CHIEDE

A SPONTANEOUS CHILD LANGUAGE CORPUS OF SPANISH[1]

Marta Garrote Salazar, Antonio Moreno Sandoval
Autonomous University of Madrid

1. CHIEDE

The spontaneous child language corpus, CHIEDE, is made up of around 60.000 words. About a third of the whole corpus is comprised of child language and the rest of adult speech. The main feature of CHIEDE is the spontaneity of interactions: texts are recordings of communicative situations in their natural context. The resource is presented in different formats: an orthographic transcription, an automatic phonological transcription, an XML tagged version and the text-sound alignment. We also provide results obtained through statistical methods, of data from the annotated texts.

CHIEDE fulfills all the requirements of a modern spoken language corpus. It is in an electronic format, allowing the storage and manipulation of data and the interchange with other interested researchers. Its proportioned design and diversity − sex, age and communicative situation variables − guarantees that it is linguistically representative. Its presentation on a web site makes it freely available (http://www.lllf.uam.es/chiede). Finally, its classification structure allows it to be properly utilized.

1.1 Corpus design

To record the corpus CHIEDE, we entered into a collaboration with a Spanish public school, which allowed us to record children from Infant School groups (from 3 to 6 years old). CHIEDE is a transversal corpus, made up of three groups of individuals, divided by ages (from 3 to 4 years old, from 4 to 5 and from 5 to 6).

Our corpus represents two kinds of interactions: *spontaneous collective conversations*, recorded at a daily activity in classroom, and *personal interviews* in which an adult talks to a single child.

Moneglia M., Panunzi A. (eds), *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective*, ISBN 978-88-8453-518-4 (print), 978-88-8453-529-0 (online) © 2010 Firenze University Press.

CHIEDE consists of 58.163 words, in 30 texts, with 7 hours and 53 minutes of recordings and 59 child participants. Each recording is aligned with its corresponding orthographic transcription, including a header with metadata or sociolinguistic and contextual information. Apart from the audio and the text files, two other kinds of files are included: those in which an automatic phonological transcription has been carried out and those where the text appears in XML format with the morphosyntactic annotation. The files are identified with a name where the age of the child participant is specified.

Table 1. CHIEDE measurements

CHIEDE	MINUTES	TURNS	UTTERANCES	WORDS
	473´11´´	10.042	15.444	58.163

A goal of the project was to obtain a balanced corpus between participants and communicative situations. For this purpose, the same number of children of each age group (three, four and five years old) were recorded and placed into three subcorpora; each subgroup contained the same number of boys and girls. The number of recording hours and words are similar for the collective interventions and the individual ones. The total number of participants in the interviews is 24, in three groups of 8 individuals having the same age, further divided by sex. For the collective conversations, the number of participants is not as balanced, as it depended on the number of pupils at classroom. Thus, the three-year-old group was made up of 21 pupils, and the four-year-old group, 21 as well, while the five-years-old had only 17 pupils. The three groups make 59 participants.

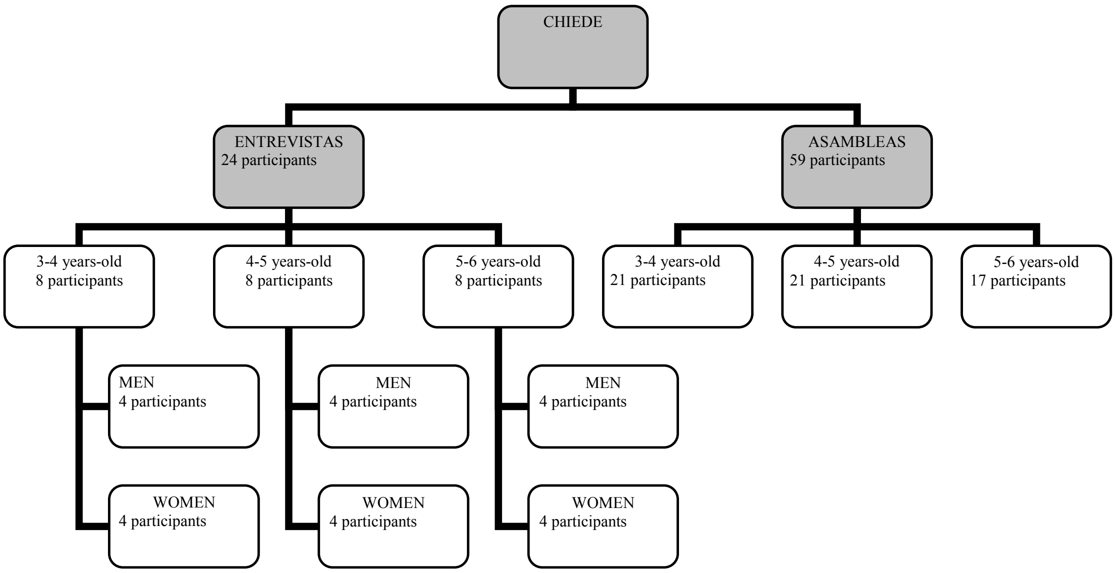

Figure 1. CHIEDE design

1.2 Format

We based our annotation criteria on the annotation system developed for the C-ORAL-ROM project, which is based upon the CHAT format. We also made consistent use of its XML schema, thus guaranteeing textual standardization throughout the corpus.

