Integrated Tools for Exploitation of a Spontaneous Speech Corpus of Spanish

Antonio MORENO-SANDOVAL

Department of Linguistics
Autonomous University of Madrid

José M. GUIRAO

Department of Software Engineering
University of Granada

Correspondence address:

Prof. Antonio MORENO-SANDOVAL
Dept. de Lingüística
Universidad Autónoma de Madrid
28049 Cantoblanco, Madrid, SPAIN

1. Introduction

This paper presents a query system to a corpus of spontaneous speech, concretely the Spanish C-ORAL-ROM corpus (Cresti and Moneglia eds. 2005; Moreno *et al*. 2005). The corpus consists of 181 transcripted sessions, in different registers and communicative situations. With over 42 hours of recorded data and almost 500 speakers, the corpus has 312,000 tokens (words) of 21,000 different types (lemmas).

The system will be accessed through a web page, although it will eventually be also available as an independent application. The tools consist of three main components:

1. A *metadata query system*, that retrieves texts with features specified in the search. Those features include sociolinguistic information about the speakers, as well as contextual and thematic information about the recording.

2. A *concordancer of text and sound*: the system looks for words or multi-words expressions in all the texts and retrieves every “utterance” where the searched string appears along with the original sound fragment (in mp3). This way the user can hear the original source, not only its transcription.

3. A *morphological processor of Spanish*, based on a broad-coverage lexicon, which provides all the possible analyses for a given wordform. It can be used as a Part-of-Speech tagger for sentences in Spanish, providing the surface syntactic analysis for the sequence.

The potential uses of this tool enhance the possibilities of the original John Benjamins version published in DVD format. In particular, some examples of its application to teaching/learning Spanish as a second language, as well as to describing properties of spoken Spanish will be given.

The first section of the paper will be devoted to a short introduction to the corpus. Then, in three separate sections, the integrated tools developed by the Laboratorio de Lingüística Informática at UAM (hence, LLI-UAM) will presented. In the last section, two particular applications of these resources will be discussed.

2. The C-ORAL-ROM Corpus

C-ORAL-ROM[1] is a multimedia corpus of spontaneous speech for the main four Romance languages, French, Italian, Portuguese and Spanish (Cresti *et al.* 2002; Cresti and Moneglia 2005). The project has been funded by the European Union under the Fifth Framework Programme (IST-2000-26228) from 2001 to 2004. The consortium consisted of nine partners, including the University of Florence (co-ordinator), University of Provence Aix-Marseille 1, University of Lisboa, and Autonomous University of Madrid[2].

Each recorded session in the corpus is provided in three different versions: in orthographical transcription; with morpho-syntactic annotation; and in a multimedia version with text and sound alignment. Every text has a metadata specification with a rich information about the context (place and situation), sociolinguistic features (sex, age, education, profession), and technical details (length in time and words, acoustic quality, transcribers, revisers).

C-ORAL-ROM is compliant with the state of the art in spoken corpora. In particular, we got the written consent of every participant (in the case of individuals) or copyright holder (for media companies) to record, transcribe and issue the sound material. The acoustic quality of the recordings is essential for the application in speech technologies and language engineering. Most texts have been recorded with a DAT recorder or provided directly by the broadcasting station, always from a digital source.

In order to verify the reliability and quality of data, the corpus passed two types of evaluation. An internal validation carried out by the team itself, with at least three linguists transcribing/revising each text. Additionally, a program verified format errors, typos, badly formed xml tags, etc. Finally, the alignment of sound and text is the best guarantee that the transcription is accurate to the sound source. Besides all this, an external validation was carried out by experts from LOQUENDO and ELDA. LOQUENDO experts conducted an evaluation of the prosodic annotation (Moneglia *et al*. 2005). ELDA verified that every file has the proper size and format, the legal authorisation and all the features specified in the data sheets. As a result, content and form have been validated exhaustively.

A central issue in the corpus design was to achieve the comparability between the four sub-corpora, in order to allow contrastive studies. This goal has been achieved basically using the same design in the sampling for the four languages. The distribution of the corpus is divided between an informal sub-corpus and a formal one, approximately 150,000 words each. While the informal register is organised according to social context (private vs. public) and dialogic structure (monologue vs. dialogue-conversation), the formal register is organised according to channel (media, telephone, natural context). The only texts sub-classified by genre are media and formal in natural context.

