CORP-ORAL: Spontaneous speech corpus for European Portuguese

Tiago Freitas, Fabíola Santos

ILTEC
Lisbon, Portugal
E-mail: taf@iltec.pt, fabiola.santos@iltec.pt

Abstract

Research activity on the Portuguese language for speech synthesis and recognition has suffered from a considerable lack of human and material resources. This has raised some obstacles to the development of speech technology and speech interface platforms. One of the most significant obstacles is the lack of spontaneous speech corpora for the creation, training and further improvement of speech synthesis and recognition programs. It was in order to suppress this gap that the CORP-ORAL project was planned. The aim of the project is to build a corpus of spontaneous EP available for the training of speech synthesis and recognition systems as well as phonetic, phonological, lexical, morphological and syntactic studies. Further possibilities of enquiry such as sociolinguistic and pragmatic research are also covered in the corpus design. The data consist of unscripted and unprompted face-to-face dialogues between family, friends, colleagues and unacquainted participants. All recordings are orthographically transcribed and prosodically annotated. CORP-ORAL is built from scratch with the explicit goal of becoming entirely available on the internet to the scientific community and the public in general.

1. Introduction

CORP-ORAL is a project being developed in ILTEC under the auspices of the Portuguese Foundation for Science and Technology (FCT). The motivation for the project was borne out of evidence that research on Portuguese has suffered from a considerable lack of human and material resources in what regards speech synthesis and recognition tools development. One of the most significant obstacles is undoubtedly the lack of spontaneous speech corpora for the creation, training and further improvement of speech synthesis and recognition programs.

The aim of the project is to build a corpus of spontaneous European Portuguese (EP) available for the training of speech synthesis and recognition systems as well as phonetic, phonological, lexical, morphological and syntactic studies. Further possibilities of enquiry such as sociolinguistic and pragmatic research are also covered in the corpus design.

Such an ambitious range of research applications is only feasible with a fully accessible corpus. Therefore CORP-ORAL is built from scratch with the explicit goal of becoming entirely available on the internet to the scientific community and the public in general.

2. Background

During the last decades there has been an increase in the development of large spoken corpora such as the Corpus Gesproken Nederlans - CGN (Schuurman et al. 2004) or the British National Corpus - BNC (Burnard, 2002). Such corpora (and particularly subcorpora among them consisting of spontaneous speech) have been prepared to meet the growing demands deriving from different areas of language research and speech tools development.

Projects as large as CGN and BNC require a great deal of financial and human resources as well as extensive planning and organization. Alongside these national enterprises there has been an emergence of supranational projects such as C-ORAL-ROM (Cresti et al. 2004) which aim to establish standardized corpora for different languages. Being supported by international teams, initiatives like C-ORAL-ROM have made intensive data collection tasks possible in a relatively short time span. These initiatives comprise the advantage of allowing one to make comparative studies of specific linguistic phenomena.

CORP-ORAL was created as a tool to further expand the spoken language resources available in EP. Its main goal is to provide highly spontaneous interaction data with high quality audio. Part of this material includes recordings made in communication contexts which have not been incorporated in previous corpora. Another relevant feature of CORP-ORAL is that it is built from start to meet current metadata standards like IMDI (2003).

3. Data

The speech data consists of face-to-face dialogues recorded with a Marantz PMD670 solid state portable recorder and Beyerdynamic Opus 55.18 microphones. The audio is captured in stereo mode with each speaker recorded into a separate channel. Recordings are made in 44 Khz WAV format and have different lengths up to 96 minutes (maximum storage capacity of an individual 1 GB CF card). The dialogues take place in partially controlled environments, mostly in closed rooms in ILTEC or in one of the informants' houses.

	Time
Recording	50 hours
Ortographic Transcription	30 hours
Phonetic Transcription	1 hour

Table 1: Expected amount of speech material collected.

	Time
Recording	53 hours
Ortographic Transcription	32 hours
Phonetic Transcription	1 hour

Table 2: Amount of speech material collected.

3.1 Communication Context

The speech data collected is spontaneous in the sense that it consists of unscripted and unprompted dialogues between family, friends, colleagues and unacquainted participants. Participants are free to pursue whichever topics they feel inclined to talk about. It is not entirely spontaneous because participants are aware of the recording process and because they must review and sign a specific form where they authorize proposed conditions for data collection and publication. This may prevent them from engaging in personal topics and from using specific speaking styles.

It is noteworthy to mention that regardless of these procedural constraints the recordings collected up to now show a surprising degree of spontaneity. Speakers do in fact pursue private topics and feel comfortable to adopt different speaking styles. This results in diverse voice quality settings produced by the speakers, dramatic changes in pitch and tempo, etc.

