coded, not coded, specific]. This word changes from verb in MCA to clusterized-as-noun in HC. It is indeed close to two sets of items that contain only nouns: one which is coded [absent, plural, abstract, animate, not coded, specific], and one which is coded [absent, singular, abstract, inanimate, not coded, specific]. The common dimensions between the three sets are probably abstract for abstractness and not-coded-for-movement, which makes a verb similar to nouns from a semantic point of view.

	1	2	Cluster	1	2
NOUN	87	1188		9%	99%
VERB	843	6		91%	1%
Category	Nb of words			% of words up to 100%	

Table 5: Comparison between categories in MCA and cluster in HC for Anaé after age 2;1. Right figures correspond to absolute number of occurrences; left figures correspond to percentages.

Results for Anaé

The overall results for Anaé show that it is indeed possible to determine automatically and rather accurately whether a word could be a verb or a noun, based on the set of values of the six semantic dimensions. Results get better as Anaé grows older, which could mean that she progressively learns to organize nouns and verbs (at least semantically) or that she learns to approximate better the regularities of her input.

9. Analyses for Antoine and Théophile

The same methods were applied for Antoine and Théophile. Both children are less talkative than Anaé so the characteristics of the first period for the two boys are different. Especially for Théophile, and to a smaller extent for Antoine, most of the words produced by the children are nouns, so that a classification into two equivalent sets is not easy to implement. For Théophile, a second difference is the age reference. As the child is a slow starter in language development, the split point between the first (8 sessions) and second (12 sessions) period occurs later (at age 2;4) and the observations went up to age 2;11. The starting age is the same.

Age 1;6 to 2;0

	1	2	Cluster	1	2
NOUN	403	15		61%	94%
VERB	260	1		39%	6%

Age 2;1 to 2;6

	1	2	Cluster	1	2

NOUN	1262	406	64%	79%
VERB	695	111	36%	21%
Category	Nb of words		% of words up to 100%	

Table 6: Comparison between categories in MCA and clusters in HC for Antoine (age 1;6 to 2;0 and age 2;1 to 2;6). Left figures correspond to absolute number of occurrences; right figures correspond to percentages.

MCA and HC comparisons for Antoine up to age 2;0 and after age 2;1 are presented in Table 6. The quality of coincidence between MCA categories and HC clusters is much lower than with Anaé: 61.5% instead of 79.1% up to age 2;0, 67.4% instead of 95.6% after age 2;1. There are two explanations for this low correspondence. First there are some elements that are nouns or verbs in the adult language and are coded in the same way on the six semantic dimensions. When this happens it is impossible to characterize the words according to the noun/verb opposition and this seems to be more the case for Antoine than for Anaé. One frequent case is the proper noun vs. verb ambiguity, already present in Anaé's data, which appears because both lexical elements are coded in the same way, especially as in both cases there is movement as well as animacy. It should be noted here that animacy (but not movement) actually is a very good and reliable means to differentiate nouns from verbs, except when nouns refer to people or animals. In this case, other dimensions have to come into play.

(3) *Anaé: 1;07.03 line: 628*.
*CHI: assis !
%pho: aʃi
*MOT: assis Omer !

(4) *Anaé: 1;07.03 line: 1607*.
*CHI: Omer (.) yy !
%pho: ɔmɛ ɛlito
*MOT: qu' est+ce+qu' i(l) fait Omer ?

In example (3) and (4), the use of 'Omer' (the name of the family dog) and of 'assis' (sit) are coded in the same way in all semantic dimensions (visible - singular - concrete - animate - mobile - specific). This is not surprising as both words refer to some action performed by the dog. In this situation, it is not the potential final state that is coded, but what the dog is actually doing (coding the final state would imply making assumptions about the general knowledge of the child, which we tried to avoid in this work)[4]. The target adult noun and the target adult verb used here represent two views of the same situation: the characterization of an action performed by an animate being. In this situation, it would be necessary to use other semantic dimensions than the one coded here to distinguish between the two target words.

