Lang Resources & Evaluation (2013) 47:945–971
DOI 10.1007/s10579-012-9213-0

ORIGINAL PAPER

Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan

Juan María Garrido · David Escudero · Lourdes Aguilar · Valentín Cardeñoso · Emma Rodero · Carme de-la-Mota · César González · Carlos Vivaracho · Sílvia Rustullet · Olatz Larrea · Yesika Laplaza · Francisco Vizcaíno · Eva Estebas · Mercedes Cabrera · Antonio Bonafonte

Published online: 12 January 2013
© Springer Science+Business Media Dordrecht 2013

Abstract Literature review on prosody reveals the lack of corpora for prosodic studies in Catalan and Spanish. In this paper, we present a corpus intended to fill this gap. The corpus comprises two distinct data-sets, a news subcorpus and a dialogue subcorpus, the latter containing either conversational or task-oriented speech. More than 25 h were recorded by twenty eight speakers per language. Among these speakers, eight were professional (four radio news broadcasters and four advertising actors). The entire material presented here has been transcribed, aligned with the

J. M. Garrido · S. Rustullet · Y. Laplaza
Computational Linguistics Group (GLiCom), Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona, Spain

J. M. Garrido
e-mail: juanmaria.garrido@upf.edu

D. Escudero (✉) · V. Cardeñoso · C. González · C. Vivaracho
Department of Computer Sciences, Universidad de Valladolid, Valladolid, Spain
e-mail: descuder@infor.uva.es

L. Aguilar · C. de-la-Mota
Department of Spanish Philology, Universitat Autònoma de Barcelona, Barcelona, Spain

E. Rodero · O. Larrea
Department of Communication, Universitat Pompeu Fabra, Barcelona, Spain

F. Vizcaíno · M. Cabrera
Department of Modern Languages, Universidad de las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

E. Estebas
Department of Modern Languages, Universidad Nacional de Educación a Distancia, Madrid, Spain

A. Bonafonte
Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, Barcelona, Spain

acoustic signal and prosodically annotated. Two major objectives have guided the design of this project: (i) to offer a wide coverage of representative real-life communicative situations which allow for the characterization of prosody in these two languages; and (ii) to conduct research studies which enable us to contrast the speakers different speaking styles and discursive practices. All material contained in the corpus is provided under a Creative Commons Attribution 3.0 Unported License.

Keywords Prosodic corpus · Radio news corpus · Dialogue corpus · Spanish corpus · Catalan corpus

1 Introduction

Prosody has been in recent years the object of intense multidisciplinary research. The characterisation of intonation, stress, rhythm, speech rate, together with their specific roles in speech, their relations to other components of the grammar (such as syntactic or information structure), and their communicative uses in specific speech situations, are all the subjects of study of a wide range of disciplines, both theoretical and applied, such as Phonetics, Phonology, Syntax, Pragmatics, Discourse Analysis, Communication Sciences or Speech Technologies, for example. It has also been studied in a wide range of speech materials, from controlled, usually read, recordings (isolated sentences, news) to spontaneous data (monologues, dialogues, emotional speech). This multidisciplinary approach involves researchers with different interests, methods and theoretical assumptions. These methodological approaches can be, for example, both 'bottom-up', from an experimental perspective (for the purposes of, for example, Acoustic Phonetics, Laboratory Phonology or Speech Technology), or 'top-down', following a more functional approach, which departs from the linguistic phenomena and then leads onto the analysis of the actual prosodic realisation (see, for example, the excellent revisions of both approaches in Xu 2001 or Botinis et al. 2001). And they will increasingly involve cross-linguistic, inter-speaker and inter-style analyses in the near future.

In the field of speech technologies, the use of annotated corpora is a necessary precondition for the development of text-to-speech and speech recognition applications (Huang et al. 2001; Taylor 2009): in text-to-speech systems, for example, the recording of several hours of speech material is needed for the development of synthetic speakers, material which must then be processed and annotated with linguistic and phonetic information; in the case of speech recognition, a large amount of speech from many different speakers is also needed in order to perform a training plan of the acoustic models. In Phonetics and Phonology, however, the use of large corpora is quite rare, mainly due to the high cost, in terms of both time and resources, of the task of manual transcription and annotation of the corpora by experts. The situation is even more complex in the case of prosody, since the transcription of prosodic phenomena needs reference systems that are generally in a process of consolidation and involves a long-term phase of training and manual work by the annotators. Speech Technology is providing more and more tools that allow some processes (such as phonetic transcription) to be carried out automatically, but their output is not reliable enough to ignore manual proofing. All these factors explain why there are currently so few speech corpora annotated with prosodic information available to the scientific community, particularly with regard to the study of Spanish or Catalan prosody, and make difficult, at this time, the corpus approach to the study of prosody in all these fields. Such an approach requires a new generation of speech corpora, allowing comparative, cross-linguistic and interdisciplinary analyses. From this point of view, they should contain:

- **a substantial amount of data**, in order to allow researchers to carry out reliable statistical studies. This is important for both theoretical studies and Speech Technology applications;
- **high acoustic quality**, to allow its use with currently existing analysis techniques such as automatic phonetic segmentation or fundamental frequency estimation algorithms;
- **data coming from different speakers**, comparable if possible, with the goal of improving the existing description of the inter-speaker variation in prosodic phenomena;
- **comprehensive enough coverage of prosodic phenomena**, to guarantee its reusability beyond the goals of a specific project;
- **annotation data in a standard format** which facilitates its use with different tools;
- **annotation data at both the phonetic and the phonological level**, which offer the potential users of the corpus the possibility of either using raw data (such as F0 values), independent of theoretical frameworks, or working within the most widespread descriptive frameworks (such as ToBI), or even across models for purposes of comparison;
- **annotation data about the prosodic structure of the utterances**, to allow the study of their phonological nature, their phonetic identification and the linguistic factors which determine the organisation of utterances into prosodic units;
- **a reliable and reviewed annotation**, carried out by more than one annotator and evaluated with the most objective criteria;
- **data from more than one speaking style** in order to meet the research requirements of those interested in more spontaneous or expressive speech;
- **comparable data from more than one language**, thus making it useful for interlinguistic studies or multilingual technological applications.

There are at present some corpora for Spanish and Catalan which include prosodic annotations—the Val.Es.Co corpus for the study of colloquial Spanish (Albelda Marco 2005), the *Corpus Audiovisual Pluriligüe* (Payrató and Fitó 2005), or the C-ORAL-ROM corpus (Cresti and Moneglia 2005), for example- or even corpora specially designed for the description of prosody—among others, the Interactive Atlas of Catalan Intonation (Prieto and Cabré 2010) the *Corpus oral de parla espontània* for Catalan (Font 2006), or the AMPER (Fernández 2005) and MULTEXT (Campione and Veronis 1998) multilingual corpora, which include both Catalan and Spanish. However, they stand short from offering high quality complete products for research on prosody from a multidisciplinary and comparative approach, as they may lack one or various resources, such as speech-text alignment, phonetic transcription, or prosodic unit annotation, usually because they have not been specifically designed for the phonetic study of prosody (Val.es.Co, C-ORAL-ROM). In other cases, although phonetic, time-aligned annotation of prosodic phenomena is provided, it is given only in a theory-dependent transcription method (MoMel INTSINT, in the case of MULTEXT, for example). And none of them contains a substantial amount of data compiled with the aim of allowing researchers to carry out reliable statistical tests. This last issue turns out to be fundamental for both theoretical studies and speech technology applications, as they both require a large amount of quantitative data in order to draw and support their results and conclusions.

Pre-existing corpora developed for languages different to Spanish and Catalan may serve as models for the development of this kind of corpora. One such corpus is the Boston University Radio News Corpus (Ostendorf et al. 1995), a corpus annotated with prosodic information obtained from recordings of several radio broadcasts. For dialogues, the Buckeye Corpus (Pitt et al. 2005; 2007) or the Corpus of Spontaneous Japanese (Maekawa et al. 2000; Maekawa 2003) are good examples of corpora including phonetically transcribed (and annotated) conversational speech. Finally, it is important to mention the Map Task protocol for corpus development (McAllister et al. 1990; Anderson et al. 1991), which is designed to create corpora of conversational dialogues with a certain degree of spontaneity and naturalness while maintaining a relative degree of control over the contents of the interactions, and which is a model for the development of dialogue corpora.

This paper describes the contents and collection procedure of Glissando, a prosodic corpus for Spanish and Catalan which intends to overcome these limitations. The Glissando corpus includes more than 20 h of speech in Spanish and Catalan, recorded under optimal acoustic conditions, orthographically transcribed, phonetically aligned and annotated with prosodic information at both the phonetic and phonological levels. The prosodic information covers both the phonetic/acoustic domain (intensity, duration and F0, codified under different systems such as MoMel or Bézier, among others) and the phonological/functional domain (prosodic phrasing, ToBI labels, prominence). It has been designed considering as remote references the Boston University Radio News Corpus, the Buckeye Corpus and the Map Task corpus. For this reason, Glissando is actually made of two subcorpora: a corpus of read news (hereafter the ‘news subcorpus’) and a corpus of dialogue material which is further subdivided into a subcorpus of informal conversations (the ‘informal dialogues corpus’), and a set of three task-oriented dialogues, covering three different interaction situations (the ‘task dialogues corpus’). This structure, as well as the high number of speakers who recorded the corpus (28 per language, between professional and non-professional), makes the Glissando corpus especially suitable for inter-speaker and inter-style prosodic analyses.