Each file contains metadata and a transcription. The metadata includes sociolinguistic information. The transcription is orthographic and includes tags that mark disfluencies, noises, overlaps and prosodic units; it is structured by turns introduced by a three-letter code which identifies the participant. Morphosyntactic tagging is provided at a separate level. Figure 2 shows a text fragment.

```
@Title: Jorge y Marta
@File: JOR4
@Participants: JOR, child, (man, 5:0, 1, Ciudad Real)
TEA, adult, (woman, B, 3, Madrid)
@Birth_Date: 20/02/2001
@Date: 22/02/2006
@Place: Ciudad Real
@Situation: conversation in an empty classroom at school.
@Topic: daily matters
@Source: CHIEDE
@Class: informal, family/private, dialogue (child not-known adult)
@Length: 13´39"
@Words: 2097
@Acoustic_quality: A
@Transcriber: Marta
@Revisor: Ana and Marta
@Comments: JOR (middle class; birth order: 1st)

*JOR: aquí ///
*TEA: a ver si puedes /// ¿ cuántos años tienes Jorge ?
*JOR: &eh tengo -> / cuatro ///
*TEA: cuatro /// que fue tu cumple el otro día /// ¿ a que sí ?
*JOR: cinco sí ///
*TEA: ¡ah! ¿ cinco ? ¿ o cuatro ?
*JOR: bueno / hoy &cum [/] mañana cumplí cinco // pero ahora / tengo cuatro ///
```

Figure 2. CHIEDE fragment

2. The computational tool

The orthographic transcription is not sufficient in itself. At LLI-UAM, various computational tools have been developed to transform the plain format of a transcription text into a proper tagged format to match the metadata with the lexical elements and calculate the corresponding statistics. In this section we will explain this process, and later we will see how the data extraction process has been carried out.

2.1 Results of CHIEDE annotation

Once sampling, transcription, and revision have been completed, the corpus is annotated, including phonological, morphological and part of speech tags. The following statistics are gathered:

- Mean Length of Utterance (MLU): in syllables and phonemes
- Frequency of use of lemmas and word forms
- Type/token ratio: lexical diversity
- Most frequent words
- Most frequent categories

All these data allow us to describe language use and establish linguistic behavior patterns.

2.1.1 Data extracted from the morphosyntactic tagger

The morphosyntactic tagger deals with three linguistic levels: the morphological, the syntactic and the lexical ones. We used the morphological information included in the lexicon entries to obtain the different lemmas that appear in the corpus, plus the part of speech information added to each word or multiword.

Table 2 presents data related to the whole corpus, including adult language, sorted by age group:

Table 2. Word forms and Lemmas by ages

	Total words	Different word forms	Different lemmas	Word form/lemma ratio
Adults	36.905	2.910	1.804	1,61
Group 3	5.713	985	718	1,37
Group 4	6.374	1.155	839	1,38
Group 5	8.993	1.450	1.056	1,37

In this table, the first column includes the total number of words for each group, the second column, the number of different words that appear and, the third one, the number of different lemmas. The total number of words for the child subcorpus is 21,080. The last column shows the lexical diversity, that is, the ratio of word forms for each lemma. This ratio scarcely changes for the three child groups, while it does change for the adult subcorpus, as word inflection increases in adult language.

Differences can be seen between the three child groups regarding the increase of word forms and lemmas. For the three-to-four period, there is an increase of 170 word forms and 121 lemmas; for the four-to-five one, this number goes up to 295 word forms and 217 lemmas. This shows that for the first of these periods − from three to four years old − learning is slower than for the second one.

We also find that, although the number of recording hours is similar for the three groups, there is a difference of 3,219 words between the three-year-old group and the five-year-old one. Five-year-old children are already able to have a pseudo-adult conversation. This can be more clearly appreciated by inspecting the MLU in syllables and phonemes. As discussed in the next section, the MLU in phonemes is 10.29 for the three-years-old group, while for the five-years-old one it increases 3.82 points.

Table 3. Word frequency by age groups

Word frequency (10 first words)					
Age group: 3		Age group: 4		Age group: 5	
Word	Frequency	Word	Frequency	Word	Frequency
y	322	y	341	y	502
no	271	que	213	a	289
sí	204	el	206	no	274
el	162	sí	195	que	259
yo	157	no	179	el	251
un	148	a	165	sí	248
la	141	la	153	la	217
a	128	de	148	me	189
me	119	en	136	de	172
se	106	mi	131	se	171

Regarding word frequency (Table 3), it is worth pointing out the abundant use of the copulative conjunction “y”. This is due to two facts: firstly, this is the first coordination strategy that children learn; secondly, “y” have a different use apart from being a copula, and it is usually used as a discourse marker, not only relational, but also a subjectivity one: it is used by the speaker as a turn-taking strategy. The following example clearly shows this last function of “y”:

*DAI: y [/] **y** yo sé +

*TEA: ¡ qué bonito ! y a ver María ///

*DAI: **y** mi &pa +

*TEA: espera / espera a María /// a ver María ///

*MRI: mi papá no le regala / nada a mi mamá ///

%alt: (6) na

*TEA: pero bueno +
*MRI: ¬ **y** mi mamá sí ///

Apart from that, it is also worth noting the high frequency of negative and affirmative adverbs ("sí" and "no") and the use of pronouns and determiners. In short, in any spoken language corpus the most frequent elements are grammatical categories; the lexical ones appear after the first twenty positions in CHIEDE. Possibly most striking is the lack of discourse markers, as "bueno", "¿sabes?" or "¿no?", so frequent in adult spoken language.

If we do not take into account the grammatical categories, the acquisition of lexical categories as verbs, nouns or adjectives is quicker from four to five years old than from three to four (Table 4). In the second period (four to five years old), the number of new lemmas more than doubles. Thus we see that there is a great increase in lexical category acquisition during the fourth and fifth-year period.

Table 4. Lexical categories by age

		Verbs	Nouns	Adjectives
Age group: 3	Word forms	906	836	157
	Lemmas	138	313	50
Age group: 4	Word forms	937	998	141
	Lemmas	143	357	57
Age group: 5	Word forms	1324	1340	161
	Lemmas	190	439	72

The two following graphs illustrate the growth of word forms (Figure 3) and lemmas (Figure 4) for the three age groups.

