Introducing New French Child Data: Thoughts on their Gathering and Coding

Katerina PALASIS
Laboratoire BCL, UNSA, CNRS, MSH de Nice[1]

Abstract: This article browses through a number of significant steps which face an investigator while compiling a corpus for acquisitional research. It ranges from thoughts when gathering the information (the choice of the informants, their number, representativity, etc.) to its final coding in order to eventually share an efficiently constructed database with the scientific community.

Keywords: acquisition, generative syntax, spontaneous data, data representativity, linguistic and methodological variables, coding, French, subject, null subject.

1. Introduction

It is well-known that collecting, transcribing and labelling child data are time-consuming and demanding tasks. However, it is also acknowledged that a reliable and appropriately transcribed and coded corpus turns out to be worthwhile and enlightening. For instance, thanks to the software provided by the CHILDES database (Child Language Data Exchange System)[2], a total of 30,542 spontaneous French utterances from three adults and twenty-two children between 2;5 and 4;0 have now been analysed with regard to the acquisition of the syntactic subject, with interesting outcomes (Palasis (2009b)).

However, the way from child spontaneous utterances through to acquisitional theory is treacherous and many intermediary steps must be achieved before one or several

[1] 98 bd E. Herriot, 06200 Nice, France.

[2] Cf. MacWhinney (2000a, 2000b) and the website at http: //childes.psy.cmu.edu/.

hypotheses can emerge from child recordings. Hence, the aim of this article is to account for my own experience while compiling two corpora (Corpus N°1 as analysed in Palasis (2005) and Corpus N°2 as analysed in Palasis (2009b)) by sharing thoughts on what are perceived as cornerstones in this domain. However, it is quite obvious that all the matters can not be addressed within these pages. Consequently, two particular topics have been chosen, i.e. data gathering and data coding, and since my experience stems from work with children above 2;5, I refer the reader to Morgenstern & Parisse (2007) in order to complement the picture insofar as these scholars mainly address complementary issues such as data interpretation and transcript with children between 1;0 and 3;0.

The first section of this article goes through the different issues dealt with before and while gathering the two above-mentioned corpora. Questions such as data reliability and representativity are hence addressed. The aim of these thoughts is to identify the different variables linguistic data can display in order to eliminate the undesirable ones and master the others. This methodology, in turn, provides the investigator with an empirical background that will allow him/her to forward reliable theoretical generalisations. The second step which is addressed is data coding and the two coding schemes put together for Corpus N°2 in order to obtain a fine-grained description and analysis of the syntactic notion of “subject” are detailed. These two tiers, i.e. %mor (morphosyntactic) and %err (non-target), are part of the many suggested by CHILDES.

2. What are reliable data?

“L’analyse ne vaut que ce que vaut le corpus”.[3] Indeed, there is no doubt that, when a sample is of poor quality, whatever the field of research, the subsequent analyses and generalisations are also likely to be flawed. Since linguists work from samples, i.e. corpora, this qualitative issue is also central in this particular domain. Consequently, the first question that needs to be addressed is: what does “good quality” mean with regard to

[3] “The analysis is only worth what the corpus is worth” (Dalbera (2002:94)).

linguistic data? I pin down the answer to “representativity”. However, it is stating the obvious to say that working from a representative corpus gives more breadth to the theory which then stems from it. So the crucial concern lies elsewhere. Indeed, the problem with this statement is its circularity: how can we get to know that a corpus is representative of a linguistic system as, if we knew this language so well, we would not be studying it? Such a corpus can hence not be defined empirically. Rather, its criteria must pertain to more general principles which I claim are linked to the notion of variation.