Additional relevant aspects for comparability have been addressed. The four teams agreed in using the same transcription format as well as the metadata. In both cases, the consortium developed the C-ORAL-ROM format, based on the known CHAT format, with a xml-tagged version, which guarantees easy interpretation through the corresponding DTD.

The methodology for the alignment of sound and text was also agreed before starting the project, as a basic feature for the corpus. One of the members of the consortium, WinPitch France, was the technological provider for the alignment. WinPicth Corpus, the tool developed for such a goal, was used for the four teams to segment the signal and synchronise it with the corresponding transcription. This tool is also used to listen to the recordings along with its transcription.

In order to search and retrieve information from the morpho-syntactic annotated version, another program has been incorporated, Contextes, developed by Jean Véronis.

WinPitch and Contextes are included in the DVD published in *Studies in Corpus Linguistics* series by John Benjamins, in the Spring 2005. From our point of view, those tools present some limitations for further exploitation of the corpus[3]:

1. A combined query of text and sound cannot be made. Currently, one has first to search in the text with Contextes, and then one must run WinPitch for listening the speech fragment selected in the textual search.
2. There is no way of searching in the metadata. In the DVD version, the user can read the metadata for each text, but cannot look for information automatically. However, to query for particular features through the corpus is a very useful and requested functionality.

Due to these limitations, the LLI-UAM decided to implement a set of tools that solved these problems. A morpho-syntactic processor (GRAMPAL) developed previously in the LLI-UAM would be integrated. In the next sections we will describe these new tools.

3. Metadata Query

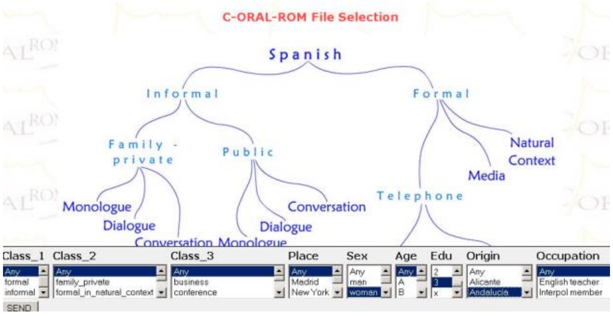
The metadata section or *header* appears at the beginning of each text, before the transcription. Most features in the header are compulsory, providing rich information on the text and its context (see Figure 1). For instance, date, place and situation of the recording. More interesting is the sociolinguistic information for every speaker, including sex, age, education, profession, dialectal origin and role in the conversation. Additional data on the duration, number of words, acoustic quality or transcribers are also codified.

```
@Title: esperando clientes
@File: efamcv07
@Participants: NUR, Nuria, (woman, B, 2, shop assistant, participant, Madrid)
             PAL, Paloma, (woman, B, 2, shop assistant, participant, Madrid)
             USE, Use, (man, B, 2, scriptwriter, participant, Madrid)
             PRI, Laura, (woman, B, 2, shop assistant, participant, Madrid)
             GUI, Guillermo, (man, A, 3, student, participant, Madrid)
@Date: 02/03/2001
@Place: Madrid
@Situation: conversation between friends at underwear shop, hidden, researcher
participant
@Topic: family and work at the shop
@Source: C-ORAL-ROM
@Class: informal, family/private, conversation
@Length: 6' 10''
@Words: 1593
@Acoustic_quality: A
@Transcriber: Manuel
@Revisor: Jesús; Guillermo, Inma and Manuel (prosody)
```

Fig.1 The header with metadata

On the structure of the header (more specifically on the DTD for the xml-tagged version) we have developed a simple query system based on pull-down menus. Some elements in the DTD have been selected as categories for the search. All the possible values found in the corpus are presented in the pull-down menu for each category. This

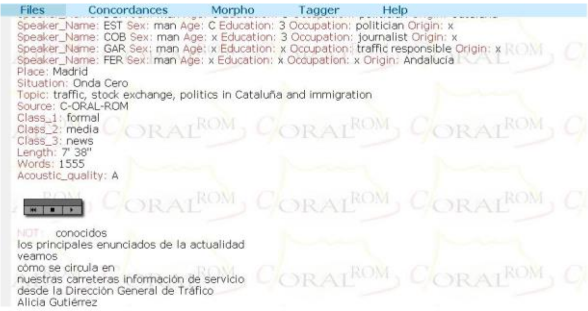
Figure 2 shows the screen capture for the query system, where the search is about how many documents contain women with higher education and from Andalucía.



