DOI: 10.17469/O2106SLI000005

EMANUELA CRESTI, LORENZO GREGORI, MASSIMO MONEGLIA, CARLOTA NICOLÁS, ALESSANDRO PANUNZI

The LABLITA Speech Resources

The LABLITA lab of the University of Florence makes available on the web three main spoken corpora: the LABLITA reference corpus of spoken Italian, the IPIC cross-linguistic database of information structure, the C-Or-DiAL Spoken Spanish corpus for teaching Spanish L2. These resources have been annotated following the Language into Act Theory (Cresti 2000) for what regards prosody and its relationship with pragmatics and information structure, and present the speech flow segmented into utterances and information units in correspondence with perceptively relevant prosodic breaks. The LABLITA corpus gives an account of the diaphasic variation of the Italian language spoken in Tuscany according to a detailed corpus design. DB-IPIC, based on a heavily annotated sub-corpus of the LABLITA corpus and comparable Spanish and Brazilian Portuguese corpora, allows the user to retrieve from corpora how information is structured in spontaneous speech, observing how information structure can vary cross-linguistically. C-Or-DiAL proposes to teachers and learners of Spanish L2 a dedicated resource for integrating speech into the learning activities.

Keywords: speech corpora, prosody, information structure, second language acquisition.

1. *Introduction*

In this paper, we present the main multilingual speech resources developed by the LABLITA lab of the DILEF Department of the University of Florence that are presently available on the web:

a. the LABLITA reference corpus of spoken Italian
b. the IPIC database of information structure in spoken Romance corpora (Italian, Brazilian Portuguese, and Spanish)
c. the C-Or-DiAL collection for teaching Spoken Spanish L2.

Paragraphs 2., 3., 4., respectively present the three resources. Despite their different focus, these corpora share a common perspective on the study of spontaneous speech. They refer to the Language into Act

Theory (L-AcT) which can be summarized as follows. The speech activity finds its origins in a mental/affective representation which is the speaker's reaction to an external input (Fagioli 1971[1]). The mental image triggers a linguistic action schema directed to the addressee, according to an embodiment process (Arbib 2012), and is conventionally codified in a pragmatic activity: *illocutionary act*, according to Speech act theory (Austin 1962). Crucially, *prosody* constitutes the interface between the pragmatic activity and the linguistic content (*locutionary act*). This frame results in a pragmatic approach centered on the speaker's activity and focusing on illocutionary force and information structure, both performed through prosodic means (Cresti 2020; Cresti & Moneglia 2018). From this premises follow a set of requirements on how corpora should be compiled and annotated for grounding corpus based linguistic research and applications.

2. *The LABLITA corpus*

2.1 General Presentation

The LABLITA corpus, which is now available online in its first public release, gathers in a single resource, published under the ORFEO platform,[1] a collection of three Italian speech corpora recorded in Tuscany between 1965 and the present days. These resources have already been partially delivered in various occasions, starting from the Corpus of Spoken Italian published by the Accademia della Crusca (Cresti 2000) (approx. 100.000 words). Additional files collected for different purposes by researchers at LABLITA has been joined to this early resource and constitutes the GRIT sub-corpus of the LABLITA collection. Then, the Italian section of the C-ORAL-ROM corpus[2] (approx. 300.000 words each; Cresti & Moneglia 2005) has been added. The above sub-corpora are integrated by the Stammerjohann corpus (Stammerjohan 1971), whose original acoustic source was made avail-

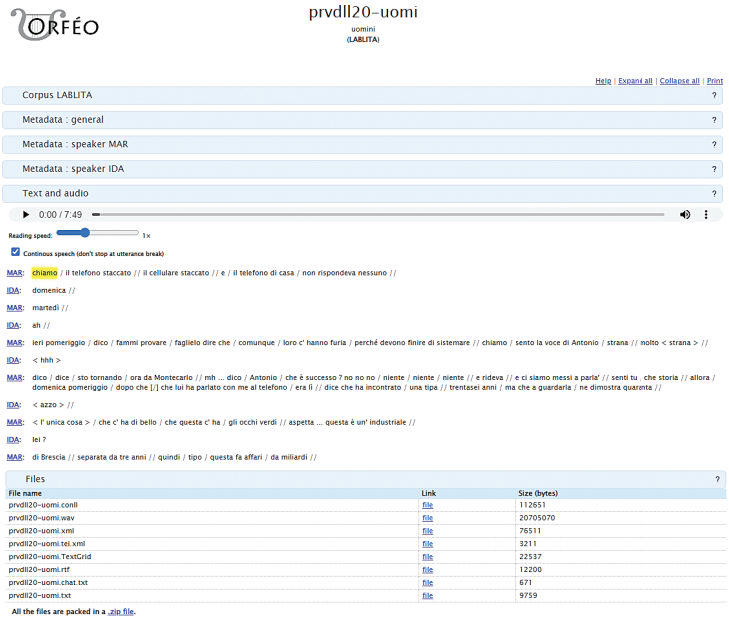
[1] We wish to thank Jeanne-Marie Debaisieux for making the Orfeo platform available to us.

[2] C-ORAL-ROM was achieved in a Project co-ordinated by LABLITA within in FP/5 of the EU. A choice of this section has been also published as a resource to be compared with the written variety for teaching Corpus Linguistic techniques (Cresti & Panunzi 2013).

able to us by the author within the FIRB project *Archivi dell'italiano orale in diacronia*. The Stammerjohan corpus, collected in Florence in 1965, is the first corpus of spoken Italian. It has been derived from 40 hours of original recordings, sampled according to the LABLITA corpus design strategy (see below). The corpus was already distributed online together with a sampling of Italian spoken in Florence during the 90' for comparative analysis of language change (approx. 100.000 words each) (Moneglia & Scarano 2008; Moneglia & Panunzi 2022).

Nowadays the LABLITA corpus comprehends approx. 700.000 token, is open and continuously updated. The acoustic signal is transcribed in CHAT/LABLITA format, which enriches the traditional CHAT transcripts with the annotation of *terminal* and *non-terminal prosodic breaks* (Cresti & Moneglia 1997). Transcripts are aligned to the acoustic source by *terminated sequences* (see below). The corpora are delivered to the users online and for downloading as a collection of files comprising Metadata, Transcription and PoS tagging in TXT fies, aliment in xml files (Winpitch and Praat formats). Figure 1 shows how sessions are presented.

