The C-ORAL-ROM CORPUS
A Multilingual Resource of Spontaneous Speech for Romance Languages.

Emanuela Cresti[1], Fernanda Bacelar do Nascimento[2], Antonio Moreno Sandoval[3], Jean Veronis[4], Philippe Martin[5], Kalid Choukri[6]

[1]LABLITA, Dipartimento di Italianistica,
Università di Firenze
Piazza Savonarola 1,50125 Firenze Italy
elicresti@unifi.it
[2]Centro de Linguistica da Universidade de Lisboa
Complexo Interdisciplinar, Av Gama Pinto, 2, 1649-003
Lisboa Portugal
fbacelar.nascimento@clul.ul.pt
[3]Laboratorio de Lingüística Informática Departemento
de Linguistica, Universidad Autonoma de Madrid
Carretera de Colmenar Viejo Km 15 Cantoblanco
28049 Madrid Spain
Sandoval@maria.lllf.uam.es

[4]Description Linguistique Informatizée sur Corpus,
Université de Provence
29, Avenue Robert Schuman13621 AIX EN
PROVENCE - Cedex 1 France
Jean.Veronis@up.univ-mrs.fr
[5]Pitch Instruments France
24, Rue Las Cases 75005 France
philippe.martin@fnac.net
[6]European Language Distribution Agency
European Language Association Agency (ELDA)
55-57, Rue Brillant-Savarin 75013 Paris France
choukri@elda.fr

Abstract

The C-ORAL-ROM project has delivered a multilingual corpus of spontaneous speech for the main romance languages (Italian, French, Portuguese and Spanish). The collection aims to represent the variety of speech acts performed in everyday language and to enable the description of prosodic and syntactic structures in the four romance languages. Sampling criteria are defined in a corpus design scheme. C-ORAL-ROM adopts two different sampling strategies, one for the formal and one for the informal part: While a set of typical domains of application is selected to document the formal use of language, the informal part documents speech variation using parameters referring to the event's structure (dialogue vs. monologue) and the sociological domain of use (family-private vs public). The four romance corpora are tagged with respect to terminal and non terminal prosodic breaks. Terminal breaks are assumed to be the more relevant cues for the identification of relevant linguistic domains in spontaneous speech (utterances). Relations with other concurrent criteria are discussed. The multimedia storage of the C-ORAL-ROM corpus is based on this principle; each textual string ending with a terminal break is aligned, through the Win Pitch speech software, to its acoustic counterpart, generating the data base of all utterances.

1. Introduction[1]

The C-ORAL-ROM project (IST 2000-26228) has delivered a multilingual corpus of spontaneous speech for the main romance languages. Four collections of Italian, French, Portuguese and Spanish corpora have been delivered by national providers. The corpus (123h35'40'') has been recorded in free situations with various technical apparatus. C-ORAL-ROM will be distributed by ELDA in a multimedia edition with the Win Pitch Corpus speech software (© Pitch France)[2]. The main features of the C-ORAL-ROM corpus regard: 1) Corpus design; 2) Metadata; 3) Multimedia format; 4) Prosodic annotation and alignment; 5) Tagging of utterance boundaries.

2. Corpus Design

C-ORAL-ROM corpora have a mid-dimension as spontaneous speech resources (300,000 words for each Language), however the collection aims to represent the variety of speech acts performed in everyday language and to enable the description of prosodic and syntactic structures in the four romance languages, from a quantitative and qualitative point of view.

One of the main characters of spoken language when compared to written language is the huge variability of the speech events according to individual characters, context of use, semantic domain. Therefore sampling criteria for the session recordings are strictly defined in a corpus design scheme in order to represent significantly the main variation parameters of the spoken domain in each 300,000 word sub-corpus (no restrictions on the number of the recorded subjects).

According with the tradition the parameters are the following (see Labov, 1966; Biber, 1998; Gadet, 1996): 1) Dialogue structure; 2) Sociological domain of use; 3) Gender; 4) Semantic domain; 5) Channel. Such parameters have been projected in a corpus design matrix. Each field has the same number of words in each corpus of the multilingual resource, ensuring their comparability:

Type	*Sub_type*	*Sub_sub_type*
Informal	Private	Dialogue and multi-dialogue
Informal	Private	Monologue
Informal	Public	Dialogue and multi-dialogue
Informal	Public	Monologue

Type	*Sub_type*	*Sub_sub_type*
Formal	**Natural context**	political speech, political debate; preaching; teaching; professional talk; conference; business; law
Formal	**Media**	talk shows; scientific press; reportage; interviews; sport; news; weather news
	Telephone	private conversations; human-machine interactions

[1] This paper was written by the project co-ordinator E. Cresti.
[2] A publication in encrypted and compressed form has been also foreseen (Cresti & Moneglia, to appear).

