An approach to corpus-based language and multimodal features in communicative exchanges at an early age for adapted hypermedia content design

Alejandro Curado Fuentes
acurado@unex.es
University of Extremadura, Spain

ABSTRACT

This paper describes the results from the analysis of English and Spanish corpora from the CHILDES database for the design of adapted hypermedia (AHS) content in English at the pre-elementary school level. In general, linguistic and paralinguistic information from selected CHILDES transcripts can contribute to the organisation of pedagogical content. In the corpus analysis, it is found that many conversational patterns in children's L1, mainly collaborative situations, present significant multimodal aspects, which are often correlated with meta-discursive items and markers. The integration of specific multimodal traits in the AHS lessons can be useful for the learners' L2 development. The use of AHS serves as a naturally resulting resource for multimodality and interactiveness in children's L2 communicative development.

Keywords: *corpus analysis, early age, language learning, collaborative exchanges, multimodality, adaptive hypermedia*.

I. INTRODUCTION

The integration of foreign languages (FL) and Information technologies (IT) in pre-elementary school (years 3 through 5) in Extremadura has led to the design of curriculum material based on surveys and observations of children's learning styles and patterns (cf. "Curriculum de Infantil", published in the Bulletin of Extremadura 2003). During these years, the FL curriculum has seemed to demand a closer look into the way children ought to learn languages. Following professional advices and methods (cf. Wintergest et al. 2003, Ellis 2004), it is found that many specific traits can be observed by analysing real situations where children communicate in their L1 (first language), "practising new words and structures in a way that sounds like a student in some foreign language classes" (Lightbrown and Spada 2006: 12). The cognitive development that takes place in the child's brain is specific and restrained to the use of cognitive skills in those particular domains. His or her "interactions are not restricted to the second language, but affect the native language as well" (Kroll et al. 2008: 109). Language form evolvement mirrors this type of cognitive development, and the same process takes place in early age FL.

To explore communicative exchanges at early age, CHILDES (Child Language Data Exchange System) is managed as a corpus provider (see Section 2 below for the database structure description). A direct and practical approach to natural language analysis is thus sought, derived from our research group's aim to design Adaptive Hypermedia System (AHS) (cf. Brusilovski 1996, 2001) lessons in pre-elementary courses (see web page in bibliography for our group GexCALL).

In this paper, the aim is to describe the main corpus-based results that determine the key linguistic and paralinguistic items in the children's situations observed, and to correlate these items with multimodal information from the corpus for the design of the L2 lessons in the AHS. Repetition and frequency are two key factors in the collaborative exchanges analysed, while the verbal and non-verbal communicative traits examined involve multimodal elements to take into account in the learning/communicative process.

II. THEORETICAL FRAMEWORK

The corpus is compiled by selecting specific directories and folders in the CHILDES database. A directory is a group of speakers from a certain country, while the folders contain the number of transcripts recorded for that directory. Table 1 displays the folders for the database directory "USA English", as children speaking English (with adults and/or other children) as L1 are a main target group.

Table 1. Folders selected for the "USA English" directory in the CHILDES database.

Corpus folder	**Age Range**	**Comments**
Bates on page 3	1;8 and 2;4	Two sessions at 1;8 and two at 2;4
Bernstein-Ratner on page 5	1;1–1;11	Mother child dyads during the earliest stages of language with play sessions
Bliss on page 7	3–10	Control participants for a study of SLI
Bohannon on page 12	Nat 2;8 and 3;0 Baxter 3;0	Interactions in a laboratory setting of different adults with two children
Brown on page 14	Adam 2;3–4;10 Eve 1;6–2;3 Sarah 2;3–5;1	Large longitudinal study of three children
Warren-Leubecker on page 74	1;6–3;1 4;6–6;2	Parent–child interactions

The folders are selected according to the age ranges and types of participants in the studies. The folder name usually corresponds to the researcher's or analyst's who recorded and transcribed the corpus. The hyper-linked page number in the corpus folder directs the user to the contents for that folder on the web, retrievable free of charge. For other languages and nationalities, CHILDES offers many other directories (e.g., English from UK, Spanish from Spain, Catalan, etc). In addition to the six folders from USA English, seven folders were chosen for Spanish from Spain, and four other folders from Bilingual speakers of English and Spanish in USA, as described below.

In all corpus-based analyses, lexical repetition and frequency are two key factors, but for child language, this premise is core not only for lexical analysis but also for the observation of communicative development and strategies, in agreement with previous works (e.g., Langacker 2000, Lightbrown and Spada 2006, Bybee 2006, 2008, Hudson 2008). This approach is feasible in children's L2 learning. For instance, Hudson (2008: 103) claims that "language is learned (...) rather than 'acquired' by the triggering of innate concepts (...) L2 can be viewed as a body of knowledge like any other, to be learned and taught by experience". This view is "controversial in linguistics" (Hudson 2008: 103), but it is held as convincing in much research.

The point is that small children, being exposed to a wide range of conversational input (i.e., Child Directed Language—CDL—cf. Buttery and Korhonen 2005, Brodsky et al. 2007), may come across similar or different linguistic/paralinguistic forms, used in collaborative situations. One example is direct request, manifested in the extended use of transitive verbs, like *want* and *like*. The correlation of "visual-spatial stimuli" conveyed with the functional and pragmatic items uttered would help to better analyse and understand the communicative exchanges (in agreement with Coventry and Guijarro-Fuentes 2008: 133).

To observe such patterns, CHILDES integrates, as mentioned above, a vast collection of recordings and transcripts. Our analysis of the data focuses on selected transcripts that are then edited with a specialised tool (called CLAN—corpus language annotator—) for the labelling of linguistic and extra-linguistic information. The amount of text in the CHILDES database is heterogeneous because it is intended for different research aims (e.g., linguistic, pedagogical, psychological, etc –cf. MacWhinney 2000). In agreement with Biber et al. (1998: 246), and Bowker and Pearson (2002: 104), text selection in this kind of database should be done according to specific purposes for empirical language study and curriculum, and by also using sub-corpora or text categories (cf. Hunston 2002, Flowerdew 2004).