Another problem with Antoine is that a division into exactly two clusters does not seem to be the most 'natural' division for his linguistic productions. We try to address this problem below (see "Reanalysis using more than two clusters").

[4] In the present work, the most important element has been that our coding choices should remain the same. We could have chosen to code a verb as what is happening or as what corresponds to the goal, but for the purposes of automatic classification, it is important that a dimension should always keep the same interpretation. Other choices are of course possible: in future work, we could also code both aspects, for example.

	Age 1;6 to 2;0				
	1	2	Cluster	1	2
NOUN	162	165		81%	99%
VERB	39	1		19%	1%

	Age 2;1 to 2;6				
	1	2	Cluster	1	2
NOUN	172	1019		15%	100%
VERB	939	5		85%	0%

Category	Nb of words	% of words up to 100%

Table 7: Comparison between categories in MCA and clusters in HC for Théophile (age 1;6 to 2;4 and age 2;5 to 2;11). Left figures correspond to absolute number of occurrences; right figures correspond to percentages.

The results for Théophile are presented in Table 7. The two periods of Théophile are very different from one another (8 sessions for the first period, 12 sessions for the second period). In the first period, Théophile produces very few verb tokens: only 10.8%. This explains success of the HC algorithm at 89.2%, which is exactly the percentage of nouns in the data. This was actually also the case for Antoine up to age 2;0 (none of the two categories contained a majority of verbs), but not for Anaé at the same age. After age 2;5, Théophile behaves much like Anaé and Antoine. Based on clusters of semantic dimensions, there are two clear categories, one that corresponds to what was coded as noun and one that corresponds to what was coded as verbs. The global result, 91.7%, is nearly as good as with Anaé (95.6%) and much better than Antoine's (67.4%).

10. Reanalysis using more than two clusters

The results for Antoine are not as good as the results for the two other children. One of the characteristics of Antoine is that in the last sessions, he has a low percentage of verb tokens (32%) and there are about half as many verbs as nouns. This is not as much the case for the other children (39.9% verbs for Anaé and 44.2% for Théophile).

One major question here is the validity of a categorization in two clusters only, especially for semantic dimensions. It is possible that a division in a higher number of clusters will offer a better categorization of the data

An interesting aspect of HC is that it is possible to let the system decide which number of clusters is an optimal division of the data. We have used this feature to find out whether it changed our results.

The computation of HC was reproduced without using the option that allows users to request for a specific number of clusters. In this case, the software suggests the best value, which was four clusters for Antoine up to age 2;0 and three clusters after age 2;1. This completely changed the distribution of nouns and verbs in the resulting clusters. The new distributions for the two age ranges are presented in Table 8.

	Age 1;6 to 2;0								
	1	2	3	4	Cluster	1	2	3	4
NOUN	271	120	12	15		55%	83%	48%	94%
VERB	223	24	13	1		45%	17%	52%	6%

	Age 2;1 to 2;6						
	1	2	3	Cluster	1	2	
NOUN	409	907	352		37%	99%	78%
VERB	701	6	99		63%	1%	22%
Category		Nb of words				% of words up to 100%	

Table 8: Comparison between categories in MCA and clusters in HC for Antoine (age 1;6 to 2;0 and age 2;1 to 2;6). Left figures correspond to absolute number of occurrences; right figures correspond to percentages.

When comparing Table 8 with the results presented in Table 6, The correspondence between the clusters and the categories noun and verb of the words in the adult language does not improve for the sessions before age 2;0 (61.7% vs. 61.5%) but it does improve as the child gets older (79.2% vs. 67.4%). An interesting feature is that two of the clusters extracted for Antoine up to age 2;0 contain nearly the same number of target language nouns and verbs (clusters 1 and 3) whereas the other clusters are clearly noun clusters. This indicates that the clustering operation is able to differentiate between nouns that are clearly different from verbs and nouns that are similar to verbs. In this respect, clusters 1 and 3 could be considered as polycategorial, at least from a semantic point of view -which means that, according to the dimensions used in the coding, all these elements have no semantic properties that differentiate them specifically as nouns and verbs.