2.1 Individual variation

Each individual has his/her own characteristics, each of us having his/her own personality which in turn is rooted in a particular era, place, culture, education, etc. Quite obviously, all these differences surface in language, whether we consider adults or children. Consequently, as far as acquisitional research is concerned, I argue that working from data uttered by a single child can present pitfalls if the scientific aim is to forward generalisations with regard to language acquisition. Indeed, the individuality of child speech is a well-established fact: De Boysson-Bardies (1996) for instance insists on this type of variation and on its importance within syntactic research. Hence, in order to compensate this individual variation, it then sounds safe to assume that working from data which come from different children reduces this risk. Cohen (1924) for instance already favours such an approach:

> Un fait observé chez l’enfant n’est bien utilisable […] que si la part originale de l’individu peut y être délimitée, ce qui ne se réalise bien que par des comparaisons nombreuses.[4]

The study of the verbal system of seventeen children between 2;3 and 3;1 (Palasis (2005)) confirms the impact of this individual variation on data analysis. Indeed, Table 1 hereunder

[4] Cohen (1924:34): “A fact observed with a child can be correctly interpreted […] only if the original part of the individual can be delimited, which can only be correctly realised with numerous comparisons”. More recently, the same kind of approach is also encouraged in Demuth (1996) for instance.

displays the occurrences for the four most uttered verbs in a particular type of sentence within this corpus, i.e. *être* 'be', *adorer* 'love', *avoir* 'have', and *faire* 'do/make'. The penultimate column illustrates that Malcolm is the only child out of the eleven present in this extract who utters the verb *adorer*. If the analysis had relied on that one child or if the statistics had been exploited globally (as shown in the second column), this particular verb would have been ranked second within these standings. However, the other columns hereunder illustrate that Malcolm's utterances are not representative of the whole group of children at least at two levels. First, Malcolm is the only child who utters the verb *adorer*. Secondly and more generally, the whole corpus points out that children within this age group only utter very few first-group verbs (5.1% of all their verbs).

Verbs	Tot	Mat	Jul	Te	Th	Ma	Ra	No	Al	Ali	Mal	Tho
être	81	1	2	1	11	4	46	1	3	10		2
adorer	14										**14**	
avoir	14				1	2	7	1	1	1	1	
faire	13		1	1	4		6			1		

Table 1: Verbal occurrences (extract, Corpus N°1)

Consequently, I argue that collecting data from different children is an important methodological step towards representativity. On the other hand, the above data also illustrate that this method is insufficient and that a *a posteriori* filtering out step must be added in order to master and eliminate possible individual characteristics. As far as the above-mentioned corpus is concerned, Malcolm's utterances of *adorer* were hence eliminated of the final and general statistics on the verbal system.

Moreover, further reasons imply that quantity *per se* does not guarantee representativity. Indeed, gathering numerous data in a random fashion can be counterproductive too as the collected samples can also include a number of other variables which can render an analysis partial or wrong.

2.2 Collective variation

Besides the above-mentioned individual variation, four types of collective variation are traditionally referred to in order to account for the differences which exist between the various linguistic systems. Indeed, variation can be diachronic (as far as acquisition is concerned, intergenerational evolutions can be considered), geographic (e.g. French in France vs. French in Quebec), diastratic (e.g. the influence of the different parental socioprofessional categories), and diaphasic (e.g. when a child speaks with a non-familiar adult vs. when he/she speaks with a family member or with another child). If a set of data used as one whole corpus cumulates two or more different sources of variation, it then makes it particularly difficult to identify which characteristics of the data are to be ascribed to which type of variation. This situation can then lead to a dead end in terms of analysis. Let's illustrate this claim with a tentative comparison of two child corpora.

For her study on the acquisition of wh-questions in French, Crisma (1992) relies on the data of one child, namely Philippe.[5] Some of his questions are repeated in (1) hereunder.

(1) Some of Philippe's questions:
(a) *Où* il est le fil ? "where is the thread"
(b) *Où* elle est la petite aiguille ? "where is the little needle"
(c) *Où* elle est maman ? "where is mummy"
(d) *Où* elle est la voiture grosse voiture ? "where is the big car"

With these French data, Crisma (1992) aims at testing the "clausal truncation" hypothesis as forwarded by Rizzi (1992) for English according to which a child who omits a subject in an utterance (i.e. the [Spec, IP] position) also omits all the elements which could appear above IP in the same clause (e.g. the [Spec, CP] position which hosts the moved wh-elements). In other words, if IP is not projected, nothing is projected above IP. As far as French is concerned, Crisma (1992:117)

[5] Data from Suppes, Smith & Léveillé (1973) available on the CHILDES database.

concludes: “As the data clearly show, the co-occurrence of a wh-element and a null subject is impossible”.