This paper is organized as follows: Sect. 2 presents the design of the corpus (contents and speakers); Sect. 3 describes the recording protocol and the technical means used; Sect. 4 summarizes the contents of the corpus, including the annotation of segmental and suprasegmental information; Sect. 5 presents a preliminary evaluation of the corpus, to show the capabilities of the Glissando corpus for the multidisciplinary study of Prosody; finally, Sect. 6 presents some conclusions.

2 Selection of contents and speakers

This section is devoted to the description of the design procedure of the corpus, and is organized in three subsections: the first one describes the collection, selection and modification of the news material for the news corpus; the second one explains the design of the speaker's interactions for the task and informal dialogues; and the third one presents the speaker selection procedure.

2.1 Selection of news items

At the beginning of the design process, the option of using recordings of real news broadcasts for the news corpus, obtained directly from a radio station, was considered. However, in order to have more control on the acoustic quality and contents of the corpus, and to keep the same recording conditions as for the dialogues corpus, it was finally decided to make 'ad hoc' studio recordings of actual news by professional speakers, simulating a 'radio news announcer' condition, at the university premises. For this reason, the design tasks in the case of this subcorpus were oriented to collect, select and prepare the news texts that the speakers would have to read in the recording sessions.

The final goal was to prepare two different sets of texts for the two subcorpora defined for this corpora: the prosodic subcorpus, which had to be designed considering prosodic criteria; and the phonetic subcorpus, whose main aim was to complement the first subcorpus by providing a full phonetic coverage in the target language. Each subcorpus would allow to obtain about half an hour of speech for every recorded speaker.

Greedy algorithms have been frequently used in a variety of studies in corpus selection, such as van Santen and Buchsbaum (1997) and Nagorski et al. (2002). In the building of the Glissando corpus, the main aim to use these algorithms was to optimize the prosodic and phonetic variability of the final corpus. In languages such as Spanish or Catalan, there are linguistic variables that can be predicted from the text and that can have an influence on the intonation patterns used by the readers. Among these variables, Garrido (1996) and Escudero and Cardeñoso Payo (2007) propose the position and the length of the intonation units, as well as the location of the stressed syllables. Greedy algorithms are useful then to balance the number of times the different prosodic units appear, as it was for the selection of the news subcorpus texts.

The procedure established to obtain the prosodic and phonetic sets involved several steps. First a collection of real news texts was collected as base material for the selection of the final texts. This base corpus (the 'mother' corpus) contained texts from a variety of news in Spanish which have been kindly offered by the *Cadena SER* Radio Station.[1] This mother corpus was translated to Catalan to have an input set completely parallel in both languages. The mother corpus, as well as the

[1] http://www.cadenaser.com.

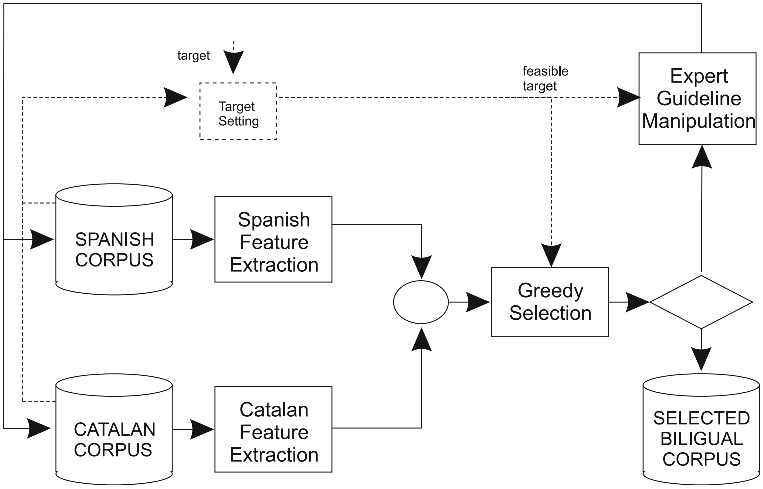


Fig. 1 Scheme of the iterative strategy to combine greedy algorithms with the expert guided manipulation (from Escudero et al. 2010c)

details related to the algorithms, are described in Escudero et al. (2010b). After the automatic analysis of the candidate texts, a selection task which determined the set of texts from the original corpus that would best meet the specified selection criteria (prosodic or phonetic, depending on the subcorpus) was carried out, by using a greedy algorithm. Thus, an iterative process was applied which involved the correction of the results of the greedy algorithm by means of an *Expert Guideline* system and the subsequent re-application of the greedy algorithm so as to get a smaller sample. This was necessary, since it was observed that there are certain types of words, such as proparoxitones (words bearing stress in the ante-penultimate syllable), which occur less frequently in non controlled texts. The original texts were manually modified several times to introduce some elements (words, punctuation marks) that would improve the phonetic or prosodic coverage of the final corpus; after each manual modification, a new automatic selection of candidates was carried out by using the greedy algorithm. With this process, a wider representation of the less frequent types of prosodic units was obtained, even though their number of occurrences never equals that of the most frequent units. The iterations ended when a sufficient coverage of the considered selection factors was achieved. This selection procedure was carried out in parallel for Spanish and Catalan texts, in order to obtain a parallel corpus in both languages. Figure 1 illustrates this process.

By this method, 72 news texts (36 for the prosodic and 36 for the phonetic corpus), were selected per language. Every set contained the same texts in both languages. Considering, as previous tests showed, that the reading of each text would last about 1 min, this would ensure the expected half an hour of read material per subcorpus and speaker. See Escudero et al. (2009) to have a detailed overview of the greedy algorithm and the Expert Guideline system, and Escudero et al. (2010c) for more details on the number of prosodic units in the original corpus and in the selected corpus for each language.

2.2 Design of dialogue scenarios

The dialogue subcorpus consists of two subsets, which are distinguished by the communicative situation in which the dialogues are set: informal dialogues and task-oriented dialogues. Their design procedure, different for each subset, is explained in the following subsections.

2.2.1 Informal dialogues

The subcorpus of informal dialogues was designed as a set of recordings of conversations between people who have some degree of familiarity with each other. The goal was to record a speaking style which corresponds to natural communicative situations and which will then allow the study of a large variety of linguistic phenomena (see Eskénazi 1993; Hirschberg 2000).

In order to obtain a speech corpus with a high degree of naturalness, as defined by the fact that participants are not limited by a formal situation and thus cease to self-monitor their speaking style, we intended to follow the model of the Buckeye Corpus of conversational speech, developed by researchers in Psycholinguistics of the Ohio State University (http://buckeyecorpus.osu.edu/ Pitt et al. 2005; 2007). The Buckeye Corpus contains a series of interviews in which very general topics of conversation are proposed in order to elicit a significant amount of unmonitored speech for each of the 40 speakers. To do this, the sessions were conducted as sociolinguistic interviews, and became essentially monologues. The interviewer's voice is not considered in the corpus.

Differently to the Buckeye corpus, where the role of the interviewer is well defined, in the Glissando corpus each conversation is maintained by a pair of speakers which have stable relations of friendship or work. A simple script was given to each pair of speakers, pointing out the nature and order of several questions to be addressed during the conversation. For the rest, the speakers were free to guide the dialogue along their own interests and intentions. The dialogue was started from the question *Do you remember how you met each other?* and the script included suggestions on how to regain conversation when it was almost to be exhausted: *Have you made any trip together?*, *Do you share any hobbies? Have you ever got angry with your mate?*

The final corpus is composed by 6 conversations and 12 speakers per language. Each conversation lasted about 10–15 min approximately and reached a good amount of naturalness, since the speakers were familiar with each other and they could talk about common interests (work, study, travel).

2.2.2 Task-oriented dialogues

The goal of this subcorpus was to collect a set of recorded interactions between two speakers oriented to a specific goal in the domain of information requests. In each conversation, one of the speakers plays the role of instruction-giver and the other, the role of instruction-follower.

Three types of interactions were designed: (a) travel information, (b) information request for an exchange university course, and (c) information request for a touristic route.

(a) Travel information is the most formal task, since the scenario consisted in a telephone-like conversation between an operator and a customer who wants information on prices and schedules of a specific route.

(b) Information request for an exchange university course takes place between a school's administrative officer that provides information on the possibilities for a course at a foreign university and a student who requests for it. The person who gives information pretends to be a member of the staff of an international office that provides information on stays abroad, while the person seeking the information, who initiates the conversation, assumes the role of a Humanities student who wants to go the following year to Paris to take some elective courses. The information available to the participant with the role of employee is in some cases more extensive than needed to answer the request, it is organized in a different way and, therefore, has to be selected, while in other cases, it is insufficient. Moreover, while the employee has mainly academic information, the student is also interested in issues of everyday life (sports, social life). They are also induced to talk about different academic subjects of similar pronunciation.

(c) The information request for a tourist route is a type of interaction inspired by the Map Task (McAllister et al. 1990; Anderson et al. 1991). Nevertheless, the description of the situation and the type of task are different. In the Map Task corpus, subjects are required to cooperate in order to reproduce on the follower's map the route printed on the giver's map, and the success of the communication is quantified by the degree of coincidence of both routes. In this case, however, one of the speakers plays the role of somebody who is planning a trip to the Greek island of Corfu, and calls a colleague who has lived for 5 years in Greece, in order to request for specific information concerning the route on the island. There is no specific route to reproduce; there is only an initial and a final point of the trip, and some places to visit on the way. This interaction was designed as the least formal of the three, because in this case both speakers are supposed to be work mates with a certain degree of familiarity.