The variation parameters, adopted in C-ORAL-ROM for the representation of the spoken language universe, have also been tested in other recent large spoken corpora collections (Dutch Corpus), and already sketched at LREC (Cresti et. al., 2002).

As for the Dutch Corpus the above matrix tends to define the crossing over between the communicative event's structure parameters (Dialogue/Conversation vs. Monologue), the sociological context of use (Private vs. Public) and the channel (Broadcast vs. Natural context)[3].

However C-ORAL-ROM adopts two different sampling strategies, one for the representation of spoken language in formal contexts and one for the informal part. Only in the formal part the genre (or domain of application) of the recorded sessions is strictly defined in a closed list of sub-sub types; this information is not provided for the informal part

While it is natural to assume the existence of a series of closed situations in which, normally, in a certain social-historical context, the formal use of language is preferred, the same does not hold for the universe of informal speech. To feature typical contexts of use is a specific, marked trait of formal speech. For this reason, it can be effectively identified by listing its most typical contexts of use.

On the contrary the set of situations where informal language is used is open, and its domain cannot be represented by a list of typical contexts of use. No context is more typical than another.

This is not so obvious. For example the Dutch corpus tries to define, as specifically as possible, the contexts of collection of monologues and dialogues, by determining the number of words for each genre (e.g Business transaction, Picture description, Interview, Face-to-face conversation, Telephone).

By observing *a posteriori* the sampling strategies used by these two collections, which are both aimed towards the documentation of spontaneous speech and the comparability of data belonging to various corpora, the concurrence of two criteria appears evident. It is a fact that, by strictly defining the genres, a higher degree of data comparability can be reached. However, the downfall of this practice is the *a priori* exclusion of significant spheres of informal speech, where genre characterizations still remain largely unexplored[4]. This choice is meaningful for formal speech only.

We can conclude that C-ORAL-ROM's corpus sampling, by not defining explicitly the genres and domains of use of speech, guarantees, in theory, the possibility of occurrence, in the collection, of any significant genre. In other words, no genre has zero probability of occurrence.

3. Meta-Data

The definition of a complete set of metadata for each session allows the validation of each corpus with respect to the corpus design of the resource. For each session a rich series of metadata is delivered in CHAT and IMDI format ensuring multitask exploitation of the resource for linguistics and Human language technologies. Metadata comprise relevant information regarding: 1) Participants (sex, age, education, profession, role in the recording event, geographical origin); 2) Recording session (topic, recording situation, location and date of recording, length, acoustic quality, number of words transcribed, recording condition); 3) Copyright and revision of transcripts.

4. Prosodic annotation and alignment

Corpora are transcribed in standard textual format (CHAT) and are completely tagged with respect to prosodic breaks and simultaneously parsed in relevant linguistic domains. *Terminal* and *non terminal breaks*, are discriminated through perceptive judgments and reported in the transcripts. Moreover the relation between such prosodic cues and the identification of the linguistic relevant domain in spoken language (utterance) (See. Miller & Weinert, 1998; Quirk et al., 1985; Biber et al., 1999; Cresti, 2000) is highlighted.

In C-ORAL-ROM the selection of utterances (Austin, 1962) within the speech flow goes hand in hand with the representation of prosody. It is assumed that each utterance has a profile of *terminal intonation* (Karcevsky, 1931; Crystal, 1975).Therefore the presence of terminal breaks turns out as the main cue for the detection of the reference unit of spontaneous speech: each prosodic unit ending with a terminal break is considered an utterance. Given the relevance of prosodic tagging delivered in C-ORAL-ROM, the level of inter-annotator agreement has been evaluated by an external institution (Danieli et al., in this volume).