For our corpus selection we mainly aim to be able to contrast socio-linguistic traits in English and Spanish, and also different backgrounds, so as to enrich contexts "in the extent to which their linguistic characteristics may be similar" (Biber 2006: 12), and to seek/identify differentiation from other groups (Nortier 2008: 38). Thus, three categories of recordings are needed: Native English speakers in USA, Native Spanish in Spain, and Spanish used as the dominant language in bilingual contexts of USA.

CHILDES includes many examples of multimodal references in the transcripts, as Tokowicz and Warren (2008: 228) explain: "CHILDES is particularly useful for investigating questions about the kind of input a learner receives, as it provides large samples of actual input". Children's production is not thus the only scope in the analysis, but also their different types of context (CDL annotations) in the learning process (cf. Robinson and Ellis 2008: 501). In the CHILDES texts analysed, multimodality occurs in the form of direct visual references during the conversations that the participants are sharing and interacting with, e.g., drawing objects, animals or people, playing with cards, toys, etc. There are also some auditory references that are considered multimodal (e.g., onomatopeias for animals and things, e.g., mooing, mewing, knocking, thumping, and thundering).

The conversations in the corpus tend to develop spontaneously, as the children participate in games and tasks, reacting to instructions, questions and feedback. Annotating and classifying this word usage appropriately can help to make observations of communicative procedures. Carter (2004: 76) refers to "the creation of fictional worlds and imaginative entry to those worlds (...) regarded as essentially the domain of the growing and developing child". These socially bonding elements in the tasks connect worlds and words: "For example, the speakers use each other's words, employ parallel syntactic forms and generally pattern question and answer replies in such a way as to suggest high degrees of affective connection and convergence" (Carter 2004: 101). Lexical and grammatical usage result from these connections, i.e., "cognitive development, including language development, arises as a result of social interactions" (Lightbrown and Spada 2006: 47).

Lexical repetition is quite important in the process. The quantitative view of the data establishes the fieldwork for classification and contrastive study. Lower lexical frequency can be also relevant in the situations observed (Bybee 2008: 231), as the qualitative examination of the data leads to "observation and awareness of what happens" (McCarthy 1998: 59); for example, some repetitions overlap due to "language-in-action collaborative tasks (...) seen as practical and goal-facilitating" (McCarthy 1998: 59). The processing of the linguistic items, when done in a learning-based context, tends to be positive for the enhancement of "communicative competence" (Fulcher and Davidson 2007: 38).

In our study, as stated above, the double-fold research question is whether there are distinctively frequent and widely used linguistic-discursive items in the corpora, and then whether these items can be correlated statistically with multi-modal references in the corpora. The results should be valuable as important verbal and non-verbal information to include in the AHS lessons, items that the learners should master to move across units. Section 3 below will describe the corpus-based analysis done to obtain the most salient (frequent and distributed) linguistic-discursive information. Section 4 then explains how this categorised information is correlated with relevant multi-modal items, pointed out in the corpora. Section 5 includes a description of the inclusion of such salient linguistic and multi-modal data in the AHS lessons, giving some examples. Finally, some conclusions on the most important findings in the study are included.

III. THE CORPUS-BASED STUDY

The conversations were selected from the CHILDES folders according to age and nationality, and whether they suited the situational/communicative purposes of the research. Figure 1 gives a general view of the corpus sources and folders selected. Some texts from years other than 3 to 6 (e.g., 0 to 2, and 7) were included for contrastive aims.

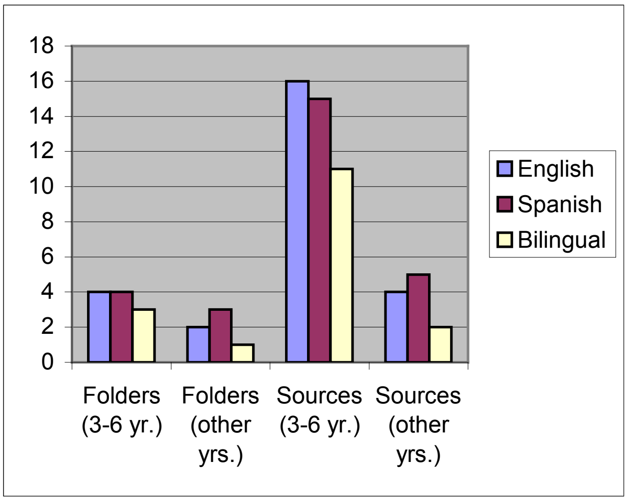

Figure 1. Number of sources and folders in the corpus.

The transcripts were edited to annotate speakers' names and adults' input as co-textual and directed, i.e., as CDL input. Some common directing strategies by adults are questions, commands, prompts, pauses, connectors, and tags. Some other annotations were also made for the identification of characteristic linguistic-discursive items, examined below.

The three categories (English, Spanish, and Bilingual) total 6,077,574 words. Most transcripts include recording sessions that last an average of one hour and 20 minutes. The high repetition of words leads to a low lexical density, measured as distinct words per 1,000 running tokens (Standardised type-to-token ratios). Native English has the highest degree of word repetition, as seen in Figure 2, whereas the highest lexical densities found are for Spanish five- and six-year olds.

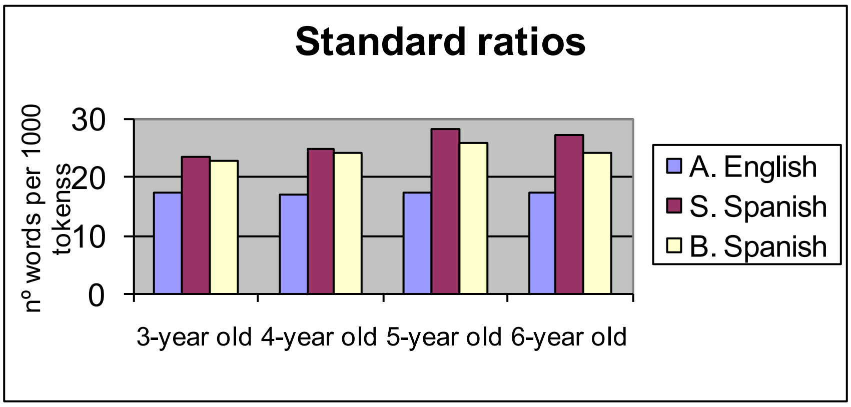

Figure 2. Contrastive view of standardised type-to-token ratios in corpus.