Within my own first study of the so-called “null-subject” phenomenon in French (Palasis (2005)), I attempted to compare Philippe’s questions with questions from Corpus N°1 in order to test Rizzi’s hypothesis against a different set of French data. Consequently, I used the 30 wh-questions uttered by one of the children of this corpus, namely Raphaël. Some of his questions are repeated in (2) hereunder.

(2) Some of Raphaël’s questions:

(a)	Il est *où* ?	“he is where”
(b)	*Où* il est la toute ?	“where it is all of it”
(c)	Elle est *où* ?	“she is where”
(d)	Je le mets *où* celui-là ?	“I put it where that one”

Wh-elements can either remain in their base position (as in (2a) above) or move to [Spec, CP] (as in (2b)). Philippe and Raphaël are about the same age, nevertheless their utterances illustrate the two possibilities with regard to wh-placement, i.e. systematic movement for Philippe vs. the *in situ* strategy for Raphaël (only 7 questions out of his 30 imply movement, 6 of which display *pourquoi* ‘why’; *où* ‘where’, *quoi* ‘what’, and *comment* ‘how’ overwhelmingly remain *in situ*). Consequently, trying to falsify the hypothesis according to which wh-movements and null subjects are incompatible is an impossible task when working from Raphaël’s data. Indeed, the purported incompatibility between wh-elements in [Spec, CP] and the so-called child “null subject” phenomenon can not be studied with Raphaël since his three null-subject questions do not display a wh-movement to [Spec, CP], as illustrated in (3) hereunder.[6]

(3) Raphaël’s “null subject” wh-questions:

(a)	Est *où* cachée ?	“is where hidden”
(b)	Est *pourquoi* ? (x2)	“is why”

His utterances hence neither infirm nor confirm the above-mentioned hypothesis. So, on the one hand, we have Philippe’s

[6] Note however that (3a) seems to display some kind of wh-movement since the wh-element does not surface in its base position after the participle.

data – with wh-movement – that confirm Rizzi's hypothesis and, on the other hand, we have Raphaël's data – which display the opposite, *in situ* strategy – whose study is pointless with regard to this particular issue. Nevertheless, the discrepancy between these two children arouses curiosity.

Indeed, it then becomes interesting to identify the reasons of such a difference with regard to wh-placement between Philippe and Raphaël. However, comparing these two sets of data turns out to be vain as several explanations are possible and no particular conclusion can be reached. Firstly, since these two corpora were collected in 1971 and 2003 respectively, their difference with regard to the wh-elements could be ascribed to diachronic (intergenerational) variation since the *in situ* strategy is often associated with current oral French. Secondly, the difference between Philippe and Raphaël could also be ascribed to diastratic or diaphasic variation. The accumulation of several variables can hence be considered as parasitic insofar as these different variables cloud the issue. Moreover, comparing only two children brings us back to the first matter addressed in this section with regard to speech individuality. Consequently, I claim that diachronic variation must be avoided within a corpus and that diastratic and diaphasic variation must be counterbalanced by a large number of informants in order to forward as representative as possible a survey.

As far as geographic variation is concerned, it is established that French in France substantially differs from Belgian, Swiss, or Canadian French. Indeed, this geographic variation is illustrated at all levels since phonology, morphology, syntax, and lexicology are concerned. With regard to syntax, Auger (1994) for instance describes the interrogative marker *–tu* as specifically Canadian and De Cat (2005) illustrates the different interrogative strategies in Belgium, Canada, and France with regard to subject-verb inversion and insertion of *est-ce que*. This is why I do not work from data which come from different French-speaking countries and that it is claimed that geographic variation also represents a parasitic factor within a corpus.

2.3 Methodological variation

On top of these linguistic sources of variation, methodological diversity can also arise between corpora. Indeed, many factors can interfere in the way data are collected, transcribed, coded, and analysed: the era, the country, the theoretical background, the matter of the study, the type of investigation, the equipment, the persons in contact with the children, the protocols of research, the transcription conventions, etc. However, all these differences seem quite easy to identify and hence to cancel. It then sounds quite straightforward to postulate that data collected by one person or one team will not present these kinds of differences.