Figure 1 - *Screenshot of a session of the LABLITA Corpus in Orfeo*



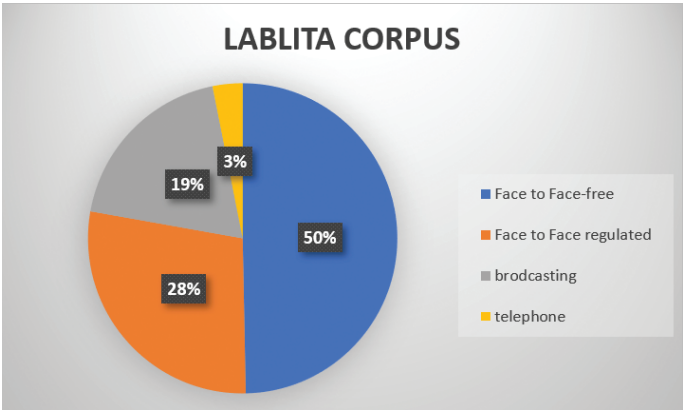
2.2 Sampling criteria and representativeness

The overall purpose of the Corpus is to provide an adequate data set for the study of "spontaneous speech", documenting its main variations. We consider "spontaneous speech" any speech production conceived and performed within a direct interactive relationship between speakers, in which language conception and language performance occur simultaneously.

The LABLITA corpus collects 422 samples of continuous speech (between 1500 and 3000 words). The corpus is proposed as a reference corpus of spoken Italian. To this end it is intended to provide a significant representation of the linguistic choices (at the lexical, syntactic, prosodic, and pragmatic levels) that characterize spoken Italian. Its specific goal is to be representative of the diaphasic variation, which is reflected in the corpus design.[3]

The variation of the corpus is limited from the diatopic point of view to speakers located in Florence and Tuscany. Diachronic variation is documented since 1965. Diastratic variation is not balanced but includes more than 1000 different speakers whose main metadata are delivered (education, sex, age, profession, and origin). Diamesic variation includes, in addition to face-to-face interactions, also a telephone sampling and a collection of broadcasting.[4] Figure 2 illustrates that more then 2/3 of the corpus document the direct interactive language usage (face to face).

Figure 2 - *Distribution of samples by Channel in the LABLITA corpus*



[3] Gestural aspects are also studied at LABLITA on the basis of new collections of audio-video recordings (Cantalini & Moneglia 2020; Cantalini, in this volume).
[4] Provided by Teche-RAI for the C-ORAL-ROM corpus.

Face-to-face speech constitutes the most relevant part of the collection, and is sampled on three dependent levels, which constraint the intersubjective interaction governing the speech activities:

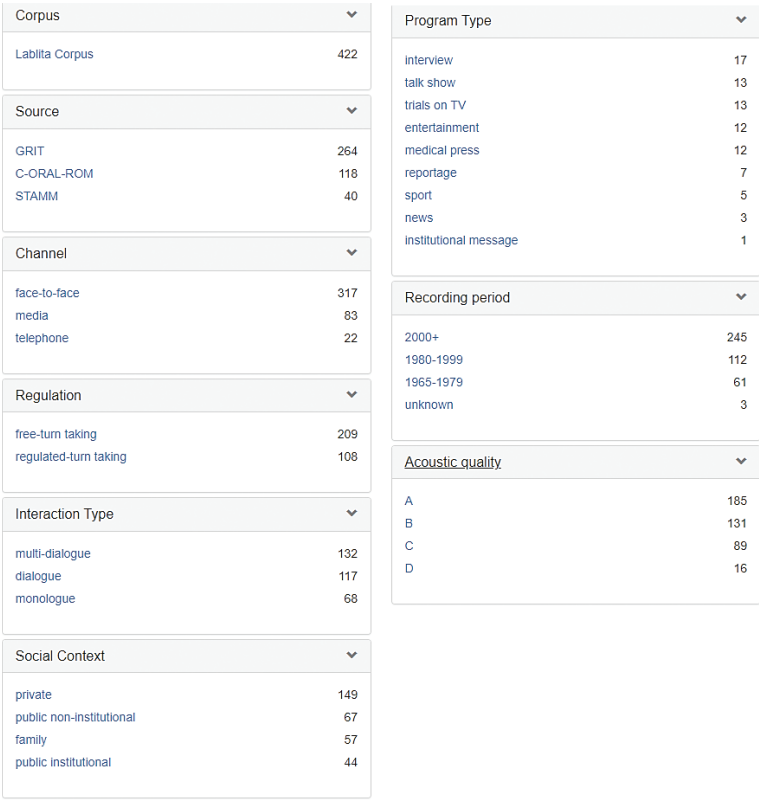
1. the top level distinguishes contexts in which the turn-taking is free or on the contrary, undergoes explicit or implicit rules. This is a way of selecting samples characterized by supervised and formal speech, although no judgment regarding the linguistic style has been given. Free turn-taking contexts (informal) record 2/3 of samples, so ensuring representativeness to the basic spontaneous speech usage which is informal and unsupervised.
2. the second level regards the relationships between speakers that are determined by the social context in which the interaction takes place. The corpus distinguishes four contexts which can in principle influence the affective relation among speakers, and, therefore, their linguistic behaviour: a) family, b) private life, c) public life, d) public life in institutional contexts. This social variation in the corpus design is intended also to give the probability of occurrence to many different possible activities of everyday life in which the linguistic interaction takes place, varying for task, relevance, and goals. The overall strategy is to privilege, from a quantitative point of view, sampling in family and private life.
3. For each of the above social context three types of linguistic interactions are documented: a) monologue;[5] b) dialogue between two speakers; c) multi-dialogue.

   The broadcasting contains a variation of formats, covering a large typology of the programs which occupy the public space on the TV, which despite the emergence of the new media, still represents a context where speech has a high impact in contemporary society.