Sentence and word lengths also provide interesting contrastive data. While the bilingual context produces the longest sentences (especially in 4- and 5-year old contexts—up to 90 words for the longest—), Native English speakers use short sentences (an average of 18 words in 3-year olds' contexts). Words tend to have similar lengths.

Word frequency is contrasted with the type of speaker and age level involved. Word lists are arranged in detailed consistency lists (DCL),[1] and then run with the concordance software. The four age divisions produce four different word lists for each of the three nationalities. Table 2 is an example with the 20 most frequent and dispersed items in the DCLs. The American English DCL presents the highest rate of word repetition; this aspect is coherent with its lower lexical density. The Bilingual DCL presents Spanish words as the most frequent and widespread data.

Table 2. Frequency- and range-based analysis by using DCLs
(words are taken as transcribed from the oral texts).

American English (monolingual)		Spain's Spanish (monolingual)		Spanish/English (Bilingual)	
Word	TOTAL	Word	TOTAL	Word	TOTAL
You	30921	A	25204	No	3485
I	27118	No	23096	A	3468
A	23615	Que	19932	Y	3209
Be	23388	La	16372	Que	2843
The	20701	El	16303	El	2162
It	20222	Es	13580	La	2010
What	16925	Se	12636	Sí	1723
To	15343	Qué	12477	Es	1609
Do	14944	De	10391	Eh	1482
That	14056	Sí	10365	Aquí	1386
Dem	10622	Éh	8511	Lo	1272
Not	9415	Lo	7069	Un	1261
And	8774	En	6673	De	1226
Go	8507	O	6071	Se	1191
This	7871	Me	5999	Me	1111
In	7848	Aquí	5951	Cómo	1078
No	7597	Está	5317	Te	1076
On	7351	Mira	5298	Ya	1047
One	7227	Los	5201	Está	946
Have	7128	Mí	4610	Yo	889

Short words (i.e., with few graphemes) repeat the most, being used in dynamic interpersonal exchanges. In many cases children produce such utterances without repeating or emulating adults' words. The age-located instances of children's personal use without intervening adults (i.e., non-CDL) demonstrate that there is a period when particular expressions are uttered individually (e.g., *I want ta go*, or *a mí no me gusta*, both at the 4-year level). This production autonomy hints at the existence of an in-built lexicon in the child's cognitive system (e.g., "go", "want", "like", "gustar", etc.), in agreement with Buttery and Korhonen (2005), Hudson (2008), and Coventry and Guijarro-Fuentes (2008), among others.

Interpersonal language is common in all the contexts, and the children's utterances reflect every day words and worlds, i.e., common semantic-pragmatic references to activities and actions done in collaboration with adults and/or other children. An example is the great reliance made on third person references by the Spanish-speaking children, paralleled by the first and second person forms preferred by the English speakers. Long stretches of conversation tend to take place in the Spanish and Bilingual contexts, with a consequent production of longer sentences, and the exchanges are shorter and more dynamic in English.

For the inspection of these linguistic-communicative traits in the categories, various tables have been built. An example is Table 3, where the comparison is made between 3- and 4-year old levels in the American English context. Linguistic and paralinguistic information is recorded to check if there is age- or nationality-based variation. For instance, one difference at age 4 is that questions are not only posed by adults but also quite often by the child. In turn, at age 3, the adults ask most questions to direct the collaborative exchanges. Thus, to introduce children to simple every day words and sentences may constitute, together with attractive audio-visual stimuli, a sound pedagogical path (in agreement with Hudson 2008, and Coventry and Guijarro-Fuentes, 2008, among others).

Table 3. Items arranged according to age level within a nationality category.

3 and 4		
Freq.	**Field – Year 3**	**Field – Year 4**
1	Do you have... / would you like (CDL) / where did you ... (CDL) / what else did you... (CDL) / why don't you... (CDL) / what do you call... (CDL)	I don't (want) / I don't see (no birds) / I'm finished
2	I don't know / I don't think you (CDL) / I want to (go) / I going to / I don't want to / I want some (more) / mommy, I want (a)	you have to / mommy, you... / how you do it / how do you do it / where you going
3	Chug a chug a chug / make a (dog) (CDL) / make a (plane) /	it looks like a / dis is a / I never heard of a / it's gonna be a
4	Oh yeah? Oh look it	what does it say / you turn it /
5	what kind of... (CDL)	I like to / would you like to (CDL)
6	Play with (+TOY)	what is dis / what is that (CDL)

The other type of table is built by contrasting the statistically significant clusters found at similar relative frequency levels. Table 4 lists the frequent pragmatic forms analysed according to nationality (with added age levels when the expression is distinctly used).

Table 4. Frequency-based expressions according to nationalities, derived from DCL data.

American English (monolingual)	Spain's Spanish (monolingual)	Spanish/English (Bilingual Latin American in USA)
I don't know	*A ver si*	*Y ya está*
I'm goin(g) to (5 & 4 years)	*A lo mejor*	*Y lo pone en*
Mommy, you… (all)	*No sé qué es* (6 & 5 years)	*Y luego* (6 & 5 years)
I'm not gonna (5 years)	*Es que como no…* (6 years)	*Me voy a* + verb (all)
I want ta go (4 years)	*Porque no* + verb (6 & 5 years)	*No me acuerdo* (all except 6 years)
You want to…? (4 & 3 years)	*A mí no me gusta* (6, 5 & 4 years)	*No se puede*
I'm gonna (6 years)	*Mira lo que* + verb (4 years)	*Me parece que* (4 years)
You have to	*Pues creo que*	*Sí es eso*
You open it	*Lo tienes que*	*Y yo también* (all)
I not going to (3 years)	*Y luego* (5, 4 & 3 years)	*Mamita, el de…* (3 years)

A salient feature is the verb *go* in the progressive form (e.g., *be* + *going to* or *be* + *gonna*). It is found that these structures are produced by children at age 4 and above, but not earlier. This observation coincides with the findings in Goldberg and Casenhiser (2008) from a CHILDES selection of two year olds' transcripts, where mothers use *go* in 39 percent of the [subject + verb + object] structures recorded. The pattern is also common in adults' speech with three-year old children, but these children do not use it autonomously in the collaborative exchanges.