The multimedia storage of the C-ORAL-ROM corpus is based on this principle; each textual string ending with a terminal break is aligned through the speech software Win Pitch Corpus (© Pitch France) to its acoustic counterpart, generating the data base of all utterances in the resource, that is one of the main novelties of the resource (roughly 120,000 in the multilingual corpus). This ensures a natural and meaningful text / sound correspondence for both prosodic modeling, speech act theory and corpus based studies of spontaneous speech.

The alignment is defined on two levels: 1) at textual level, allowing access to the acoustic signal from the text; 2) at wave level, where the sound source and the textual information corresponding to each speaker are displayed on independent layers following the time axis.

As a consequence of this, C-ORAL-ROM's alignment simultaneously allows the appreciation of textual properties and the direct analysis of speech: (real-time fundamental frequency tracking, spectrographic display; re-synthesis of prosodic parameters), see figures below.

5. The identification of utterance limits in spontaneous speech

Although the concept of utterance is in actual fact generally recognized as the product of the speech performance (Biber et al., 1999), its definition is far from obvious. The definition of utterance may be linked to syntactic-semantic properties, thus enabling its identification through a clause, or propositional structure (a C-Unit in the Longman Grammar's lexicon). Alternatively, a practical equivalence has often been

[3] The *formality* variation parameter however was not explicitly considered in the Dutch corpus that refers to the same variation through the distinction between Scripted and Unscripted events.

[4] From a practical point of view, the inclusion in the corpus sampling of vague categories such as "face to face conversations" can make the two criteria equivalent.

proposed between the utterance and the linguistic sequence between two silences (see. TEI guidelines). Such criteria may be considered preferable to the prosodic criterion adopted in C-ORAL-ROM, as they are more objective, whereas the prosodic criterion may be considered arbitrary as it is based on perception.

In the following we will point out: a) that the syntactic structure appears frequently underdetermined in spontaneous speech, and that its definition is rather a function of prosodic cues; b) that the timing of an utterance is not linguistically significant, as it is, at the same time, too weak and too strong to determine the utterances boundaries in spontaneous speech corpora.

C-ORAL-ROM offers a piece of oral material which clearly demonstrates both facts. The following dialogic exchange between the beautician and the client who is about to undergo a depilation of the legs, allows the evaluation of the difficulties encountered by both criteria. Let consider the bare transcription (words only) accompanied by the essential contextual information:

*EST: *o vieni dai* [come on then]
%act: the beautician invites her client to begin the hair removal
*CLA: *a patire* [to suffer]
*EST: *no ascolta qui sopra sì* [no listen up here yes]
%act: the beautician gets closer to the leg to be depilated
*CLA: *qui sì* [here yes]

The third and the fourth turns are verb-less. Their syntactic structure is underdetermined by the actual syntactic data and may be compatible with many possible interpretation. The following punctuation highlights eight possible word grouping in the first case and three in the second:

No. Ascolta. Qui? Sopra? Si.	[No. Listen. Here ? At the top ? Yes]
No. Ascolta qui sopra. Si	[No. Listen here . Yes]
No. Ascolta. Qui sopra, si.	[No. Listen. Up here, yes]
No, ascolta . qui. Sopra si.	[No, listen, here. Yes on top .]
No, ascolta . Qui sopra, si.	[No, listen. Up here, yes]
No. Ascolta. Qui. Sopra si.	[No. Listen. Here. Yes, on top]
No, ascolta. Qui. Sopra si.	[No, listen. here. Yes, on top]
No. Ascolta, qui sopra. Si	[No. Listen . (what about) up here ? Yes]
Qui si.	[Here, ok]
Qui, si.	[Here, yes]
Qui. Si	[Here. Yes]

All word groupings, that correspond to distinct possible utterance boundaries, are consistent with the pragmatic context. Therefore, neither the syntactic nor the contextual information are sufficient to determine the actual structure of the previous turns.

On the contrary these turns are not ambiguous in speech. Once the information bared by terminal and non terminal breaks is perceptively recovered the reference units can be determined with precision. The speech act label in the dependent tier may help the reader to achieve the proper interpretation:

*EST: o vieni / dai // [come on then]
%ill: invitation
*CLA: a patire // [to suffer]
%ill: ironical assertion
*EST: no // ascolta / qui sopra ? sì //
[no // listen / (what about) up here ? yes //]
%ill: reassurance (1) question, introduced by a conative (2) self answer (3)
*CLA: qui ? sì //
[here ? yes]
%ill: question (1) answer(2)

In other words, the prosodic structure is the index which determines the choice of the possible structure for both turns (the last one of both sets), not the reverse.