Class_1	Class_2	Class_3	Place	Sex	Age	Edu	Origin	Occupation
Any	Any	Any	Any	Any	Any	2	Any	Any
formal	family_private	business	Madrid	man	A	3	Alicante	English teacher
informal	formal_in_natural_context	conference	New York	woman	B	x	Andalucia	Interpol member

Fig. 2 A sample of search

From the results of the query, one can retrieve directly documents that contain relevant data for the query. The access to each file allows to visualise the complete metadata and to listen the complete recording. In addition, when the cursor is placed on a word, its morpho-syntactic information is displayed in a pop window – see Fig. 3.



Speaker_Name: EST Sex: man Age: C Education: 3 Occupation: politician Origin: x
Speaker_Name: COB Sex: man Age: x Education: 3 Occupation: journalist Origin: x
Speaker_Name: GAR Sex: man Age: x Education: x Occupation: traffic responsible Origin: x
Speaker_Name: FER Sex: man Age: x Education: x Occupation: x Origin: Andalucia
Place: Madrid
Situation: Onda Cero
Topic: traffic, stock exchange, politics in Cataluña and immigration
Source: C-ORAL-ROM
Class_1: formal
Class_2: media
Class_3: news
Length: 7' 38''
Words: 1555
Acoustic_quality: A

NOT: conocidos
los principales enunciados de la actualidad
veamos
cómo se circula en
nuestras carreteras información de servicio
desde la Dirección General de Tráfico
Alicia Gutiérrez

Fig. 3 Full document view after the search

4. Concordancer with Sound and Morphological Information

In the Benjamins edition, each recording/text appears in three files that must be consulted independently:

1. The orthographic transcription,
2. The morpho-syntactic annotation of each word, in format MULTEXT.
3. The sound and text alignment.

For the textual versions (the transcripted and the annotated), the user must run Contextes. The multimedia files only can be consulted and searched with WinPitch[4].

Our version incorporates all the possible searches in one tool:

1. a word or a multiword expression can be looked up,
2. a morpho-syntactic tag can be searched
3. in all the retrievals, the searched expression is shown in the context of the utterance and the corresponding speech fragment in mp3 is provided.

The following additional information is supplied:

- The sociolinguistic characteristics of the speaker in each example.
- The morpho-syntactic information of each word in the utterance.

In Figure 4, an example of a multiword expression search can be seen. To the right the information about the speaker is displayed (in red). When locating the cursor upon any word, a pop window displays its morpho-syntactic features, as it occurs in the case of the complete document retrieval. Pressing in the button to the right the corresponding speech fragment will be listened to.

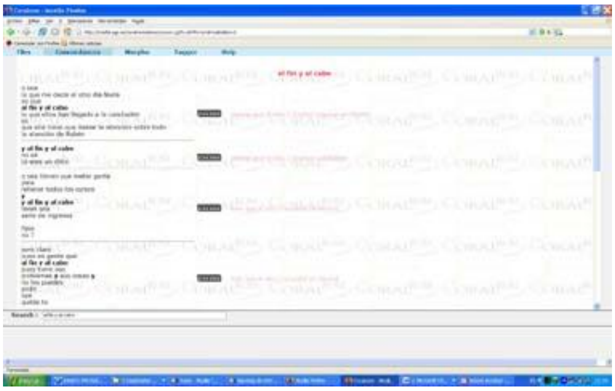


Fig. 4 Search retrieved for a multi-word expression

5. Morphological Analyzer and POS Tagger

For the morphological analysis we use GRAMPAL, designed by Moreno (1991) as a part of his Ph.D. dissertation. GRAMPAL is based on a rich lexicon of over 45,000 lexical units (see Table 1), and morphological rules to expand the analysis or the generation to over 500,000 inflected words. The system has been successfully used in language engineering applications as ARIES (Goñi *et al*. 1997) as well as in linguistic description (Moreno and Goñi 2002). GRAMPAL was originally developed for analysing written texts, but during the C-ORAL-ROM project it was extended and adapted to spoken Spanish. In particular, a PoS tagger and an unknown word recogniser were integrated in the tool (Moreno and Guirao 2003). The most complete and updated information on GRAMPAL can be read in Moreno and Guirao (in press), including an extensive evaluation of the tool.