2.4 Outcome

To summarise so far, it has been argued that some sources of variation must absolutely be prohibited within a corpus if the aim is to provide broad generalisations on language acquisition. These sources of variation have been identified as being diachronic, geographic, and methodological. On the other hand, other types of variation are necessary within a set of data in order to guarantee its representativity. Indeed, it has been mentioned that individual variation (which in turn can include diastratic and diaphasic variation) can also represent a pitfall. Consequently, I favour an approach which takes a broad number of informants into account, since, as illustrated with Malcolm in Section 2.1, this methodology enables to identify peculiarities and filter them out.

Nevertheless, these conclusions are insufficient since many questions still need to be addressed, e.g. how many children to interview, how to select them in order to be representative, etc.? At this stage, the issue could then be summarised as follows: what kind of child group could be linguistically representative of a defined geographic area at a precise moment in time? And quite a satisfactory answer to this issue was eventually found. Indeed, it was considered to work from an already-constituted group, homogeneous in age, time, and space, and often regarded as generally representative of the adult population, i.e. an entire kindergarten class. Consequently, Corpus N°2, as analysed in Palasis (2009b), stems from interviews with a total of twenty-two children from a first-year class in a French *maternelle*. These children together with their two caretakers and myself were audio-recorded and video-taped in school over a period of seven months, as detailed in Table 2 hereunder.[7]

Language	L1 French (spontaneous)
Place	South of France
Dates	From Nov. 2006 to June 2007
Longitudinal study	7 months (to be continued)
Sessions	13
Intervals between sessions	Min.:10 days; Max.: 40 days
Children	22
Ages first session	Between 2;5.5 and 3;4.24
Ages last session	Between 3;0.13 and 4;0.1
Adults	3
Recordings	Audio and video
Transcript tier	Orthographic
Coding tiers	Morphosyntactic (%mor) Non-target (%err)
Utterances (total)	30542
Utterances (children/adults)	17500/13042

Table 2: Database specifications (Palasis (2009b))

Additionally, all these recordings were transcribed and coded by one unique person along one single transcript and coding protocol, i.e. CHAT (Codes for the Human Analysis of Transcripts) as provided by the CHILDES database. Hence, these facts guarantee further homogeneity at the methodological level.

2.5 Towards an alternative to the competence vs. performance dichotomy

It is well-known that generative grammar does not favour such an approach based on language observation, i.e. "performance".

[7] These children were recorded and filmed over a total period of three years however Corpus N°2 in Palasis (2009b) only displays the data corresponding to the first year.

A statement such as the one found in Chomsky (1965:4) illustrates this rationale:

> Observed use of language may provide evidence but surely cannot constitute the actual subject matter of linguistics, if this is to be a serious discipline.

On the other hand, the relative absence of data within generative research can be considered as one of the drawbacks to this theory:

> […] generative grammar has produced many explanatory hypotheses of considerable depth, but is increasingly failing because its hypotheses are disconnected from verifiable linguistic data. Issues of frequency of usage are by design made external to matters of syntax, and as a result categorical judgments are overused where not appropriate, while a lack of concern for observational adequacy has meant that successive versions have tended to treat a shrinking subset of data increasingly removed from real usage.[8]

After having read the above extract, one could imagine that Manning (2003) rather favours corpus linguistics such as the approach chosen by Tomasello (2003) with regard to acquisition. However, this is not the case either:

> On the other side, corpus linguistics […] or "usage-based models of grammar" […] has all the right rhetoric about being an objective, falsifiable, empirical science interested in the totality of language use, but is failing by largely restricting itself to surface facts of language, rather than utilizing sophisticated formal models of grammar, which make extensive use of *hidden structure* (things like phrase structure trees and other abstract representational levels).[9]

As far as Tomasello (2006:5-6) is concerned, he claims that generative grammar is "more adult-centered" than usage-based

[8] Manning (2003:296).