Figure 3 illustrates the fields of the Corpus design as they are presented in the Orfeo platform. Corpus design fields can be the object of selection by the user.

[5] Many sessions that have been classified as "monologue", record the participation of more than one speaker, but in this case, there is always a dominant actor occupying the turn-taking with a long stretch of speech.

Figure 3 - *Number of samples of the LABLITA Corpus within the corpus design structure*



The user can access the corpus and make searches restricting to the type of context of his interest, compare the language usage of a given context with others, or according to the type of research, can select only files of a given period or acoustic quality.[6]

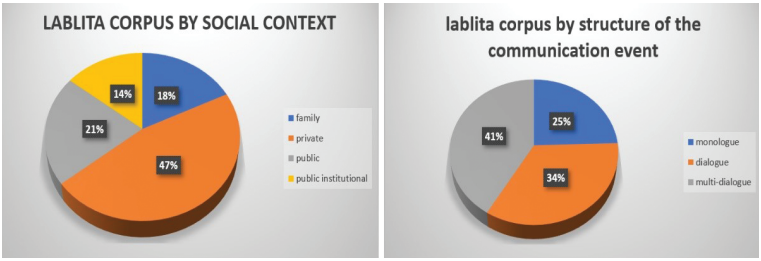
The representativeness of the corpus follows from the intersection of the above parameters. Corpus sampling gives a reasonable proportion to each field in the variation according to the probability of occurrence in the everyday life of the population (Maruyama, in this volume).

[6] Although speaker metadata are available, they cannot be used for selecting speakers of a certain type in the present release.

As regards the social domain of the interaction, half of the face-to-face sub-corpus records events occurring in the private life and about 20% in family interactions, so giving peer relationships larger space with respect to the public and institutional contexts, which occur less frequently and in which the language activity might be more controlled.

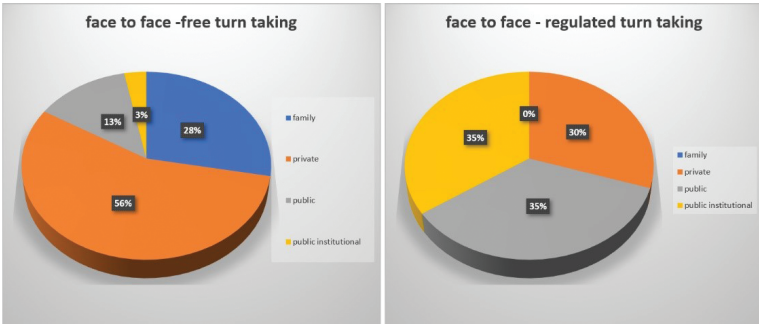
The parameter "structure of the communicative event" allows recording of dialogues between two speakers and multi-dialogues where many speakers interact within a given situation. These are sampled in a similar percentage and cover most of the corpus, being by far the most frequent contexts of usage. However, the corpus also gives a large space (1/4 of the corpus) to monologic performances, so also documenting the linguistic structures that can have some probability of occurrence only in narratives, conferences, explanations, political speech, preaching, etc.

Figure 4 - Social *and* Structural *variations of speech events in the LABLITA corpus*



Finally, when considered face to the social domain, the distinction between contexts in which the spoken interaction foresees free or regulated turn taking shows interesting feature of the Italian socio-linguistic situation. In family life the turn taking is always free and regulation does not occur, while in public institutional context regulation is almost a requirement.

Figure 5 - *The intersection between* Regulation of turn taking *and* social context *in the LABLITA corpus*



The representativeness of the universe of spontaneous speech can be ensured when in principle all event types have at least some probabilities of occurrence, and for this reason, each parameter in the corpus design should be represented by speech events of many different types (Cresti & Panunzi 2013; Maruyama, in this volume). This objective is reached as a function of an implementation strategy that on one side tries to fulfil the various parameters of the corpus design in adequate proportion and on the other to avoid over-representation of a single event type, ensuring variation of task, goal, and genre in the samples gathered within every single field. The following is a picture of the event type variation which can be found in the LABLITA corpus.

When filling the *public-institutional* parameter we sampled monologues, dialogues, and multi dialogues in various contexts of the public life: in administrative and political contexts, in religious life, in cultural life, in economic life. For instance, sampling regards meetings in public companies, interventions at the district council, sermons during a Sunday Mass, harangue, and interrogations by the Public Prosecutor in the trial, university lectures, papers in a conference, professional explanations, narratives by children or teachers and oral test at school, rallies or political meetings during the election campaign, and many other contexts.

When considering *family life*, the corpus records typical monologues or dialogues regarding whatever topic occurring in this context, for instance: memories of old family members, life stories, conversations at dinner, interactions while preparing food, chat between relatives, planning of feast events, hated discussions, problems with

the neighbour, reproach to children, play sessions with children, storytelling to children, plans for vacations or future activities and explanations on how to manage the house affairs, driving school to young relatives, chat in the car, discussions regarding shows, politics or while browsing a photo album, criticisms to teen-agers, etc.

When looking to the *private life* the corpus also comprises a large variety of context and topics: honeymoon stories, travel stories and adventures said to friends, interviews to many professionals on their work (doctor, ceramist, knitter, projectionist, professors, actors, nurse, railway worker, soldiers, partisans, politicians, film directors, etc), psychological interview, professional explanations in various trades, chat at lunch with colleagues, chat among bartenders, instructions to domestic workers, sail of financial products by promoters, professional interactions between mechanics, plumbers, masons, electricians, chefs etc., private lessons, work discussions with colleagues in the office, dialogues among musicians of a band, chat with the beautician while doing a depilation, with the hairdresser while she is styling, discussions among friends on various topics, dinners with friend, chat among teenagers, plans for participating to a competition, chat while driving, in the train, in the park.

There is no internal balance among context types. The collection strategy is to get random samples according to the recording opportunities.