Figures show that in spontaneous speech, verb-less contexts, where the syntactic structure is underdetermined, as the above, appear in around 30% of utterances (38% according to Longman Grammar). More specifically the statistic measurements on the C-ORAL-ROM corpus show that verbless-utterances are 38.1% in Italian, 24.1% in French; 37.23% in Spanish and 36.57% in Portuguese .

The "from silence to silence" criterion is the most common in present approaches to multimedia spoken language archives, probably because the automatic recognition of pauses in the speech flow is a quite easy task to be pursued given the actual technologies. As a matter of fact it sound also reasonable. The utterance boundaries frequently occur together with significant wave interruptions. For example, in the first example considered, after "no" there is an interruption of around 600 ms (in yellow in Fig.1) that accompanies the beginning of another utterance.

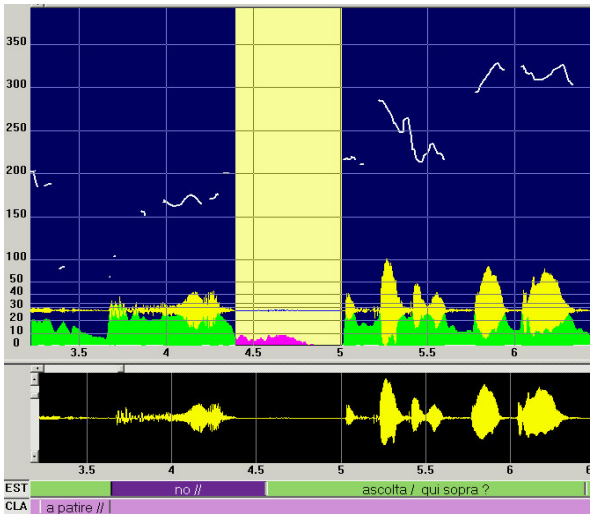

Fig. 1

The relevant fact from both a theoretical and technical point of view is that utterances may also occur with no need for pauses (too strong criterion) and on the contrary the occurrence of a pause is not a sufficient cue to infer the conclusion of an utterance (too weak criterion).

For example the speech flow of the fourth turn corresponds to two distinct speech acts (respectively marked in blue and yellow in Fig. 2), but there is no pause separating the two speech events. On the contrary the perception of prosody appears to be sensible to the 20hz discontinuity occurring at the start of the second utterance.

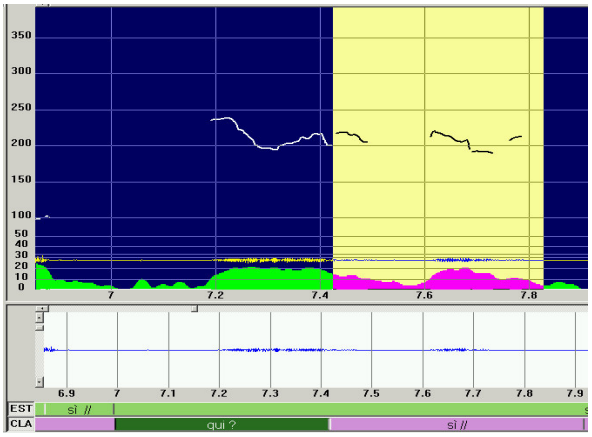

Fig. 2

In spontaneous speech the reverse cases are also frequent. A perceptively relevant prosodic break may be accompanied by a pause even if the break does not mark the end of the utterance. The following, according with perception of prosodic movement is a typical utterance with Topic – Comment structure taken from the same dialogue. The first element has prefix intonation and is perceived as non concluded, while the second string is concluded ('t Hart et al., 1990):

*EST: ... lei / prima veniva tutte le settimane //
[she / used to come every week once //]