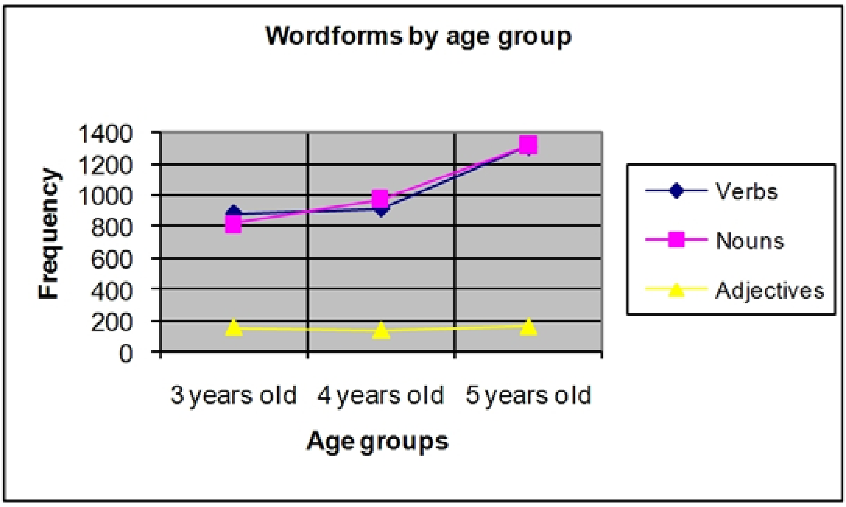

Figure 3. Word forms by age group

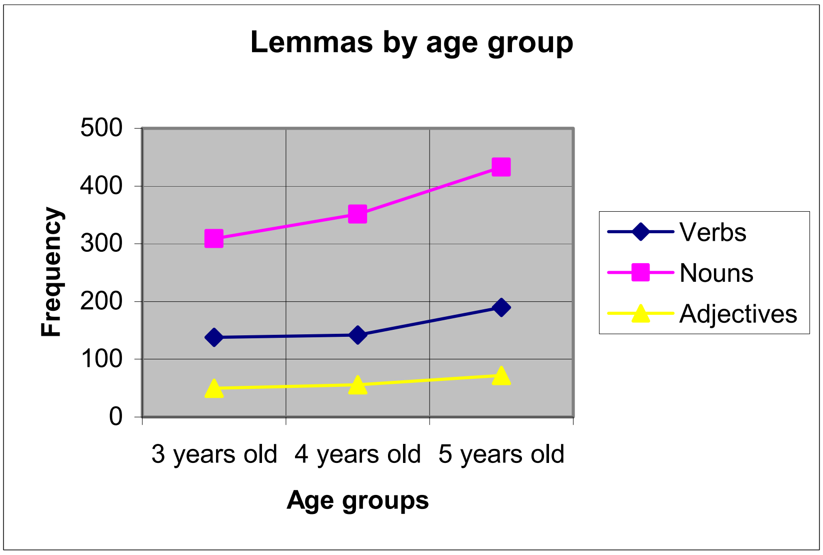

Figure 4. Lemmas by age group

The last thing we are going to deal with in this section is the analogy errors typical of child language. This phenomenon takes place during the first years of acquisition, when the child has developed adequate skills in morphology and inflection. Once the general morphological rules are learnt, the child tends to apply them without discriminating, for instance, between regular and irregular forms. These analogy errors have been always one of the main arguments of those who reject behaviorist theories of learning. They argue that if the child simply reproduces what he listens, this kind of errors would not be possible; on the contrary, it is easier to find an explanation for the analogy errors if we consider that the child is able to learn the morphological rules of his/her language and apply them in a creative way, innovating and not imitating. In fact, it is interesting how, after the adult correction, the child generally does not rectify and goes on keeping his proposal.

In CHIEDE, which totals 21.080 words, there are 31 analogy errors, that is, 0.15% of the whole[2]. If we reduce the list of word forms to lemmas, we can see that there is a total of 17 specially problematic: “acordar”, “cerrar”, “conducir”, “decir”, “escribir”, “hacer”, “ir”, “leer”, “mentir”, “morder”, “poder”, “poner”, “querer”, “ser”, “soltar”, “tener”, “traer”. According to our data, this phenomenon is more frequent in four-year-old children, being reduced in five-year-old ones. So it seems to be that it is at the age of five when children start to understand the irregular inflection of Spanish verbs. In our corpus, the most persistent irregular verb form is “hació/hizo”.

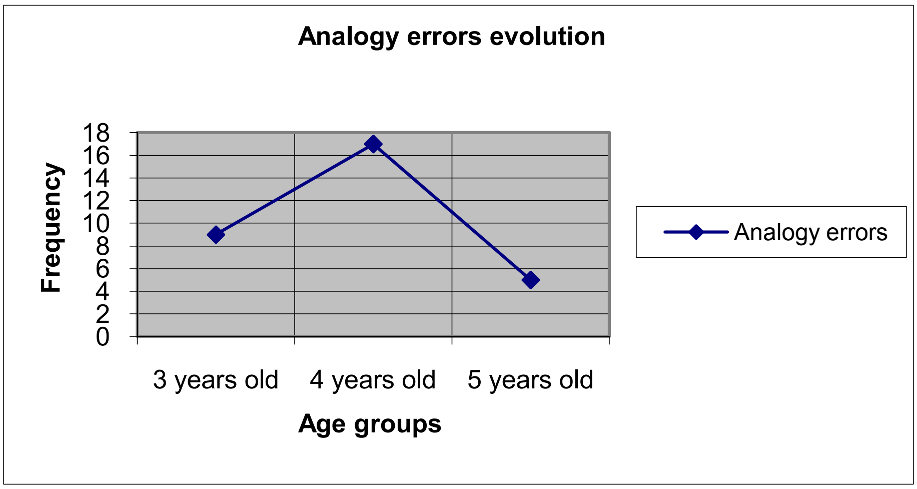

Figure 5. Analogy errors evolution

2.1.2 Data extracted from the phonological transcription

Psycholinguistics, in the attempt to explain the acquisition process of a first language, usually resorts to different measurement patterns. Childhood development stages are calculated, measuring the number of words that comprise the lexicon of a particular age, the number of phonemes they handle, or the most frequent syllables.