In Spanish, children after the age of 4 begin to explain ideas in longer clauses (e.g., *es que como no…*). The same holds true for bilingual children after age 4, when they state more opinions (e.g., *me parece que…*). This fluidity is not detected earlier. Slobin (2000) refers to an example of this lengthy statement usage in Spanish as a "richer imagery" for movement clauses when places are described (Cadierno 2008: 254). Again, the implications for EFL in our pre-elementary context point to the need for verbal simplification and audio-visual stimuli to formulate ideas.[2] In addition, significant vocative expressions and personal preferences/inclinations form a major feature of interpersonal oral discourse in collaborative tasks (Koester 2006: 86), by which children often ask concrete things in the transcripts in all languages, and use negative forms (e.g., *not*, *don't*, *no*, etc) in significant pragmatic functions (e.g., stating likes and dislikes, lack of interest, or being told by adults what they cannot do).

Both linguistic-discursive variation and similarity can be inferred from the relative frequency data analysis. To confirm or refute such observations, a quantitative examination of age-based and nationality-based features should come from a key item computation based on variance and standard deviation. These two parameters can work as a sort of statistical yardstick with which to compare the dispersion of scores around given means (cf., Bachman 2004). The top 60 expressions from each age category can establish means from which variance and standard deviations are calculated. Next, the age categories are run in pairs to contrast the information (e.g., year 3 with 4, 3 with 5, and so forth). This comparison enables the calculation of t-values, which then indicate the degrees of statistical probability that two age categories may have for the use of similar or different linguistic features.[3]

Table 5 displays the three most salient features or dimensions measured in the English and Spanish corpora: 1. Interpersonal (use of first and second person pronouns, vocative words, and commands); 2. Declarative (demonstrative pronouns and adjectives, third person statements, and expressions for preferences and dislikes); and, 3. Markers (discourse connectors, interjections, and gambits). The bilingual category is excluded here because we want to focus on the monolingual data to be extrapolated to the Spanish monolingual learners' context alone. In addition, to my knowledge, a large general bilingual corpus for the comparative analysis is not available.[4]

Table 5. Probability statistics for three discourse features examined in the children's speech.

Nationality/ Age comparison	Interpersonal	Declarative	Markers
American English			
3 <> 4	,4583	,0057	,0593
3 <> 5	,0003	,4923	,5289
3 <> 6	,4660	,2085	,0002
4 <>5	,0000	,0311	,0968
4 <> 6	,0252	,0003	,0000
5 <> 6	,5989	,0629	,0062
Spain's Spanish			
3 <> 4	,3617	,1213	,9714
3 <> 5	,7595	,0052	,1917
3 <> 6	,9027	,0794	,0398
4 <>5	,4110	,9072	,2047
4 <> 6	,3279	,5768	,0434
5 <> 6	,7979	,2432	,4016

Usage probability derives from the calculation of t-scores for each pair set, and these scores have different degrees of freedom. Fisher's and Yates' Table III (Bachman 2004: 336) provides the critical values of t according to such degrees of freedom. A score equal to or over 0.5 would mean that the difference between the two items is due to chance. In Table 5, few contrasted items are different due to chance: 11.2 percent of the cases in American English and 33.4 percent in Spanish. For English, such a distinction is acute (22.2 percent more than for Spanish), i.e., there are markedly objective differences between age levels.

In the English conversations, the age 4-level appears as the recorded period at which a wider use is made of all three discursive dimensions. Needless to say, this difference should not be interpreted as a sign of little or irrelevant linguistic use in the other age categories. Quite the opposite, this information reveals the time when children are most likely to use certain items that characterise overall pre-elementary age conversation in collaborative exchanges.

The score differences can also point to pair set proximity for certain age levels. In other words, the different speakers may produce a similar proportion of discourse features. For example, in American English, age 3 comes quite near year 4 in the use of interpersonal statements (cell 3 <> 4 in Table 5). The production of discourse markers is as significant at age 5 as it is at age 3 (3 <> 5), and the proportion of interpersonal statements is similar at years 5 and 6 (cell 5 <> 6).

IV. MULTIMODAL FEATURES

The data from the linguistic analysis can be correlated with the various visual-spatial stimuli and auditory features that prompt, direct, and/or engulf the conversations. This correlation should form a better image of linguistic and paralinguistic items (cf. Coventry and Guijarro-Fuentes 2008). The spontaneous fictional, imaginative worlds that develop in the conversations are the speakers' own, enhanced by their interaction with other children and adults in playful and collaborative tasks, while cognitive development unfolds as a result (cf. Lightbrown and Spada 2006). The multi-modal items are projected in a learning context, and contribute to fostering "communicative competence" (cf. Fulcher and Davidson 2007).

Table 6 displays the percentages of the correlated multimodal features in the three salient dimensions. Obviously enough, there may exist other types of linguistic-discursive items that include multimodal references in the transcripts. Our concern is only with the significant features drawn from the quantitative analysis because we want to apply the most relevant communicative traits to the learning/pedagogical process.

Table 6. Percentages in the correlation of dimensions with multimodality in the two corpora.

Corpus	**Interpersonal**	**Declarative**	**Markers**
English	10	36	54
Spanish	18	35	47

Most multimodal information (e.g., 54 percent in the English corpus) is correlated with short phrases and gambits that convey the use of markers and meta-discursive items. These gambits include (in English) uptakers like "Ok" and "there", starters like "now" and "then", and appealers such as "isn't it?" or "ok?" (based on a classification by Thomas 1983). A common example is the use of *There* (by both adult and child) to signal transition and progress. In Spanish, the percentage for markers is a bit lower but still the majority, with a similar proportion for declarative statements, but a slightly higher percentage for interpersonal items with multimodal information than in English.