Although the majority of the recordings consists of conversation, there are some in which the speakers naturally assume an interviewer-interviewee relationship. Since the initial planned amount of 30 hours of spontaneous conversation has been obtained, the recording of different genres is now encouraged as standard procedure: interaction in videogame playing context (already tested), map task dialogues, role playing, etc.

3.2 Participants

Participants recorded are male and female speakers of standard EP with ages between 17 and 74 years. It is a requirement that all participants be born and raised in the vicinity of Lisbon. Even though there is some debate over the geographical delimitation of EP, Lisbon was chosen because it is the most populated city in the country and it is the main base of national radio and TV stations.

Up to now there are approximately 60 speakers taped, with several degrees of education ranging from 9th grade to PhD.

3.3 Ortographic Transcription

Words uttered by the participants in the recordings are transcribed in orthographic format as this is the preferred text annotation method for Portuguese (Bacelar do Nascimento 2005, Freitas in press).

The orthographic transcription of all conversations is made using ELAN software. An excerpt of such transcription is depicted as screenshot in Figure 1 (end of the paper). ELAN was chosen because it can export annotations to different file formats so that they can be processed by other linguistic analysis applications such as Praat without any loss of information (particularly without losing the information about time/speech alignment).

Each ELAN annotation file comprises a total of five tiers. These are primarily split in:

- orthographic transcription (OT)
- prosodic annotation (PA)

The OT string is used for transcribing not only the speech but also disfluencies and breathing periods, laughter and all sort of linguistic or paralinguistic noises made with the vocal tract.

The PA string is a copy of OT without the laughter and breathing marks and with additional linguistic information regarding prosodic structure. Prosodic boundaries are defined by the transcriber according to his or her own perception.

Prosodic breaks can be of two kinds:

- terminal (marked with one slash)
- non-terminal (marked with two slashes)

This prosodic annotation system conforms to the criteria established for the previous C-ORAL-ROM project, which are published in Moneglia et al. (2005).

Besides these two levels of transcription for each speaker, there is a fifth string which is used to inscribe all phenomena external to the conversation such as background noises (including third speakers' speech, noisy cars passing by, etc.) and any kind of noise in the recording file itself.

The orthographic transcription of recorded speakers can be made concurrently or one at a time according to transcribers' preference. Previous tests revealed no significant differences using one or the other method. In any case, all tiers are defined initially and encoded into the ELAN transcription file.

The fact that speakers are recorded into separate channels facilitates the transcription of overlapped speech, allowing the transcriber to check the speech of each participant individually. In all recorded stereo files there is some sound leaking from one channel to another but the leaked samples are never loud enough so as to prevent overlapped speech portions from being easily verified.

After an audio file is transcribed it undergoes a non-destructive revision process. It is non-destructive in the sense that original tiers are not altered: they are copied into new revision tiers and only these can be changed. This procedure prevents information loss, and furthers the potential for linguistic enquiry resulting in information which can be used for instance to compare between different interpretations and different prosodic boundary perceptions of the transcriber and the reviser. Therefore transcribers can become potential objects of study, which adds even further possibilities in terms of the analysis of the corpus without the need of supplementary human

resources.
The background tier is the only one to be revised directly and thus the only one submitted to a destructive process. Such procedure was adopted because changes introduced in this tier were not deemed relevant enough for future analysis.
Because the recordings are annotated by different transcribers, the revision process is carefully planned so that the amount of researchers assigned that task is reduced to a minimum. This directive is enforced in an attempt to make the data more homogenous.

4. Phonetic Transcription

Phonetic transcription is made in Praat. This allows for the full integration of orthographic and phonetic tiers in ELAN transcription files. An excerpt of the phonetic transcription is depicted as screenshot from Praat in Figure 2 (end of the paper).
For this process, the stereo audio files are first converted into mono. Here again, it is particularly felicitous to have channel separation recordings because there is no information loss in the conversion.
Each transcriber is assigned a portion of a mono audio file. The audio file is then opened in Praat and three annotation tiers are created in a TextGrid file connected with the audio.
The three strings consist of:

- Seg tier: segment string
- Pal tier: word tier
- Obs: transcriber observations tier

In the Seg tier, each phonetic segment is aligned with the corresponding section in the spectrum. After individual segments are identified, these are grouped as word sequences in the Pal tier. Finally, the Obs tier is reserved for comments of the transcriber regarding different aspects of the transcription process which are considered relevant. Comments can be inserted relating to:

- particular types of phonation
- particular types of intonation
- unusual articulations
- procedural difficulties

Although these transcriptions are made separately for mono files, they are later assembled in master TextGrid files combining a total of six tiers with the phonetic transcription of the entire conversation segment by segment.