Endings lexicon		**Stems lexicon**		**Multi-words lexicon**	
		N	25.426	ADV	507
Noun morphs	4	ADJ	11.290	PREP	349
Verb morphs	179	V	10.568	C	91
		ADV	189	INT	70
		P	109	FOREING-W	30
		MD	74	Q	24
		PREP	40	ADJ	15
		C	26		
		POSS	26		
		REL	16		
		ART	5		
Total	**183**		**47.769**		**1086**

Table 1 Number of entries in the GRAMPAL lexicons

The morpho-syntactic annotation of the whole corpus has been one of the most important contributions of C-ORAL-ROM, since it has allowed making the listing of *lemmas* and the count by Part-of-Speech categories, which provides an added value to the empirical studies extracted from the four corpora. In particular, each subcorpus comes with the frequency list of lemmas and forms, which is a significant contribution to lexical studies of the most frequent units in spoken Italian, French, Spanish and Portuguese.

Category	**Main Tag**	**Number of different subtags per category**
Noun	N	5
Proper Noun	NPR	1
Adjective	ADJ	9
Article	ART	4
Possessive	POSS	1
Demonstrative	DEM	1
Quantifier	Q	1
Pronoun	P	9
Relative Pron.	PR	1
Verb	V	46
Auxiliary verb	AUX	46
Preposition	PREP	1
Adverb	ADV	1
Conjunction	C	1
Discourse Marker	MD	1
Interjection	INTJ	1

Table 2 Spanish C-ORAL-ROM tagset

Table 2 shows the tagset used by GRAMPAL, consisting of 129 tags, including one for each Spanish verb wordform (46) and each auxiliary verb wordform (46). This information is annotated by the program as output of the analysis. In the Benjamins edition, the annotation was delivered in the MULTEXT format, that is, all the information in an atomic tag. We prefer, however, to display it in the form of feature structures, which allows retrieving partial information in a more flexible way. We have developed an xml version that can express the retrieval of a given tag as a feature structure.

The annotated version of the Spanish C-ORAL-ROM or whatever other corpus tagged by the LLI-UAM is a process in which first an automatic tagging with GRAMPAL was produced and then expert human annotators revise and correct the tagged corpus. Guirao and Moreno (2004) describe a tool for aiding human annotator in the revision.

In the current version, the analyser and the PoS tagger are split. The former can be used for observing which analyses GRAMPAL provides for a given string. The tagger can be used for verifying which analysis is chosen, after disambiguation, by the program. From the evaluation, we estimate that there are 5% of errors, one out of twenty words.

In the future, an automatic phonological transcriptor will be included, which will provide a standard phonological transcription from the orthographic one.

6. Application of the Tools on the Corpus

The C-ORAL-ROM corpus, in the first place, and the computer tools for searching, secondly, seem to us essential resources to investigate the properties of spoken Spanish. As linguists, we are interested especially in this topic: the empirical studies of the spoken language. But we cannot forget the applied side, and in particular, the interest of the spontaneous speech for teaching/learning languages. Traditionally, second language teaching has been based almost exclusively on written texts or written reformulation of conversations. There are very few examples of the use of real conversations in second language education, due to its complexity, but mainly to its lack of “grammaticality” and the typical presence of digression, interruptions, improvisation, etc. In addition, the spontaneous speech texts do not have in consideration the level of difficulties for the language learner, so that very different grammatical structures can be found in the same text.

6.1 Describing Spontaneous Speech Spanish Properties

The LLI-UAM team has conducted several studies on the empirical data, with the help of computer programs. This is a summary of the findings.

Transcription problems with spoken language

.

The transcription of spontaneous recordings presents different difficulties depending on the type of text. To be able to determine which communicative interactions cause more difficulties in the transcription is a useful information for the compilation of new spontaneous speech corpora: the human resources are limited and the estimation of the degree of difficulty is essential to calculate the costs and to design the recordings sampling.

For that reason, the LLI-UAM team conducted the following experiment (González *et al.* 2004), presented in the Workshop on Compiling and Processing Spoken Corpora at the LREC-2004. We started from the hypothesis that the transcription problems are related to the occurrence frequency of two classes of typical linguistic phenomena in the spontaneous speech:

- **Production features**, such as *fragmented words, supports* and *retractings*
- **Interaction features**, such as the *number of turns* or the *overlapping*.