Most experts place the phonological system acquisition period from nine months to four years old. However, most research on child language does not consider individuals older than 36 months, and the child language description stops when the child reaches the age of three years old, though it is held that the acquisition process continues until the beginning of puberty. That is the reason why we consider our corpus of interest for the scientific community, as we provide data belonging to children from three to six years old.

In Table 5, we present the phonological data of the three age groups of our corpus. The total number of phonemes is 75,535, and the MLU in phonemes is 12.72 per turn. It is striking in Table 5 the fact that at the age of three, the child has already acquired the complete Spanish phonological system.

The table we present does not show the phonemes' order of appearance, since they have been already acquired by the children that participate in the corpus, but rather their frequency of use. It is known that children tend to use the phonemes they better know, while they avoid those which are harder. In the frequency table, we can see how the phonemes in the latest rows are the least frequent in Spanish, and therefore it is normal that their frequency of use is lower than that of the most usual phonemes.

However, the numbers increase as children ages do. This argues that the linguistic acquisition process is still active from three to five years old, and that research on first language acquisition must not stop at the age of 36 months.

Apart from the phonological data extracted from CHIEDE, we have also added a fourth column that includes the same information from C-ORAL-ROM (Moreno et al. 2006). Although the data are similar for both child and adult language, we can see, especially in the first positions of the table, a higher similarity between the five-year-old group and the adult one than between the last one and the three and four-year-old group.

To appreciate more clearly the child phonological system – at least the one presented by the individuals that participated in our corpus – we present five independent graphs. Table 5, reported in Appendix to this text, includes the vocals frequency and their evolution from one group to another.

The following graphs show the consonants frequency according to their manner of articulation: plosive, nasal, liquids and fricatives (and the affricate tʃ).