The results for Antoine after age 2;1 show yet another distribution. Some clusters are nearly only composed of nouns. In clusters 1 and 3 the difference between nouns and verbs is less marked, but cluster 1 is more of a verbal cluster and cluster 3 more of a nominal cluster.

	Child	Anaé	Antoine	Théophile
First sessions (younger children)	Two clusters	79.1%	61.5%	89.1%
	Optimal number (Nb of clusters)	78.5% (8)	61.7% (4)	89.1% (6)
	% of nouns in the data	66%	62%	88%
	% of verbs in the data	33%	38%	12%
Last sessions (older children)	Two clusters	95.6%	67.4%	91.7%
	Optimal number (Nb of clusters)	92.8% (6)	79.2% (3)	86.9% (4)
	% of nouns in the data	60%	67%	58%
	% of verbs in the data	40%	33%	44%

Table 9: Summary of the global results for all children

We performed again the HC procedure for Anaé and Théophile, leaving the number of clusters open. The results are summed up in Table 9. Results were never better with more than two clusters, but the difference between the first and second lines, and between the third and fourth lines of results in table 9 was rather small. More importantly, the results were computed fully automatically, without the knowledge that there are exactly two categories to be found.

The results obtained have to be compared to a baseline. Random choice between noun and verb is not the highest baseline because it does not take into account the fact that nouns are more frequent than verbs. This is expressed in the percentages of nouns and verbs that appear in Table 9 (lines % of nouns and % of verbs). Our results are not better than this baseline for the two children with the lowest language development (Antoine and Théophile)

when they are young (top part of Table 9). The results are clearly better than the baseline, especially for Anaé and Théophile, when the children are older (bottom part of Table 9).

11. General discussion

The results obtained show that it is possible to use the regularities of the child's noun / verb distribution in the adult target language using only purely semantic dimensions. Moreover, as soon as their language is more developed, and as they grow older, our results are much better than chance. This means that there is a means for children -who might or might not use the type of strategy that we used- to learn to associate forms and meaning and to generalize this association into something like the classical grammatical categories of the adult. One major difference between our work and Pinker's (1984) is that we used characteristics that are cognitive rather than linguistic and that do not presuppose the existence of grammatical innate linguistic knowledge. The cognitive characteristics our coding dimensions refer to may be innate, or based on even more primitive innate abilities that appear during development, but do not correspond to grammatical knowledge only.

While remaining completely agnostic as to what characteristics might be innate, our results are compatible with the cognitive linguistics approach of Langacker (1987) and Golberg (2006) in which language is based on general cognitive abilities. Another interesting result of the present analysis, which we did not clearly obtain in Rossi and Parisse (2012), is that a developmental effect could be evidenced. The identification of nouns and verbs thanks to regular association with certain semantic dimensions was better for all children when they were older. This is coherent with the usage-based hypothesis and is contrary to the generativist view, where grammatical knowledge appears suddenly and does not follow a developmental trend.

Let us now try to answer more precisely the two questions raised in the introduction, starting with the second one: How do children come to grasp the categories of adult language? The key feature of the data that we presented here is that a gradual organization of meaning takes place during language development. Added to the rich and varied analyses that showed gradual development of forms – for French-speaking children, see for example Veneziano et al. (1990), Veneziano & Sinclair (2000), Bassano (2000) –, they suggest that forms and meaning develop together gradually. The child uses her knowledge about the world together with the language data she learns to gradually organize linguistic data into categories. This results in form- meaning relationships. Our analyses and results have only been concerned with early semantics, covering some of the real-world characteristics that are relevant to children's early linguistic productions. Our assumption has been that these semantic dimensions might come into play early in the child's language development. Our results were better for older children than for younger children, which would be an argument in favor of such a developmental hypothesis. However, our data do not contain information about the form of the production (for example, the existence of grammatical markers, or the quality of the phonological form), so it cannot determine where and when parameters of form (for example the syntactic knowledge) might be brought in the picture.