[9] Manning (2003:296).

theories of acquisition, which he contrastively defines as "more child-centered" approaches. These different statements illustrate that there is a deep methodological dichotomy between these two approaches, which in turn seems to hinge upon one main contrast, i.e. formal, symbolic-based theories such as generative grammar do not stem from the observation of many empirical facts whereas less symbolic theories, e.g. connectionism, emerge from broad data observation. Hence, as far as scientific conclusions are concerned, abstraction seems to manage without reality and *vice versa*.[10] Moreover, this mutual exclusion is not just a current trend in linguistics; rather it is deeply rooted in theory since it brings us back to the seminal Chomskyan "competence" vs. "performance" and Saussurian "langue" vs. "parole" dichotomies.

Although theoretically essential and indispensable, it is argued that such dichotomies could be slightly qualified in order to avoid the mutual exclusion they entail. Indeed, I claim that both could be associated within a unique scientific reasoning thus taking up a third, intermediary stance which combines the formal, generative model of grammar with the observation of many empirical facts. More precisely, it is argued that abstract, theoretical hypotheses can arise from language observation.[11] The close scrutiny of Corpus N°2 for instance leads to split the Pro-drop Parameter into two morphosyntactic parameters and to argue that the so-called "null-subject" phenomenon is generated by principles of economy applied by the child to his/her linguistic system.[12] However, in the first part of this article, it was also argued that not all kinds of data can lead to such conclusions and that a corpus must meet certain criteria which were subsumed with one term, i.e. representativity. Therefore, such data can not be strictly related to "performance" anymore since, as mentioned above, these linguistic facts are controlled and filtered. One

[10] Also see Scheer (to appear:chapter 13) for a historical overview on this opposition.

[11] Also see Oliviéri (2009, 2010) on this matter with regard to syntax and dialectology.

[12] See Palasis (submitted) and Palasis (2010) respectively for further details.

then obtains an intermediary level of access which includes a part of abstraction. Such an intermediary tier reminds us of the intermediary representational level within the Government & Binding framework, i.e. S(hallow)-Structure, insofar as this level was not the deepest in the representation but it was not the surface level either since it displayed abstract items such as traces.

3. Coding child data

The second step addressed in this article is the importance of data coding. Indeed, one obvious peculiarity of a large corpus is the high number of its utterances which makes it impossible to handle manually or by memory. Consequently, in order not to lose the benefit of the quantity, the information must then be processed in an efficient manner so that any aspect of the data can be extracted and analysed. Within the CHAT files, once the investigator has transcribed his/her data on a main tier, as illustrated in (4a) hereafter, he/she can then choose from a broad variety of dependent tiers which represent different coding domains, e.g. phonology, morphosyntax, syntax, intonation, etc. It is hence expected that each investigator will select one or two tiers from this list according to his/her domain(s) of research. The study of the acquisition of the syntactic subject has led to favour two particular tiers, i.e. %mor and %err, which are illustrated in (4b) and (4c) hereunder and which are then described in the following two sections.

(4) The transcript (Alan, 2;8.18):
(a) Main tier: et moi je suis [*] fait un escargot .
(b) First coding: %mor: conj|et,
pro:ton:dg:nom|moi&1S,
pro:cli:d:nom|je&1S,
v:aux|être&PRES&1SV [*],
v:mdllex|faire&PP&_MASC&_SING,
det|un&MASC&SING,
n|escargot&_MASC .
(c) Second coding: %err: suis = ai $LEX $SUB $AUX

3.1 The morphosyntactic coding tier (%mor)

The first coding tier is morphosyntactic (%mor): it displays information on the nature of the items (nouns, pronouns, verbs, etc.) as well as information on various features such as person, gender, or case for instance.

More specifically, particular attention was devoted to the coding of the subject clitic and non-clitic items in this database, as illustrated in Table 3 hereafter for the first person forms *je* and *moi*. Indeed, these different codes show that these items were coded differently depending on the following conditions: (i) they are either clitic or strong pronouns (lines a and b vs. c, d, e, and f), (ii) they appear on their own (lines a and c) vs. along with a coindexed item (line b: ‘d’ in the code stands for ‘doubling’), (iii) the coindexed element is at the left of the clitic (line d: ‘g’ stands for *gauche*) vs. at its right (line e: ‘d’ for *droite*), or both (line f).