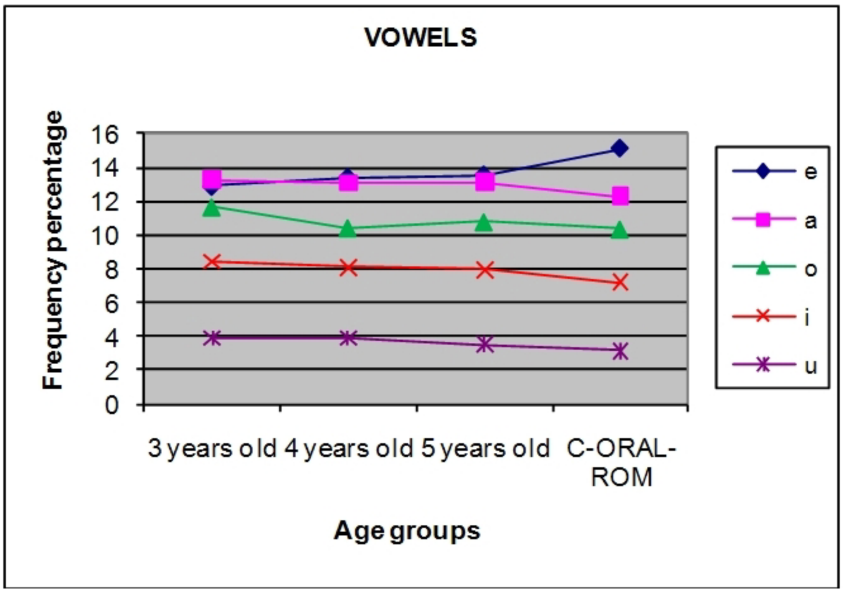

Figure 6. Vocals by age group

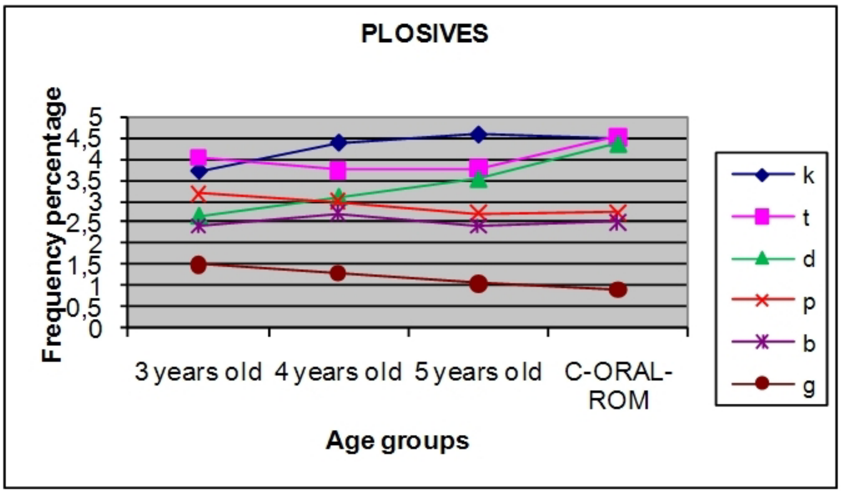

Figure 7. Plosives by age group

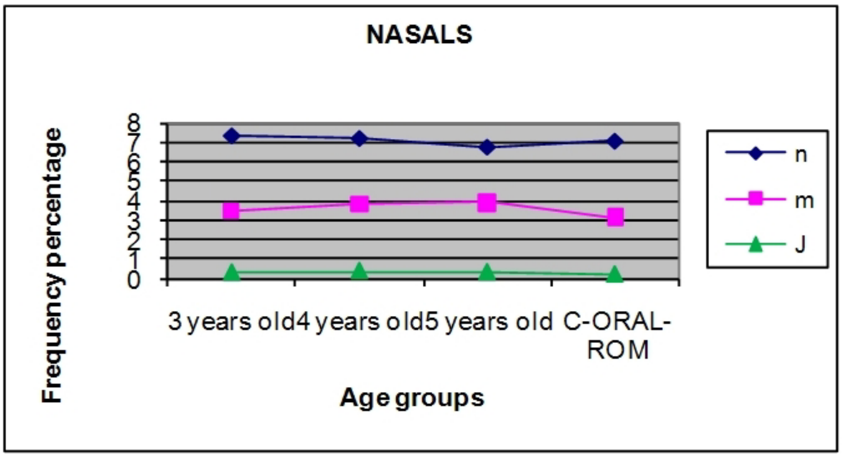

Figure 8. Nasals by age group

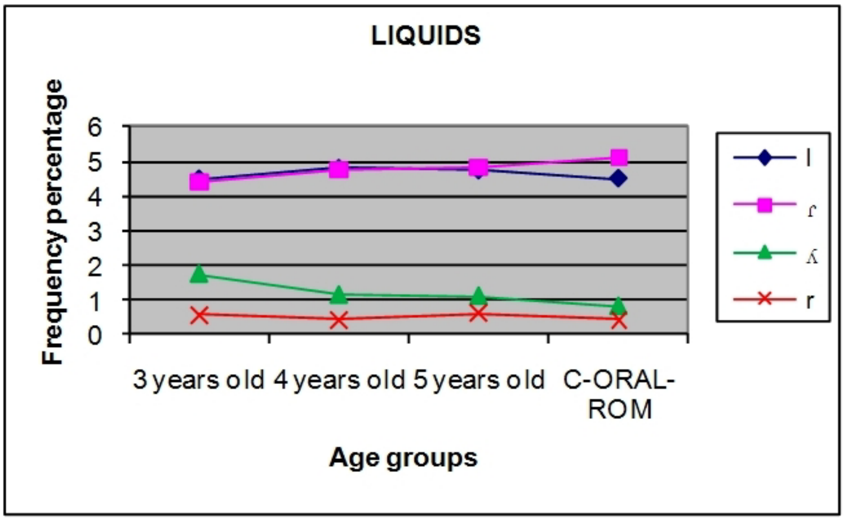

Figure 9. Liquids by age group

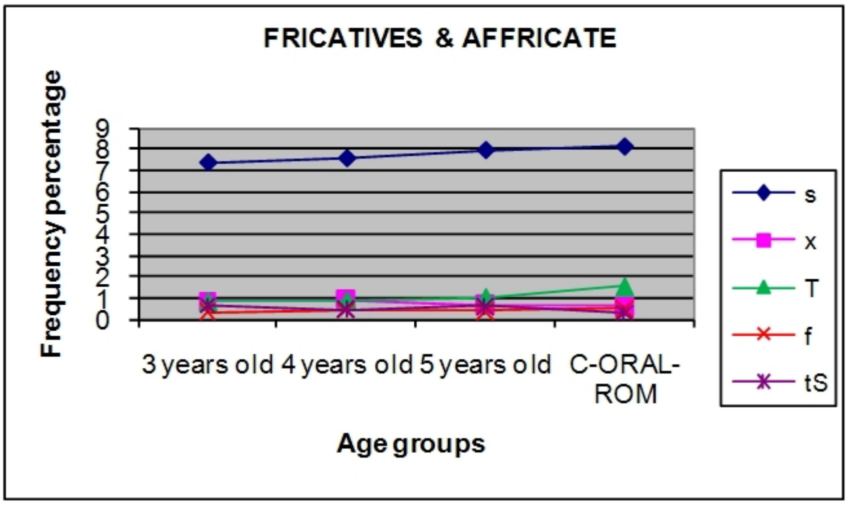

Figure 10. Fricatives and the affricate tʃ by age group

Another possibility that offers the automatic phonological transcriber is the segmentation of words into syllables. In this way, we can quickly and reliably know the total number of syllables that comprise our corpus, which ones are those syllables and which is their frequency of use. The total number of syllables is 35,086 and the MLU, 5.91. With these data, we can easily calculate the Mean Length of Utterance in phonemes and syllables for each age group. In the following table, we present the exact figures and the increase percentage from three to five years old. Figure 11 shows this MLU increase.

Table 6. Mean Length of Utterance

	Mean Length of Utterance	
	Phonemes	Syllables
Age group: 3	10.29	4.88
Age group: 4	13.57	6.26
Age group: 5	14.11	6.49

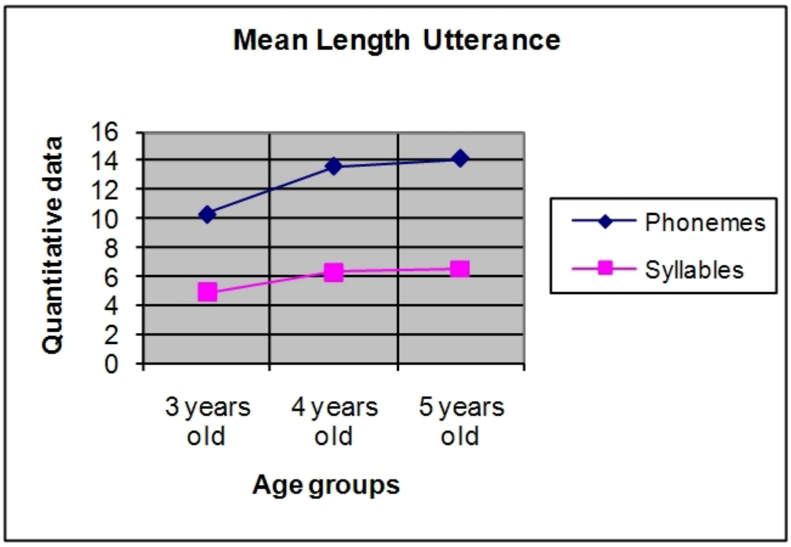

Figure 11. MLU increase

3. Using xml-tagged corpus for relating meta-data linguistic features

The original annotation of CHIEDE has been designed to take into account a wide range of phenomena, including the acoustic ones (prosodic marks, noises, etc.) that can be used by the speech technology community.