2.3 Corpus annotation

The overall size of the corpus is conspicuous but not enormous (about 700,000 words), its main added value is that transcripts are aligned to the acoustic source by pragmatic units (utterance). The utterance is the linguistic counterpart of a speech act (Austin 1962); i.e. the minimal linguistic entity that can be pragmatically interpreted (Biber *et al.* 1999; Cresti 2000) and the utterance boundaries in the speech flow identify the domain of relevance of linguistic relations (*reference units* according to Izre'el *et al.* 2020). For this reason, the relevance of the corpus for linguistic studies can be evaluated not only in terms of the number or words but also considering the number of pragmatic units achieved by speakers in their language performance (105,000 reference units). Within this set, 86,000 reference units belongs to face-to-face interactions. According to Language into Act Theory (L-AcT,

Cresti 2000), speech comprehends two types of reference units: *utterance*, as the counterpart of a simple speech act, representing approximately 90% of entries, and *stanza*, the linguistic counterpart of a flow of thought (Chafe 1994), that can be composed of many short utterances linked by prosody the one to the other. Stanzas are most frequent in monologues and when the spoken interaction is formal (Saccone 2020).

The segmentation of speech into reference units is not a trivial matter (Raso et al. 2020). The LABLITA corpus adopts the L-AcT framework which stresses the strict correlation between pragmatic activities and prosodic cues. More specifically L-AcT provides an operative method for speech segmentation based on the perception of prosodic cues ('t Hart *et al.* 1990). The continuous flow of speech is primarily segmented by perceptually relevant prosodic interruptions to which competent speakers of a language assign *terminal value* (Izre'el *et al.* 2020). The identification of the terminated sequences (TS) allows the parsing of speech into autonomous speech activities which can be a reasonable object of linguistic analysis.[7] Terminated sequences can be in turn segmented into prosodic units (PU) by prosodic breaks, which are perceived with a *non-terminal* value. PUs corresponds in the L-AcT framework to *information units* (IUs) (Cresti 2000; Cresti & Moneglia 2018; Moneglia & Raso 2014).

For instance, the following is the word-by-word transcription of the first dialogic turn of the conversation among girls presented in Figure 1:

*MAR: chiamo il telefono staccato il cellulare staccato e il telefono di casa non rispondeva nessuno

Considering the possible syntactic relations among words in this sequence, the structure turns out highly undetermined. Some of the possible interpretations are suggested below through punctuation and capital letters: it may be a unique structure where constituents stand in asyndetic coordination relation (a), it can be parsed in three utterances (b), or in five utterances, as in (c):

[7] The agreement regarding perception of terminal breaks has been evaluated in C-ORAL-ROM from the L-AcT perspective (Danieli et al 2004) and also independently of this theoretical background (Panunzi et al 2020).

(a) Chiamo, il telefono staccato, il cellulare staccato e il telefono di casa non rispondeva nessuno.
(b) Chiamo. Il telefono staccato. Il cellulare staccato e il telefono di casa non rispondeva nessuno.
(c) Chiamo il telefono: staccato! E il cellulare: staccato! E il telefono di casa non rispondeva nessuno.

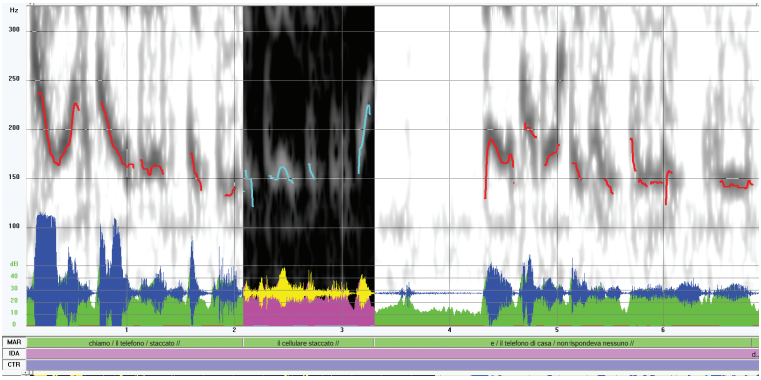
However, when considering prosody, the structure is not underdetermined, since the perception of terminal prosodic breaks ("//") defines the boundaries of the sequence, and restricts the possible interpretations of three specific speech acts corresponding to three TSs:

*MAR: chiamo / il telefono / staccato // il cellulare staccato // e il telefono di casa / non rispondeva nessuno //

The annotation files of the LABLITA corpus are delivered in Praat and Winpitch files where the transcript is shown aligned to the acoustic source. The user can verify that each stretch of speech marked with a terminal break is an independent utterance and can study the linguistic relations established within each of them and their prosodic correlations. For instance, the second utterance of the turn is a *nominal utterance* performed in one PU corresponding to one IU, while the third utterance is an *anacoluthon* performed in two PUs conveying a *Topic* / *Comment* information pattern.

In Figure 6 the text / sound alignment is displayed in a screenshot of the Winpitch software, which distributes the transcription in independent layers for each speaker. The program displays F_0 track, spectrogram, and intensity of each utterance, ensuring full and direct exploitation of corpus data for phonetic and linguistic analysis.

Figure 6 - *Text sound alignment in Winpitch*



3. *DB-IPIC: a Database for Information Patterning Interlinguistic Comparison*

DB-IPIC is an online resource, that allows to browse and perform complex searches on spoken corpora annotated following the L-AcT principles. It is designed to host corpora with prosodic segmentation and information structure and allows to study linear relations between PUs and IUs within each reference unit (TS) type marked by a terminal prosodic break (*utterances* and *stanzas*). In addition to it, DB-IPIC supports token-level annotation of parts of speech and lemmas.