These scenarios have been selected due to their interest for both speech technology dialogue systems (automatic travel information systems, machine learning systems and touristic guides, respectively) and linguistic studies that investigate the effect of the change of communicative conditions on the speech of a given speaker. A relationship of cooperation was established since both speaker and listener were involved in the completion of the task and wanted to achieve it with the maximum communicative success possible. It is an example of intentional speech, similar to other kinds of intentional speech found in natural contexts, but obtained in a laboratory environment. Interestingly, there are different degrees of formality motivated by the content and by the role played by each speaker.

The design of each interaction involved the collection of the information (real in all cases) that the giver should have available to answer the asker, and the definition of the protocols that both participants should have to follow during the interaction. These protocols (described in detail at Escudero et al. 2010a) were provided to the participants prior to recordings so that they would become familiar with them. Figure 2, which depicts the graph facilitated to the instructions-giver to solve the travel task, serves as an example of one of those protocols.

All conversations were planned to simulate a telephone call, because of the special interest of this scenario for spoken dialogue systems design and evaluation. It was decided also that the participants would alternate their role (instruction-giver or instruction-follower) along the three interactions. In order to avoid long silences or unnatural hesitations, both informants read separately both the information needed to solve each task and the protocols before the recording started in order to become familiar with each scenario.

2.3 Speaker selection procedure

The selection of speakers that participated in the recording sessions received special attention in the design process of the Glissando corpus. This process involved both a careful selection among the initial candidates, considering their linguistic and

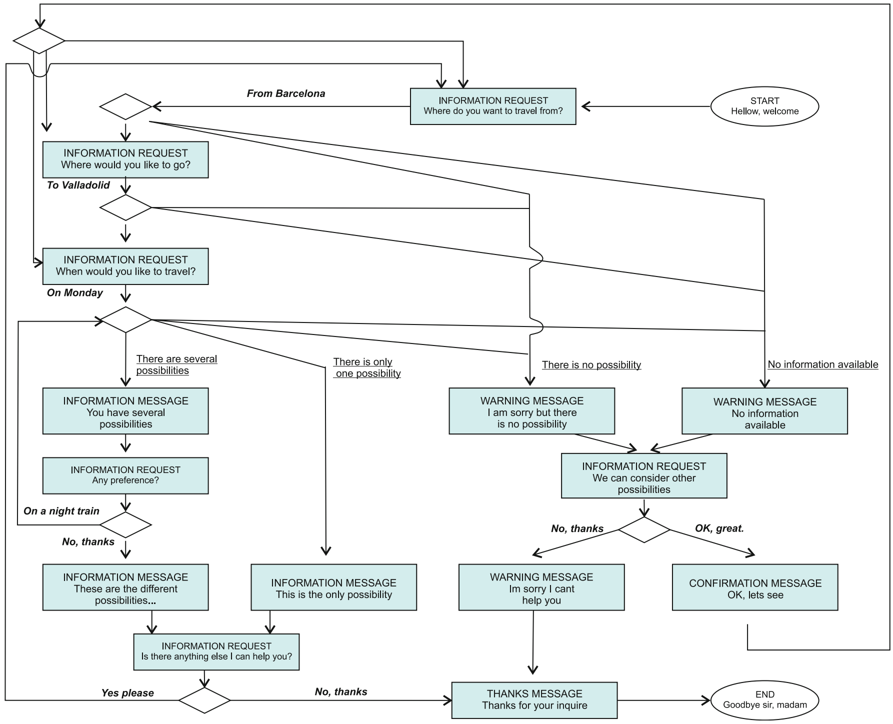


Fig. 2 Interaction graph that assists the speaker in the *travel information dialogue*

professional background, and the grouping of the selected speakers into categories, defining their contribution to the recordings. During the design process, it was decided that two types of speakers would be used for the recordings: professional (for the news subcorpus) and non-professional (for the dialogue subcorpus). This distinction was drawn because several studies have noticed that they have different speaking styles (Strangert and Gustafson 2008). The number of professional speakers was set to eight, four of them having a 'news broadcaster' and four an 'advertising' profile. Television and radio broadcasters receive very little speech training, and moreover, the very nature of news discourse imposes a supposed objectivity that inevitably ends up in all presenters speaking exactly the same way. Journalists—including radio news broadcasters- tend to deliver information which is characterised by prosodic signals of persistent emphasis which is repeated by most speakers. Furthermore, live coverage of the news makes impossible for the journalist to make any kind of amendments (de-la-Mota and Rodero 2011). Because they often work with strong time pressure, they usually care less about speech and other linguistic aspects (Rodero 2006). This stands in sharp contrast with advertising professionals, who are not only more trained in speech delivery skills but are also far better paid. The fact that they record only one text in each session allows these professionals to rehearse the commercial message, which causes them to be more careful about prosodic features. One further advantage is that the suggestive texts used in advertising facilitate a richer prosodic realization (Rodero 2007). Taking into account these diverse features of radio communication, a corpus has been built that joins these two models together: news are read by both radio news broadcasters and advertising professionals with prosody training with the aim of characterising, analysing and comparing each intonational pattern.

As far as dialogues were concerned, the number of informal conversations to be collected was established in six, which would require 12 different speakers, and, in the case of the task-oriented dialogues, the number of pairs to be recorded was fixed in 12 (24 different speakers). Table 1 summarizes these figures. This gave initially a total of 42 different speakers, between professional and non-professional, to be selected. To reduce this large number of speakers, it was decided to ask some of them to participate in the recordings of more than one subcorpus: four professional speakers would also record the task-oriented and informal dialogue subcorpora, and all the non-professional speakers involved in the collection of the informal dialogues would also participate in the task-oriented dialogues recordings. These different types of participation defined a set of speaker categories which are

Table 1 Number and typology of speakers required for each subcorpus

	Radio news broadcasters	Advertising professionals	Non professional speakers	Total
News	4	4		8
Task-oriented dialogues	2	2	20	24
Informal dialogues	2	2	8	12

explained in detail in Sect. 2.3.3. In addition, by using this method, the corpus would contain also some speech material uttered by the same speaker in different styles, which will enlarge the capabilities of the corpus for a future use in inter-speaker comparisons.

2.3.1 Sociolinguistic background of the speakers

In the process of informant selection, non-standard dialectal varieties that might influence the speaker's prosody in the corpus were avoided. Thus, the variety of Spanish spoken in Valladolid was used for both the dialogues and radio news reading, since this Castilian accent is representative of standard European Peninsular Spanish (Penny 2000). The speakers should meet the following requirements: (i) they must have lived in Valladolid for a relative long period of time; and (ii) Spanish must be the language they use on a regular basis. Likewise, Central Catalan was the form chosen for the dialogues and for reading the news. Once more, the speakers had to meet the requirements aforementioned. A special distinction was drawn in this case between those who had Catalan as their mother tongue and those who learned it as a second language.

Detailed questionnaire forms, which included questions about their linguistic background (place of birth, mother tongue of the parents, etc.) and about their use in different situations, were designed to gather information from each individual. The pre-selected subjects had to take this test to assess their relationship to the language, in terms of competence and performance.

Finally, the same number of male and female speakers was sought so that the variable gender was also balanced. In the case of the dialogue pairs, it was also intended to have a balance between male-male, female-female and male-female combinations.

2.3.2 Speaker categories

The final number of speakers to be recruited for each of the types considered (news professional, advertising professional, and non-professional) was set to 28, considering the categories described in Table 2. This organization allowed to have speakers participating in the recordings of all three (A category), two (C category) and only one (B and D categories) subcorpora. Following this distribution, the news task was only performed by professional speakers (four radio announcers and four advertising speakers), while dialogues were recorded by both professional and non-professional speakers (10 non-professional, one radio professional and one advertising professional pairs in the case of task-oriented dialogues; four non-professional, one radio professional and one advertising professional couples in the case of informal dialogues). It is important to mention that all the speakers participating in the informal dialogues task had to be colleagues, or friends.

Table 2 Speaker categories defining the task in which they participated (news, informal dialogues and task-oriented dialogues)

Category	Tasks	Speaker types	Number of speakers
A	News (prosodic)	Professional radio	2
	Informal dialogue	Professional advertising	2
	Task-oriented dialogue		
B	News (prosodic + phonetic)	Professional radio	2
		Professional advertising	2
C	Informal dialogue	Non-professional	8
	Task-oriented dialogue		
D	Task-oriented dialogue	Non-professional	12
Total			28

2.3.3 Speaker recruitment

As for the news professional type, radio speakers with large experience in the field were contacted by members of the Department of Communication at Universitat Pompeu Fabra (Barcelona), and their sociolinguistic background evaluated as described in Sect. 2.3.1. The speakers finally chosen were, in the case of Spanish, two male and two female radio news presenters working at the *Cadena SER Radio Station* in Valladolid, with more than 10 years of experience in the field. Their age range was between 41 and 49. The two male and two female Catalan speakers came from Catalunya *Ràdio*, *RAC1* and *Ràdio Estel*, which are among various radio stations that broadcast in Catalan. In this case, the age range was larger: between 26 and 66. A similar procedure was followed for the advertising candidates: all eight Spanish and Catalan advertising professionals finally chosen were renowned radio and dubbing voices in their respective languages, speakers of the chosen dialect, and with an active use of the language in their personal and professional lives. Their ages ranged from 34 to 46, in the case of the Spanish speakers, and from 38 to 49, in the case of Catalan.