Our aim in this experiment was searching for relevant lexical units in two subcorpora: one of adult language and another of child language. The first step consists on the segmentation of each speech turn in utterances to prevent wrong-formed word groups. This task is similar to tokenization in written language corpora.

The utterance segmentation is also necessary to delimit the context which the morphosyntactic tagger uses to disambiguate.

A computational program generates a new tagged corpus with a single tag: UNIT (utterance), with attributes for *speaker*, *startTime* and *endTime*. In Figure 12, we can see the result of this process of XML conversion. The numbers stand for the sound alignment times, expressed in milliseconds. In this way, each utterance is limited, identifying its corresponding speaker.

The next step is the morphosyntactic annotation from the XML file. The morphosyntactic tagger procedure is the following (Guirao et al. 2006):

- Unknown word detection.
- Lexical processing: the program splits the fused words (amalgams and verbs with clitics).
- Multi-word recognition: through a lexicon.
- Single word recognition.
- Unknown word recognition.
- Disambiguation phase 1: a feature-based Constraint Grammar resolves some of the ambiguities.
- Disambiguation phase 2: a statistical tagger (TnT tagger, Brants 2000) resolves the remaining ambiguous and unknown words.

```
<UNIT speaker="JOR" startTime="0" endTime="4.482">aquí </UNIT>
<UNIT speaker="TEA" startTime="4.482" endTime="7.655">a ver si puedes </UNIT>
<UNIT speaker="TEA" startTime="7.655" endTime="9.246"> ¿ cuántos años tienes Jorge ?
</UNIT>
<UNIT speaker="JOR" startTime="9.246" endTime="12.459">&eh tengo -> / cuatro </UNIT>
<UNIT speaker="TEA" startTime="12.459" endTime="13.131">cuatro </UNIT>
<UNIT speaker="TEA" startTime="13.131" endTime="14.267"> que fue tu cumple el otro día
</UNIT>
<UNIT speaker="TEA" startTime="14.267" endTime="14.817"> ¿ a que sí ? </UNIT>
<UNIT speaker="JOR" startTime="14.817" endTime="15.667">cinco sí </UNIT>
<UNIT speaker="TEA" startTime="15.667" endTime="16.411">¡ah! </UNIT>
<UNIT speaker="TEA" startTime="16.411" endTime="17.09"> ¿ cinco ? </UNIT>
<UNIT speaker="TEA" startTime="17.09" endTime="17.755"> ¿ o cuatro ? </UNIT>
<UNIT speaker="JOR" startTime="17.755" endTime="23.601">bueno / hoy &cum [/] mañana
cumplí cinco // pero ahora / tengo cuatro </UNIT>
```

Figure 12. XML Conversion

The final result after the tagger revision is a XML file where the text is morphosyntactically analyzed:

```
<Text>
<p>
<f h="JOR" st="0.0" et="4.482" id="1">
```

```
<sf t="enu" id="1-1">
<w cat="P" lem="aquí" id="1-1-1"> aquí </w>
</p>
<p>
<f h="TEA" st="4.482" et="7.655" id="2">
<sf t="enu" id="2-1">
<w cat="MD" lem="a ver" id="2-1-1"> a ver </w>
<w cat="C" lem="si" id="2-1-2"> si </w>
<w cat="V" lem="poder" tie="pres_ind" num="sing" per="2" id="2-1-3"> puedes </w>
</p>
<p>
<f h="TEA" st="7.655" et="9.246" id="3">
<sf t="int" id="3-1">
<w cat="PUNCT" lem="¿" id="3-1-1"> ¿ </w>
<w cat="P" lem="cuántos" gen="masc" id="3-1-2"> cuántos </w>
<w cat="N" lem="año" gen="masc" num="plu" id="3-1-3"> años </w>
<w cat="V" lem="tener" tie="pres_ind" num="sing" per="2" id="3-1-4"> tienes </w>
<w cat="NPR" lem="Jorge" id="3-1-5"> Jorge </w>
<w cat="PUNCT" lem="?" id="3-1-6"> ? </w>
```

Thus, each word in the corpus can be related to the speaker. The file keeps in the header all the socio-contextual information, being possible to create as many subcorpora as different features appear in the header – an adult language subcorpus, a child language subcorpus, etc. After the division into subcorpora, it is possible to calculate the occurrences (tokens) for each lexical unit (types). The procedure can be applied to any type of linguistic information that had been annotated in the corpus.

3.1 Extracting word clusters

If we calculate the statistics on each unit directly, the result would not be correct, as the pluri-verbal lexical elements (that is, idioms) would not be included in the count. Frequent discourse markers like “por ejemplo”, “o sea” or “es decir” would not appear if we work on lexical units made up of a single word. To solve this problem, it has been created an idioms list by categories, including nominal compounds (“fin de semana”). Each idiom is considered a lexical unit, equivalent to a single word.

3.2 Applying the statistics of surprise

To identify distinctive words, lemmas, or categories of a given subcorpus we have used the log-likelihood ratio test proposed by Dunning (1993). This method does not assume normal statistical distributions of units in a corpus. Instead, the log-likelihood ratio λ assumes a binomial distribution more appropriate for rare but distinctive words. In addition, this test does not need balanced subcorpora for comparison.

This method has been successfully applied for finding collocations (Dunning 1993) and terms (Daille 1994). In order to test the method to find distinctive units in specified domains, we can work on two hypotheses:

Two registers (or subcorpora) show no difference in distinctive units (*Null hypothesis*).

i. For a given subcorpus, we can find out distinctive units (*Alternative hypothesis*).
ii. We applied the test to two well-defined subcorpora: adult and child language. Results are shown in Tables 7 and 8.