Forms	Codes & Examples	
j(e) ‘I’	a	pro:cli:nom|je&1S *j'* ai perdu l(e) chien . “*I* have lost the dog”
	b	pro:cli:d:nom|je&1S attends *je* l' ai *moi* . “wait *I* have it I_{strong}”
moi ‘$\mathrm{I}_{\mathrm{strong}}$’	c	pro:ton:nom|moi&1S *moi* 0 [*] veux ça les cartes . “I_{strong} want that the cards”
	d	pro:ton:dg:nom|moi&1S *moi j(e)* suis jaune ! “I_{strong} *I* am yellow!”
	e	pro:ton:dd:nom|moi&1S *j'* ai perdu *moi* . “*I* have lost I_{strong}”
	f	pro:ton:dg:nom|moi&1S and pro:ton:dd:nom|moi&1S *moi j'* ai pas encore fini *moi* . “I_{strong} *I* have not yet finished I_{strong}”

Table 3: The %mor coding of *je* and *moi*

Similar codes were also applied to the five other persons, hence accounting for a total of thirteen different clitics, i.e. *je, tu, il,* expletive *il, elle, on, ce, ça, nous, vous,* polite *vous, ils,* and *elles* whether these elements appear on their own (a total of fourteen different codes since expletive *il* appears preverbally as well as postverbally) or are coindexed with another item (a total of thirteen codes).[13]

3.2 The error coding tier (%err)

As illustrated with line c in the above table, young children's utterances include a number of non-target structures and elements. The database consequently also displays an "error" line (%err) which labels all the non-target utterances symbolised with [*] on the main tier and for which an accurate and levelled coding system was also devised. In order to be of interest to acquisitionists whatever their field of research, all the different types of non-target utterances were tagged on the main tier and further described on %err. Hence, this database provides the investigator with a total of 130 different codes which pertain to syntax ($SYN) and morphology ($MOR) as well as phonology ($PHO) and lexicology ($LEX). Each of these main codes is then complemented with one or more indications which describe the phenomena as precisely as possible, as illustrated in (5) hereafter.[14]

(5) Some of the 130 code combinations available on %err:
(a) Determiner omission:
NOE: oh la [: y+a] 0 [*] oiseau dedans !
"oh there's 0 bird inside!"
%err: 0 = un $SYN $LOS $DET $INDEF
(b) Gender substitution on a subject clitic:
ZOE: xxx moi il$_i$ [*] a pris ça à moi Noémie$_i$.
"me he$_i$ [*] has taken that from me Noemie$_i$"
%err: il = elle $MOR $SUB $AGA $PRO

[13] See Palasis (2009b:228 et sqq) for the detail.

[14] The whole list together with illustrations are available in Palasis (2009b:372-377).

(c) Unexpected liaison:
NOE: les mamans zarrivent [*] .
%err: zarrivent = arrivent $ALL $ADD

(d) Addition of a consonant applying perseveration:
ZOE: i(l) va arriver le Papa_Nonël [*] [: Noël] .
“he will arrive Father Christmas”
%err: Nonël = Noël $PHO $ADD $PER $CON

(e) Substitution of an auxiliary:
ALA: et moi je suis [*] fait un escargot.
%err: suis = ai $LEX $SUB $AUX

As far as the nominative clitic omission is concerned, Table 4 hereafter displays the total range of codes available. The examples in (6) further illustrate each possibility.