This would lead to the conclusion that the more frequent these phenomena were in a spoken interaction, the more time and effort needed by the linguist in the process of transcription. This hypothesis is intuitive and recognised by different researchers. The novelty of the experiment was in that those phenomena were tagged in the transcription and, therefore, a complete count can be done automatically. Our task was to analyze the results to verify empirically and quantitatively the hypothesis. In order to evaluate the hypothesis we decided to create two scales of transcription difficulty:

1. **Degree of formality**: we assumed that on one end of the scale we find the private recordings, as the most complex ones, and on the other end the formal texts. In general, the more formal the speech is, closer to the written standards is. Media texts are located in the meddle, since one can find examples of informal register (sports and talk shows) as well as very formal speech (scientific programs and political interviews).
2. **Number of speakers**: this parameter only applies to the informal subcorpus. The more participants in the recording, the more complex the transcription is.

We can represent graphically both scales in the following Figure:

Degree of formality
informal media formal
+ difficult -------------------------------------- - difficult

Number of speakers
conversation dialog monolog
+ difficult --- - difficult

The procedure for quantifying the parameters was first to choose a series of features tagged in the corpus. For the degree of formality, we selected fragmented words, vocalic supports and retracting. For the number of speakers, overlapping, number of turns and speed rate. A program automatically calculated the average number of words between tags of the same feature, except for the speed rate where the count was the average number of words per second. This way, the higher the number of words per feature, the less important the feature is for that particular type of text (formal, informal or media). In other words, since we assume that those features are problematic for transcription, less frequency implies less difficulty.

Let us see an example to illustrate the method. Consider the overlapping rate: the count shows that an overlapping takes place each 54.4 words of average in informal texts, whereas in formal texts is every 317.7 words. That is to say, the overlapping in the informal texts is approximately six times more frequent that in the formal ones.

Thus we applied the procedure to the 6 selected features in the two scales and obtained the following conclusions:

1. The production features (fragmented words, vocalic supports and retractings) are much more frequent in the formal texts that in the informal register, contrary to the hypothesis. This can be explained because speakers in a formal situation tend to take care of their expression and dedicate more time to produce their utterances. However, in informal texts the necessity of communication seems to be more important than producing a correct and precise speech.
2. The interaction features (words per turn, overlapping and speed rate) behave according to the intuitive hypothesis: they are more frequent in informal texts, where the communication is much faster that in the formal ones. The speed of production explains that the turns are shorter but more frequent, and also that there are continuous overlapping and more words per second.

As a main conclusion of the study, we can say that the more influential features in the transcription difficulty are those of interaction, whereas the production features are not a problem for the transcriptor. A secondary conclusion is that if we combine both types of characteristics, the less problematic texts to transcribe are those of media, since although locating themselves in the middle of the scale, they are always closed to the simplest cases. Indeed, the professional speakers do not produce many vocalic supports or fragmented words, whereas they speak as fast as the speaker in informal texts.

Forms, lemmas and part-of-speech distribution in spoken Spanish.

The C-ORAL-ROM book includes a chapter to compare the statistics extracted from the annotated data. The tables and the graphs also appear in the DVD accompanying the corpus so that researchers can have easy access to the data. The following parameters have been computed:

1. Average length of utterances measured in words and by types of text.
2. Average length of the turns, in words and by types of text.
3. The speed rate, measured in words per second.
4. The average length of the tone units, in words.
5. The fragmentation of the speech, understood as the average number of words between two interruptions.

An interesting observation is that, in general, Portuguese, Spanish and Italian are closer as far as these measures are concerned than French. The data are not sufficient to draw significant conclusions but they illuminate tendencies for future investigations.

In our opinion, the most remarkable of the material published in the book are the lists, ordered by frequencies, of the forms and the lemmas in each corpus. These inventories could be elaborated from the morpho-syntactic annotation. They are a landmark in the study of the lexicon of the four languages, for the spontaneous speech variant, by the reliability and representativeness of the sampling. As a example, we will write down some data from the Spanish inventory:

Number of different forms	21536
Number of different lemmas	19315
Number of different Noun lemmas	4583
Number of different Verb lemmas	1779
Number of different Adjective lemmas	2023

Table 3 Basic figures for lemmas and forms in the Spanish corpus

Relating linguistic units to socio-contextual information

Guirao *et al*. (2006) present the application of the log-likelihood test (Dunning 1993) in order to find the most distinctive lexical and grammatical items for a given socio-contextual feature, related to the speaker (sex, age, education, etc.) or the context (dialogue, media, telephone, etc.). To this end, we make use both of the morpho-sytactic tagged version and the metadata for every text in the corpus.