5. Availability

This corpus will be made integrally available on the Internet in two different ways:

- Spock, a spoken corpus access tool
- IMDI database

Spock is a lightweight web-based application built by Maarten Janssen providing easy, online access to time-aligned corpora. It is currently in development stage. A preliminary version for browsing CORP-ORAL is accessible at:

http://www.iltec.pt/spock/

The corpus will also be made available in browsable full file format by accessing the IMDI database. Users interested in obtaining the corpus for research purposes will be given access privileges.

6. Acknowledgements

CORP-ORAL is an FCT funded project grant reference POCTI/LIN/60019/2004. Gratitude must be expressed towards CORP-ORAL consultants for their valuable input throughout different stages of the project: Miguel Oliveira Jr., Maria do Céu Viana, Maria Fernanda Bacelar do Nascimento and Fernando Martins.

7. References

Bacelar do Nascimento, M. F. et al. (2005). The Portuguese corpus. In Cresti and Moneglia, editors, *C-ORAL-ROM: integrated reference corpora for spoken Romance languages*. Amsterdam, Benjamins,

Bael, C.P.J. et al. (2004). On the Usefulness of Large Spoken Language Corpora for Linguistic Research. In *Proceedings of LREC 2004, 4th International Conference on Language Resources and Evaluation*. Lisbon, Portugal: ELRA, pages 2135-2138.

Burnard, L. (2002) Where did we go wrong? a retrospective look at the British National Corpus In *Teaching and learning by doing corpus analysis*, ed. B. Kettemann and G. Markus. Amsterdam: Rodopi, pp 51-71.

Freitas, T. (in press). Recolha e transcrição de corpora orais. In *Perspectivas sobre a oralidade*. Santiago de Compostela, Consello da Cultura Galega.

IMDI (2003). *IMDI Metadata Elements for Session Descriptions*. MPI Nijmegen.

Moneglia, M. et al. (2005). Evaluation of consensus on the annotation of terminal and non-terminal prosodic breaks in the C-ORAL-ROM corpus. In Cresti and Moneglia, editors, *C-ORAL-ROM: integrated reference corpora for spoken Romance languages*. Amsterdam, Benjamins, pages 257-276.

Schuurman, I. et al. (2004). Linguistic annotation of the Spoken Dutch Corpus: if we had to do it all over again. In *Proceedings of LREC 2004, 4th Internacional Conference on Language Resources and Evaluation*. pages 57-60.

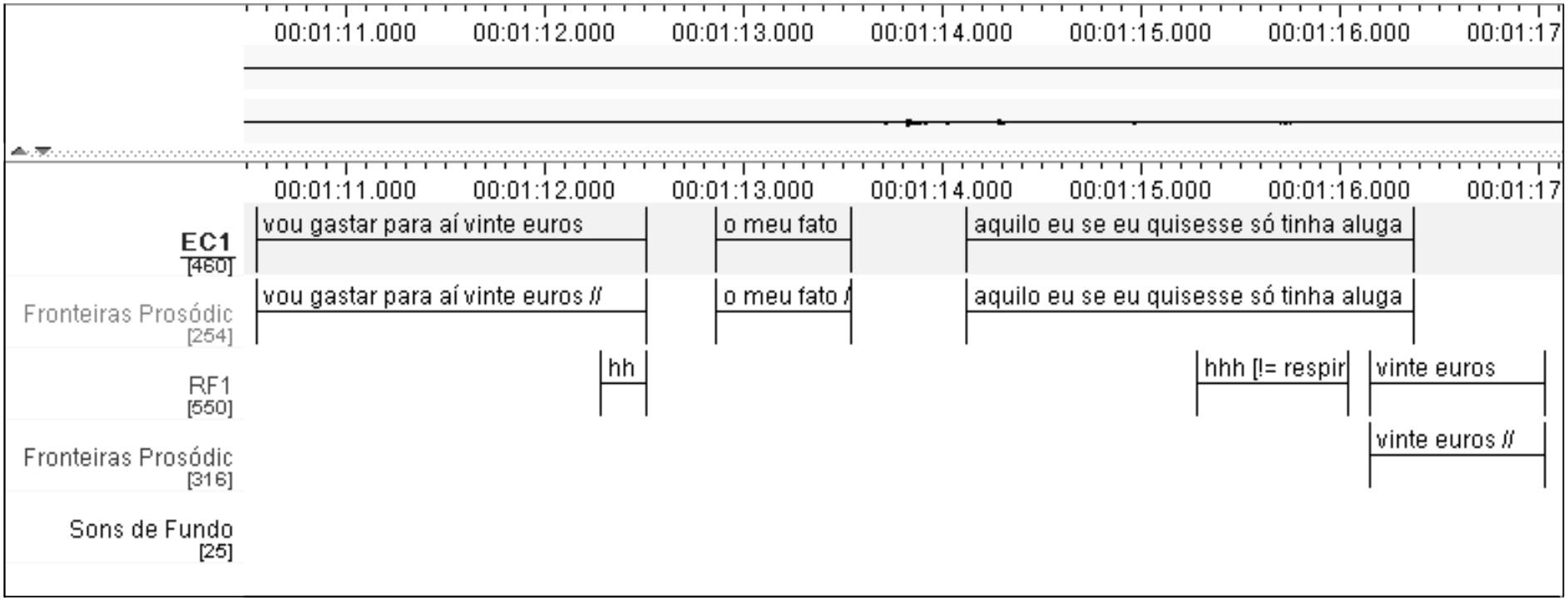

Figure 1: Fragment (presented in the form of an ELAN screenshot) of the orthographic transcription of conversation #10, 70-79 seconds.