%err levelled codes	Persons	Missing elements
$SYN $LOS $PRO $SUBJ	unidentified	?
$SYN $LOS $PRO $SUBJ $1S	1	je
$SYN $LOS $PRO $SUBJ $2S	2	tu
$SYN $LOS $PRO $SUBJ $3S	3 ref.	il, elle
$SYN $LOS $PRO $SUBJ $IMPRS	3 expl.	il
$SYN $LOS $PRO $SUBJ $DEM	3 dem.	ce
$SYN $LOS $PRO $SUBJ $3P	6	ils, elles

Table 4: Coding of the subject clitic omissions

(6) Examples of %err coding: the nominative clitic omissions:

(a) TOM: 0 [*] veux [=? faut] remett(r)e ça .
“(?) want to put that back”
%err: 0 = ? $SYN $LOS $PRO $SUBJ

(b) MAX: euh 0 [*] sais pas .
“(I) don’t know”
%err: 0 = je $SYN $LOS $PRO $SUBJ $1S

(c) ALA: 0 [*] as vu j' ai rangé .
“(you) have seen I have cleaned up”
%err: 0 = tu $SYN $LOS $PRO $SUBJ $2S

(d) TOM: 0 [*] veut pas manger .
“(he) doesn’t want to eat”
%err: 0 = il $SYN $LOS $PRO $SUBJ $3S

(e) CEL: 0 [*] faut pas la casser .
“(one) mustn’t break it”
%err: 0 = il $SYN $LOS $PRO $SUBJ $IMPRS
(f) EMA: 0 [*] est des feuilles .
“(these) are leaves”
%err: 0 = c' $SYN $LOS $PRO $SUBJ $DEM
(g) EKT: 0 [*] [?] mangent [?] des herbes !
“(they) eat grass”
%err : 0 = ils $SYN $LOS $PRO $SUBJ $3P

3.3 Computerized Language Analysis (CLAN)

The two coding tiers described above then enable the investigator to submit very accurate requests to the CLAN programs (Computerized Language Analysis) in order to extract data. Moreover, the combination of both these tiers furnishes the researcher with efficient tools that allow him/her to access a broad range of phenomena directly in two concomitant ways. First of all, a host of characteristics in the child linguistic system can be grasped by analysing the frequency or co-occurrence of any of the items coded on the %mor tier (types of nouns, verbs, adjectives; which grammatical persons are uttered; which gender, number, or tenses are favoured; mastery of negation, etc.). Secondly, the %err coding then provides direct information on all the non-target utterances which bear on the selected matter, whether the utterances result from addition, substitution, or loss of an item and whether the mechanism resorts to phonology, morphology, syntax or the lexicon, as illustrated in (5) above.

4. Conclusions

By and large, as far as language acquisition is concerned, two different types of approaches are usually contrasted. Indeed, formal linguistics, e.g. generative grammar, are usually opposed to usage-based theories, e.g. connectionism, following two main criteria: (i) theoretical representations (abstract vs. actual) and (ii) observed use of data (scarce vs. broad). This article has argued in favour of a third, intermediary approach which associates formal and corpus linguistics insofar as I aim at formalising abstract, theoretical hypotheses within the latest generative framework using what can be considered as “broad” corpora from a generativist point of view. However, going through the different variables linguistic data can display (individual, collective, and methodological), it has also been claimed that quantity *per se* does not guarantee quality which is pinned down to representativity (Section 2). Moreover, quantity then requires efficient data processing. Section 3 hence details the morphosyntactic (%mor) and non-target (%err) coding grids applied to Corpus N°2 (Palasis 2009b) in order to furnish the investigator with as many holds as possible on the data, whether coarse or fine-grained.

The amount of data (30,542 utterances) together with the conditions under which this information was collected, transcribed, and coded provide the researcher with what is assumed to be a sound foundation stone for theoretical investigation. I hence feel that crossing these two coding tiers on information gathered from twenty-two different children has already allowed to shed new light on various long-standing matters such as the child null-subject phenomenon (Palasis (2010)) and the status of the nominative clitics in oral French (Palasis (2009a)). However, many other issues remain to be studied in order to forward an in-depth study and formalisation of the French child grammar which I name “spontaneous” French (Palasis (submitted)).

References

Auger J. (1994). *Pronominal Clitics in Québec Colloquial French: A Morphological Analysis*. PhD Dissertation, University of Pennsylvania.

Chomsky N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: The MIT Press.

Cohen M. (1924). "Sur les langages successifs de l'enfant", *L'acquisition du langage oral et écrit* 31 (1993), 33-43.