The annotation of the multimodal references is done semi-automatically. The frequency-based features are automatically extracted from the concordance (e.g., all the annotated lines with the interpersonal label, or all the CDL lines from a given age period where more declarative statements are recorded). The key is to observe examples to which the previous quantitative analysis can hint and direct. Sample 1 is an excerpt of an extracted concordance for age 4 in the English corpus according to the condition "declarative" (produced and received by the child), to be later assigned multimodal features.

```
Concordance          Set     Tag    Word    No.                    File
1.    should we use this time ?  *DECL:   3.867
      c:\texts\childr~2\english\blissu~1\4-norj~1.cha
2.    called yolk . den, this be *DECL   1.913
      c:\texts\childr~2\english\blissu~1\4-nort~1.cha
3.    I looked at this and it goes just like *DECL    2.324
      c:\texts\childr~2\english\blissu~1\4-norj~1.cha
4.    I'm gonna take this up like a ball  *DECL  3.327
      c:\texts\childr~2\english\blissu~1\4-nort~1.cha
5.    we'll take this spatula and use it    +CDL   *DECL  2.067
      c:\texts\childr~2\english\blissu~1\4-nort~1.cha
6.    this is the hard part    *DECL  1.727
      c:\texts\childr~2\english\blissu~1\4-norj~1.cha
7.    this one is hard  *DECL 945
      c:\texts\childr~2\english\blissu~1\4-nort~1.cha
```

Sample 1. Excerpt of a concordance to be added multimodal labels.

In Sample 1, multimodality can be annotated with metadiscourse features in some lines (e.g., lines 3, 4 and 5). However, the rest of the lines may be harder to interpret. In such cases, it is useful to go to the transcripts where the amount of dimensions with possible multimodal traits is greater (e.g., the Bliss folder for age 4, according to the file name appearing in Sample 1). This qualitative examination may illustrate and aid the overall analysis.

The following dialogue excerpt (Sample 2) includes a mother and her four-year-old child. The presence of the three linguistic-discursive dimensions described is high. The conversation is part of a collaborative task where short exchanges of information take place in the form of direct questions/answers, commands, markers, and meta-discursive items. Such items have been annotated within brackets, and the presence of multimodality is highlighted.

*MOT:	want to take it apart first ?	[interpersonal question]
*CHI:	**right here** +...	[marker / metadiscourse / production]
*MOT:	how do you get it out ?	[interpersonal question]
*MOT:	how do you get the pieces out ?	[interpersonal question / repetition]
*MOT:	**like this** ? +	[question / metadiscourse / repetition]
*CHI:	yeah .	
*MOT:	ok .	[answer / marker]
*CHI:	are ya gonna talk to it without the puzzles out of it ?	[interpersonal question / production]
*MOT:	yeah .	
*MOT:	<you can just put> [//] why don't you put a piece and then I'll put a piece .	[interpersonal command / question]
*CHI:	ok .	

*MOT:	**this looks like Mickey's head .** +	[declarative / naming]
*MOT:	**is that his head** ? +	[question / repetition]
*CHI:	yep .	
*MOT:	ok .	[answer / marker]
*CHI:	**there .** +	[metadiscourse / production]
*MOT:	now it's your turn .	[interpersonal prompt]
*CHI:	um .	
*MOT:	ok.	[answer / marker]
*CHI:	**there.** +	[metadiscourse / production]
*CHI:	it’s your turn .	[interpersonal prompt / repetition / production]

Sample 2. Conversational excerpt (*MOT—mother— / *CHI—four-year-old child—).

The multimodal elements of communication with the child are visual in Sample 2. Most are connected with the child’s own production of metadiscourse, while both directing and being directed in the conversation. In turn, the items chosen by the adult are declarative, pointing to specific objects and drawings.

In the Spanish corpus, as mentioned, the interpersonal stage is more significant at age 4, while age 5 goes first in the use of markers (see Table 5 above). It would seem then that the young speakers of Spanish tend to move into discursive interactions a bit more slowly (at age 5) than their English counterparts. In Sample 3, this tendency can be observed. The girl is five years and 6 months old, and is able to answer with clear information, establishing a rapport based on discourse identities with the observer, through which the child is already claiming her position in the socio-cultural/ educational scale (cf. Koester 2006: 6).

*OBS:	a ver # me dices como te llamas .	[interpersonal question]
*CRI:	Cristina Perez Perez .	
*OBS:	Cristina Perez Perez ?	[question / repetition]
*OBS:	oye que estabas haciendo ahora en clase ?	[marker / interpersonal question]
*CRI:	estaba escribiendo y pintando .	
*OBS:	y que estabas escribiendo y pintando ?	[interpersonal question / repetition]
*CRI:	escribiendo en el cuaderno azul .	[answer / declarative / production]
*OBS:	si # oye y que es el cuaderno azul ?	[marker / interpersonal question / repetition]
*CRI:	uno que tiene cuadrados rojos y lo voy a terminar .	[answer / declarative / production]
*OBS:	si y que te ha dicho la sor # que lo haces bien ?	[marker / interpersonal question]
*CRI:	si .	
*OBS:	**y tambien pintas en ese ?** +	[marker / metadiscourse / question]
*CRI:	&=afirma .	
*OBS:	y que pintas ?	[marker / interpersonal question]
*CRI:	pin [/] pinto cuadros .	
*OBS:	de muchos colores de que colores .	[answer / question / repetition]
*CRI:	rojo # marron # amarillo # rosa # morado y # y verde .	

*OBS: **hala # +** si te sabes todos los colores . [marker / interpersonal statement]
*CRI: sí, verde de la vaca [answer / declarative / production]
*OBS: ah sí, y y que mas hace la vaca? [marker / interpersonal question]
*CRI: mm . [answer / production]
*OBS: la vaca hace muu sí y que más pintas?+ [interpersonal statement / repetition /
prompt]

Sample 3. Conversational excerpt (*OBS—adult observer— / *CRI—Cristina, five-year-old child—).