The resource is composed of an XML database, and a web interface for corpus querying. The database contains speech transcriptions, each one enriched with metadata and all the levels of annotation; data related to each session is embedded in a single XML file, according to a specific XML model, exemplified by the excerpt in Figure 7:

Figure 7 - *XML model for linguistic annotation in DB-IPIC*

```
<turn speak="EDO">
        <term_seq num="1" type="utt">
          <tone_unit inf="COM">
                    <word lemma="guardare" pos="VER:fin">guarda</word>
                    <word lemma="chi" pos="WH">chi</word>
                    <word lemma="ci" pos="ADV">c'</word>
                    <word lemma="essere" pos="VER:fin">è</word>
                    <break type="nonterminal">/</break>
          </tone_unit>
          <tone_unit inf="ALL">
                    <word lemma="nonna" pos="NOUN">nonna</word>
                    <break type="terminal">//</break>
          </tone_unit>
        </term_seq>
</turn>
```

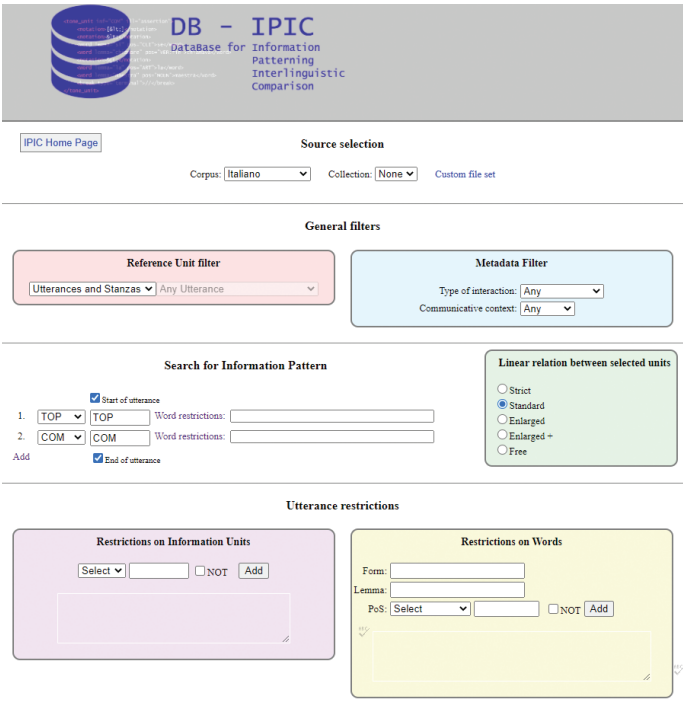
DB-IPIC online interface (Figure 8) is a PHP web application that provides a user-friendly way to extract information from the database. With this tool it's possible to query a corpus at different levels, according to the logical structure of the data set. DB-IPIC can operate on five levels:

1. data source: it is possible to query a whole corpus or to specify a subset of sessions; different corpora can be managed in DB-IPIC
2. metadata: sessions can be filtered by their properties, specifying the communicative context (familiar or public) and the interaction type (monologue, dialogue, or conversation)
3. informational patterns: the user can select the TSs by specifying their IU pattern
4. information units: it's possible to search TSs containing or not containing specific IUs independently of their informational pattern
5. words: users can refine their search by including or excluding words with a specific form, PoS, or lemma.

DB-IPIC currently contains an Italian corpus of 74 texts (124,735 total words) chosen from the Informal section of Italian C-ORAL-ROM (Cresti & Moneglia 2005), and 3 small comparable corpora (mini-corpora) of Italian (IT), Spanish (ES) and Brazilian Portuguese (BP). These last two are derived from C-Or-DiAL (see

§ 4) and from C-ORAL-BRASIL (Raso & Mello 2010), while the small Italian is a subset of the main Italian corpus. These mini-corpora have a similar size (30,000 to 40,000 words) and the same design; all of them are manually annotated with prosodic and information structure annotation according to the L-AcT tagset and definitions of IU types (Moneglia & Raso 2014)[8]. This allows the use of DB-IPIC for studying how information is structured by prosody in each language corpus and to perform cross-linguistic comparisons between spoken Italian, Spanish and Brazilian Portuguese. For instance, Figure 8 is a screenshot of the IPIC query interface presenting the search for *Topic* / *Comment* utterances in the Italian sub-corpus.

Figure 8 - *DB-IPIC search interface*



Query results page (Figure 9) displays the list of utterances that match the search criteria. In the corpus 687 TS (on 5117 utterances) show

[8] PoS-tagging and lemmatization are derived through automatic tools.

a *Topic* / *Comment* structure. Each line contains utterances transcription, prosodic and information structure annotation, PoS-tagging and lemmatization. Moreover, a direct access to audio source is available and downloadable in MP3 format. Results can be also exported in CSV format to be analysed through a spreadsheet software.

Figure 9 - *Screenshot of data returned by DB-IPIC*

Beside specific queries on information patterning and morpho-lexical fillings, DB-IPIC can be used for deriving comparative distributional data with respect to the represented languages. One of the main comparative data is reported in Table 1., which contains the percentage of the different type of reference units of speech according to L-AcT theory (Utterances Vs Stanzas) in the three mini-corpora. Data show that the percentage of Utterances and Stanzas is almost constant in the three languages, given that stanzas are about 9% of the total TSs.

Table 1 - *Utterance vs Stanzas in IT, ES and BP mini-corpora*

	IT	%	ES	%	BP	%
Utterances	5117	90,36%	5898	91,51%	5046	91,55%
Stanzas	546	9,64%	547	8,49%	466	8,45%
Total units	5663		6445		5512	

If we consider only the Utterances, we can derive data regarding the complexity of the information patterns. In Table 2, for instance, we distinguished Simple Utterances (see examples 1a-1c), in which there is only the Comment unit, and Complex Utterances (see examples 2a-2c), in which other information units occur.

(1a) *ART: le quattro componenti son queste //COM (ifamdl04, 47)
the four components are these //

(1b) *PAC: no sé muy bien que hay //COM (efamdl04, 43)
I don't know very well what is there //

(1c) *FLA: seu dinheiro tá caindo hhh //COM (bfamdl01, 510)
your money is falling [out of your pocket]' //

(2a) *MAR: più di così /TOP 'un arriva /COM caro //ALL (ifamdl19, 339)
more than that / it doesn't arrive / dear //

(2b) *PIL: no /PHA es que ese /TOP era una estrategia de la defensa //COM
no / it is that this / was a defence strategy

(2c) *BEL: uhn /TMT talvez na parte maior /COM não //PHA (bfamdl02, 194)
hm / maybe in the bigger part / no //

Table 2. shows that, with respect to this measure, Italian and Spanish present similar values, while Brazilian Portuguese records a higher percentage of Simple Utterances. This can be interpreted as a different overall strategy in structuring the spoken production, as already observed in specific comparative studies that exploited the DB-IPIC data (Panunzi & Mittmann 2014).