Non-professional speakers were recruited among college students of communication, with some training in radio and TV broadcasting, assuming that this fact would give a more coherent profile to all (professional and non-professional) speakers. Also, it is widely assumed that they are more willing to participate in projects related to their future careers. They were recruited at various university departments in Valladolid (for Spanish) and Barcelona (for Catalan). During the pre-selection process more than a hundred of these non-professional speakers showed interest in participating in the project. After sociolinguistic evaluation of the candidates, using the same questionnaire as for the professional candidates, subjects finally selected for the recordings in Spanish were all journalism students between 19 and 24 years old. The selected Catalan speakers were communication students aged between 18 and 23. A detailed description of the speakers profile is provided in Escudero et al. (2010a).

Table 2 summarises the exact number of speakers needed for each sub-corpus: dialogues and news reading. The total number of non-professional speakers could be reduced thanks to a reassignment of the functions, as explained in the previous section.

A unique speaker ID label was assigned to every selected speaker, that was later used to identify their recordings across subcorpora. Each label includes a number, unique for each speaker in the corpus, and some letters indicating their gender (m for male, f for female) and profile (r for professional radio, a for professional advertising, and s for students) information. So for example, the label m05a identifies the speaker number 5 (male, advertising professional), and f37s refers to speaker number 37 (female, student).

3 Recording of the corpus

3.1 Recording sessions

The recording sessions differed depending on the category of the speakers involved. For the A category speakers, they were as follows: first, both speakers of the couple read the prosodic news corpus; then they performed the task-oriented dialogues together, and finally they completed the informal dialogue task. Category B speakers had two-part sessions, one for the prosodic news corpus and another one for the phonetic news corpus. Sessions involving category C speakers included also two parts, the first one for the task-oriented dialogues and the second one for the informal dialogue. Finally, category D sessions included only the recordings of task-oriented dialogues. For all four categories, speakers were paid for their contribution.

In the news sessions, the speakers were told to read the proposed news texts as if they were on the air. However, unlike in real radio broadcast, they were asked to repeat their reading if they had noticeable reading mistakes.

Task-oriented dialogue sessions were split in three blocks, one for each of the proposed situations (travel, university and tourism, in this order, from more to less formal situation). Before the start of each block, the experimenter explained the participants' role in the dialogue, and gave them the paper sheets containing the information they needed to play their role (train and bus timetables in the travel dialogue; information about courses and activities in different French universities, for the university dialogue; and some tourist flyers and Corfu maps for the tourist condition). After a quick review of this information, participants could ask to the experimenter all the questions they might have before the start of the recording. In these sessions, a panel was placed between both speakers to avoid direct eye-contact and simulate the telephone condition, so they could hear each other but not give information through gestures.

Finally, in the informal dialogue sessions, participants were first informed by the experimenter about the goal of the task, and about the initial question they should answer to start their conversation. Once the conversation started, they could speak freely for about 10–15 min, with no intervention of the experimenter, unless both speakers stopped talking, in which case the controller proposed a new question or topic. In this condition, both speakers sat face-to-face, as in normal conversations.

3.2 Recording setup

Recordings took place at two different premises: soundproof rooms at the Audiovisual Media Service of the University of Valladolid for the Spanish recordings, and at the Communication Campus of the Universitat Pompeu Fabra, in Barcelona, for Catalan. In Valladolid, recordings were made on a Marantz PMD670/W1B and a Marantz PMD560 recorders, using a Mackie CR1604-VLZ mixer, at a sampling frequency of 44 KHz. In Barcelona, the Sony Vegas program running on a PC with a RME Hammerfall HDSP 9652 soundcard, and a Yamaha 02R96 mixer with ADAT MY16AT cards, were used for recordings, at a sampling frequency of 48 KHz.

All the recordings were made using two microphones for each speaker: a fixed directional one in front of them (Neumann TLM103 P48 in Valladolid; AKG C 414 B-ULS in Barcelona), and a headset wireless one (Senheisser EW100-G2, both in Barcelona and Valladolid). Headset microphones were used to ensure that the distance between the speaker's mouth and the microphone was kept constant throughout the recordings, making the energy registration reliable for prosodic analyses. The signal from both microphones has been included in the corpus, so the user can choose which one to analyse depending on their research interest: signals from the fixed microphones show a higher overall quality, although sometimes energy differences can be noticed depending on the distance of the speaker to the microphone; this problem is avoided with the headset microphone signals, but sometimes some bursts are heard due to air impacts. In dialogue recordings, each speaker used different microphones in order to have separate recordings of the speech of each participant, so as to minimise as much as possible the overlapping of signals. A laringograph (Laryngograph Processor, from Laryngograph Ltd) was also used to record the glottal activity in some of the news recordings (those of the category B speakers). This signal can be used to detect the glottal closure instants and to get an accurate pitch estimation. In total, four synchronous channels (six if the laryngograph was included) were recorded.

Recordings were stored on wav files, one per signal (one wav for the fixed microphone, one for the headset microphone and one for the laringograph, if any). In the case of dialogue recordings, stereo wav files were created, including the signal of each speaker's microphone. Then, two stereo wav files were obtained for each dialogue, one for the fixed microphones and one for the headset microphones.

4 Corpus structure and contents

4.1 News subcorpus

Table 3 lists the features of the news subcorpus. Two groups of speakers can be distinguished: those who read 36 news items (only the *prosodic subcorpus*), that is, category A speakers (f11r, m12r, m09a, m10a, f01r, m04r, f02a and m05a, those who also participated in the dialogue recordings, as it can be observed in Table 3); and those who read 72 news items (*prosodic subcorpus* and *phonetic subcorpus*)

Table 3 Contents of the news subcorpus

Speaker Id	Speaker type	Gender	Language	# News items	Duration
f11r	Radio	Female	sp	36	30′ 53″
m09a	Advertising	Male	sp	36	30′ 59″
m10a	Advertising	Male	sp	36	30′ 42″
m12r	Radio	Male	sp	36	32′ 24″
m14r	Radio	Male	sp	72	55′ 44″
f13r	Radio	Female	sp	72	1h 3′ 55″
f15a	Advertising	Female	sp	72	1h 28′ 20″
f16a	Advertising	Female	sp	72	1h 7′ 18″
Total time (SP)					6h 40′ 19″
f01r	Radio	Female	ca	36	30′ 16″
f02a	Advertising	Female	ca	36	32′ 30″
m04r	Radio	Male	ca	36	28′ 12″
m05a	Advertising	Male	ca	36	28′ 20″
f06r	Radio	Female	ca	72	1h 4′ 55″
f07a	Advertising	Female	ca	72	1h 8′ 3″
m03r	Radio	Male	ca	72	1h 3′ 25″
m08a	Advertising	Male	ca	72	1h 7′ 21″
Total time (CA)					6h 23′ 6″

that is, category B speakers. The amount of speech collected varies for each speaker type: about half an hour for the speakers of the first group, and approximately 1 h for the speakers of the second group. About six and a half hours of news speech were collected in total per language.

4.2 Dialogue subcorpus

Table 4 shows the features of the dialogue subcorpus, both informal and task-oriented, showing the total amount of speech per pair. More than 12 h of dialogue have been recorded: almost 5 h and 45 min in the case of Spanish, and about 6 h and 45 min for Catalan.

All the dialogues are currently available in two versions: ‘complete’, in which each dialogue has been stored in a single stereo wav file, as described in Sect. 3; and ‘turns’, in which each talk turn within the dialogue has been segmented and stored in separate mono wav files. Table 5 lists both the different time duration and the number of user turns per dialogue.

A close look at the data of the task-oriented dialogues allows to observe that the length of the dialogues varies noticeably, ranging from more than 17 min—in the case of speakers *m47s-f48s* in the *travel information dialogue*—to just 4 min and 6 s (speakers *f19s-m20s* in the *university information dialogues*). The particular duration of each dialogue appears to depend not just on the speakers but also on the nature of the task. Such a contrast evidences that, even though the speakers’ activity

Table 4 Features of the dialogue subcorpus

Speaker Id	Speaker type	Gender	Language	# Dialogues	Duration
f11r-m12r	R–R	F–M	sp	4	22′ 55″
f19s-m20s	S–S	F–M	sp	4	30′ 33″
f21s-f22s	S–S	F–F	sp	4	37′ 52″
f23s-f24s	S–S	F–F	sp	4	35′ 29″
f29s-m30s	S–S	F–M	sp	3	24′ 54″
f31s-m32s	S–S	F–M	sp	3	14′ 29″
f33s-f34s	S–S	F–F	sp	3	23′ 15″
f35s-f36s	S–S	F–F	sp	3	18′ 56″
m09p-m10p	A–A	M–M	sp	4	41′ 52″
m17s-m18s	S–S	M–M	sp	4	48′ 14″
m25s-m26s	S–S	M–M	sp	3	16′ 11″
m27s-m28s	S–S	M–M	sp	3	29′ 09″
Subtotal			sp	42	05h 43′ 55″
f01r-m04r	R–R	F–M	ca	4	54′ 57″
f02p-m05p	A–A	F–M	ca	4	44′ 05″
f37s-f38s	S–S	F–F	ca	4	35′ 13″
f39s-m40s	S–S	F–M	ca	4	37′ 04″
f49s-m50s	S–S	F–M	ca	3	33′ 47″
f53s-f54s	S–S	F–F	ca	3	26′ 22″
m41s-f42s	S–S	M–F	ca	4	35′ 11″
m43s-m44s	S–S	M–M	ca	4	36′ 06″
m45s-f46s	S–S	M–F	ca	3	24′ 25″
m47s-f48s	S–S	M–F	ca	3	29′ 01″
m51s-f52s	S–S	M–F	ca	3	21′ 17″
m55s-m56s	S–S	M–M	ca	3	23′ 56″
Subtotal			ca	42	06h 41′ 29″

Speaker type can be radio broadcasters (R), advertising speakers (A) and non-professional speakers (S)

was guided by specific protocols, the subjects were relatively free when it came to task-solving.