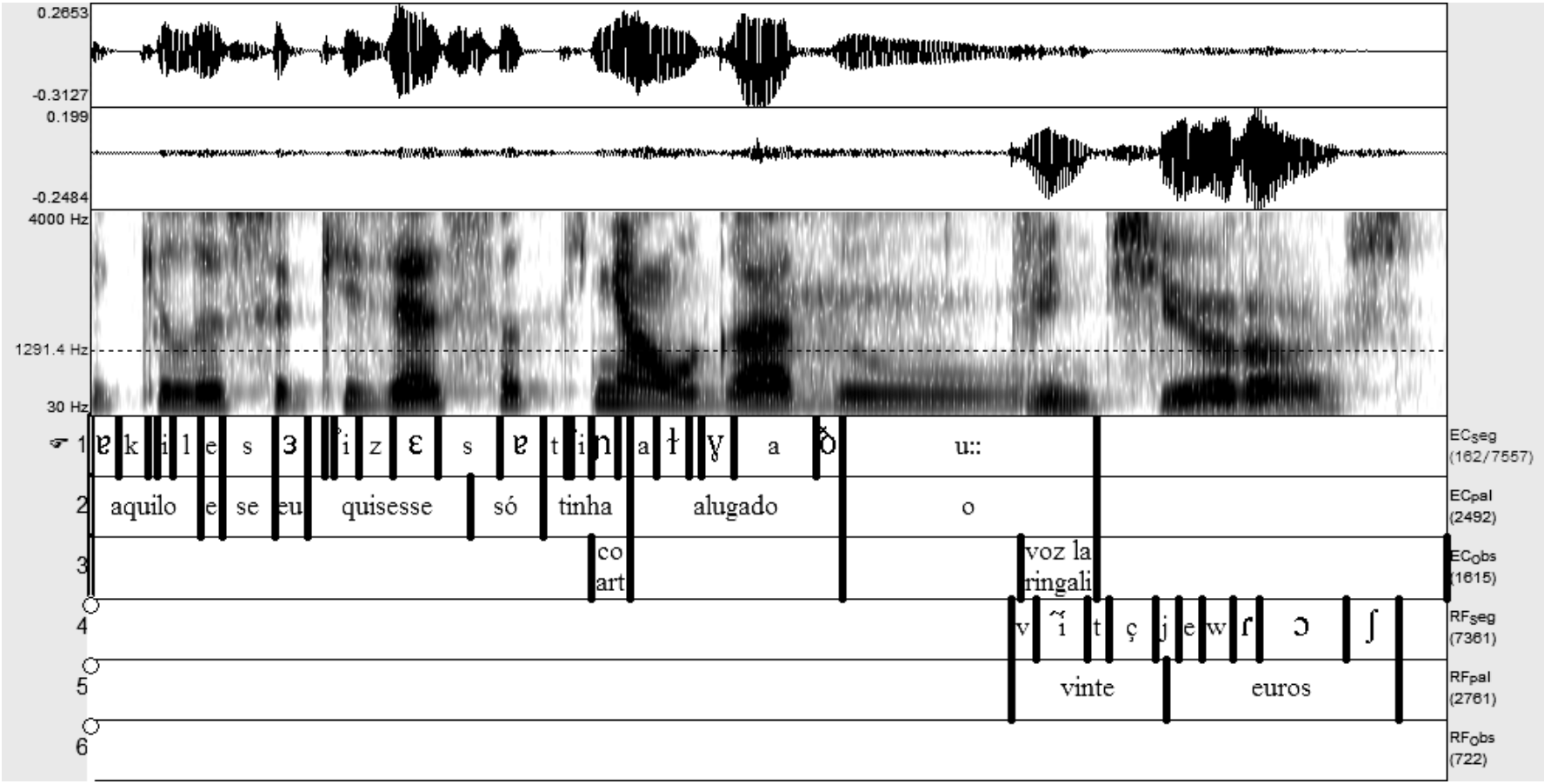

Figure 2: Fragment (presented in the form of a Praat screenshot) depicting the phonetic transcription of conversation #10, 74-77 seconds.