Table 2 - *Simple vs Complex Utterance in IT, ES and BP mini-corpora*

	IT	%	ES	%	BP	%
Simple Utterances	3491	68,22%	4073	69,06%	3840	76,10%
Complex Utterances	1394	27,24%	1627	27,59%	1089	21,58%
Uncomplete Utterances[9]	232	4,53%	198	3,36%	117	2,32%
Total Utterances	5117		5898		5046	

4. *C-Or-DiAL*

C-Or-DiAL is a spoken corpus of Spanish created for teaching purposes and published online (Nicolás Martínez 2012). The aim of

[9] The class of Uncomplete Utterance contains mostly interrupted sequences in which there is not any Comment unit.

C-Or-DiAL is to provide learners of Spanish L2 with a didactic resource freely accessible on the web that allows the study of the language through an explicit reflection on its spoken variety.

The corpus is integrated with linguistic annotation and metadata, and is delivered as a didactic material. Teachers can select texts by difficulty level, according to learning needs of their classes. The corpus includes 240 short audio recordings (approximately 3 minutes each) with the corresponding transcripts for 120,000 words and is presented in Spanish through a user-friendly web interface.

Figure 10 shows in a screenshot the file index of C-Or-DiAL. From this page, the user can select the sessions of his interest and access the audio and text files. Each field contains higher-level metadata orienting the user to select the sessions to be used in a class.

Figure 10 - *Index of the C-Or-DiAL corpus*

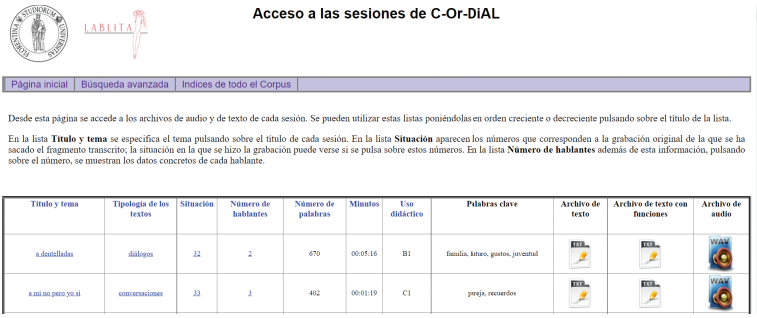


Figure 11 - *Higher-levels metadata of C-Or-DiAL files*

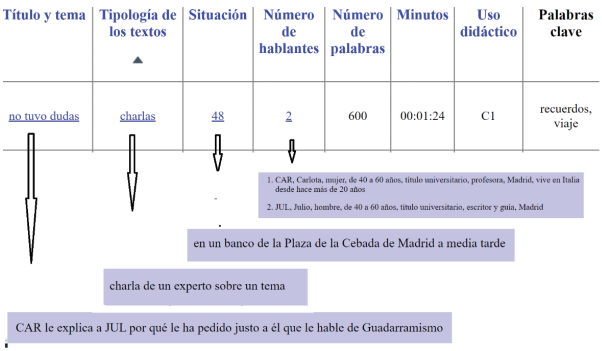
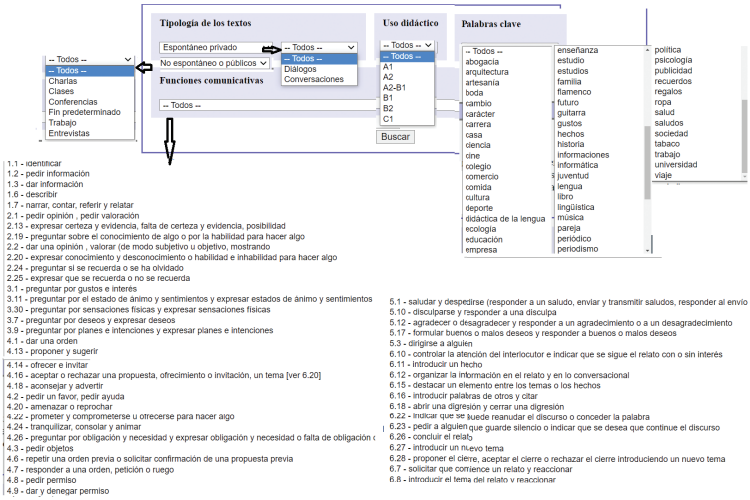


Figure 11 illustrates in detail the kind of information available for each session. Each recording has a *Title*, a *Genre* of text (chat, conference, dialogue, multi-dialogue, interview, talks at work), a *Situation* where the communicative event occurs, and *Speakers* metadata. The last four fields show the length and keywords of the recording, and the suggested level according to the European frame.

An advanced search interface is also available to teachers and learners to make decisions on what might be the better oral source for their activities. To this end, the user can select any of the fields present on the page. The system will return all sessions that have the required characteristics. Figure 12 illustrates the parameters governing the possible choices. For instance, the choice can be oriented to *informal private* sessions, which allow selecting among *dialogues* or *multi-dialogues,* alternatively to *formal public* speech, where a good lot of professionals have been recorded. In this case, the user can select sessions among: Speeches, Classes, Conferences, Interviews, etc. The user can also explore the database by restricting his search by level of difficulty of the text or selecting a topic through the keywords.

Moreover, the main annotation of the corpus, which is specifically significant for learning activities, regards the communicative functions which are instantiated in the spoken interactions. These functions (listed in Figure 12) have been identified and annotated in the transcripts according to the repertory published within the *Plan Curricular Cervantes*, which is the main background framework for Spanish L2.