In order to identify the distinctive units we applied the Dunning test, an statistical method that, contrary to the common fashion, does not assume the normal distribution (the bell-shaped) of the units in a corpus. Instead, the log-likelihood ratio assumes a binomial distribution more appropriate for rare but distinctive words. In addition, this test does not need balanced corpora for comparison. The method consists of selecting a sub-corpus (for instance, the male speakers) and comparing it against the rest of the corpus (in this case, the female speakers). Units that appear only in the selected sub-corpus but not in the complementary set are given a higher value.

We conducted experiments on different units and registers. In particular, we took words and multi-words; lemmas and part-of-speech tags, and related them to well-defined domains (meteo news, professions) and general registers (formal vs. informal;

male vs. female). As example, Table 4 gives the most distinctive verb lemmas in the male and female sub-corpora.

Female	**Male**
ir	escuchar
decir	recordar
saber	aparecer
venir	llegar
decir	contemplar
mirar	caminar
comprar	intentar
gustar	amar
quedar	juntar
contar	superar

Table 4 The 10 most distinctive verbs for male and female subcorpora

Frequency dictionaries of phonemes and syllables.

Our most recent work on the corpus has been to extract phonemic and syllabic inventories for Castilian Spanish[5]. These lexicons have been developed based on a automatic transcriptor whose input has been revised and assessed by linguists. The inventories include absolute frequencies of occurrence of the different phonemes and syllables, and have been contrasted with a similar inventory extracted from a comparable written corpus, finding evidence that previous frequency dictionaries of Spanish, based mainly on written texts, do not provide an accurate description of spontaneously spoken Spanish. This is specially marked in the most frequent syllables, since in spoken language are those corresponding to the most frequent words: *a, que, de, es* … In written texts, among the top 10 syllables are *ta, do, na,* and *ma,* none of them is an isolate word.

Another empirical relevant finding is that the first 100 syllables represent more than the 80% of the spoken corpus. The first 650 syllables cover more than the 99%.

We estimate that in Spanish there are no more than 1,500 different syllables. Preliminary results will be published in Moreno *et al*. (in preparation).

6.2 Teaching/Learning Spanish as a Second Language

With the purpose of evaluating the applicability of C-ORAL-ROM to research and education in the fields of phonetics, language technologies and the Romance languages, the Department of Language Projects and Technologies of the Instituto Cervantes developed a data base with information on potential users of corpus, whose experience in the field gives special relevance and reliability to their opinions about the applicability of this tool.

Some of the potential uses in the field of teaching/learning Spanish as a Second Language provided by the consulted experts are:

- It is a basic material for Computer Assisted Language Learning (CALL).
- It is an important material for self-learning.
- It is a useful material to practice and develop spoken skills.
- It is an essential material for a communicative approach
- It is useful to practice with different language registers and communicative situations
- It is useful to practice with discourse markers, syntactic structures, pragmatic strategies and pronunciation (including use contexts for allophones).
- It is useful to study vocabulary and lexical expressions.
- It is useful to analyse the variable use of the pronominal subject from a pragmatic point of view.

- It is a multimedia material that contributes to increase the interest of the students

The comments and suggestions from the experts can be summarised in the following conclusions with respect to the potential that this corpus has:

1. C-ORAL-ROM has many direct applications to foreign language teaching, as well as a great potential to be improved by means of new developments. It partly resolves the lack of samples that can be used for the study and practice of the spoken language.
2. Nevertheless, it is necessary to facilitate the task to the teachers and students, providing guides and suggestions about how using the tool, as well as to improve the access to the examples and the information linked to them. A new search engine to query the corpus was needed[6].

We are preparing a version for Spanish as Second Language, where texts and examples have been selected by hand, according to general criteria of the specialists in the field. In particular, the most usual language and communicative structures found in textbooks are selected to be used as a teaching complement to the current materials. A first version has been presented in a workshop on spoken skills in teaching Spanish, at the XVII International Congress of the Asociación para la Enseñanza de Español como Lengua Extranjera (ASELE), in September 2006.