Concerning the informal dialogues, the total amount of recorded speech was 2 h, 15 min and 21 s (see Table 5). No specific length was imposed either on the speakers' dialogues this time, as reflected in the different duration registered, ranging from 5 min and 44 s for the speaker pair *f11r-m12r* to 16 min for the speaker pair *f19s-m20s*.

4.3 Corpus transcription and annotation

After recording, the whole corpus has been annotated with several levels of linguistic information, all relevant for the study of prosody. At the current state of the corpus, the following levels are available, all of them time-aligned with the speech signal:

Table 5 Features of the dialogue subcorpus in Spanish and Catalan

Speakers Id	trd		tod		und		fcd	
	Duration	Turns	Duration	Turns	Duration	Turns	Duration	Turns
Spanish								
f11r-m12r	5′ 40″	164	5′ 48″	88	5′ 41″	117	5″44″	140
f19s-m20s	5′ 54″	163	4′ 31″	144	4′ 06″	106	16′ 00″	454
f21s-f22s	12′ 35″	318	7′ 41″	197	5′ 49’″	99	11′ 45″	366
f23s-f24s	11′ 38″	314	6′ 55″	130	9′ 30″	197	7′ 25″	209
f29s-m30s	10′ 23″	251	7′ 22″	176	7′ 08″	192	0′ 00″	0
f31s-m32s	5′ 28″	120	4′ 36″	110	4′ 24″	92	0′ 00″	0
f33s-f34s	9′ 02″	221	7′ 19″	156	6′ 54″	171	0′ 00″	0
f35s-f36s	7′ 27″	159	7′ 09″	139	4′ 19″	83	0′ 00″	0
m09a-m10a	10′ 11″	258	9′ 15″	253	9′ 09″	218	13′ 15″	365
m17s-m18s	11′ 38″	291	14′ 16″	318	9′ 42″	230	12′ 36″	337
m25s-m26s	5′ 30″	164	5′ 20″	102	5′ 20″	65	0′ 00″	0
m27s-m28s	10′ 43″	191	9′ 52″	185	8′ 32″	138	0′ 00″	0
Catalan								
f01r-m04r	13′ 11″	407	16′ 27″	446	15′ 12″	434	10′ 06″	344
f02a-m05a	12′ 39″	484	10′ 06″	335	11′ 09″	486	10′ 10″	399
f37s-f38s	7′ 30″	186	8′ 36″	206	6′ 32″	156	12′ 34″	423
f39s-m40s	8′ 52″	161	6′ 17″	122	9′15″	174	12′ 38″	328
f49s-m50s	16′ 03″	460	9′ 02″	203	8′ 41″	174	0′ 00″	0
f53s-f54s	9′ 54″	233	5′ 32″	151	10′ 55″	252	0′ 00″	0
m41s-f42s	7′ 10″	179	7′ 55″	196	6′ 30″	142	13′ 33″	444
m43s-m44s	9′ 02″	229	7′ 33″	219	10′ 05″	284	9′ 25″	381
m45s-f46s	6′ 53″	190	9′ 53″	236	7′ 38″	204	0′ 00″	0
m47s-f48s	17′ 32″	621	6′ 40″	165	4′’48″	112	0′ 00″	0
m51s-f52s	5′ 27″	139	8′ 08″	187	7′ 40″	175	0′ 00″	0
m55s-m56s	7′ 44″	189	10′ 13″	183	5′ 59″	100	0′ 00″	0

Travel information dialogue is trd, *tourist information dialogue* is tod, *university information dialogues* is und, and *free conversational dialogue* is fcd

(a) the orthographic transcription of the recordings;
(b) the phonetic transcription;
(c) the syllable segmentation, with indication of the stressed ones;
(d) the annotation of minor prosodic breaks (defining minor prosodic units);
(e) the annotation of major prosodic breaks (defining major units or breath groups)

The annotation of this amount of information in such a large corpus is a huge task, which could not be faced by manual means within the terms of the project. For this reason, these annotations (with the exception of the orthographic transcription) have been obtained automatically using different tools, although they are being reviewed manually by expert annotators.

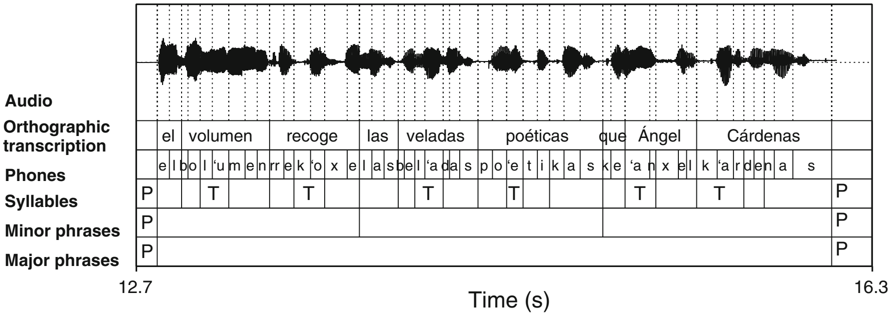


Fig. 3 TextGrid and waveform corresponding to the utterance "*el volumen recoge las veladas poéticas que Ángel Cárdenas*", spoken by a female professional speaker (Spanish prosodic subcorpus, text 1). TextGrid tiers include word orthographic transcription, phonetic transcription, syllable segmentation and annotation (T labels indicate stressed syllables), minor phrases segmentation and major phrases segmentation. The label P in the tiers indicates a pause segment

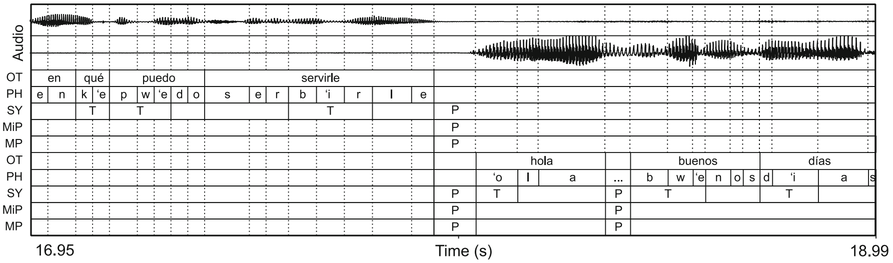


Fig. 4 TextGrid and waveform corresponding to two turns in the transport dialogue performed by the two radio professional speakers (Spanish task-oriented subcorpus). OT stands for orthographic transcription, PH for phones, SY for syllables, MiP for minor phrases and MP for major phrases. As in Fig. 3, T labels mark stressed syllables, and P labels, pause segments

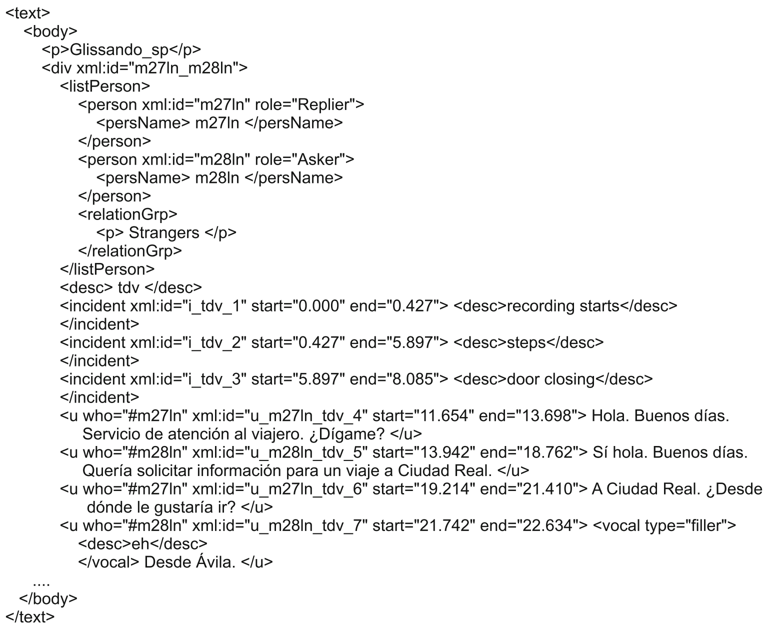
These annotations have been stored as Praat TextGrid files (Boersma and Weenink 2012), in which each level has been included in a separate Tier, as it can be observed in Figs. 3 (for the news corpus) and 4 (for the dialogue corpus; here, the annotations corresponding to both speakers are included in the same file).

Other types of prosodic annotation, such as ToBI, MoMEL, or the ones produced by MelAn (Garrido 2010) or the tool described in Escudero and Cardeñoso Payo (2007) and Escudero et al. (2002) are planned to be included in future public versions of the corpus, however at present they have been kept for use of the participants in the project, or are still in progress of development.

In addition, raw values for F0 and intensity have been calculated and stored in text files for the whole corpus.