Figure 12 - *Advanced search interface of C-Or-DiAL*



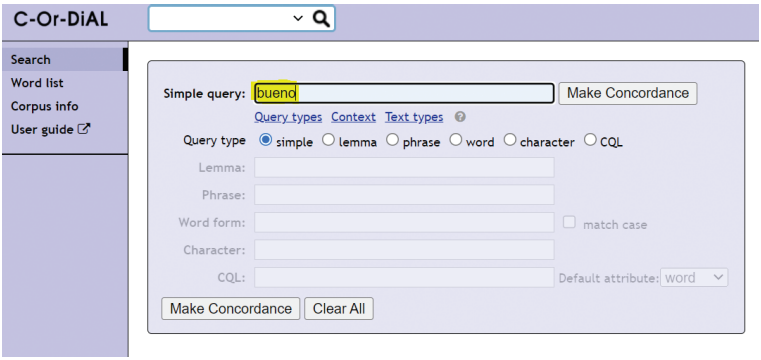
Transcripts are in the CHAT/LABLITA format and bear the systematic annotation of the terminal and non-terminal prosodic breaks marking the utterance boundaries. Transcripts are headed by the full set of session metadata. Files are in txt format; they can be downloaded also in a richer version, comprising the annotation of communicative functions codes, as in the stretch of dialogue in Figure 13:

Figure 13 - *Transcripts in C-Or-DiAL*

```
*BRI: ¿ 1.2 pero ponemos Madrid Cuneo o Madrid el destino que [/] qué queréis ver el xxx 1.2 ?
*PED: no tú [/] tú / 3.9 para que nos vuelva [/] para que nos vuelva a dar esto mismo y ahora vamos por el mapa y lo vamos viendo ya 3.9 ¿ entiendes ? 4.1
pones Madrid / dirección Purchena / es lo mismo / para que sea exactamente / para que nos salga lo mismo que esto / Purchena 19 Madrid 4.1 //
*CAR: 2.2 es que Arles / < que es más bonito >2.2 /
*PED: 4.1 < y abajo > pones Cuneo / Italia 4.1 // < xxx >
*CAR:/ 1.3 < lo más bonito > de esta zona Chavela es la [//] es el paisaje de la Camargue 1.3 //
*PED: 1.3 no Cuneo es abajo / Cuneo [/] Cuneo es abajo / eso es la &esta / la dirección nos va a llevar a Cuneo centro pero Cuneo hhh 1.3 ...
*CAR: 4.13 Cuneo 4.13 //
```

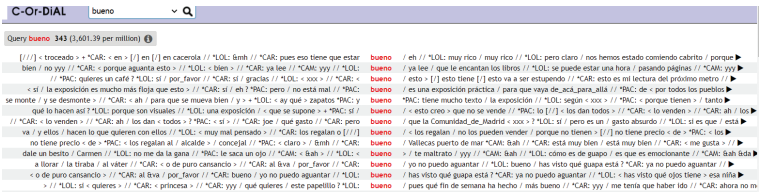
Although the alignment files of the corpus were not delivered, the acoustic source of each utterance is available to the user through a concordances search, exploiting standard corpus linguistics methods. The corpus has been implemented within the NoSketchEngine interface, where keywords can be searched through a standard form (Figure 14). Occurrences are returned (Figure 15) and can be listened in the context of each utterance in which they occur. The user can search for tokens, lemmas, or phrases in all corpus or in sub-corpora.

Figure 14 - *Query interface in C-Or-DiAL*



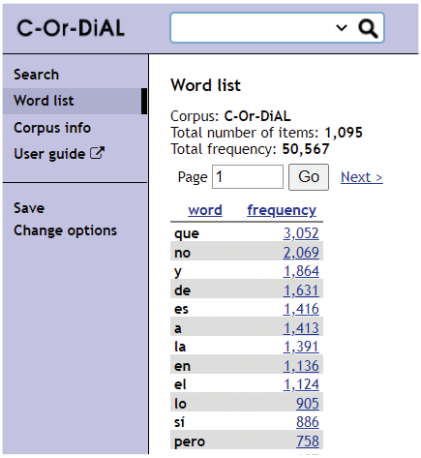
He can appreciate for instance what is the acoustic and distributional difference between one expression used as a Discourse Marker, (*bueno*) in the second row of the screenshot in Figure 15, and the same expression used as an Adjective (in the first row).

Figure 15 - *Keyword in context with the audio file of the utterance in which they occur*



The user can also generate frequency lists by word, lemmas, and PoS (Figure 16).

Figure 16 - *Screenshot of the Frequency lists*



C-Or-DiAL can contribute to the development of teaching materials and practices in several ways: it may implement classical teaching methods with examples derived from real speech but can also promote the development of new teaching method, where speech is at the core of learning activities (Nicolás Martínez et al. 2016). For instance, after careful listening to the audio and the control of comprehension with the help of transcription we can move on to listening and speaking exercises, and only afterward do we pass to the study of vocabulary and grammar. The practice of transcribing small audio fragments from C-Or-DiAL is fruitful at the A1-A2 levels. Transcription with tags can be done at intermediate levels, to learn how to analyze speech, focusing on the relevance of prosodic patterning, and dialogic strategies. The specific lexicon used in spontaneous speech can be the object of focused study with students of medium and high level.

References

Arbib, Michael. 2012. *How the brain got language*. Oxford: Oxford University Press.

Austin, John. 1962. *How to do things with words*. Oxford, Oxford University Press.

Biber, Douglas & Johansson, Stig & Leech, Goffrey & Conrad, Susan & Finegan, Edward. 1999. *The Longman Grammar of Spoken and Written English*. London and New York: Longman.

Cantalini, Giorgina. 2022. Corpus multimodale annotato per lo studio della gestualità co-verbale nel «parlato-parlato» e nel «parlato-recitato». (in this volume)

Cantalini, Giorgina & Massimo Moneglia. 2020. The Annotation of Gesture and Gesture / Prosody Syncronization in Multimodal Speech Corpora. *JOSS Journal of Speech Science*, 9. 1-24.

Chafe, Wallace. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: The University of Chicago Press.

Cresti, Emanuala. 2000. *Corpus di Italiano Parlato*. Firenze: Accademia della Crusca.

Cresti, Emanuela. 2020. The pragmatic analysis of speech and its illocutionary classification according to the Language into Act Theory. In Izre´el, Shlomo & Mello, Heliana & Panunzi, Alessandro & Raso, Tommaso (eds), *In search of basic units of spoken language: A corpus-driven approach*, 181-219. Amsterdam: John Benjamins.