The variation that our results point to could help answer the first question: What is the nature and range of children's categories? First, as a result of young children's empirical approach to categorization, the categories obtained can be different from one child to another. And on the whole, semantic categories are not necessarily dividing lines between nouns and

verbs. In some cases (especially for a human performing an action), there are true ambiguities between noun and verb semantic dimensions. This could lead to polycategorial categories, both at the individual level and/or at the level of language. Our answer to the first issue raised in the introduction would then be that there is not always a clear-cut difference between nouns and verbs, not only in child speech, but generally also in the language.

A number of open questions remain to be addressed. First, what the child says is not independent of what the adults say to the children, and especially what is said in the same recording session. The correspondence between the child's word use and the adults' one is linked to the context and is part of a common ground on which these words revolve. Also, what is the exact nature of the interaction between learning semantic categories and learning syntactic categories, not to mention other domains such as phonetics and pragmatics? Further study of the type presented here might help answer these issues. It is unlikely that the six dimensions that we studied are the only ones that create a correspondence with adult categories (nouns vs. verbs). Other dimensions probably exist, and other sets of dimensions might also account for the data. Indeed, in some cases, such as imperative verbs and proper nouns, our semantic dimensions have proved to result in an ambiguous classification. Even the best learning algorithm could not distinguish between noun and verb categories in this case. This would indicate that for these words (imperative verbs and proper nouns), either other semantic notions apply, or regularities of form do not go hand in hand with regularities of meaning.

So other dimensions could be added as there is no limit in the number of dimensions of the type of statistical processes we used. Also, it is possible that the effect of a perceptual dimension depends on the child's cultural setting and the physical characteristics of the situation?

An issue that could be raised is the adequacy of our choice of semantic characteristics. We tried to choose characteristics that could reasonably be learned and applied without access to language. Some characteristics might be more linked to what is considered as a noun or a verb in traditional grammars. For example, animacy or number might apply more basically to the arguments of verbs (often nouns) than to the verbs. However, young children have access to animacy and number before they start to talk, so they have to use animacy and number for all language material before they start to process nouns and verbs differentially. Also, number is an efficient classifier in at least some cases, so it is a valuable knowledge, but it has to be complemented by other dimensions. This explains why it is necessary to perform multivariate analyses and not univariate ones.

Our results suggest that although some semantic characteristics might be ambiguous in their relationship to word categories, it is possible for form-meaning correspondences to stabilize and help children categorize the world they live in -according to the way language categorizes this world in a given cultural setting.

Finally, our results do not necessarily demonstrate that the semantic dimensions we chose to use are the only ones at work or even the best ones. It demonstrates that using multiple semantic dimensions could be a fruitful direction in the understanding of young children's language development. Besides, it stresses the importance of using multivariate analysis in the study of linguistic data. Our previous work (Rossi & Parisse, 2012), using only univariate analyses did not confirm our initial predictions about noun/verb differentiation nor revealed any early language development. The multivariate analysis on the contrary validates these predictions, and even though much work in this domain still needs to be done, our results have taken us one step further by showing how children's early nouns and verbs were clearly distinguished according to clusters of semantic dimensions. This confirms the power

of statistics as a tool for linguistic analysis –an intuition that linguists have had for a long time: 'Statistical considerations are essential to an understanding of the operation and development of languages' (Lyons, 1968:98). The fact that no definite conclusions can be drawn should not come as a surprise, as language is inherently a multidimensional knowledge. Our analyses should therefore remain as varied as possible, taking into account both the necessity of developing powerful and robust statistical methods and the irreducible part that remains hidden to statistics –e.g. children's errors as evidence for language development.