Discourse markers are quite common in this case. Their use reproduces an analysed aspect of discourse, the "interpersonal and the textual functions" (Ädel 2006: 17). The observer motivates the child's responses and actions by relying on many discourse markers, and leads her to demonstrate her knowledge. The interaction is also done through direct questioning/answering turns. Sound and visual items are pointed out by the researcher in this case (CDL).

Undoubtedly, together with the age variable, such independent (socio-cultural) variables entail proportional differences in the dimensions described. The corpus-based information may work as positive feedback for children's EFL teaching/learning at early age. The communicative items pinpointed may differ not only depending on the type of topics and collaborative tasks being carried out, but also on whether the children must interact with familiar adults, unfamiliar people, teachers, or other children. In the corpus, the participants exchange information and communicate by activating socio-cultural variables (e.g., what the situation is like, who the other speakers are or what they represent, what they must use the lexical item for, etc). In this way, in social, cultural and educational contexts, communication is at least aided in its processing thanks to much visual-spatial input data favoured (much in CDL form).

V. TEACHING IMPLICATIONS

The most salient verbal and non-verbal information in the corpora serves to lead the selection of linguistic items and the design of audio-visual resources for the AHS (Adaptive Hypermedia System) lessons. The material and the different access channels to knowledge, e.g., verbal, visual, repetitions, gestures and interaction, etc., can be defined and specified for the EFL activities in the hypermedia form, attempting to adapt to the child's learning preferences and demands. Thus, as described below, the AHS course contains audio-visual material that includes colourful characters and units, but also adequate means of access and interaction at the age levels. These devices in the system challenge the learners' communicative competence by leading them through a three-phase approach in the situations: Introduction of topic, Interaction/ Reinforcement, and Evaluation. The verbal skills to be tested include both recognition and production of corpus-based lexical items, whereas the non-verbal skills include their reception and activation of frequent audio-visual elements, taken from the corpora.

In particular, each lesson runs on a specific topic and set of tasks/activities with which children are familiar at that age level. The units contain key forms of exchange and language derived from the analysis of the CHILDES transcripts. For example, the simple and concise sentences with everyday words imitate the generally short and clear functional-pragmatic items examined. The contrasted Spanish and Bilingual material can also give insights of similarities and variation to take into account for the sequencing of the pedagogical content.

For instance, in unit 1, "greetings and introductions" (Table 7), the characters use many declarative statements with first and second person pronouns; this input works as basic reference material at age 3.

Table 7. Linguistic and conceptual units in the AHS lessons.

Concepts	**3**	**4**	**5**	**Linguistic content**	**3**	**4**	**5**
UNIT 1: Greetings and introductions							
Simple descriptions	X	X	X	Personal pronouns/ declarative statements	X	X	X
Greetings/ introductions	X	X	X	Prepositions / interpersonal questions		X	X
UNIT 2: The family							
Simple descriptions of people and objects	X	X	X	Third person pronouns/ possessive pronouns		X	X
Family members	X	X	X	These is/are		X	X
UNIT 3: The house							
Simple descriptions of objects and people	X	X	X	Common and proper nouns / It is …		X	X
Specific Vocabulary; numbers		X	X	To have / To be going to		X	X
UNIT 4: The toys							
Feelings (love, hate …) and likes (I like …)	X	X	X	Direct questions: Are you…? / What is this?		X	X
Colours		X	X	Like/ Dislike	X	X	X
UNIT 5: The food							

Types of food / meals	X	X	X	Wh/ open questions Interrogative pronouns		X	X
Daily routines (wash one's hands, have breakfast…)		X	X	To be / to be going to		X	X
UNIT 6: The school							
Actions (read, jump, run)		X	X	Adjectives Comparative and superlative		X	X
Sizes and shapes / numbers		X	X	Commands (Make… / Don't make…)	X	X	X
UNIT 7: The holidays							
Space /time orientation (up, down, near ...)	X	X	X	Can/Could Would you like …			X
Sensations, states of mind (happy, bored, I am cold…)	X	X	X	Do/does Yes/no questions		X	X

At age 3, written words are kept to a minimum and the focus is placed on the general pictures / characters pointed out, while at later years, more details are shown (see an example in Figure 3). The main verbal difference in this case is the larger number of proper and concrete nouns for years 4 and 5. In the children's interaction with the AHS input, attractive audio-visual and multimedia stimuli must accompany the verbal content. Information technology (IT) suitability for early age education is the result of implementing key aspects for motivation, adaptability, and friendliness.

Figure 3 illustrates how such ideas can guide the design of activities that integrate the computer input/output devices for specific recognition (the captions in Figure 3 are sound files in the AHS). By recognising pictures with sounds, the young learner may communicate with key language in the topic or situation, which demands some specific knowledge. In this case, the nouns are more specific for parts of the face (Unit 5). The content is here made available after the second level (age 4), in agreement with corpus-based information about noun use after that age. Thus, the L2 progress parallels L1 development.

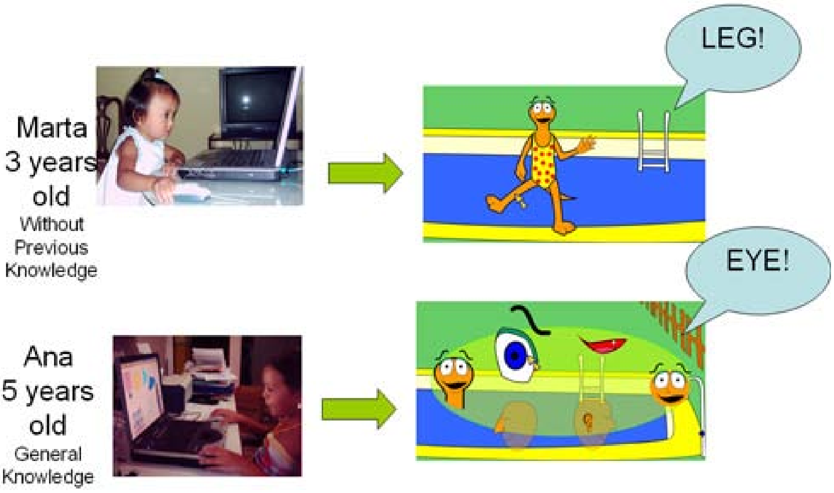

Figure 3. Example of hypermedia-based identification.

