The Glissando project

Building and exploitation of a large bilingual annotated corpus for multidisciplinary prosodic analysis and applications[1]

Juan María Garrido Almiñana[1], David Escudero Mancebo[2], Lourdes Aguilar Cuevas[3]

[1]Computational Linguistics Group (GLiCom), Universitat Pompeu Fabra, Barcelona
juanmaria.garrido@upf.edu
[2]Advanced Computer Environments and MultiModal Interaction Systems group (ECA-SIMM), Universidad de Valladolid
descuder@infor.uva.es
[3]Group of Prosodic Studies (GrEP), Universitat Autònoma de Barcelona
lourdes.aguilar@uab.cat

Abstract. This paper presents the Glissando project, whose goal is the design, collection and annotation of a speech corpus for prosodic studies in Spanish and Catalan, and its joint exploitation for different research goals by the involved research groups.

Keywords. Speech corpora, Prosody, Spanish, Catalan.

1 Introduction

The aim of this paper is to introduce the Glissando project, a join initiative of three research groups interested in the analysis of Spanish and Catalan Prosody from different perspectives: the Computational Linguistics Group (GLiCom) at Pompeu Fabra University, in Barcelona; the Advanced Computer Environments and MultiModal Interaction Systems (ECA-SIMM) research group at the University of Valladolid; and the Group of Prosodic Studies (GrEp) at the Barcelona Autonomous University. The goal of the project is the design, collection and annotation of a speech corpus suitable for their research interests, as well as its joint exploitation for different research goals. This project, started in 2009, is being funded by the *Plan Nacional de I+D* of the Spanish Government (coordinated projects FFI2008-04982-C03-01 and FFI2011-29559-C02-00).

At the current state of the project, a first version of the corpus is already available for public use, including orthographic and phonetic transcription, and different types

[1] 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).

of prosodic annotation. First exploitation results have also been obtained. Next sections describe the main features of this corpus, and outline the active exploitation research lines.

2 The Glissando corpus

Glissando is a speech corpus specially designed for the analysis of Prosody from different perspectives (Phonetics, Phonology, Discourse Analysis, Speech Technologies, comparative studies). It is actually made of two parallel corpora, **Glissando_sp** (Spanish) and **Glissando_ca** (Catalan), designed following the same criteria and structure. It is also representative of three different speaking styles: news reading, task-oriented dialogue and informal conversation. The design characteristics of the Glissando corpus a first-quality material for the experimental analysis of Spanish and Catalan Prosody:

- Good acoustic quality: Glissando has been recorded at professional studios, using high-quality microphones and recording conditions that guarantee the best quality.
- Transcribed and annotated: the Glissando corpus has been orthographically and phonetically transcribed; in addition, it includes several levels of annotation (currently prosodic unit segmentation and F0 annotation).
- High number of speakers: Glissando includes recordings of 28 different speakers per language, both professional and non-professional.
- More than 20 hours of speech have been recorded per language, an adequate size for corpus-based and experimental studies, and also for technological applications.

2.1 Contents

Both Glissando_sp and Glissando_ca are made up of three subcorpora, representing three different speaking styles: **news** (studio recordings of news readings), **task dialogues** (studio recordings of interactions between two speaking oriented to a concrete goal, the collection of information about a topic) and informal dialogues (studio recordings of informal conversations between two speakers). Table 1 summarizes the contents of all three subcorpora in both languages.

The news subcorpus was recorded using a selection of real news texts kindly provided by the *Cadena SER* radio station. These news texts were previously modified in order to obtain an adequate coverage of the target prosodic and segmental considered criteria. News texts were read in studio by eight professional speakers per language: for radio speakers and four dubbing/advertising speakers. The subcorpus includes two subsets: **'prosodic' news** (36 texts chosen according to prosodic criteria, such as the location of stressed syllables in lexical words or breath group length) and **'phonetic' news** (36 texts chosen according to segmental criteria, such as phone frequency). The design of this corpus is partially inspired in the Boston University Radio News Corpus [1].

Task dialogues subcorpus is made up of a set of interactions between two people (about 10 minutes long) oriented to the consecution of a goal: one of the speakers (*'asker'*) asks for information and the second one (*'giver'*) tries to give it to him, if he has the requested information available. It is inspired in the Map Task [2] protocol. Each pair of speakers performed three tasks: **transport** (planning a business travel between Ávila and Ciudad Real, by train or bus, for a concrete date), **university** (planning an Erasmus stay in Paris) and **tourist** (planning a tourist travel to Corfú). The subcorpus was recorded by twelve pairs of speakers (24 speakers in total): one of radio professional speakers, one of dubbing/advertising professional speakers, and 10 student pairs.