References

Bates, E., & MacWhinney, B. (1982). Functionalist approaches to grammar. In E. Wanner & L. R. Gleitman (Eds.), *Language acquisition - the state of the art*. New York: Cambridge University Press.

Bassano, D. (2000). Early development of nouns and verbs in French: exploring the interface between lexicon and grammar. *Journal of Child Language*, 27(3), 521–559.

Bernal, S., Lidz, J., Millotte, S., & Christophe, A. (2007). Syntax Constrains the Acquisition of Verb Meaning. *Language Learning and Development*, 3(4), 325–341.

Bird H, Lambon Ralph MA, Patterson K, Hodges J. (2000). The rise and fall of frequency and imageability: noun and verb production in semantic dementia. Brain and Language, 73 (1), 17-49.

Black, M., & Chiat, S. (2003). Noun-verb dissociations: a multi-faceted phenomenon. *Journal of Neurolinguistics*, 16(2-3), 231–250.

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

Bloom, L. (1991). *Language development from two to three*. Cambridge: Cambridge University Press.

Bourdieu, P. (1984). *Distinction: A Social Critique of the Judgement of Taste*. Harvard University Press.

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

Breedin, S., Saffran, E. M., & Coslett, H. B. (1994). Reversal of the concreteness effect in a patient with semantic dementia. *Cognitive Neuropsychology*, 11(6), 617–660.

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

Budwig, N. (1995). *A developmental-functionalist approach to child language*. Mahwah, NJ: Lawrence Erlbaum.

Casasola, M., Bhagwat, J. and Ferguson, K. (2006). Precursors to verb learning: infant's understanding of motion events. In: K. Hirsh-Pasek and R.M. Golinkoff (eds), *Action meets word: how children learn verbs.* New York: Oxford University Press, pp. 160–190.

Chapman, R. S. (2000). Children's Language Learning: An Interactionist Perspective. *Journal of Child Psychology and Psychiatry*, 41(1), 33–54.

Clark, E. V. (1982) The young word-maker: A case study of innovation in the child's lexicon. In: E. Wanner and L. R. Gleitman (eds), *Language Acquisition: The State of the Art*. Cambridge: Cambridge University Press, pp. 390–425.

Clark, E. V. (1987). The principle of contrast. In MacWhinney, B. *Mechanisms of Language Acquisition*. Routledge, 1987, pp. 1-33.

Croft, William. (2001). *Radical construction grammar: Syntactic theory in typological perspective.* Oxford: Oxford University Press.

Croft, W., & Cruse, D. A. (2004). *Cognitive Linguistics*. Cambridge University Press.

Desagulier, G. (in press). Visualizing distances in a set of near synonyms: rather, quite, fairly, and pretty. In D. Glynn & J. Robinson (Eds.), *Polysemy and Synonymy : Corpus Methods and Applications in Cognitive Linguistics*. Amsterdam: John Benjamins.

Divjak, D., & Gries, S. T. (2006). Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory*, 2(1), 23-60.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(05), 429–448.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In: S. A. Kuczak. (ed.), *Language Development, vol. 2*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gelman, S. A. and Gottfried, G. M. (1996). Children's causal explanations of animate and inanimate motion, *Child Development*, 67:1970–1987.

Gentner, D. (2006). Why verbs are hard to learn. In: Hirsch-Pasek, K. and Golinkoff, R. (eds.) *Action meets word: How children learn verbs*. New York: Oxford University Press.

Gleitman, L.R. (1990) The structural sources of verb meanings. *Language Acquisition* 1, 3–55.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Goldberg, A. E. (2003). Constructions: A new theorical approach to language. *Trends in Cognitive Sciences,* 7(5), 219–224.

Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Golinkoff, R. M., Hirsh-Pasek, K., Mervis, C. B., Frawley, W. & Parillo, M. (1995). Lexical principles can be extended to the acquisition of verbs. In: M. Tomasello and W. Merriman (eds.), *Beyond names for things: Young children's acquisition of verbs* (pp. 185-222). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hockett, C.F. (1960). The Origin of Speech. *Scientific American*, 203:88–111.