Cresti, Emanuela & Moneglia, Massimo. 1997. L'intonazione e i criteri di trascrizione del parlato adulto e infantile. In Bortolini, Umberta & Pizzuto, Elena (a cura di), *Il progetto CHILDES: strumenti per l'analisi del linguaggio parlato*, vol. II, 57-90. Pisa: Edizioni del Cerro.

Cresti, Emanuela & Moneglia, Massimo (eds). 2005. *C-ORAL-ROM. Integrated reference corpora for spoken romance languages*. Amsterdam: John Benjamins.

Cresti, Emanuela & Moneglia, Massimo. 2018. The illocutionary basis of Information Structure. Language into Act Theory. In Adamou, Evangelia & Haude, Katharina & Vanhove, Martine (eds.), *Information structure in lesser-described languages: Studies in prosody and syntax*, 359-401. Amsterdam: Benjamins.

Cresti, Emanuela & Panunzi, Alessandro. 2013. *Introduzione ai corpora dell'italiano*. Bologna: Il Mulino.

Danieli, Morena & Garrido, Juan Maria & Moneglia, Massimo & Panizza, Andrea & Quazza, Sivia & Swerts, Marc. 2004. Evaluation of Consensus on the Annotation of Prosodic Breaks in the Romance Corpus of Spontaneous Speech "C-ORAL-ROM". In Lino, Maria Teresa & Xavier, Maria Francisca & Ferreira, Fátima & Costa, Rute & Silva, Raquel (eds.), *Proceedings of the 4th LREC Conference*, vol. 4, 1513-1516, Paris.

Fagioli, Massimo. 1972[1]. *Istinto di morte e conoscenza*. Roma: L'Asino d'oro. (Trad. Eng. *Instinct of death and knowledge*, L'Asino d'oro 2021).

't Hart, Johan & Collier, René & Cohen, Antonie. 1990. *A Perceptual Study on Intonation. An Experimental Approach to Speech Melody*. Cambridge: Cambridge University Press

Izre'el, Shlomo & Mello, Heliana & Panunzi, Alessandro & Raso, Tommaso (eds.). 2020. *In Search of Basic Units of Spoken Language: A Corpus-Driven Approach*. Amsterdam: Benjamins.

MacWhinney, Brian.2000. *The CHILDES Project: Tools for Analyzing Talk. 3rd Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Maruyama, Takehiko. 2022. Designs and Analyses of Japanese Speech Corpora. (in this volume)

Moneglia, Massimo. 2005. The C-Oral-Rom resource. In Cresti, Emanuela & Moneglia, Massimo (eds), *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*, 1-70. Amsterdam: Benjamins.

Moneglia, Massimo & Raso, Tommaso. 2014. Notes on Language into Act Theory (L-AcT). In Raso, Tommaso & Mello, Heliana (eds), *Spoken corpora and linguistic studies*, 468- 495. Amsterdam: Benjamins.

Moneglia, Massimo & Scarano Antonietta. 2008. Il Corpus Stammerjohann. Il primo corpus di italiano parlato, in rete nella base dati di LABLITA. In Pettorino, Massimo (ed.), *Atti del Congresso Internazionale "La comunicazione Parlata", Napoli 2006*, 1699-1739. Napoli: Liguori.

Moneglia, Massimo & Panunzi, Alessando. 2022. Micro-Diachronic Corpora for Measuring the Lexical Change of Spontaneous Speech in Florence Compared to Standard Italian. *Langages*, 226. 41-54.

Nicolás Martínez, Carlota. 2012. *C-Or-DiAL Corpus Oral Didáctico Anotado Lingüísticamente*. Madrid: Liceus.

Nicolás Martínez, Carlota & Hernández Toribio, María Isabel. 2016. *Del oído al habla*. Barcelona: Octaedro.

Panunzi, Alessandro & Gregori, Lorenzo. 2012. DB-IPIC. An XML database for the representation of information structure in spoken language. In Panunzi, Alessandro & Raso, Tommaso & Mello, Heliana (eds), *Pragmatics and prosody. Illocution, modality, attitude, information patterning and speech annotation*, 121-127. Firenze: Firenze University Press.

Panunzi, Alessandro & Mittmann, Maryualê. 2014. The IPIC resource and a cross-linguistic analysis of information structure in Italian and Brazilian Portuguese. In Raso, Tommaso & Mello, Heliana (eds.), *Spoken Corpora and Linguistic Studies*, 129-151. Amsterdam: Benjamins.

Panunzi, Alessandro & Gregori, Lorenzo & Rocha, Bruno. 2020. Comparing annotations for the prosodic segmentation of spontaneous speech: Focus on reference units. In Izre'el, Shlomo & Mello, Heliana & Panunzi, Alessandro & Raso, Tommaso (eds.), *Search of Basic Units of Spoken Language. A corpus-driven approach*, 403-431. Amsterdam: Benjamins.

Raso, Tommaso & Mello, Heliana. 2012. *C-ORAL-Brazil I, Corpus de referência do português brasileiro falado informal*. Belo Horizonte: EDITORAufmg.

Raso, Tommaso & Teixeira, Bárbara & Barbosa, Plinio. 2020. Modelling Automatic Detection of Prosodic Boundaries for Brazilian Portuguese Spontaneous Speech. *JOSS Journal of Speech Science*, 9. 105-128.

Saccone, Valentina. 2020. La Stanza nella Teoria della Lingua in Atto: Un'analisi sintattica. *CHIMERA: Revista De Corpus De Lenguas Romances Y Estudios Lingüísticos*, *7*. 55–68.

Stammerjohann, Harro. 1970. Strukturen der Rede. Beobachtungen an der Umgangssprache von Florenz. *Studi di Filologia Italiana* XXVIII. 295-397.

C-ORDIAL< http://lablita.it/app/C-Or-DiAL/>

C-ORDIAL <http://lablita.it/C-Or-DiAL/run.cgi/first_form>

IPIC: <http://www.lablita.it/app/dbipic/>

LABLITA Corpus, http://corpus.lablita.it/>

Praat <http://www.fon.hum.uva.nl/praat/>

sketchengiNE <https://www.sketchengine.eu/>

TreeTagger <https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

WinPitch <https://www.winpitch.com/>