Finally, informal dialogue subcorpus is made up of informal conversations between two people with a previous knowledge of each other. The dialogue always started in the same way, remembering how both speakers first met. However, after this common starting point, conversations could derive to any other topic. This task was carried out by 6 pairs of speakers (12 people): one of radio professional speakers, one of dubbing/advertising professional speakers, and 4 student pairs. This subcorpus was designed taking as general reference the Buckeye corpus of conversational speech [3].

Corpus	**Contents**	**Speakers**	**Size**	
			Spanish	Catalan
News	36 'prosodic' texts	4 radio professional 4 dubbing/advertising professional	6 hours 40 minutes	6 hours 23 minutes
	36 'phonetic' texts	2 radio professional 2 dubbing/advertising professional		
Task-oriented dialogues	12 'transport' interactions 12 'university' interactions 12 'tourist' interactions	2 radio professional 2 dubbing/advertising professional 20 students	4 hours 37 minutes	5 hours 33 minutes
Informal dialogues	6 conversations	2 radio professional 2 dubbing/advertising professional 8 students	1 hour 6 minutes	1 hour 8 minutes

Table 1. Content summary of the Glissando corpus

2.2 Recordings

Recordings took place in two different premises: soundproof rooms at the Audiovisual Media Service of Valladolid University for the Spanish recordings, and at the Communication Campus of the Pompeu Fabra University, 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 along the recordings, making the energy registration reliable for prosodic analyses. In dialogue recordings, each speaker used different microphones, to have separate recordings of the speech from each participant, so as to avoid 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 laryngograph, 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.

2.3 Speakers

As already mentioned, 28 different speakers participated in the recordings of each language subcorpus, 8 professional (4 radio speakers and 4 dubbing/advertising speakers) and 20 undergraduate students. Radio speakers came from the *Cadena SER* radio station in Valladolid, in the case of Spanish, and from different radio stations in Barcelona (*Catalunya Ràdio*, *RAC1*, *Ràdio Estel*) in the case of Catalan. Dubbing/advertising speakers were all active professional in the field in Valladolid, in the case of Spanish, and Barcelona, in the case of Catalan speakers. The most general profile of the student speakers was last-year undergraduate students of Communication (at *Universidad de Valladolid* in the case of Spanish; at *Universitat Pompeu Fabra* and *Universitat Autònoma de Barcelona* in the case of Catalan), although some other student profiles were also accepted in some cases.

The selected speakers were all native speakers of the target language, speakers of the Castilian variety in the case of Spanish, and of Central Catalan in the case of Catalan.

All the speakers were organized into four categories, defining their contribution to the corpus recording. Table 2 describes these categories. According to this organization, some speakers (those of the A category, all of them professional) participated in the recordings of al three subcorpora; some others in recorded two subcorpora (those of the C category, all of them students); and finally, the last two groups participated only in the recordings of one subcorpus (the news subcorpus, in the case of the B category, and the task oriented dialogues, in the case of D category). By this distribution, we ensured speaker comparability among all three speaking styles, at least for the professional profiles.

2.4 Orthographic transcription

In addition to the recordings, the corpus includes also the corresponding orthographic transcriptions of the collected materials. In the case of the news subcorpus, the original texts read by the speakers were revised in order to adapt their contents to the actual reading of each speaker. The resulting transcription is stored in plain text files. Dialogues were all transcribed manually, applying the criteria proposed by TEI [4]. The obtained transcription was stored in xml files, which include, in addition to the transcription itself, additional information about the organisation of the dialogues into speech turns, and the annotation of the paralinguistic events.

2.5 Annotation

The annotation of the corpus is stored in TextGrid Praat files [5]. Five levels of annotation are currently available, stored in separate tiers of the TextGrid file and time-aligned with the speech signal: orthographic transcription (word by word), phonetic transcription (using the SAMPA phonetic alphabets for Spanish [6] and Catalan [7]), syllable boundaries, intonation group boundaries, and breath group boundaries.

A first version of this annotation was obtained automatically, by using several transcription, annotation and alignment automatic tools. Orthographic and phonetic transcriptions were obtained from the output of the phonetic segmentation tool embedded in the Cereproc's Voice Creation Kit. The versions of this tool for Spanish and Catalan were developed jointly by Cereproc and the Speech and Language Group of *Fundació Barcelona Media*, with the participation of members of GLiCom [8]. Prosodic unit segmentation (syllables, intonation groups, breath groups) was obtained using SegProso, an automatic annotation tool developed at GLiCom.