4.3.1 Orthographic transcription

Since the news were read by speakers, it was only necessary to modify the original text to adapt it to what the speaker actually said. The output of this review was a set



```
<text>
  <body>
    <p>Glissando_sp</p>
    <div xml:id="m27ln_m28ln">
      <listPerson>
        <person xml:id="m27ln" role="Replier">
          <persName> m27ln </persName>
        </person>
        <person xml:id="m28ln" role="Asker">
          <persName> m28ln </persName>
        </person>
        <relationGrp>
          <p> Strangers </p>
        </relationGrp>
      </listPerson>
      <desc> tdv </desc>
      <incident xml:id="i_tdv_1" start="0.000" end="0.427"> <desc>recording starts</desc>
      </incident>
      <incident xml:id="i_tdv_2" start="0.427" end="5.897"> <desc>steps</desc>
      </incident>
      <incident xml:id="i_tdv_3" start="5.897" end="8.085"> <desc>door closing</desc>
      </incident>
      <u who="#m27ln" xml:id="u_m27ln_tdv_4" start="11.654" end="13.698"> Hola. Buenos días.
          Servicio de atención al viajero. ¿Dígame? </u>
      <u who="#m28ln" xml:id="u_m28ln_tdv_5" start="13.942" end="18.762"> Sí hola. Buenos días.
          Quería solicitar información para un viaje a Ciudad Real. </u>
      <u who="#m27ln" xml:id="u_m27ln_tdv_6" start="19.214" end="21.410"> A Ciudad Real. ¿Desde
           dónde le gustaría ir? </u>
      <u who="#m28ln" xml:id="u_m28ln_tdv_7" start="21.742" end="22.634"> <vocal type="filler">
        <desc>eh</desc>
        </vocal> Desde Ávila. </u>
    ....
  </body>
</text>
```

Fig. 5 Example of xml coding of the orthographic transcription or a task-oriented dialogue in Spanish

of txt files, in plain text (UTF-8) format, each one containing the actual transcription of a news text for a given speaker.

In the case of dialogues, it was necessary to transcribe all of them manually by listening to the recordings. This task was performed in two steps: first a raw transcription was made, turn by turn, on TextGrid files, with two tiers containing the turns transcription for each speaker, time-aligned with the signal, and a third tier for the time-aligned annotation of non-linguistic, external events occurring during the conversation; then, from these hand-made TextGrid files, xml files were generated automatically for each dialogue, containing the orthographic transcription of each turn, their time-alignment with the speech signal, some additional tags indicating truncated or mispronounced words or the presence of paralinguistic and non-linguistic events, and a header with the basic information about the speakers and the task performed. The TEI conventions (Sperber-McQueen and Burnard 1994) were used as standard reference for the coding of these informations. Table 7 presents a list of the tags used for the annotation of fillers (vf) and non-linguistic events (vn), and Fig. 5 includes the header and the transcription of some turns of the Spanish task-oriented dialogues, as examples of this coding.

4.3.2 Phonetic transcription and alignment

Once the orthographic transcription of both news and dialogues was available, the entire corpus was processed to obtain automatically the phonetic transcriptions of the texts, and the alignment of the phone symbols with the signal. These two tasks were carried out using an automatic transcription and segmentation tool kindly provided by the *Speech and Language Group of Barcelona Media Centre d'Innovació*, research partner of GLiCom. This tool is the result of a collaboration between Barcelona Media and Cereproc Ltd to develop the Spanish and Catalan modules for the Cerevoice text-to-speech system (Garrido et al. 2008). This tool allowed to generate, for each input wav file, a TextGrid containing two tiers, the first one for the orthograhic transcription of the text (word by word), and the second one for the phonetic transcription, both aligned with the speech signal. The phonetic transcription was generated using the SAMPA phonetic alphabets for Spanish[2] and Catalan.[3] For the segmentation of dialogues the 'turns' version was used (one file per turn), so initially one TextGrid file per turn was generated.

4.3.3 Prosodic units segmentation

After orthographic and phonetic transcription, three more tiers were added to the existing TextGrids to annotate the boundaries of three types of prosodic units: syllables, minor and major prosodic breaks. Minor and major prosodic breaks are intended to be theory-independent labels to name two types of prosodic units with a long tradition in prosodic studies: major units are defined here as portions of utterance ended by a pause, silent or not ('breath-groups', in some frameworks); and minor groups have been defined as portions of an utterance with a 'complete' intonation contour, that is, ending with an F0 movement perceived by listeners as terminal, irrespective of the presence or absence of a pause after it ('intonation unit' or 'intermediate phrase' in some theoretical frameworks like Beckman et al. 2005).

This annotation was carried out by means of SegProso, a tool for the automatic annotation of prosodic boundaries from an aligned phonetic transcription developed by the GLiCom group at Pompeu Fabra University.

4.3.4 Intonation annotation

In addition to the raw acoustic data (F0, intensity) and the segmentation in prosodic units (syllables, minor and major groups), specific intonation annotation is being carried out, although it will not be available at the first public version of the corpus. This annotation will include ToBI labels, but also other types of annotation used by the groups involved in the project, such as MelAn (Garrido 2010) or Bézier (Escudero and Cardeñoso Payo 2007).

Intense research on the automatic annotation of corpora using ToBI labels has been carried out in parallel to the development of the corpus (Escudero et al. 2012; Gonzalez-Ferreras et al. 2012; Escudero et al. 2011a, b). This research has led a first automatic annotation of prominences using the ToBi-framework conventions, as illustrated in Fig. 6.

Also, the corpus has been partially annotated using MelAn, an automatic tool for the annotation of intonation inspired in the IPO model (Garrido 2010). This

[2] http://www.phon.ucl.ac.uk/home/sampa/spanish.htm.

[3] http://liceu.uab.es/~joaquim/language_resources/SAMPA_Catalan.html.

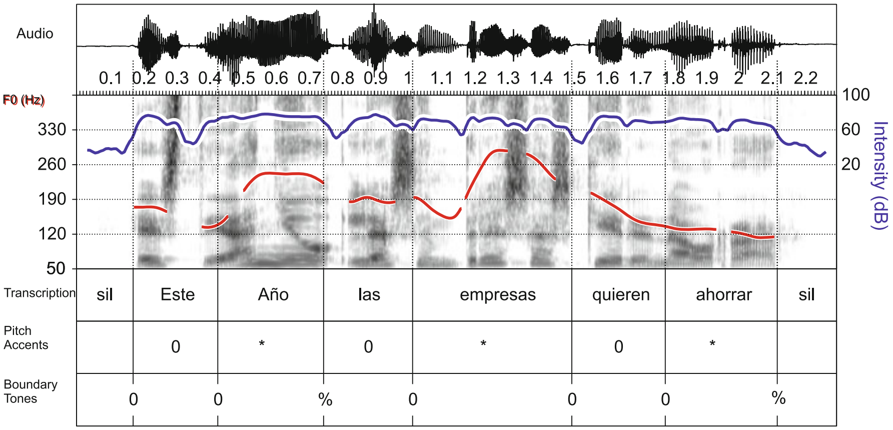


Fig. 6 Sample TextGrid file with the automatic prominence annotation for a Spanish news subcorpus file. *Blue line* (higher) indicates energy evolution, and *red line* (lower) represents the F0 contour. Words showing prominence according to the tool are marked in the *Pitch Accents* tier with a *asterisk* symbol. Automatic boundary tone detection output is included in the last tier. (Color figure online)

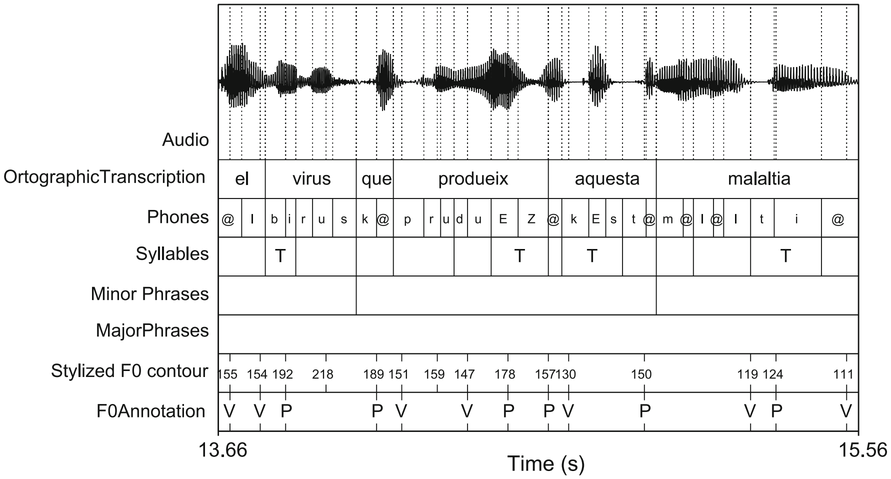


Fig. 7 Example of automatic intonation annotation using MelAn in the Catalan news subcorpus

annotation allows to keep raw F0 values corresponding to the relevant inflection points in the F0 contours, and their annotation in terms of 'peaks' (P) and 'valleys' (V), as illustrated in Fig. 7.

5 Preliminary evaluation

This section includes the results of some preliminary analysis of specific features of the collected corpus, as a sample of its capabilities for prosodic analysis. Several prosodic features are compared across speakers and styles (F0 contours, F0 register, speech rate, pause duration, breath group length). Finally, an example of the possibilities of the corpus for the description of Spanish and Catalan dialogues is also given.

5.1 Inter-speaker variability: F0 contours

Inter-speaker prosodic analysis is one of the possible uses of the Glissando corpus. As a sample of the inter-speaker variety of the Glissando corpus, Fig. 8 shows a representation of the F0 contours corresponding to the same sentence in the Spanish new subcorpus, uttered by four professional speakers who recorded it. Differences among the different F0 contours are easily observable, both in shape and duration.