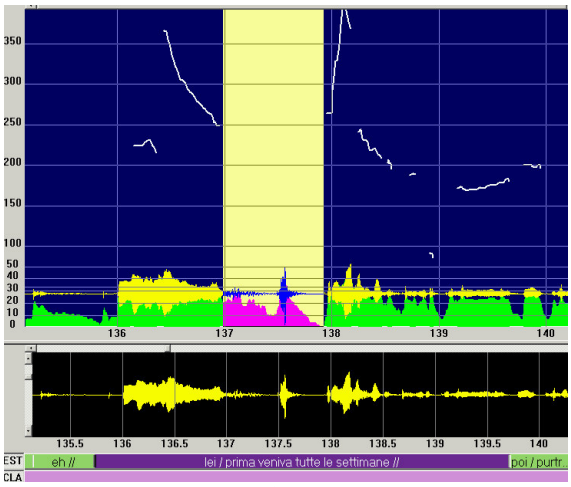

Fig. 3

The Topic unit is separated from the Comment unit by a pause of around a second (in yellow). Following the silence to silence criterion, the sequence will be wrongly considered a sequence of two distinct utterances. Therefore the concept of utterance as a sequence between two silences does not match the concept determined on a prosodic basis. It is at the same time too weak and too strong a notion.

C-ORAL-ROM can provide a quantitative measurement of the incidence of both kind of noises that may emerge in the application of the silence to silence criterion. The French corpus C-ORAL-ROM has been very nicely tagged with both the temporal and the perceptual criteria. Pauses of more then 200 ms. have been detected automatically in the speech flow and annotated in the transcripts. At the same time the corpus has also been tagged with respect to all terminal and non terminal prosodic breaks, perceived by the expert operators who transcribed and tagged the corpus. On the basis of the results of this double tagging, we recorded that around 63% of sequences ending with a terminal break are accompanied by a pause, while 37% of sequences ending with a terminal break do not bear a pause. Quite a big under-extension.

On the other side it is also extremely relevant to note that around 42% of breaks that have been considered non terminal are also accompanied by a pause. A dramatic over-extension .

The prosodic strategy proposed by C-ORAL-ROM to identify utterance boundaries in spoken corpora is at the same time reliable and easy to be pursued.

6. References

Austin, L.J. (1962). How to do things with words. Oxford: Oxford University Press.

Biber, D. (1988). Variation across speech and writing. Cambridge: Cambridge University Press.

Biber, D., S. Johansson, G. Leech, E. Finegan (Eds.) (1998). Corpus linguistics: investigating language structure and use. Cambridge: CUP.

Biber D., S. Johansson, G. Leech, S. Conrad, E. Finegan (Eds.) (1999). The Longman grammar of spoken and written English. London: Longman.

CHAT http://childes.psy.cmu.edu/manuals/CHAT.pdf

Cresti, E. (2000). Corpus di italiano parlato, vol. I- II, CD-Rom, Firenze: Accademia della Crusca.

Cresti E.; Moneglia M., Bacelar F., Sandoval A.M., Veronis J., Martin PH., Choukri, K., Mapelli V., Falavigna D.; Cid, A. (2002). The C-ORAL-ROM Project. New methods for spoken language archives in a multilingual romance corpus. In Proceedings of LREC 2002, vol. 1 (pp. 2--10). Paris: ELRA.

Cresti, E., Moneglia, M. (Eds.) (to appear). C-ORAL-ROM. Amsterdam: John Benjamins.

Crystal, D. (1975). The English tone of voice. London: Edward Arnold.

Danieli, M., Garrido, J. M., Moneglia, M., Panizza, A., Quazza, S., Swerts, M. (in this volume). Evaluation of consensus on the annotation of prosodic breaks in the romance corpus of spontaneous speech "C-ORAL-ROM".

Dutch Corpus http://lands.let.kun.nl/cgn/edesign.htm

Gadet, F. (1996). Variabilité, variation, varieté: le Français d'Europe. French Language Studies, 6, 45--58.

't Hart, H., Collier, R., Cohen, A. (1990). A perceptual study on intonation. An experimental approach to speech melody. Cambridge: CUP.

IMDI http://www.mpi.nl/IMDI/

Karcevsky, S. (1931). Sur la phonologie de la phrase. In Travaux du Cercle Linguistique de Prague, IV.

Labov, W. (1966). The social stratification of English in New York City. Washington D.C.

Miller, J., Weinert, R. (1999). Spontaneous Spoken language . Oxford: Clarendon Press.

Quirk, R., Greenbaum, S., Leech, G. Svartvik, J. (1985). A comprehensive Grammar of the English Language. London: Longman.

TEI http://www.tei-c.org/Guidelines2/

Winpitch, http://www.winpitch.com