Category	Tasks	Speaker type	Number
A	News (prosodic) Informal dialogue Task-oriented dialogue	Professional (radio) Professional (dubbing/advertising)	2 2
B	News (prosodic+ phonetic)	Professional (radio) Professional (dubbing/advertising)	2 2
C	Informal dialogue Task-oriented dialogue	Student	8
D	Task-oriented dialogue	Student	12
Total			28

Table 2. Speaker categories for each language

2.6 Public version

A first version of the corpus has been made accessible for public access at the web page of the project (http://veus.barcelonamedia.org/glissando/), under a Creative Commons license, which allows its free use, modification and distribution if the original source is mentioned. Previous registration is required to access the corpus. This public version includes all the collected recordings, their orthographic transcription and the 'automatic' version of the annotation files. Tasks-oriented and informal dialogues are available in two versions: **'Complete'**, in which each dialogue recording is included into a single file; and **'Turns'**, which includes separate wav files for every speaking turn in a dialogue. Table 3 describes the available contents for each corpus unit in the different subcorpora.

2.7 Ongoing improvements

There are currently two ongoing tasks in the project oriented to improve the current version of the Glissando corpus: the manual revision of the transcription and annotation of the corpus obtained by automatic means, and the addition of new levels of annotation.

The revision task is aimed to detect and fix possible errors in the transcription and annotation generated by the automatic tools. All the tiers included in the time-aligned

annotation TextGrid (word, phone, syllable, intonation group and breath group boundaries) are being revised by linguists specially trained for this task. Most frequent errors found in the already corrected material include the insertion of wrong pause boundaries, mismatches between the theoretical word and phonetic transcription (the ones derived from the orthographic transcription) and actual realisations of the speakers in order to find and correct, and false detections of intonation group boundaries. At the current stage of the work, all the news subcorpus and a small part of the dialogues subcorpora of Glissando_sp (Spanish), and a subset of the news subcorpus (5 speakers out of 8) in the case of Glissando_ca, have been already revised.

Special attention is also being paid to the addition of new annotation tiers to the corpus. One of the main goals established at the beginning of the project was to include in Glissando annotation several types of intonation standards, including the most used ones, such as ToBI [9] or MoMel [10], in order to offer researchers working on different prosodic analysis frameworks a corpus suitable for their research interests, and also to attempt comparisons between some of these systems. Intense research on the automatic annotation of corpora using ToBI labels has been carried out in parallel to the development of the corpus [11] [12] [13] [14]. This research has led to a first automatic annotation of prominences using the ToBi-framework conventions, and it is planned to apply these results to the full automatic ToBI annotation of the Glissando corpus. Some efforts have also been done in the definition of the ToBI label set and labelling conventions. In addition, the corpus has been partially annotated using MelAn, an automatic tool for the annotation of intonation inspired in the IPO model [15]. This 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).

Finally, some advances have been done in the task of including other types of labelling to the corpus, not directly related to prosodic annotation, but relevant for prosodic analyses. This is the case, for example, of speech act labelling: a pilot annotation test, using a first list of speech act labels, has been done with a sample dialogue in Spanish; also, a pilot version of a tool for the automatic for detection of speech acts in text has been developed.

Future tasks to be started in the near future in this area include the annotation of the corpus with morphosyntactic information, and the annotation of focalised elements, including the adaptation or development of automatic tools to support these annotations.

3 Corpus exploitation

As already mentioned, the Glissando corpus was designed to fulfill the research interest of the groups involved in its development, different but with common points:

- Modelling of Spanish and Catalan Prosody, based on the analysis of large amount of data (phonetic modelling; phonological modelling, within the ToBI framework; analysis of the relation between phonetic and phonological patterns).
- Experimental comparison of prosodic annotation systems

- Development of prosodic models for Text-to-Speech applications
- Experimental analysis of the prosodic features of the dialogue speech in Spanish and Catalan
- Experimental analysis of the prosodic features of news readings by professional speakers
- Inter-linguistic comparison of prosodic features
- Development of teaching materials

Exploitation tasks are in a very initial phase: some work has been carried out in the area of phonetic modeling of F0 contours in Spanish dialogues, in the development of prosodic models for Text-to-Speech, and in the description of phonetic features of Prosody in news reading in Spanish and Catalan. A preliminary description of the F0 patterns used by the Spanish professional speakers in the task-oriented and informal dialogues is given in [16], and first results about the F0 patterns used by professional speakers in news readings in Catalan have been recently described in [17]. Also, some work has been done in the description of the prosodic features of news material by professional speakers, and in the description of the paralinguistic elements used in dialogues, although the obtained results, still preliminary, have not been published yet. And new research tasks, related to the different research areas detailed before, are expected to be started in the next future.

4 Conclusions

To date, the Glissando project has already offered some useful results, mainly in the form of a high-quality speech annotated material that is starting to be exploited by the involved research groups. However, this corpus may also be of interest for other research groups interested in the description of Spanish and Catalan Prosody from different perspectives.