Crisma P. (1992). "On the Acquisition of Wh-Questions in French", *Geneva Generative Papers* 0, 1-2: 115-122.

Dalbera J.-P. (2002). "Le corpus entre données, analyse et théorie", *Corpus* 1, 89-104.

De Boysson-Bardies B. (1996). *Comment la parole vient aux enfants. De la naissance jusqu'à deux ans*. Paris: Editions Odile Jacob.

De Cat C. (2005). "French Subject Clitics are not Agreement Markers", *Lingua* 115, 9: 1195-1219.

Demuth K. (1996). "Collecting Spontaneous Production Data", *in* McDaniel D., McKee C. & Cairns H.S. *Methods for Assessing Children's Syntax,* Cambridge, MA: The MIT Press, 3-22.

MacWhinney B. (2000a). *The CHILDES Project. Tools for Analyzing Talk, Third Edition. Volume I: Transcription Format and Programs*. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney B. (2000b). *The CHILDES Project. Tools for Analyzing Talk, Third Edition. Volume II: The Database*. Mahwah, NJ: Lawrence Erlbaum Associates.

Manning C.D. (2003). "Probabilistic Syntax", *in* Bod R., Hay J. & Jannedy S. *Probabilistic Linguistics,* Cambridge, MA: The MIT Press, 289-341.

Morgenstern A. & Parisse C. (2007). "Codage et interprétation du langage spontané d'enfants de 1 à 3 ans", *Corpus* 6 "Interprétation, contextes, codage", 55-78.

Oliviéri M. (2009). "Syntactic Parameters and Reconstruction", *in* Kaiser G.A. & Remberger E.-M. *Null-subjects, Expletives and Locatives in Romance,* Konstanz: Konstanz Working Papers of the Department of Linguistics, 27-46.

Oliviéri M. (2010). "From Dialectology to Diachrony", *in* Upton C. & Heselwood B. *Proceedings of Methods XIII: Papers from the 13th International Conference on Methods in Dialectology,* Frankfurt: Peter Lang, 42-52.

Palasis K. (2005). *Problèmes d'acquisition et le Paramètre du Sujet Nul*. Mémoire de Master, Université de Nice-Sophia Antipolis.

Palasis K. (2009a). "On the Variable Morpho-Syntactic Status of the French Subject Clitics", *in* Kaiser G.A. & Remberger E.-M. *Null-subjects, Expletives and Locatives* *in Romance,* Konstanz: Konstanz Working Papers of the Department of Linguistics, 47-62, http://ling.uni-konstanz.de/pages/publ/arbeitspapiere.html.

Palasis K. (2009b). *Syntaxe générative et acquisition: le sujet dans le développement du système linguistique du jeune enfant*. Thesis, Université de Nice-Sophia Antipolis.

Palasis K. (2010). "Principles of Economy within the Child Speech: when the Nominative Clitic does not Surface", poster presented at *The Romance Turn IV. Workshop on the Acquisition of Romance Languages*, Tours, 25-27 August 2010.

Palasis K. (submitted). "From Spontaneous to Normed French: *il* and *ne* as Conclusive Indicators of two Differently Parameterized Grammars", *in Selected Proceedings of Going Romance 23,* Amsterdam: John Benjamins.

Rizzi L. (1992). "Some Notes on Linguistic Theory and Language Development: The Case of Root Infinitives", *Geneva Generative Papers* 0, 2: 102-114.

Scheer T. (to appear). *A Lateral Theory of Phonology. Volume 2: How Morpho-syntax Talks to Phonology. A Survey of Extra-phonological Information in Phonology since Trubertzkoy's Grenzsignale*. Berlin: Mouton de Gruyter.

Suppes P., Smith R. & Léveillé M. (1973). "The French Syntax of a Child's Noun Phrases", *Archives de Psychologie* 42, 207-269.

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

Tomasello M. (2006). "Acquiring Linguistic Constructions", *in* Kuhn D. & Siegler R. *Handbook of Child Psychology: Cognition, Perception and Language, Volume I,* Hoboken, NJ: John Wiley & Sons, 255-298.