5.2 Cross-style variability: mean F0, speech rate, pause duration

The Glissando corpus has also been designed for cross-style studies, by including speech of three different speaking styles, in some cases from the same speakers. Figure 9 presents some data about cross-style variation in the mean F0 in two different advertising professional speakers, m09a and m10a. These data show significant differences among styles within the same speaker. So, for example, mean F0 register of speaker m10a along the news subcorpus is clearly different to the one in dialogues: higher for news (mean 104 Hz) and in general lower in dialogues, with differences among dialogue types (mean 83 Hz in the case of transport dialogue). A Student t-test applied to the data showed that these differences are statistically different ($p = 2.2.e - 16$; <0.05).

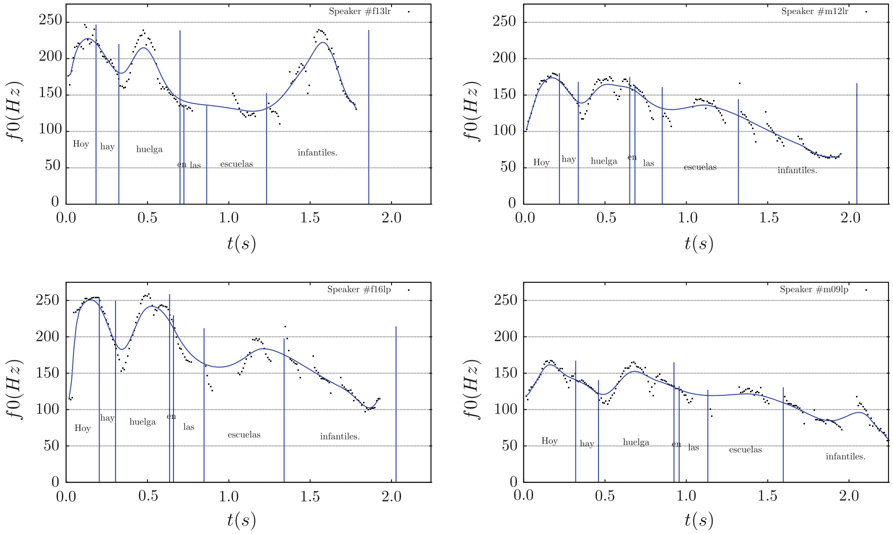


Fig. 8 F0 contours corresponding to the Spanish sentence of the news corpus "*Hoy hay huelga en las escuelas infantiles*", uttered by four different professional speakers. *Left column* female speakers. *Right column* Male speakers. Common time scale

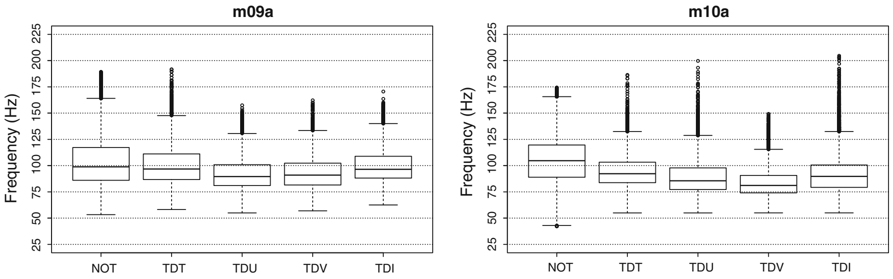


Fig. 9 Mean F0 values for speakers m09a and m10a in five different speaking conditions: news (NOT), tourist dialogues (TDT), university dialogues (TDU), transport dialogues (TDV) and informal dialogues (TDI)

Table 6 Mean values for speech rate (SR) in words/minute, pause duration (MPD) and words between pauses (MNWIP) across five different speaking conditions in four professional speakers (f11r, m12r, m09a and m10a)

Variable	Speaker	News	Dialogue t	Dialogue v	Dialogue u	Dialogue i
SR	f11r	242.8	223.4	216.7	229.3	228.0
	m12r	214.1	217.5	199.7	205.8	211.9
	m09a	221.8	233.2	246.8	225.2	245.4
	m10a	222.9	226.7	215.1	203.6	238.2
MPD	f11r	493.8	78.6	117.2	95.3	125.1
	m12r	558.4	99.4	128.9	93.8	154.5
	m09a	466.1	89.0	77.1	91.8	95.0
	m10a	494.6	91.2	87.9	104.2	81.8
MNWIP	f11r	5.5	20.0	15.7	39.5	10.0
	m12r	7.5	10.3	10.0	11.2	5.6
	m09a	6.8	6.3	7.1	8.3	7.6
	m10a	7.1	8.6	6.1	6.2	6.4

Table 6 contains cross-style and inter-speaker comparisons for three different variables: speech rate (measured in words per minute), pause duration, and breath group length (measured in words between pauses). A look to the data reveals, for example, that differences in speech rate seem to be more speaker dependent than style dependent: speaker f11r shows speech rates between 216.7 and 242.8 words per minute, whereas speaker m12r speaks considerably slower, with speech rates ranging between 199.7 and 214 wpm. However, pause duration does show clear cross-style differences: pauses are clearly longer in news reading than in dialogues, independently on the speaker. Finally, length of breath groups shows again differences in the behavior of the speakers; one speaker (f11r) has clearly longer mean lengths in all four types of dialogue than in news reading; two speakers (m09a and m10a) show similar mean lengths across all five conditions; and one speaker (m12r) shows a more complex pattern, with longer groups in task-oriented dialogues and shorter groups in news reading and informal dialogues.

Table 7 List of TEI-inspired annotation tags used for the transcription of fillers (vf) and non-linguistic events (vn), and number of occurrences in the Glissando corpus

Event	Description	Spanish				Catalan			
		tod	trd	und	fcd	tod	trd	und	fcd
[*vf* − 01]	Filler “ah”	8	3	8	8	10	4	8	9
[*vf* − 02]	Filler “eh”	171	243	325	58	215	277	330	39
[*vf* − 03]	Filler “mmm”	157	113	126	73	289	226	268	92
[*vf* − 04]	Filler “mmm mmm”	57	67	79	15	49	64	62	2
[*vn* − 01]	Kiss	0	0	0	1	0	0	0	2
[*vn* − 02]	Yawn	0	3	0	0				
[*vn* − 03]	Humming	0	0	2	0	1	1	1	0
[*vn* − 04]	Click	86	74	45	39	211	62	70	100
[*vn* − 05]	Mumbling	1	1	3	0	3	0	0	3
[*vn* − 07]	Grumbling	0	0	0	0	0	0	1	1
[*vn* − 09]	Laugh	80	6	3	136	58	18	16	199
[*vn* − 11]	Whistle	1	0	0	0				
[*vn* − 12]	Sigh	0	1	0	2	0	0	1	0
[*vn* − 13]	Blow	5	3	2	0	30	5	13	32
[*vn* − 14]	Cough	0	0	1	2	3	1	0	1
[*vn* − 15]	Clear throat	5	5	8	15	5	12	1	6
[*vn* − 16]	Breath	35	67	97	19	12	8	20	40
[*vn* − 17]	Grinding	0	2	1	0	0	1	0	0
[*vn* − 18]	Hiccup	0	0	0	0	0	0	0	1

Travel information dialogue is trd, tourist information dialogue is tod, university information dialogues is und, and free conversational dialogue is fcd

5.3 Prosody in dialogue: fillers and non-linguistic elements

Another possible application of the Glissando corpus is the analysis of the prosodic properties of dialogues using a large set of data. One of these properties is the use of fillers and non-linguistic elements, such as laughs or coughs, which are not present in the news subcorpus. Apart from the intrinsic interest in the description of these type of elements in dialogues, some studies (Adell et al. 2012, for example) have attempted to insert fillers and non-linguistic elements in synthetic utterances to improve their naturalness in person-machine dialogues. The Glissando corpus contains a wide variety of such elements, as shown in Table 7, and seems to be a perfect material to attempt an in-depth analysis.

6 Conclusions

In the present paper the development procedure and main features of the Glissando corpus have been presented. This corpus offers high-quality annotated material to researchers approaching the study of Spanish and Catalan prosody. The Glissando corpus differs from other similar corpora in as much as it provides both a larger volume of information and a wider coverage of real-life communicative situations. Additionally, the news selected for the news subcorpus guarantees the presence of all types of prosodic units that might be helpful and also offers a good phonetic coverage. As for the dialogue subcorpus, different types of problem-solving dialogues were collected: map-task (tourist information), information service, and telephone travel reservations. Free conversation was also included in the corpus.

We made sure to recruit informants with different professional skills; thus, in addition to the distinction between professional and non-professional speakers, there are also radio news broadcasters and advertising actors in the former category. The sociolinguistic variables 'dialect' and 'professional status' have been controlled in the corpus, with speakers from Valladolid as representative of Standard European Spanish, and students from Barcelona as fairly good examples of Central Catalan.

The corpus contains more than 12 h of read speech—news—and another 12 h of high-quality studio recordings of dialogues which have been transcribed, aligned with the acoustic signal and prosodically annotated. The xml transcription of dialogues, including speech turns and disfluencies, offers also a way for the study of dialogue phenomena only from the text version, without the need of the speech material.

From the viewpoint of the technical conditions, there is one fact that should be assessed positively: two-channel high quality recordings were performed by using two kinds of microphones, a wireless headset microphone and a fixed desktop microphone. The use of these two microphones permits the researcher to add the acoustic parameter intensity to the prosodic models resulting from the study of this corpus. Also, the voice of each speaker in dialogues has been saved in a different channel, so that it is easier to analyze automatically.