These improvements will be made public in future releases of the corpus, available from the project website.

					News	Dialogues ('complete' version)	Dialogues ('turns version)
Unit					News text	Dialogue	Speech turn
Contents	Speech signal	Format wav (16.000 Kz)	Fixed microphone signal	.fix.wav	X (Mono)	X (Stereo)	X (Mono)
			Headset microphone signal	.wir.wav	X (Mono)	X (Stereo)	X (Mono)
			Laryngograph signal	.lar.wav	X (Mono)		
	Orthographic transcription	Text format (UTF-8)	Text	.txt	X		
		XML format (UTF-8)	Text enriched with information about speech turns and paralinguisitc events	.xml		X	
	Time-aligned annotation	TextGrid Praat format	Tiers containing different levels of annotation	.TextGrid	X	X	X
	Intensity values	Intensity Praat format	Intensity values corresponding the the fixed microphone signal	.fix.Intensity	X	X	X
			Intensity values corresponding the the headset microphone signal	.wir.Intensity	X	X	X
	F0 values	Fixer Pitch de Praat	F0 values corresponding the the fixed microphone signal	fix.Pitch	X	X	X
			F0 values corresponding the the headset microphone signal	.wir.Pitch	X	X	X

Table 3. Units, contents and files available in the public version of each language subcorpus

5 References

1. Ostendorf, M., Price, P., Shattuck, S.: The Boston University Radio News Corpus. Tech. rep., Boston University (1995)
2. Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H.S., Weinert, R.: The hrc map task corpus. Language and Speech (24), pp. 351 – 366 (1991)
3. Pitt, M., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., Fosler-Lussier, E.: Buckeye corpus of conversational speech (2nd release) [www.buckeyecorpus.osu.edu] Columbus, oh: Department of psychology, Ohio state University (distributor) (2007)
4. Sperber-McQueen, C, Burnard, L.: Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford: Text Encoding Initiative (1994)
5. Boersma, P., Weenink, D.: Praat: doing phonetics by computer [Computer program] http://www.praat.org/ (2012)
6. http://www.phon.ucl.ac.uk/home/sampa/spanish.htm

7. http://liceu.uab.es/ joaquim/language resources/SAMPA Catalan.html
8. Garrido, J. M., Bofias, E., Laplaza, Y., Marquina, M., Aylett, M., Pidcock. Ch.: The CEREVOICE speech synthesiser. In: Actas de las V Jornadas de Tecnología del Habla (Bilbao, 12-14 noviembre 2008), pp. 126-129 (2008)
9. Beckman, M,, Hirschberg, J., Shattuck-Hufnagel, S.: The original ToBI system and the evolution of the ToBI framework. In Jun SA (ed): Prosodic Typology: The Phonology of Intonation and Phrasing, New York: Oxford University Press, pp 9–54 (2005)
10. Hirst, D., Espesser, R.: Automatic modelling of fundamental frequency using a quadratic spline function. Travaux de l'Institut de Phonétique d'Aix, 15, pp. 71-85 (1993)
11. Escudero, D., Aguilar, L., Vanrell, M., Prieto, P.: Analysis of inter-transcriber consistency in the Cat ToBI prosodic labeling system. Speech Communication, 54, pp. 566–582 (2012)
12. González, C., Escudero, D., Vivaracho, C., Cardeñoso, V.: Improving automatic classification of prosodic events by pairwise coupling. IEEE Transaction on Audio, Speech and Language Processing p (in press)
13. Escudero, D., Aguilar, L., Ferreras, C.G., Vivaracho, C., Cardeñoso, V.: Cross-lingual English Spanish tonal accent labeling using decision trees and neural networks. In: Travieso-González C.M., Hernández, J.B.A. (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, 7015, pp 63–70 (2011)
14. Escudero, D., Vivaracho, C., González, C., Cardeñoso, V., Aguilar, L.: Analysis of inconsistencies in cross-lingual automatic ToBI tonal accent labeling. In: Habernal, I,, Matousek, V. (eds): Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings, Springer, Lecture Notes in Computer Science, 6836, pp 41–48 (2011)
15. Garrido, J. M.: A tool for automatic f0 stylisation, annotation and modelling of large corpora. In: Speech Prosody 2010, Chicago (2010)
16. Garrido, J. M., Rustullet, S.: Patrones melódicos en el habla de diálogo en español: un primer análisis del corpus Glissando. Oralia: Análisis del discurso oral, 14, pp. 129-160 (2011)
17. Garrido, J. M.: GLISSANDO. Un corpus anotat per a l'anàlisi de la prosòdia del català i del castellà. Descripció i primers resultats d'explotació. Workshop d'entonació del català, juny 2012 (2012)