7. Conclusions and future work

C-ORAL-ROM is a resource of multiple applications in the research, the education, and the language technologies. We have shown some examples in the different fields and will continue working in new tasks, as the extension of the methodology to the compilation of a corpus of child speech.

Among the more urgent tasks we have the conclusion of the phonological and syllabic transcription, as well as the publication of a version of C-ORAL-ROM for Spanish as a Second Language, with the selection of the most appropriate texts, approximately half of original corpus.

Acknowledgements

The authors’ work has been supported by a grant of the Spanish Ministry of Education and Science under the contract TIN2004-07588-C03-02.

REFERENCES

Cresti, E. and Moneglia M. (eds) (2005). *C-ORAL-ROM Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: John Benjamins.

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

González, A. De La Madrid, G., Alcántara, M., De La Torre, R, and Moreno, A. (2004). Orality and difficulties in the transcription of a spoken corpus. In *Proceedings of the Worshop on Compiling and Processing Spoken Corpora. LREC-2004*. Lisboa.

(

Goñi, J. M., González, J.C. and Moreno, A. (1997). ARIES: a lexical platform for engineering Spanish processing tools. In Natural Language Engineering, 3 (4), 317-345.

Guirao, J. M. and Moreno, A. (2004). A Toolbox for tagging the Spanish C-ORAL-ROM corpus. In *Proceedings of the Worshop on Compiling and Processing Spoken Corpora, LREC-2004,* Lisboa.

Guirao, J. M., Moreno, A., González, A., De La Madrid, G. and Alcántara, M. (2006). Relating linguistic units to socio-contextual information in a spontaneous speecho corpus of Spanish. In *Corpus Linguistics around the World*. Amsterdam: Rodopi.

Moneglia, M., Fabbri, M., Quazza, S., Panizza, A., Danieli, M., Garrido, J.M., and Swerts, M. (2005): Evaluation of consensus on the annotation of terminal and non-terminal prosodic breaks in the C-ORAL-ROM Corpus. In Cresti and Moneglia (eds), *C-ORAL-ROM Integrated Reference Corpora for Spoken Romance Languages*, 257-275, Amsterdam: John Benjamins.

Moreno, A. (1991). *Un modelo computacional basado en la unificación para el análisis y generación de la morfología del español*. Ph.D. Thesis, Universidad Autónoma de Madrid.

Moreno, A. and Goñi, J. M. (2002). Spanish Inflectional Morphology in DATR. *Journal of Logic, Language and Information*. 11, 70-105.

Moreno, A. and Guirao, J. M. (2003). Tagging a spontaneous speech corpus of Spanish. In *Proceedings of RANLP 2003*. Borovets, Bulgaria.

Moreno, A. and Guirao, J. M. (in press). Morpho-syntactic Tagging of the Spanish C-ORAL-ROM corpus: methodology, tools and evaluation. To be published in Kawaguchi, Zaima, Takagaki and Usami (eds) *Linguistic Informatics: Contributions of Linguistics, Applied Linguistics, Computer Sciences*. Series UBLI vol. V. Amsterdam: John Benjamins.

Moreno, A., De la Madrid, G., Alcántara, M., González, A., Guirao, J. M. and De la Torre, R. (2005). The Spanish corpus. In Cresti and Moneglia (eds), *C-ORAL-ROM Integrated Reference Corpora for Spoken Romance Languages*, 135-161, Amsterdam: John Benjamins

[1] C-ORAL-ROM stands for Corpus ORAL ROMance.

[2] The project home page is: http://lablita.dit.unifi.it/coralrom/index.html

[3] WinPitch is surely very sophisticated in features for acoustic and experimental phonetics for most users of C-ORAL-ROM. On the other hand, Contextes is a simple concordancer, which does not include statistical tools or complex searches, as other popular programs such as WordSmith.

[4] Note that in the version published by Benjamins all the files are encrypted to be used only with the tools provided. Therefore, those files may not be analysed with other programs.

[5] For the American dialects, the phonological system and the frequency of the units may be significantly different.

[6] This comment was made before we started working in the present tools. In fact, this suggestion was the initial motivation for the current developments at the LLI-UAM.