Table 7. Distinctive word forms in adult language

WORDS	ADULTS (36.905)	CHILDREN (21.080)	DUNNING
qué	1.123	108	510.29
te	743	59	373.43
a ver	371	23	207.58
bien	304	14	189.00
ah	270	18	146.32
claro	231	15	126.53
tú	264	27	113.88
has	184	9	112.02
tu	197	14	103.64
cómo	249	27	103.26

Table 8. Distinctive word forms in child language

WORDS	CHILDREN (21.080)	ADULTS (36.905)	DUNNING
mi	334	24	524.66
yo	417	166	300.54
sí	647	428	255.77
me	423	248	198.53
tengo	130	28	141.16
Candi	67	9	88.55
porque	150	86	71.99
un	431	424	71.25
padre	60	13	64.86
he	79	27	64.09

Results confirm the alternative hypothesis and the suitability of the Dunning test for the task. Most of the top 10 lemmas in both domains have a low occurrence, but all are typical terms in their domain.

3.3 Preliminary results

Our aim was to show a range of possibilities for applying this method to information extraction from a corpus. By the moment, we present incomplete data; currently, there exists a disproportion of social and register features regarding the linguistic ones. Our intention is to enlarge the corpus later.

In this paper, the linguistic phenomena taken into account are words and idioms, phonemes and categories.

Below, we present the Dunning test results for the two subcorpora: adult and child language. The first one is made up of 36,905 words and the second, 21,080. Tables 9 and 10 show the distinctive categories in each subcorpus.

Table 9. Distinctive categories in adult language

CATEGORIES	ADULTS (36.905)	CHILDREN (21.080)	DUNNING
MD	1.731	449	264.71
P	6.564	2.739	234.8
INTJ	524	81	162.52
V	6.450	3.167	59.07
AUX	1.278	522	44.88

Table 10. Distinctive categories in child language

CATEGORIES	CHILDREN (21.080)	ADULTS (36.905)	DUNNING
POSS	453	360	127.55
N	3.174	4.419	110.27
ADV	1.861	2.428	96.97
Q	1.739	2.497	42.94
NPR	910	1.242	33.33
C	2.184	3.403	19.83
PREP	1.773	2.786	13.63
ART	1.338	2.068	13.29

These results make us interpret the following:

- In adult language, contrary to what happens in child language, elements like discourse markers (DM) or interjections (INTJ) are highly frequent. Both elements belong to the pragmatic level and require higher linguistic skills.
- While in adult language verbs (V) are the element that guides the speech, children from three to six years old base their speech on nouns (N).

- Possessive pronouns (POSS) are the most distinctive element in child language. According to J. Piaget (1965), until seven years old, child language is characterized by being egocentric, that is, it is a simple accompaniment of the action and the child does not have any other perspective than his/hers. If we have a look back at Table 8, we can see that some of the most common words in child language are "mi", "yo" or "me".
- In child language, categories like conjunction (C), preposition (P) and article (ART) are distinctive. In particular, the high occurrence of conjunctions in this subcorpus is caused by the frequent use by children of the copulative conjunction "y", explained in section 2.1.1.
- Finally, in Table 10 the category proper noun (NPR) appears as distinctive of child language. This is a consequence of the context of situation, that is, a school context where children constantly demand the teacher's attention, calling his/her name.

Apart from categories, the Dunning test has been also applied to the phonological level. The orthographic transliterations obtained from the audio recordings are automatically transcribed in IPA. In this way, the texts can undergo the same process explained in section 2.1 to extract the phonological information.

Table 11. Distinctive phonemes in adult language

PHONEMES	ADULTS (136.721)	CHILDREN (77.240)	DUNNING
t	6.408	2.949	90.86
b	4.197	1.914	63.6
e	19.938	10.342	58.27
k	6.851	3.352	49.62
s	11.382	5.924	28.72

Table 12. Distinctive phonemes in child language

PHONEMES	CHILDREN (77.240)	ADULTS (136.721)	DUNNING
ʎ	1.024	1.108	128.15
i	6.277	9.635	82.59
p	2.265	3.176	72.57
m	2.928	4.348	55.2
o	8.490	14.247	16.88
u	2.872	4.667	13.39
r	407	571	12.71
g	957	1.461	12.66
x	618	940	8.54

The most interesting thing about the results is the distinctive character of phonemes ʎ and r in child language. The first one (ʎ) can be the result of the high frequency of the personal pronoun "yo" in the egocentric language of children. Both phonemes are liquid, curiously the last phonemes that children acquire in the learning process —together with fricatives— due to the difficulty that involves their place of articulation (Anula 1998).

4. Conclusions and future work

In this paper we have presented a spontaneous child language corpus, CHIEDE, made up of 60,000 words, of which a third part correspond to child language. The main contribution of this work is the creation of a linguistic resource that is still in short supply. The research on language acquisition must be based upon the direct observation of reality. With CHIEDE, we provide a wide sample of child language in both, audio and text format. Moreover, texts are enriched by phonological and morphosyntactic annotation, from which information relating to these linguistic levels can be automatically extracted.

Future work will address the following issues:

- Increase in the size of the corpus, not only in number of words, but also in the number of participants and communicative situations.
- Carrying out qualitative researches from the quantitative data. The different phonological and morphosyntactic phenomena can be an object of study for future researches.

We have also proved the significance of the Dunning test as a method for the validation of psycholinguistic hypothesis in spoken language, as well as for determining a register typology. This test correlates linguistic to socio-contextual data applying the Statistics of Surprise. For this task, it is necessary to have an annotated corpus and the use of XML. The preliminary results are promising and had not been shown for Spanish before. However, it is rather premature to extract conclusions and interpretations for these data, as the corpus size is clearly insufficient.