Therefore, multimodality varies across the different units and levels. The use of gambits such as “Ok”, and “There” for age 3, or others, like “Great”, and “this is good” at later years, is recurrent to confirm that something has been done right (together with pop-up multimedia effects of flowers and applause, medals, trophies, etc). Other expressions, e.g., “Nope”, “Oops”, and “That’s not it”, underline mistakes, accompanied by pictures of tomatoes, eggs, or raindrops, and disapproval effects like booing, mumbling, etc.

Socio-cultural traits are equally important for the AHS design. These factors correspond to main ideas gathered in surveys and questionnaires (cf. Cumbreño et al. 2006). The characters, for instance, are the result of most children’s preferences; even the choice for colour is based on direct observation of children’s drawings in some schools. The topics (“the family”, “the house”, “food”, etc) are taken from most teachers’ material selections in the teaching curriculum, but they also agree with the type of situations explored in CHILDES (e.g., playing with toys, counting things in the house, naming animals, etc). Figure 4 shows a sequence for a basic oral exchange between some characters, with captions included here but, obviously, not in the lessons. The elephant is chosen as a “less smart” animal for the playful excuse of linguistic repetition and knowledge confirmation.

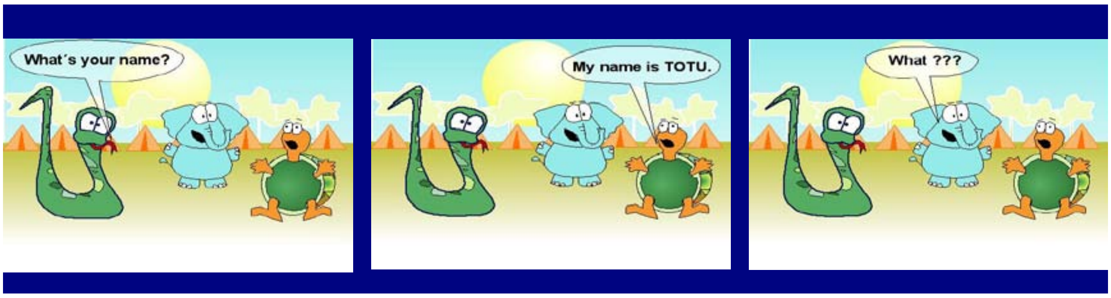

Figure 4. A sequence of basic interaction in the AHS presentation unit.

VI. CONCLUSIONS

The corpus-based analysis has served as an engine for linguistic-discursive content identification. It is found that the young EFL students' learning context can benefit from the examination of linguistic, paralinguistic, and multimodal input in the exchanges. The evaluation phase of the AHS system is currently going on in various schools of Extremadura, and the overall results already point to significant vocabulary gain and phrase production at the basic levels of simple direct questions and answers, personal statements, object identification, and declarative knowledge.

Another significant finding is that the teachers find that the AHS interactive lessons are flexible and useful to adapt to age levels in terms of both verbal (e.g., vocabulary, sentences) and non-verbal (e.g., cursor, mouse buttons) skills. This is a key educational challenge for children's EFL learning via the AHS lessons. The adaptation involves the effective understanding and use of English words and phrases without translation into L1, the use of concise lexical constructions taken from real conversations, and the control and command of multimodality via pictorial and sound media.

It is also concluded that the salient linguistic/paralinguistic traits observed in the corpus have positive effects on the identification of productive content for communication. In the case of children from age 3 to 5, distinguishing age period-based input data is quite relevant to determine key content and preferred ways of interaction (e.g., a focus on everyday words, the use of concise statements, importance of context-based references, familiar socio-cultural aspects, collaborative interaction, and so forth). The hypermedia distribution of the content enables the easy-to-follow process, while the intelligent tutor in the AHS directs the students to the appropriate learning stages and levels.

Notes

[1] Detailed consistency lists (DCL) are the result of combining frequency and range across the corpora. Therefore, the order of the items is listed not only according to their higher frequency but also to their wider distribution over the texts in the given corpus.

[2] It is found in most examples that the bilingual speakers use many words in the sentences, including abstract thinking in their conversations (e.g., telling opinions about topics, people, games, etc); in contrast, the excerpts checked for the other two categories reflect this abstract level less intensively, and probably focus on more everyday references (naming of things, people, animals, etc). This general observation cannot be investigated further at this point, but may be left open for possible contrastive probing.

[3] This classification is based on a keyness-based measurement of the items in relation to other corpora frequency lists (*The British National Corpus* [2001], and the *Spanish Web Corpus* [Sharoff 2006]), each having more than 100 million words.

[4] The only bilingual corpus found contains literary texts and is intended for code-switching study (Callahan 2004). Needless to say, the code-switching phenomenon is beyond the scope of this research.

REFERENCES

Ädel, A. 2006. *Metadiscourse in L1 and L2 English*. Amsterdam/Philadelphia: John Benjamins.

Bachman, L. 2004. *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

Biber, D. 2006. *University Language. A Corpus-based Study of Spoken and Written Registers*. Amsterdam/Philadelphia: John Benjamins.

Biber, D., Conrad, S. and Reppen, R. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bowker, L. and Pearson, J. 2002. *Working with Specialized Language. A Practical Guide to Using Corpora*. London: Routledge.

British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. <http://www.natcorp.ox.ac.uk/>

Brodsky, P, Waterfall, H.R. and Edelman, S. 2007. "Characterizing motherese: On the computational structure of child-directed language". In McNamara, D.S. and J.G. Trafton (Eds.) *Proceedings of the 29th Cognitive Science Society* Conference. Austin, TX: Cognitive Science Society, 833-838. 20 August 2010

<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.5957&rep=rep1&type=pdf>

Brusilovsky, P. 1996. "Methods and techniques of adaptive hypermedia". *User Modeling and User Adapted Interaction*, 6 (2-3), 87-129.