Husson, F., Lê, S., & Pagès, J. (2011). *Exploratory Multivariate Analysis by Example Using R*. CRC Press.

Langacker, R. (1987). *The Foundations of Cognitive Grammar: Volume I: Theoretical Prerequisites*. Stanford University Press.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. Chicago: University of Chicago Press.

Lyons, J. (1968). *Introduction to Theoretical Linguistics.* Cambridge University Press.

Macnamara, J. (1984). *Names for things: A study in human learning*. Cambridge, MA: Bradford.

MacWhinney, B. (1999). *The Emergence of Language*. Mahwah, NJ: Lawrence Erlbaum Associates.

Maratsos, M. P. (1976). *The Use of Definite and Indefinite Reference in Young Children: An Experimental Study of Semantic Acquisition*. Cambridge: Cambridge University Press.

Morgenstern, A. & Parisse, C. (2012). The Paris Corpus. *Journal of French Language Studies*, 22, 7-12.

Nelson, K. (1973). Structure and strategy in learning to talk. *Monographs of the Society for Research in Child Development*, 38(1-2, Serial No. 149).

Ninio, A. (2006). Kernel vocabulary and Zipf's law in maternal input to syntactic development. *BUCLD 30: Proceedings of the 30th Annual Boston University Conference on Language Development*, 423–431.

Ochs, E. & Schieffelin, B. (1984). Language acquisition and socialization: Three developmental stories. In R Shweder & R. LeVine, *Culture theory: Mind, self, and emotion*. Cambridge: Cambridge University Press

Olguin, R., & Tomasello, M. (1993). Twenty-five-month-old children do not have a grammatical category of verb, *Cognitive Development*, 8, 3, 245-272.

Payne, T. E. (1997). *Describing Morphosyntax: A Guide for Field Linguists*. Cambridge: Cambridge University Press.

Pinker, S. (1984). *Language Learnability and Language Development*, Cambridge, MA, USA: Harvard University Press.

Pinker, S. (1994). How could a child use verb syntax to learn verb semantics? *Lingua*, 92:377-410.

Rakison, D. H. (2005). Developing knowledge of motion properties in infancy. *Cognition*, 96: 183–214.

Rossi, C., & Parisse, C. (2012). Categories in the making: Assessing the role of semantics in the acquisition of noun and verb categories. *Journal of French Language Studies*, *22*, 37–56.

Slater, A. (1989). Visual Memory and Perception in Early Infancy. In: A. Slater, and G. Bremner, (eds.) *Infant Development.* London: Lawrence Erlbaum Associates.

Slobin, D.I. (2001). Form-function relations: how do children find out what they are? In Bowerman, M. & Levinson, S. *Language Acquisition and Conceptual Development*, pp. 406-449. Cambridge University Press.

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

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

Tomasello, M. (2004). What kind of evidence could refute the UG hypothesis? *Studies in Language*, 28, 642-44.

Tomasello, M., & Olguin, R. (1993). Twenty-three-month-old children have a grammatical category of noun. *Cognitive Development*, 8, 4, 451-464.

Veneziano, E. (2003). The emergence of noun and verb categories in the acquisition of French. *Psychology of Language and Communication*, 7(1), 23–36.

Veneziano, E., Parisse, C. and Delacour, A. (2010). The Early Comprehension of Noun-Verb Distinction in French. An Experimental Method. *International Conference on Infant Studies 2010*, Baltimore, USA, Mars 2010.

Veneziano, E., & Sinclair, H. (2000). The changing status of "filler syllables" on the way to grammatical morphemes. *Journal of Child Language*, 27(3), 461–500.

Veneziano, Edy, Sinclair, H., & Berthoud-Papandropoulou, I. (1990). From one word to two words: repetition patterns on the way to structured speech. *Journal of Child Language*, 17, 633–650.

Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: the featural and unitary semantic space hypothesis. *Cognitive Psychology*, 2004. 48(4): p. 422-88.

Zingeser, L.B. & Berndt, R.S. (1990). Retrieval of nouns and verbs in agrammatism and anomia, *Brain and Language*, 39(1):14-32, ISSN 0093-93.

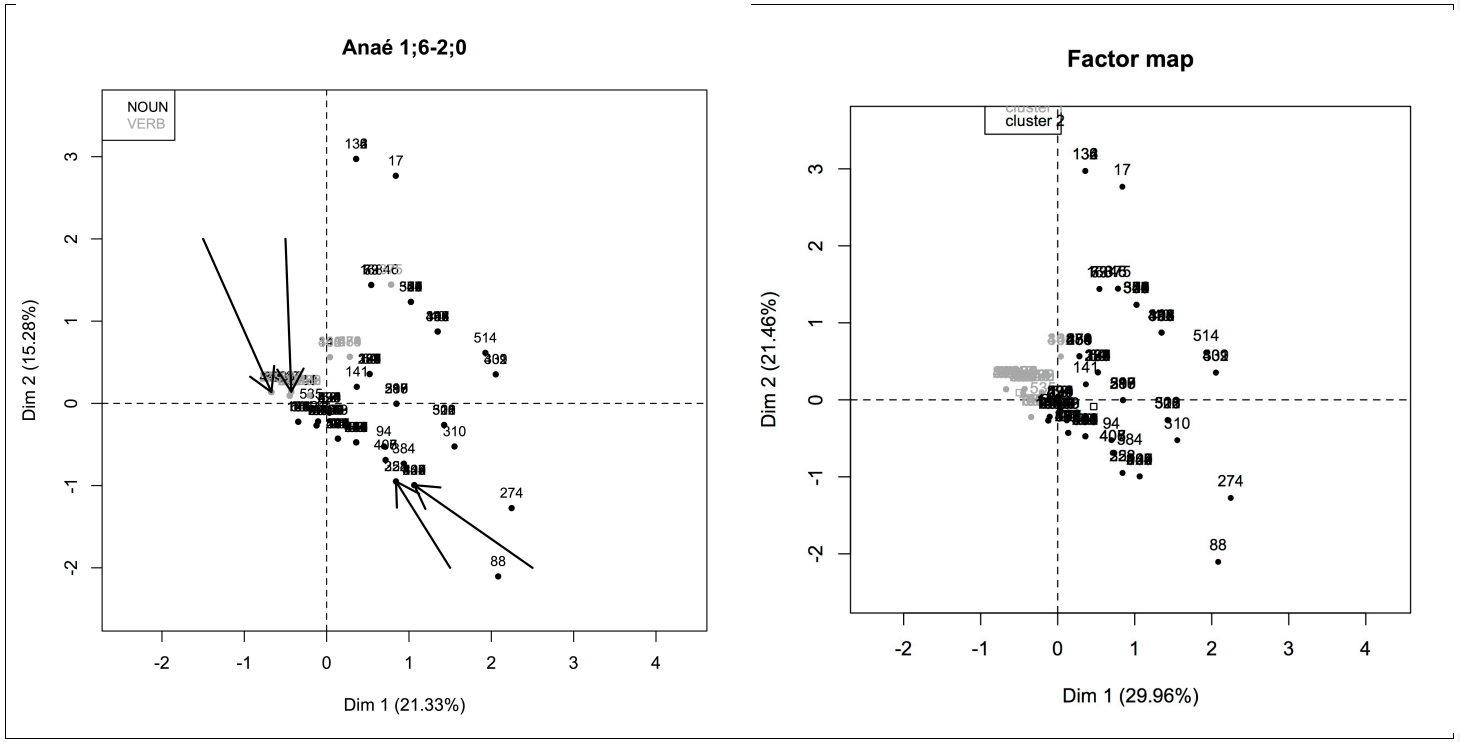

Figure 1: MCA and HC analyses for Anae at age 1;6 to 2;0

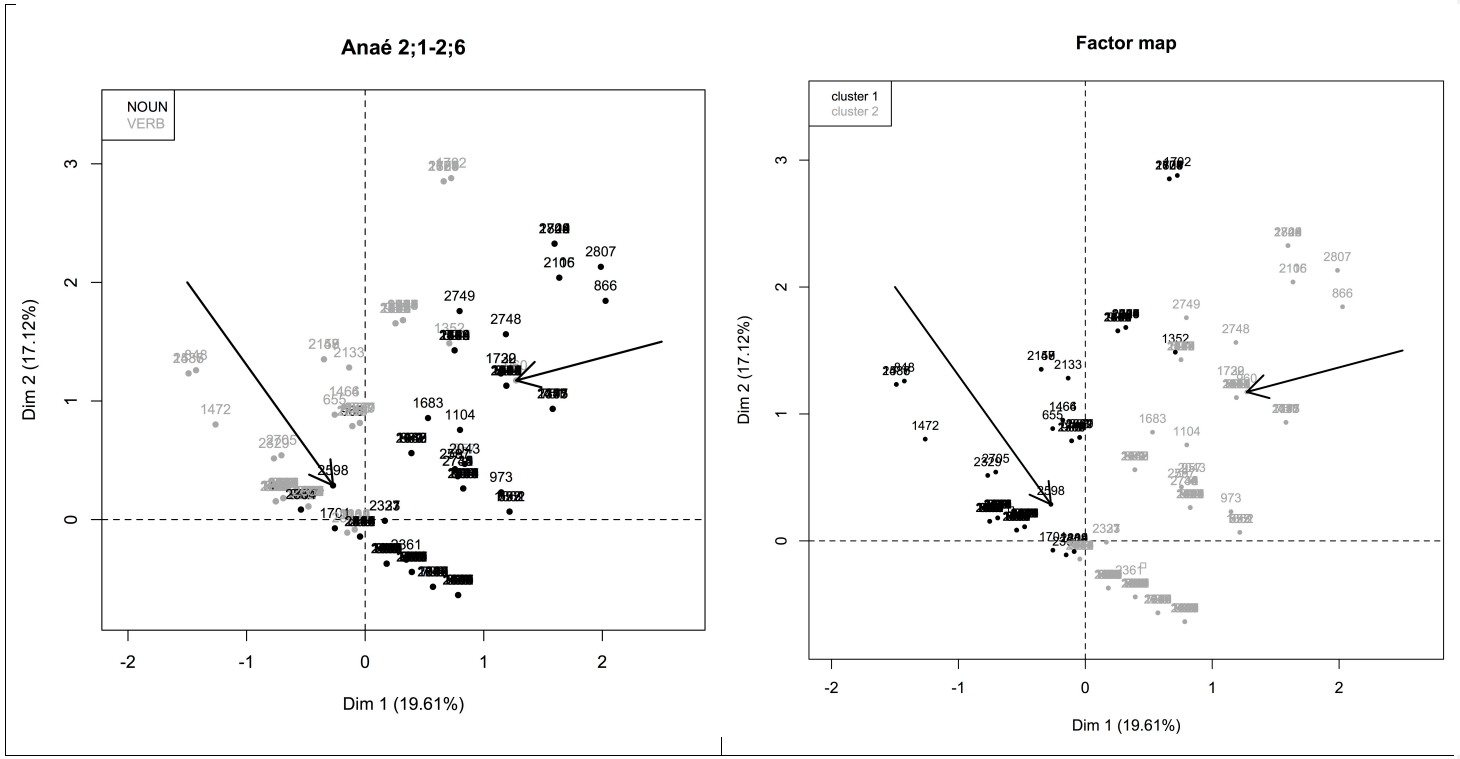

Figure 2: MCA and HC analyses for Anae at age 2;1 to 2;6