Notes

[1] This research has been supported by the Madrid Regional Government under the contract MAVIR (S-0505/TIC/0267) and the Spanish Government under the project BRAVO-RL (TIN2007-67407-C03-02).

[2] We have taken into account only verb forms, as, because of their high irregularity, they are one of the most problematic issues in learning Spanish.

References

Anula, A. 1998. *El abecé de la psicolingüística*. Madrid: Arco-Libros, D.L.

Biber, D. 1988. *Variation across speech and writing*. Cambridge: CUP.

Biber, D. 1995. *Dimensions of register variation*. Cambridge: CUP.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finnegan. 1999. *Longman grammar of spoken and written English*. London: Longman.

Cresti, E., F. Bacelar do Nascimento, A. Moreno, J. Veronis, Ph. Martin, K. Choukri 2002. The C-ORAL-ROM project. New methods for spoken language archives in a multilingual romance corpus. In M. Gonzàlez Rodriguez and C. Paz Suàrez Araujo (eds), *Proceedings of LREC 2002*. Paris: ELRA, 2-9.

Daille, B. 1994. Combined approach for terminology extraction: lexical statistics and linguistic filtering. PhD diss., Paris 7.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19, 1: 61-74.

Garrote, M. 2008. *CHIEDE. Corpus de habla infantil espontánea del Español*. PhD diss., Universidad Autónoma de Madrid.

Guirao, J.M., A. Moreno Sandoval, A. González Ledesma, G. De La Madrid and M. Alcántara. 2006. Relating linguistics units to socio-contextual information in a spontaneous speech corpus of Spanish. In A. Wilson, D. Archer and P. Rayson (eds), *Corpus linguistics around the world*. Amsterdam: Rodopi, 101-113.

Labov, W. 1966. *The social stratification of English in New York City*. Washington: Center for Applied Linguistics.

Miller, J. and R. Weinert 1999. *Spontaneous spoken language*. Oxford: Clarendon.

Moreno, A., D. T. Toledano, N. Curto and R. de la Torre. 2006. Inventario de frecuencias fonémicas y silábicas del castellano espontáneo y escrito. In L. Buera, E. Lleida, A. Miguel and A. Ortega (eds), *Actas de las IV Jornandas de Tecnologías del Habla*. Zaragoza: Universidad de Zaragoza, 77-80.

Moreno, A. 2002. La evolución de los corpus de habla espontánea: la experiencia del LLI-UAM. In A. Rubio Ayuso (ed.), *Actas de las II Jornadas en Tecnologías del Habla*. Granada: Universidad de Granada.

Moreno, A. and J.M. Guirao 2003. Tagging a spontaneous speech corpus of Spanish. In N. Nicolov, R. Mitkov, G. Angelova and K. Boncheva (eds), *Proceedings of Recent Advances in NLP (RANLP-2003)*. Amsterdam: John Benjamins, 292-296.

Piaget, J. 1965. *El lenguaje y el pensamiento en el niño*. Buenos Aires: Paidós.

Uchimoto, K. 2002. Morphological analysis of the spontaneous speech corpus. In Shu-Chuan Tseng (ed.), *Proceedings of Conference of Computational Linguistics (COLING 2002)*. Taipei, Taiwan, 1298-1302.

Appendix

Table 5. Phonemes frequency by age groups.

Age group: 3			Age group: 4			Age group: 5			C-ORAL-ROM (Adults)		
Phoneme	Frequency		Phoneme	Frequency		Phoneme	Frequency		Phoneme	Frequency	
	Absolute	Relative		Absolute	Relative		Absolute	Relative		Absolute	Relative
a	2616	13,29	e	3190	13,44	e	4360	13,58	e	188196	15,12
e	2539	12,90	a	3108	13,09	a	4215	13,13	a	152664	12,27
o	2300	11,68	o	2484	10,46	o	3472	10,81	o	129208	10,38
i	1656	8,41	i	1922	8,10	s	2551	7,94	s	100881	8,11
n	1447	7,35	s	1792	7,55	i	2550	7,94	i	89799	7,22
s	1444	7,34	n	1709	7,20	n	2159	6,72	n	87775	7,05
l	884	4,49	l	1139	4,80	ɾ	1553	4,84	ɾ	63702	5,12
ɾ	869	4,41	ɾ	1130	4,76	l	1523	4,74	t	56287	4,52
t	793	4,03	k	1041	4,39	k	1472	4,58	l	56107	4,51
u	765	3,89	u	927	3,91	m	1258	3,92	k	55863	4,49
k	732	3,72	m	914	3,85	t	1214	3,78	d	54284	4,36
m	681	3,46	t	885	3,73	d	1134	3,53	m	39278	3,15
p	625	3,18	d	739	3,11	u	1134	3,53	u	39146	3,14
d	524	2,66	p	709	2,99	p	871	2,71	p	34135	2,74
b	477	2,42	b	644	2,71	b	777	2,42	b	31126	2,50
ʎ	344	1,75	g	308	1,30	ʎ	361	1,12	θ	18940	1,52
g	296	1,50	ʎ	274	1,15	g	342	1,06	g	11359	0,91
x	166	0,84	x	221	0,93	θ	330	1,03	ʎ	10356	0,83
θ	165	0,84	θ	210	0,88	x	214	0,67	x	7681	0,62
tʃ	125	0,64	tʃ	106	0,45	tʃ	200	0,62	f	6217	0,50
r	111	0,56	r	101	0,43	r	199	0,62	r	5236	0,42
f	68	0,35	f	95	0,40	f	126	0,39	tʃ	3744	0,30
ɲ	58	0,29	ɲ	89	0,37	ɲ	98	0,31	ɲ	2427	0,19
Total	19685	100,00		23737	100,00		32113	100,00		1244411	100