Brusilovsky, P. 2001. "Adaptive hypermedia". *User Modeling and User-Adapted Interaction,* 11 (2-3), 87-110.

Buttery, P and A. Korhonen. 2005. "Large scale analysis of verb subcategorization differences between child directed speech and adult speech". Verb Workshop 2005, *Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*. Saarland University. 20 August 2010 <http://www.cl.cam.ac.uk/~pjb48/ButteryKorhonen.pdf>

Bybee, J. 2006. "From usage to grammar: The mind's response to repetition". *Language*, 82, 711-733.

Bybee, J. 2008. "Usage-based grammar and second language acquisition". In Robinson, P. and N.C. Ellis (Eds.) *Handbook of Cognitive Linguistics and Second Language Acquisition*. London: Routledge, 216-236.

Cadierno, T. 2008. "Learning to talk about motion in a foreign language". In Robinson, P. and N.C. Ellis (Eds.) *Handbook of Cognitive Linguistics and Second Language Acquisition.* London: Routledge, 239-275.

Callahan, L. 2004. *Spanish/English Code-Switching in a Written Corpus*. Amsterdam: John Benjamins.

Carter, R. 2004. *Language and Creativity. The Art of Common Talk*. London: Routledge.

CHILDES. Child Language Data Exchange System. 12 August 2010 <http://childes.psy.cmu.edu/>

Coventry, K.R. and Guijarro-Fuentes, P. 2008. "Spatial language learning and the functional geometric framework". In Robinson, P. and N.C. Ellis (Eds.) *Handbook of Cognitive Linguistics and Second Language Acquisition.* London: Routledge, 114-138.

Cumbreño Espada, A.B., Rico García, M., Curado Fuentes, A. and Domínguez Gómez, E. 2006. "Developing adaptive systems at early stages of children's foreign language development". *ReCALL Journal*, 18 (1), 45–62.

Curriculum de Infantil. 2003. Diario Oficial de Extremadura. 11 April 2006 (pp. 4505-4515) <http://doe.juntaex.es/>

Ellis, R. 2004. *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Flowerdew, L. 2004. "The argument for using English specialized corpora to understand academic and professional language". In Connor, U.and T.A. Upton (Eds.) *Discourse in the Professions. Perspectives from Corpus Linguistics*. Amsterdam: John Benjamins, 11-36.

Fulcher, G. and Davidson, F. 2007. *Language Testing and Assessment*. London: Routledge.

GexCALL group. Extremadura's Group for Computer Assisted Language Learning. 12 September 2010 <http://gexcall.unex.es>

Goldberg, A.E. and Casenhiser, D. 2008. "Construction learning and second language acquisition". In Robinson, P. and N.C. Ellis (Eds.) *Handbook of Cognitive Linguistics and Second Language Acquisition.* London: Routledge, 197-225.

Hudson, R. 2008. "Word grammar, cognitive linguistics, and second language learning and teaching". In Robinson, P. and N.C. Ellis (Eds.) *Handbook of Cognitive Linguistics and Second Language Acquisition.* London: Routledge, 89-113.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Koester, A. 2006. *Investigating Workplace Discourse*. London: Routledge.

Kroll, J.F., Gerfen, C. and Dussias, P.E. 2008. "Laboratory designs and paradigms: Words, sounds, and sentences". In Wei, L. and M.G. Moyer (Eds.) *The Blackwell Guide to Research Methods in Bilingualism and Multilingualism*. Oxford: Blackwell Publishing, 108-131.

Langacker, R. 2000. "A dynamic usage-based model". In Barlow, M. and S. Kemmer (Eds.) *Usage-based models of language*. Stanford: CSLI, 1-63.

Lightbrown, P.M. and Spada, N. 2006. *How Languages are Learned*. Oxford: Oxford University Press.

MacWhinney, B. 2000. *The CHILDES Project. Tools for Analyzing Talk*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

McCarthy, M. 1998. *Spoken Language & Applied Linguistics*. Cambridge: Cambridge University Press.

Nortier, J. 2008. "Types and sources of bilingual data". In Wei, L. and M.G. Moyer (Eds.) *The Blackwell Guide to Research Methods in Bilingualism and Multilingualism*. Oxford: Blackwell Publishing, 35-52.

Robinson, P. and Ellis, N.C. 2008. "Conclusion: Cognitive linguistics, second language acquisition and L2 instruction — Issues for research". In Robinson, P. and N.C. Ellis (Eds.) *Handbook of Cognitive Linguistics and Second Language Acquisition.* London: Routledge, 489-546.

Sharoff, S. 2006. *Spanish Web Corpus.* 12 August 2010 <http://trac.sketchengine.co.uk/wiki/Corpora/SpanishWebCorpus>

Slobin, D.I. 2000. "Verbalized events: A dynamic approach to linguistic relativity and determinism". In Neimeier, S. and R. Dirven (Eds.) *Evidence for Linguistic Relativity*. Amsterdam/Philadelphia: John Benjamins, 107-138.

Thomas, J. 1983. "Cross-cultural pragmatic failure". *Applied Linguistics*, 4, 91-112.

Tokowicz, N. and Warren, T. 2008. "Quantification and statistics". In Wei, L. and M.G. Moyer (Eds.) *The Blackwell Guide to Research Methods in Bilingualism and Multilingualism*. Oxford: Blackwell Publishing, 214-231.

Wintergest A.C., Decapu, A. and Vrena, M.A. 2003. "Conceptualizing learning style modalities for ESL/EFL students". *System*, 31 (1), 85-106.

Received July 2010

Cite this article as:

Curado Fuentes, A. 2010. "An Approach to corpus-based language and multimodal features in communicative exchanges at an early age for adapted hypermedia content design". *Language Value*, 2 (1), 27-50. Jaume I University ePress: Castelló, Spain. http://www.e-revistes.uji.es/languagevalue.

ISSN 1989-7103