The corpus has been designed to cover the research needs of the groups involved in the project, but its possibilities for future research are numerous. It has proven to be useful, for example, for the automatic prosodic description and modelling of intonation (Garrido and Rustullet 2011). And some other pilot studies have shown its interest for the study of disfluencies, and the differences in the reading style of radio and advertising professionals. It will be also very useful for sure for other types of comparative studies, such as cross-lingual (Spanish–Catalan) or inter-style (read-dialogue).

The corpus is available to the scientific community via the project website http://veus.barcelonamedia.org/glissando, from which it is accessible online once a noncommercial use license agreement (Creative Commons licence) is accepted.

Acknowledgments This work has been partly supported by the National R&D&I Plan of the Spanish Government (FFI2008-04982-C03-01/FILO, FFI2008-04982-C03-02/FILO and FFI2008-04982-C03-03/FILO projects).

References

Adell, J., Escudero, D., & Bonafonte, A. (2012). Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. *Speech Communication*, *54*(3), 459–476.

Albelda Marco, M. (2005). Sistemas de transcripción de los corpus orales del español. In M. Carrió (Ed.), *Perspectivas interdisciplinares de la linguística aplicada*, vol 2. Asociación Española de Lingüística Aplicada, pp. 381–388.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., et al. (1991). The hrc map task corpus. *Language and Speech, 24*, 351–366.

Beckman, M., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S. A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 9–54). New York: Oxford University Press.

Boersma, P., & Weenink, D. (2012). Praat: Doing phonetics by computer [computer program]. version 5.3.09, retrieved 10 march 2012 from http://www.praat.org.

Botinis A., Granstrom, B., & Mobius, B. (2001). Developments and paradigms in intonation research. *Speech Communication, 33*(4), 263–296.

Campione, E., & Veronis, J. (1998). Multext: A multilingual prosodic database. In *Proceedings of ICSLP 98*, vol. 7 (pp. 3163–3166).

Cresti, E., & Moneglia, M. (2005). *C-ORAL-ROM. integrated reference corpora for spoken romance languages*. John Benjamins Studies in Corpus Linguistics 15.

de-la-Mota, C., & Rodero, E. (2011) La entonación en la información radiofónica. In *El Estudio de la prosodia en España en el siglo XXI: perspectivas y ámbitos* (pp. 159–176). Annex de Quaderns de Filologia, Facultat de Filologia, Universitat de València.

Escudero, D., & Cardeñoso Payo, V. (2007). Applying data mining techniques to corpus based prosodic modeling. *Speech Communication 49*(3), 213–229.

Escudero, D., Cardeñoso, V., & Bonafonte, A. (2002). Corpus based extraction of quantitative prosodic parameters of stress groups in Spanish. In *Proceedings of ICASSP 2002*, vol. 1 (pp. 481–484).

Escudero, D., Aguilar, L., Bonafonte, A., & Garrido, J. (2009). On the definition of a prosodically balanced copus: Combining greedy algorithms with expert guided manipulation. *Revista de la Sociedad Española de Procesamiento del Lenguaje Natural, 43*, 93–102.

Escudero, D., Cardeñoso, V., Vivaracho, C., Aguilar, L., de-la-Mota, C., Garrido, J., et al. (2010a). *Proyecto glissando: Grabación de corpus prosódico de noticias y diálogos en español*. Tech. Rep. IT-DI-2010-3, Departamento de Informática, Universidad de Valladolid.

Escudero, D., Garrido, J., Aguilar, L., Bonafonte, A., González, C., & Rodero, E. (2010b). *Glissando project: Bilingual Spanish and Catalan corpus radio news text contents selection*. Tech. Rep. IT-DI-2010-2, Departamento de Informática, Universidad de Valladolid.

Escudero, D., Gonzalez-Ferreras, C., Garrido, J. M., Rodero, E., Aguilar, L., & Bonafonte, A. (2010c). Combining greedy algorithms with expert guided manipulation for the definition of a balanced prosodic Spanish–Catalan radio news corpus. In *Proceedings of Speech Prosody 2010*.

Escudero, D., Aguilar, L., Ferreras, C. G., Vivaracho-Pascual, C., & Cardeñoso-Payo, V. (2011a). Cross-lingual English Spanish tonal accent labeling using decision trees and neural networks. In C. M. Travieso-González & J. B. A. Hernández (Eds.), *Advances in nonlinear speech processing—5th international conference on nonlinear speech processing, NOLISP 2011*, Las Palmas de Gran Canaria, Spain, November 7–9, 2011. *Proceedings, Springer, Lecture Notes in Computer Science*, vol. 7015, (pp. 63–70).

Escudero, D., Vivaracho-Pascual, C., González-Ferreras, C., Cardeñoso-Payo, V., & Aguilar, L. (2011b). Analysis of inconsistencies in cross-lingual automatic tobi tonal accent labeling. In I. Habernal & V. Matousek (Eds.), *Text, speech and dialogue—14th International conference, TSD 2011, Pilsen*, Czech Republic, September 1–5, 2011. *Proceedings, Springer, Lecture Notes in Computer Science*, vol. 6836 (pp. 41–48).

Escudero, D., Aguilar, L., Vanrell, M., & Prieto, P. (2012). Analysis of inter-transcriber consistency in the Cat_ToBI prosodic labeling system. *Speech Communication, 54*, 566–582

Eskénazi, M. (1993). Trends in speaking styles research. In *Proceedings of Eurospeech 1993*, vol. 1, pp. 501–505.

Fernández, A. (2005). Aspectos generales acerca del proyecto internacional AMPER en España. *Estudios de Fonética Experimental, XIV*, 13–27.

Font, D. (2006). Corpus oral de parla espontània. Gràfics i arxius de veu. Biblioteca Phonica 4.

Garrido, J. (1996). *Modelling Spanish intonation for text-to-speech applications*. PhD thesis.

Garrido, J. (2010). A tool for automatic F0 stylisation, annotation and modelling of large corpora. In *Speech Prosody 2010*, Chicago.

Garrido, J., & Rustullet, S. (2011). Patrones melódicos en el habla de diálogo en español: un primer análisis del corpus Glissando. *Oralia: Análisis del discurso oral, 14*, 129–160.

Garrido, J., Bofias, E., Laplaza, Y., Marquina, M., Aylett, M., & Ch, P. (2008). The Cerevoice speech synthesiser. In *Actas de las V Jornadas de Tecnología del Habla (Bilbao).*

Gonzalez-Ferreras, C., Escudero-Mancebo, D., Vivaracho-Pascual, C., & Cardenoso-Payo, V. (2012). Improving automatic classification of prosodic events by pairwise coupling. *Audio, Speech, and Language Processing, IEEE Transactions on, 20*(7), 2045–2058.

Hirschberg, J. (2000). A corpus-based approach to the study of speaking style. In H. Horne (Ed.), *Prosody: Theory and experiments. Studies presented to Gosta Bruce* (pp. 335–350). Berlin: Kluwer Academic Publishers.

Huang, X., Acero, A., & Hon, H. W. (2001). *Spoken language processing: A guide to theory, algorithm and system development*. Prentice Hall PTR.

Maekawa, K. (2003). Corpus of spontaneous Japanese: Its design and evaluation. In *Proceeding of ISCA and IEEE workshop on spontaneous speech processing and recognition*, pp. 7–12.

Maekawa, K., Koiso, H., Furui, S., & Isahara, H. (2000). Spontaneous speech corpus of Japanese. In *Proceeding of the 2nd international conference of language resources and evaluation*, Vol. 2 (pp. 947–952).

McAllister, J., Sotillo, C., Bard, E., & Anderson, A. (1990). *Using the Map Task to investigate variability in speech*. Occasional paper.

Nagorski, A., Boves, L., & Steeneken, H. (2002). Optimal selection of speech data for automatic speech recognition systems. In *ICSLP*, pp. 2473–2476.

Ostendorf, M., Price, P., & Shattuck, S. (1995). *The Boston University Radio News Corpus*. Tech. rep., Boston University.

Payrató, L., & Fitó, J. (2005). Corpus audiovisual plurilingüe. Tech. Rep. 35, Universitat de Barcelona.

Penny, R. (2000). *Variation and change in Spanish*. Cambridge: Cambridge University Press.

Pitt, M., Johnson, K., Hume, E., Kiesling, S., & Raymon, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication, 45*, 89–95.

Pitt, M., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., et al. (2007). *Buckeye corpus of conversational speech (2nd release).* [http://www.buckeyecorpus.osu.edu]. Columbus, OH: Department of Psychology, Ohio State University (distributor).

Prieto, P., & Cabré, T. (2010). (coords.). *The interactive atlas of Catalan intonation*. http://prosodia.upf.edu/atlesentonacio/index-english.html.

Rodero, E. (2006). Analysis of intonation in news presentation on television. In *Proceedings of ExLing*.

Rodero, E. (2007). Characterization of a proper news presentation in the audiovisuals messages. *Estudios del mensaje periodístico*, pp. 523–542.

van Santen, J. P., & Buchsbaum, A. L. (1997). Methods for optimal text selection. In *Proceedings of Eurospeech 1997*, pp. 553–556.

Sperber-McQueen, C., & Burnard, L. (1994) *Guidelines for electronic text encoding and interchange*. Chicago and Oxford: Text Encoding Initiative.

Strangert, E., & Gustafson, J. (2008). What makes a good speaker? Subject ratings, acoustic measurements and perceptual evaluations. In *INTERSPEECH'08*, pp. 1688–1691.

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge: Cambridge University Press.

Xu, Y. (2001). Speech prosody: A methodological review. *Journal of Speech Sciences, 1*(1), 85–115.